UNIVERSITY OF CALIFORNIA
Los Angeles

Nature-inspired Metaheuristics for Biostatistical and Biomedical Research

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Elvis Han Cui

2024

ABSTRACT OF THE DISSERTATION

Nature-inspired Metaheuristics for Biostatistical and Biomedical Research

by

Elvis Han Cui
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2024
Professor Gang Li, Co-Chair
Professor Weng Kee Wong, Co-Chair

This dissertation addresses complex biostatistical challenges through the application of nature-inspired metaheuristics. As the name suggests these algorithms are often motivated by animals' behavior and natural processes. A salient feature of nature-inspired metaheuristic algorithm is that they provide flexible and robust strategies for solving tackling all types of optimization problems and can solve optimization problems that traditional methods cannot. Interestingly, they rarely come with rigorous proofs of convergence to the global optimum, but they frequently do so or get close to the optimum in practice. Codes for these algorithms are widely available in different format and platforms, and they are easy to implement and use. Consequently, nature-inspired metaheuristic algorithms are popular and are increasingly used across disciplines. There are many such algorithms, and to fix ideas, we focus on two such algorithms, Particle Swarm Optimization (PSO) and Competitive Swarm Optimizer with Mutated Agents (CSO-MA), and demonstrate their utility and effectiveness for tackling several types of biostatistical applications.

The primary contribution of this research is the development and applications of these algorithms to solve a range of biostatistical problems. They include solving challenging optimization problems to improve accuracy in statistical inference for single-cell RNA sequencing

data analysis (Chapter 2), parametric and non-parametric statistical estimation (Chapter 3), and finding more efficient and realistic experimental designs in toxicology (Chapter 5 and Chapter 6). In addition, the dissertation introduces an innovative semi-parametric Bayesian model (DPMIV) for interval-censored and doubly-censored data (Chapter 4).

The applications and results showcased in the dissertation not only highlight the adaptability of metaheuristics to tackle a diverse set of biostatistical problems but also open up new avenues for future research in statistical methodology and its applications in biomedicine.

The dissertation of Elvis Han Cui is approved.

Donatello Telesca

Michael D. Collins

Weng Kee Wong, Committee Co-Chair

Gang Li, Committee Co-Chair

University of California, Los Angeles

2024

To my parents, the whole extended family, and professors

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Throughout the six years I spent in the Department of Biostatistics at the University of California, Los Angeles (UCLA), I have been privileged to receive support, encouragement, and guidance from many individuals. This dissertation would not have been possible without their invaluable contributions.

First, I would like to express my heartfelt gratitude to my advisors and committee chairs, Dr. Weng Kee Wong and Dr. Gang Li, for their exceptional mentorship throughout my Ph.D. journey. They not only taught me how to engage in collaborative research, methodological development, and academic writing but also shared their experiences, both as graduate students and as professors. Their advice and support have inspired and encouraged me to become a professional researcher.

I am also deeply thankful to my dissertation committee members. In particular, Dr. Michael Collins provided invaluable suggestions on the applications of optimal design, and Dr. Donatello Telesca served as a great academic advisor, sharing numerous materials to further my learning. A special thanks goes to Dr. Jessica Li from the Department of Statistics, whose profound thinking in statistics has been a constant source of inspiration.

This dissertation has greatly benefited from the contributions of many collaborators, whose invaluable support I cannot fully express in words. Chapter 2 is a joint effort with Dr. Jessica Li and Dr. Dongyuan Song from the Department of Statistics at UCLA. Chapter 3 was developed in collaboration with Dr. Culsome Chen at Tsinghua University, Dr. Zizhao Zhang from Alibaba Group, and Dr. Yanan Li at Zhejiang University. Chapter 4 builds upon foundational work by Dr. Xuyang Lu at Adlai Nortye, whose generous provision of MCMC codes significantly expedited my research. This chapter was further enhanced by the contributions of Dr. Hua Zhou, Dr. Jin Zhou, and Ms. Aubrey Jensen. Chapter 6 was greatly enriched by the efforts of Ms. Jessica Munson, a student of Dr. Michael Collins from the Department of Environmental Health Sciences.

I am especially thankful to Mrs. Roxy Naranjo for her vital role in our department and her unwavering support throughout my Ph.D. journey. I am deeply grateful to Dr.

Sudipto Banerjee for fostering an inspiring and collaborative academic environment, and to Dr. Dorota Dabrowska for laying a strong theoretical foundation while answering my countless questions with patience. I also want to express my appreciation to Dr. Kaverisiano Youngchingski and Mr. Peterish Long, who supported me during my Master's studies, and to the Chinese Language and Literature expert Dr. Zhuang Lyeutsaon Liu, who took me on an unforgettable adventure to indulge in Google's free meals and creatively exploit capitalism's perks during a financially challenging period. Additionally, I extend my sincere thanks to Dr. Stavoros Panageas for making complex financial concepts accessible to statisticians, as well as to Dr. Yihao Li, Dr. Ruochen Jiang, Dr. Shanpeng Li, Dr. Alex Sverdlov, Dr. Mitchell Schepps, Dr. Lu Zhang, Dr. Rui Yang, Dr. Leiwen Gao, Dr. Heather Zhou, Dr. Zhenyu Zhang, Dr. Will Gertsch, and many others who offered their invaluable guidance and encouragement throughout this journey.

Finally, I would like to express my deepest appreciation to my family. I am especially grateful to my father, Dr. Donghui Wang, and my mother, Dr. Chenyang Cui, for their unwavering encouragement and support in pursuing my graduate studies abroad. Their unconditional love and belief in me have meant everything and have made this journey possible.

VITA

2014 - 2018 Bachelor of Agriculture, Colledge of Environmental & Resources Sciences, Zhejiang University, China

2018 - 2020 Master of Science, Department of Biostatistics, UCLA, USA

PUBLICATIONS

Li, Y., Li, P., **Cui, E. H.**, & Wang, D. (2021, May). Inference fusion with associative semantics for unseen object detection. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 3, pp. 1993-2001).

Collins, M. D., **Cui, E. H.**, Hyun, S. W., & Wong, W. K. (2022). A model-based approach to designing developmental toxicology experiments using sea urchin embryos. *Archives of toxicology*, 1-14.

Cui, E. H., Song, D., Wong, W. K., & Li, J. J. (2022). Single-cell generalized trend model (scGTM): a flexible and interpretable model of gene expression trend along cell pseudotime. *Bioinformatics*, 38(16), 3927-3934.

Cui, E. H., Li, B., Li, Y., Wong, W. K., & Wang, D. (2023). Trajectory-aware Principal Manifold Framework for Data Augmentation and Image Generation. *arXiv preprint* arXiv:2310.07801.

Cui, E. H., Goldfine, A. B., Quinlan, M., James, D. A., & Sverdlov, O. (2023). Investigating

the value of glucodensity analysis of continuous glucose monitoring data in type 1 diabetes: an exploratory analysis. *Frontiers in Clinical Diabetes and Healthcare*, 4, 1244613.

Cui, E. H., Zhang, Z., Chen, C. J., & Wong, W. K. (2024). Applications of nature-inspired metaheuristic algorithms for tackling optimization problems across disciplines. *Scientific reports*, 14(1), 9403.

Cui, E. H., Zhang, Z., & Wong, W. K. (2024). Optimal designs for nonlinear mixed-effects models using competitive swarm optimizer with mutated agents. *Statistics and Computing*, 34(5), 156.

Cui, E. H., Xu, H., & Wong, W. K. (2024). What is Metaheuristics? A Primer for the Epidemiologists. *arXiv preprint* arXiv:2411.05797.

Cui, E. H., Lu, X., Jensen, A., Zhou, J., Zhou, H., & Li, G. (2024). A semiparametric Bayesian approach for instrumental variable analysis with censored time-to-event outcomes. *In preparation.*

CHAPTER 1

Motivation: Opening New Insights for Statisticians

As Kai Lai Chung (1917-2009) pointed out in his monograph (Chung, 1974), "A mathematical course is not a stockpile of raw materials nor a random selection of vignettes. It should offer a sustained tour of the field being surveyed and a preferred approach to it." Similarly, a PhD dissertation in biostatistics may not be a compilation of various models and examples nor tedious and endless theoretical derivations. It shall bring new insights and provide new techniques (either conceptually or methodologically) for peers in statistics. In this dissertation, I am not aiming for such a high goal but I do hope that we can have some new horizons for either conventional or brand new statistical optimization problems. As the title suggests, such hope heavily depends on the so-called "metaheuristics".

Optimization plays a paramount role in statistics (Everitt, 2012; Lange, 2013; Rustagi, 2014). From stock marketing prediction in econometric, image classification in computer vision and allocation of coupons in Alibaba to protein structure prediction in biology, optimal design in clinical trials and variable selection in biomarker studies, everything involves optimization at different levels. Because of its great demand, there has been extensive literature on developing different optimization algorithms for approximating optimal solutions for different types of problems. Among all algorithms, this dissertation focuses on metaheuristics, a class of gradient-free algorithms popular in optimal design and engineering that is efficient for solving optimization problems with constraints and undesirable analytical properties (Talbi, 2009).

Metaheuristics is widely used in increasingly many disciplines because 1) there are many codes and packages available for users in a lot of programming languages including Matlab, R, Python, C++, etc., including examples of their very different applications to solving real

complex optimization problems. For example, particle swarm optimization (Eberhart and Kennedy, 1995) has been applied to 1) estimate infection fatality rate (IFR) in COVID-19 (Haouari and Mhiri, 2021), 2) optimize and estimate the relation between pseudotime (Trapnell et al., 2014) and gene expression values (Cui et al., 2022), 3) select the knot positions in adaptive spline regression models (Wang and Mohanty, 2010; Normandin et al., 2018; Mohanty and Fahnestock, 2021). Cuckoo search (Yang and Deb, 2009) has been applied to 1) train parameters in neural networks (Valian et al., 2013) and 2) parameter estimation problems in reliability theory (Valian et al., 2011). Genetic algorithm (Storn and Price, 1997) has been applied to find experimental optimal designs in statistics (Stokes et al., 2020). Competitive swarm optimizer and its variants has been applied to 1) find Bayesian optimal design for nonlinear mixed models applied to HIV dynamics and 2) travelling salesman problem (Zhang, 2020b).

Strangely, metaheuristics is still very underused in the mainstream of biomedical and biostatistics research. There are several reasons. First, many types of metaheuristics have not been exposed to statisticians so that statisticians do not use them often in practice (Cui et al., 2024a). Second, global convergence of many metaheuristics is only guaranteed under stringent assumptions, limiting it to the use in the statistical community (Tong et al., 2021). Third, it will be better for statisticians to have an "all-in-one" function (*lm()*, *glm()*, *geeglm()* in R) instead of having an optimization algorithm along. However, existing statistical packages do not use metaheuristics but other algorithms for finding estimates.

The dissertation mainly focuses on two particular types of metaheuristics known as particle swarm optimization (PSO) and competitive swarm optimization with mutated agents (CSO-MA) and the dissertation is organized as follows. In the rest of the chapter, we first state the main aim of the dissertation and then provide a historical review of metaheuristics, together with an introduction to several metaheuristic algorithms used in the dissertation. In chapter 2, we apply PSO to solve a constrained optimization problem in bioinformatics using single-cell RNA sequencing datasets. In chapter 3, we illustrate how metaheuristics can be applied to solve different types of estimation problems in statistics. In chapter 4, we develop a semi-parametric Bayesian instrumental variable analysis model with a customized

Markov Chain Monte Carlo (MCMC) algorithm. In chapter 5, we apply metaheuristics to solve optimal approximate design problems and in chapter 6 we propose a new sequential optimal design schema for toxicological studies. In chapter 7, we provide supplementary materials for the previous chapters.

1.1 Main Aim of the Dissertation

Main aim of the thesis is to demonstrate the usefulness and power of modern metaheuristic algorithms for solving complex biostatistics and biomedical problems that seem unsolvable until now. More specifically, we aim to solve the following problems involving metaheuristics:

- Develop an effective model for making improved inference of pseudotime to gain better insights of single cell dynamics, which is an increasingly important topic in bioinformatics.
- Demonstrate the effectiveness and usefulness of metaheuristics in parameter estimation problems including constrained and unconstrained maximum likelihood estimation, matrix completion, variable selection, etc.
- Develop a semiparametric Bayesian causal inference model for the analysis of doubly interval-censored data and methods for sampling from the nonparametric maximum likelihood estimator, and then develop an R package for users to use in practice.
- Develop a unified framework for finding 2-point optimal designs for binary regression with different link functions and then extend it to 3- or 4-point designs.
- Develop a two-stage design framework for toxicologists to perform experiments in a sequential manner, leading to a more reliable and efficient estimation of dose-response relationships.

In each of the above proposed research, we apply metaheuristic algorithms, especially PSO and CSO-MA, to solve different types of optimization problems and make recommendation on specific choice of the algorithms, including choice of tuning parameters that seem

to perform adequately for the problem at hand. In chapter 2, PSO plays a dominant role for finding the constrained optimal solution. In chapter 3, PSO and CSO-MA are applied to a series of estimation problems arising in statistics, ecology, engineering, etc. In chapter 4, PSO is applied to select single nucleotide polymorphisms (SNPs) that are highly correlated with the outcome. In chapter 5 and 6, we apply PSO and CSO-MA to find optimal designs when analytical solutions are not available or difficult to derive.

1.2 Metaheuristics

Metaheuristics is a widely-adopted, but under-appreciated term. The prefix "meta" originates from Greek with meaning "more comprehensive" or "transcending" and the word "heuristic" means "discovering or unravelling something by oneself". In literature, there are several jargons or synonym of metaheuristics, i.e., metaheuristic algorithms, nature-inspired metaheuristics, nature-inspired metaheuristic algorithms and nature-inspired algorithms. Though strictly speaking, there are subtle differences among these terminologies, many papers just use them to represent the same meaning. Hence, throughout the dissertation, we mainly adopt the term "metaheuristics" but sometimes switch to the other terminologies without further instruction. Metaheuristics is a class of approximate algorithms [1] (Archetti and Schoen, 1984) with four common characteristics: 1) it mimics a natural phenomenon from biology, physics, or ethology; 2) it has stochastic components (random perturbations); 3) it has tuning-parameters; 4) it does not require gradient information (Boussaïd et al., 2013; Korani and Mouhoub, 2021).

Based on the number of candidate solutions, metaheuristics can be classified into two main cagetories: single-solution-based (SS) algorithms and population-based (PB) algorithms. Based on the mechanism, the population-based algorithms can be further classified into two sub-categories: evolutionary computation (EC) algorithms and swarm intelligence

[1] The term "approximate" means the algorithms find nearly-optimal solutions to the problems while convergence to the global optimal is not guaranteed theoretically. In computer science, approximate algorithms have been used to solve NP-hard problems.

(SI) algorithms. For more details on this type of classification, see Gogna and Tayal (2013); Sörensen (2015); Aranha et al. (2021). The Table 1.1 provides a list of metaheuristics and we provide a description of some of them in the next subsection. For a more comprehensive and detailed review on metaheuristics, see Abdel-Basset et al. (2018); Hussain et al. (2019); Korani and Mouhoub (2021); Rajwar et al. (2023) for the general information, the monograph (Yang, 2017) and the paper on the applications of metaheuristics in deep learning (Tian and Fong, 2016).

Table 1.1: A Brief List of Metaheuristics

Algorithm	**Category** [1]	**Reference**
Stochastic Approximation	SS	Lai (2003); Robbins and Monro (1951)
Simulated Annealing	SS	Pincus (1970)
Genetic Algorithm	PB-EC	Sampson (1976)
Tabu Search	SS	Glover (1986)
Threshold Accepting	SS	Dueck and Scheuer (1990)
Ant Colony Algorithm	PB-SI	Drigo (1996); Dorigo et al. (1991)
Particle Swarm Optimization	PB-SI	Kennedy and Eberhart (1995)
Differential Evolution	PB-EC	Storn and Price (1997)
Evolutionary Programming	PB-EC	Fogel (1998)
Harmony Search	PB-EC	Geem et al. (2001)
Water Flow-like Algorithm	PB-SI	Yang and Wang (2007)
Cuckoo Search	PB-SI	Yang and Deb (2009)
Firefly Algorithm	PB-SI	Yang (2010)
Bat Algorithm	PB-SI	Yang and Gandomi (2012)
Flower Pollination	PB-SI	Yang (2012)
Competitive Swarm Optimizer	PB-SI	Cheng and Jin (2014)
Grey Wolf Optimizer	PB-SI	Mirjalili et al. (2014)
DPSO	PB-SI	Kim and Wong (2018)
NovDE	PB-EC	Xu et al. (2019)

Competitive Swarm Optimizer with Mutated Agents	PB-SI	Zhang et al. (2020)
Skip Salp Swarm Algorithm	PB-SI	Abualigah et al. (2022)
Snake Optimizer	PB-SI	Hashim and Hussien (2022)

[1]SS refers to single-solution-based, PB-EC and PB-SI refer to population-based evolutionary computation and swarm intelligence respectively.

1.3 Particle Swarm Optimization

We introduce one of the most commonly used metaheuristics, the particle swarm optimization (PSO) algorithm.

1.3.1 The PSO Algorithm

Swarm intelligence algorithms, such as ant colony (Dorigo et al., 2006), cuckoo search (Yang and Deb, 2009) and firefly algorithms (Yang, 2009), mimic the behaviour of a swarm to solve optimization problems. They are now receiving more and more interest and attention not only in the literature of mathematics, but also in econometrics, optimal design, engineering, etc (Yang, 2017). Particle swarm optimization (PSO), proposed by Kennedy and Eberhart (1995), is one of the most widely used swarm intelligence algorithms to optimize an objective function with boundary constraints. It is the main optimization tool in this dissertation and is introduced in below.

PSO tackles an optimization problem by producing a sequence of candidate solutions. Unlike the gradient descent algorithms widely used for deep learning, PSO does not require either differentiability or convexity (Boyd and Vandenberghe, 2004) of the objective function and constraints. Therefore, PSO is particularly useful when the objective function does not have desirable analytical properties (e.g., not differentiable).

PSO encodes swarm intelligence, such as bird flocking, into two simple dynamic equations

to solve optimization problems of the following form:

$$\min\ f(\boldsymbol{x}) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{S}\,,$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is a d-dimensional vector, $f(\boldsymbol{x})$ is a real-valued objective function (measurability is the only requirement), and $\mathcal{S} \subset \mathbb{R}^d$ is the search space or domain of $\boldsymbol{x}$). The algorithm starts with n candidate values of $\boldsymbol{x}$, denoted as $\boldsymbol{x}_1^0, \cdots, \boldsymbol{x}_n^0$. Each $\boldsymbol{x}_i^0$, $i = 1, \cdots n$, represents a *particle* and is initialized with a velocity vector $\boldsymbol{v}_i^0 \in \mathbb{R}^d$. Then for $i = 1, \cdots, n$, PSO iterates with the following two equations Bratton and Kennedy (2007):

$$\begin{aligned}
\boldsymbol{v}_i^{k+1} &= w\boldsymbol{v}_i^k + c_1 r_{i1}^k(\widehat{\boldsymbol{x}}_i^k - \boldsymbol{x}_i^k) + c_2 r_{i2}^k(\widehat{\boldsymbol{x}}^k - \boldsymbol{x}_i^k)\,,\\
\boldsymbol{x}_i^{k+1} &= \boldsymbol{x}_i^k + \boldsymbol{v}_i^{k+1}\,,
\end{aligned} \tag{1.3.1}$$

where $k = 0, 1, \cdots$ is the number of iterations finished, w is called the *inertia weight*, c_1 and c_2 are called the *cognitive* and *social* parameters respectively, and r_{i1}^k and r_{i2}^k are two random numbers independently generated uniformly from $[0, 1]$. Usually, w, c_1, and c_2 are set to numbers in $[0, 2]$ by users. Most importantly,

$$\begin{aligned}
\widehat{\boldsymbol{x}}_i^k &= \underset{\boldsymbol{x}\in\mathcal{A}_i}{\arg\min}\, f(\boldsymbol{x})\,,\\
\widehat{\boldsymbol{x}}^k &= \underset{\boldsymbol{x}\in\cup_{i=1}^n \mathcal{A}_i}{\arg\min}\, f(\boldsymbol{x})\,,\\
\text{where } \mathcal{A}_i &= \{\boldsymbol{x}_i^t : t = 0, \cdots, k\}\,.
\end{aligned}$$

Thus, $\widehat{\boldsymbol{x}}_i^k$ is the best position recorded by particle i up to the k^{th} iteration, and $\widehat{\boldsymbol{x}}^k$ is the best position recorded by the whole swarm up to the k^{th} iteration. The inertia weight w controls the level of a particle moving towards its last direction $\boldsymbol{v}_i^k$. The cognitive parameter c_1 represents how a particle is affected by its best known position $\widehat{\boldsymbol{x}}_i^k$. Similarly, the social parameter c_2 determines the influence of the swarm's best knowledge $\widehat{\boldsymbol{x}}^k$ on particle i. Because $\widehat{\boldsymbol{x}}^k$ is the best solution found by the whole swarm, the set of equations (1.3.1) is also called the *global best PSO* (Bratton and Kennedy, 2007).

To better understand the logic of PSO, suppose we have 10 ants starting around origin $(0, 0)$ and they are looking for food at point $(2, 2)$ (left panel of Fig. 1.1). Background colors represent distances to the food. The initial position of each ant corresponds to $\boldsymbol{x}_i^0$, and the objective function $f(\boldsymbol{x})$ is the Euclidean distance between point $\boldsymbol{x}$ and the food at $(2, 2)$. Each ant is initialized with an velocity vector $\boldsymbol{v}_i^0$ (blue arrow). After moving one step, ants re-analyze their positions and distances to the food so that (1) the best position of ant i is recorded as $\widehat{\boldsymbol{x}}_i^1$; (2) the best position of all ants is recorded as $\widehat{\boldsymbol{x}}^1$. Here the *best* position has the minimum distance to the food at (2,2). Then, each ant re-corrects its velocity according to equation (1.3.1) (middle panel of Fig. 1.1). After several iterations, all ants gather around the food and the velocity decreases to 0 gradually (right panel of Fig. 1.1). Finally, in section 2.4.6, we illustrate the premature convergence issue of PSO using a real-world dataset.

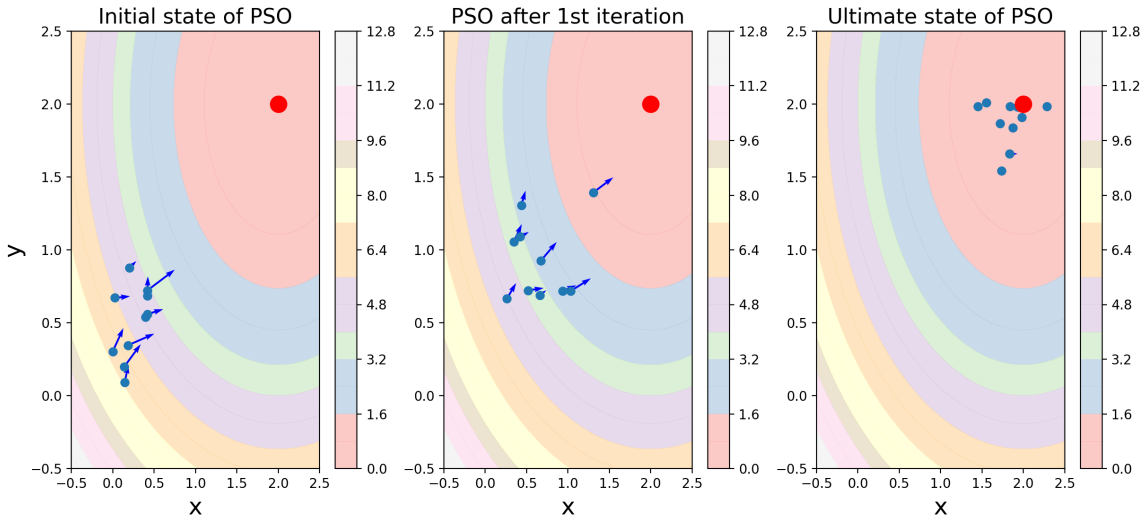


Figure 1.1: Illustration of PSO.

1.3.2 Applications of PSO in the Dissertation

PSO is one of the two dominating optimization tools throughout the whole dissertation and we briefly describe below how we apply PSO in solving problems in many different fields. Other optimization techniques used in this dissertation are Gaussian quadrature methods

(section 5.3.1), Newton-type algorithms (section 3.5), Metropolis-Hastings (MH) algorithms (section 4.3.2), Markov Chain Monte Carlo for nonparametric Bayesian statistics (section 7.3.4 and 7.3.6) and Expectation-Maximization algorithms (section 3.3 and 3.8).

In the next chapter, we demonstrate the usefulness of PSO by applying PSO in a novel way to gain insights into gene expression trends in single-cell RNA sequencing technologies (section 2.2.1). Throughout the whole chapter 3, PSO is compared with many other algorithms, either metaheuristics or gradient-based methods for tackling various interesting optimization problems across disciplines. PSO plays as a bottle opener in chapter 4 where we apply it to select SNPs as instrumental variables for downstream data analysis tasks (section 4.5.1). In chapter 5, we apply PSO to solve the optimal design for binary regression with Laplace link function, since the analytical solution is not available and the dimension of the optimization problem is unknown (section 5.3.1). Finally, PSO helps to find the proposed robust optimal designs in chapter 6 (section 6.4.2.2).

1.4 Competitive Swarm Optimizer with Mutated Agents

1.4.1 Competitive Swarm Optimizer

Competitive Swarm Optimizer (CSO) swarm-based algorithm proposed by Cheng and Jin (2015) and has proven its effectiveness for solving different types of optimization problems with various dimensions . For example, Gu et al. (2018) applied CSO to select variables for high-dimensional classification models, and Xiong and Shi (2018) used CSO to study a power system economic dispatch, which is typically a complex nonlinear multivariable strongly coupled optimization problem with equality and inequality constraints.

CSO minimizes a given continuous function $f(\mathbf{x})$ over a user-specified compact space $\mathbf{\Omega}$ by first randomly generating a set of candidate solutions. They take the form of a swarm of n particles at positions $\mathbf{x}_1, \cdots, \mathbf{x}_n$, along with their corresponding random velocities $\mathbf{v}_1, \cdots, \mathbf{v}_n$. For tackling design problems, each particle is a candidate design and upon convergence, the solution is the optimal design.

After the initial swarm is generated, at each iteration we randomly divide the swarm into $\lfloor \frac{n}{2} \rfloor$ pairs and compare their objective function values. At iteration t, we identify $\mathbf{x}_i^t$ as the winner and $\mathbf{x}_j^t$ as the loser if $f(\mathbf{x}_i^t) < f(\mathbf{x}_j^t)$. The winner retains the status quo and the loser learns from the winner. The two defining equations for CSO are

$$\mathbf{v}_j^{t+1} = \mathbf{R}_1 \otimes \mathbf{v}_j^t + \mathbf{R}_2 \otimes (\mathbf{x}_i^t - \mathbf{x}_j^t) + \phi \mathbf{R}_3 \otimes (\bar{\mathbf{x}}^t - \mathbf{x}_j^t) \tag{1.4.1}$$

$$\text{and } \mathbf{x}_j^{t+1} = \mathbf{x}_j^t + \mathbf{v}_j^{t+1}, \tag{1.4.2}$$

where $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$ are all random vectors whose elements are drawn from $U(0,1)$. The operation $\otimes$ represents element-wise multiplication and the vector $\bar{\mathbf{x}}^t$ is the swarm center at iteration t. The social factor ϕ controls the influence of the neighboring particles to the loser and a large value is helpful for enhancing swarm diversity (but possibly impacts convergence rate). This process iterates until a pre-specified stopping criterion or criteria are met.

Simulation results have repeatedly shown that CSO either outperforms or is competitive with several state-of-the-art evolutionary and swarm based algorithms, including several enhanced versions of PSO. This conclusion was arrived at after comparing CSO performance with state-of-the-art EAs using a variety of benchmark functions with dimensions up to 5000 and found that CSO was frequently the fastest and with the best quality results Cheng and Jin (2015); Mohapatra et al. (2017); Sun et al. (2016); Zhang et al. (2016).

1.4.2 Mutated Agents

Zhang et al. (2017) proposed an improvement on CSO and call the enhanced version, Competitive Swarm Optimizer with Mutated Agents (CSO-MA). After pairing up the swarm in groups of two at each iteration, the variant randomly chooses a loser particle p as an agent, randomly picks a variable indexed as q and then randomly changes the value of $\mathbf{x}_{pq}$ to either $\mathbf{xmax}_q$ or $\mathbf{xmin}_q$, where $\mathbf{xmax}_q$ and $\mathbf{xmin}_q$ represent, respectively, the upper bound and lower bound of the q-th variable. If the current optimal value is already close to the global

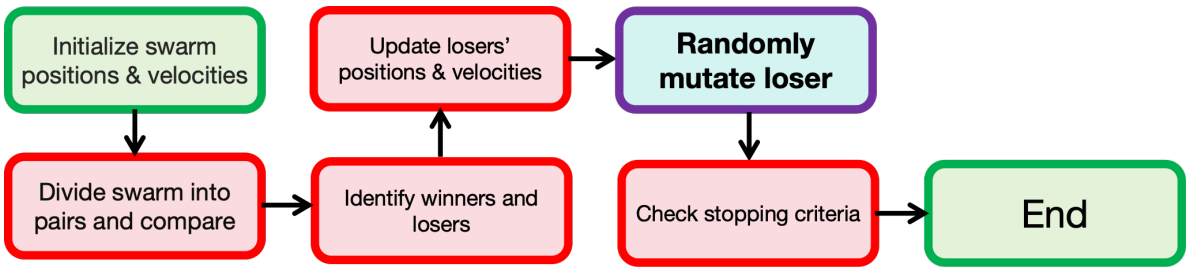


Figure 1.2: Flowchart of CSO-MA.

optimum, this change will not hurt since we implement this experiment on a loser particle, which is not leading the movement for the whole swarm; otherwise, this chosen agent restarts a journey from the boundary and has a chance to escape from a local optimum. We present the flowchart of CSO-MA in Figure 1.2. The mutation step (the box in purple) is a key feature of the CSO-MA that differentiates it from the standard CSO. The mutation is intended to increase the diversity of the solutions and prevent premature convergence to a local optimum by allowing particles to explore more distant regions of the search space.

Let n be the swarm size and let D be the dimension of the problem. The computational complexity of CSO is $\mathcal{O}(nD)$ and since our modification only adds one coordinate mutation operation to each particle, its computational complexity is the same as that of CSO. The improved performance of CSO-MA over CSO for finding optimal designs for many complex multi-dimensional benchmark functions has been validated in Zhang et al. (2020); **?**.

1.4.3 Applications of CSO-MA in the Dissertation

CSO-MA is the dominating optimization tool throughout the whole chapter 3, and is compared with many other algorithms, either metaheuristics or gradient-based methods for tackling various interesting optimization problems across disciplines. CSO-MA also plays a crucial role in finding optimal designs for toxicological experiments and other types of design problems (section 3.9 and 6.4).

CHAPTER 2

Single-cell Generalized Trend Model (scGTM)

2.1 Background and Existing Work

In recent years, single-cell RNA-sequencing (scRNA-seq) technologies are booming and have provided many valuable insights into complex biological systems, ranging from cancer genomics to diverse viral and bacterial evolution (Saliba et al., 2014). Tons of newly developed and advanced computer science, topological and statistical methods have been applied to bioinformatics and scRNA-seq data analysis including protein structure prediction, differential expression analysis and time series analysis of gene expression data (Li and Li, 2018; Rabadán and Blumberg, 2019).

Pseudotime analysis is one of the most important topics in single-cell transcriptomics. There has been fruitful work on inferring cell pseudotime (Magwene et al., 2003; Bendall et al., 2014; Trapnell et al., 2014; Shin et al., 2015; Ji and Ji, 2016; Qiu et al., 2017; Street et al., 2018; Cao et al., 2019; Mondal et al., 2021) and constructing statistical models for gene expression along the inferred cell pseudotime (Campbell and Yau, 2017; Bacher et al., 2018; Van den Berge et al., 2020; Ren and Kuan, 2020; Song and Li, 2021). Informative trends of gene expression along cell pseudotime may reflect molecular signatures in biological processes. For instance, if a gene shows a *hill-shaped* (first-upward-then-downward; Fig. 2.1b) or *valley-shaped* (first-downward-then-upward; Fig. 2.1c) trend, the hill or valley position may indicate the occurrence of some biological event. Hence, it is of great interest to have a statistical model for characterizing hill- and valley-shaped gene expression trends along cell pseudotime.

Two types of statistical methods have been developed to model the relationship between a gene's expression in a cell (or a sample) and the cell pseudotime (or the sample's physical

time, whose modeling is similar from a statistical perspective). Methods of the first type are based on statistical regression models, such as the generalized linear model (GLM) and generalized additive model (GAM), whose parameters do not have direct relevance to gene expression dynamics. Specifically, the GLM used in the Monocle3 method (Cao et al., 2019) assumes that a gene's log-transformed expected expression in a cell is a linear function of the cell pseudotime, making it unable to capture hill- and valley-shaped trends that linear trends cannot approximate well. To avoid this issue, most methods use nonparametric regression models, such as the GAM and piece-wise linear models, to capture complex trends. To name a few, Storey et al. (2005) applied basis regression; Trapnell et al. (2014) considered the GAM with the Tobit likelihood; Ren and Kuan (2020) applied the GAM with Bayesian shrinkage dispersion estimates; Van den Berge et al. (2020) proposed tradeSeq using the spline-based GAM; the more recent PseudotimeDE method by Song and Li (2021), which fixes the p-value calibration issue in tradeSeq, also uses the spline-based GAM with spline functions; Bacher et al. (2018) used a piecewise linear model, which is more restrictive than the GAM. Although these nonparametric regression methods can fit complex gene expression trends, they are prone to over-fitting if without proper hyper-parameter tuning (as we will show in Section 3), and their parameters do not directly inform the shape of a trend (e.g., hill-shaped) or carry biological meanings.

Unlike the first type, methods of the second type use models with direct relevance to gene expression dynamics, and notable methods include ImpulseDE/ImpulseDE2 (Chechik and Koller, 2009; Sander et al., 2017; Fischer et al., 2018) and switchDE (Campbell and Yau, 2017). Specifically, ImpulseDE2 estimates a gene expression trend as a double-logistic curve so it can capture non-monotone trends; however, its parameters, though having biological interpretations, do not intuitively inform the shape of a trend. In contrast, switchDE has a restrictive model and can find only monotone trends, although its parameters directly inform the shape of a trend (e.g., a gene's activation time).

With the above summary, we find a gap in the existing methods: no method can capture monotone, hill-shaped, and valley-shaped trends with biologically interpretable and trend-informative parameters. Hence, we propose the single-cell generalized trend model

(scGTM) that has three advantages over existing methods and models: (i) capturing hill- and valley-shaped trends in addition to monotone trends, (2) estimating interpretable and trend-informative parameters, and (3) allowing flexible modeling of count data. Fig. 2.1a illustrates the scGTM's four parameters for a hill-shaped trend (a valley-shaped trend has four similar parameters; a monotone increasing trend is a special case of a hill-shaped trend with the increasing part only): the maximum log expected expression μ_{mag}, the activation strength k_1, the repression strength k_2, and the change time t_0 where the expected expression stops increasing. Fig. 2.1b–c show how the scGTM fits to the two example genes (*Tmsb10* in the GYRUS dataset and *NFKBIA* in the LPS dataset; Supplementary Table S1) and reveals their hill- and valley-shaped trends.

To estimate the scGTM parameters, we apply the particle swarm optimization (PSO) algorithm (Section 2.2) for constrained maximum likelihood estimation (MLE). PSO has several advantages that make it suitable for our optimization problem: (i) it does not require the convexity and differentiability of the objective function; (ii) it can handle boundary constraints and discrete parameters without re-formulating the objective function, (iii) unlike the Newton-type algorithms used in (Trapnell et al., 2014; Wood, 2017; Campbell and Yau, 2017), it is gradient-free and thus easy to implement.

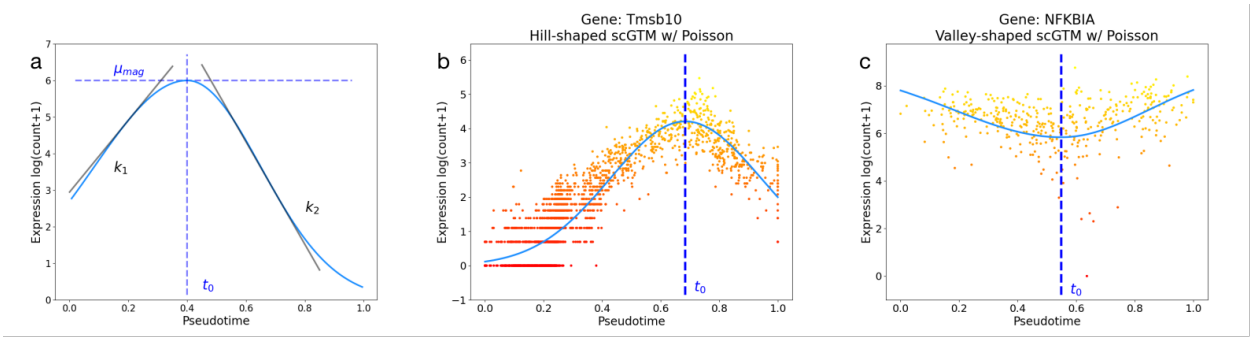


Figure 2.1: Illustration of the scGTM. (**a**) Four parameters of the scGTM in Equation (2.2.2) for a hill-shaped trend: the maximum log expected expression μ_{mag} (horizontal blue line), the activation strength k_1 (absolute value of the left tangent line's slope), the repression strength k_2 (absolute value of the right tangent line's slope), and the change time t_0 (vertical blue line). (**b**) A hill-shaped trend of gene *Tmsb10* (in the GYRUS dataset) fitted by the scGTM with counts modeled by the Poisson distribution. (**c**) A valley-shaped trend of gene *NFKBIA* (in the LPS dataset) fitted by the scGTM with counts modeled by the Poisson distribution. In b–c, the scatter points indicate gene expression levels, and the curves are the trends fit by the scGTM.

2.2 scGTM Formulation and Estimation

Let $\mathbf{Y} = (y_{gc})$ be an observed $G \times C$ gene expression count matrix, where G is the number of genes, C is the number of cells (i.e., the number of pseudotime values), and y_{gc} is the (g, c)-th element indicating the observed expression count of gene $g = 1, \ldots, G$ in cell $c = 1, \ldots, C$. We consider gene expression counts as random variables whose randomness comes from experimental measurement uncertainty, so y_{gc} is a realization of the random count variable Y_{gc}. Given a particular gene g, for notation simplicity, we drop the subscript g and denote Y_{gc} as Y_c and y_{gc} as y_c. We denote by t_c the inferred pseudotime of cell c. In the scGTM, $t_1, \ldots, t_C$ are treated as fixed values of pseudotime and serve as the covariate vector of interest.

Given t_c, the scGTM can model the count variable Y_c using four count distributions commonly used for gene expression data: the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) distributions.

For a hill-shaped gene, the scGTM is

$$Y_c \overset{\text{ind}}{\sim} F\left(\tau_c, \phi, p_c\right), \quad c = 1, \ldots, C, \tag{2.2.1}$$

$$\log(\tau_c + 1) = \begin{cases} \mu_{\text{mag}} \exp\left(-k_1 (t_c - t_0)^2\right) & \text{if } t_c \leq t_0 \\ \mu_{\text{mag}} \exp\left(-k_2 (t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}, \tag{2.2.2}$$

$$\log\left(\frac{p_c}{1 - p_c}\right) = \alpha \log(\tau_c + 1) + \beta, \tag{2.2.3}$$

where $F\left(\tau_c, \phi, p_c\right)$ in (2.2.1) represents one of the four common count distributions. The most general case is when $F\left(\tau_c, \phi, p_c\right) = \text{ZINB}\left(\tau_c, \phi, p_c\right)$ with mean parameter $\tau_c \geq 0$, dispersion parameter $\phi \in \mathbb{Z}_+ := \{1, 2, 3, \cdots\}$ and zero-inflated parameter $p_c \in [0, 1]$. As special cases, $F\left(\tau_c, \phi, 0\right) = \text{NB}\left(\tau_c, \phi\right)$, $F\left(\tau_c, \infty, p_c\right) = \text{ZIP}\left(\tau_c, p_c\right)$, and $F\left(\tau_c, \infty, 0\right) = \text{Poisson}\left(\tau_c\right)$.

Fig. 2.1a displays and shows the roles of the 4 parameters in 2.2.2 for modelling a hill-shaped trend. For a valley-shaped trend, there are four similar parameters and we note

that a monotone increasing trend is a special case of a hill-shaped trend with the increasing part only. The four parameters in the Fig. 2.1a are the maximum log expected expression μ_{mag}, the activation strength k_1, the repression strength k_2, and the change time t_0 where the expected expression stops increasing. Fig. 2.1b–c show the scGTM fits to the gene (*Tmsb10* in the GYRUS data set and another gene *NFKBIA* in the LPS data set from the Supplementary Table S1). Their trends reveal a hill- and valley-shaped trend., respectively.

In this hill-shaped scGTM, we assume that the gene's expression count Y_c in cell c has mean parameter τ_c and zero-inflation parameter p_c, and both depend on the pseudotime t_c of cell c. In (2.2.2), we link τ_c to t_c by assuming that $\log(\tau_c + 1)$ is a non-negative transformation that compresses extremely large values of τ_c using a two-part Gaussian function corresponding to $t_c \leq t_0$ and $t_c > t_0$; we choose the Gaussian function for its good mathematical properties and interpretability. We link p_c to t_c in (2.2.3) using a logistic regression, with predictor $\log(\tau_c+1)$, i.e., the logistic transformation of p_c is a linear function of $\log(\tau_c + 1)$ (with slope α and intercept β) and thus a function of t_c.

Besides $\phi \in \mathbb{Z}_+$ and $\alpha, \beta \in \mathbb{R}$, the following parameters of the hill-shaped scGTM shown in Fig. 2.1a need to be estimated for biological interpretations:

- $\mu_{\text{mag}} \geq 0$: magnitude of the hill,
  i.e., $\mu_{\text{mag}} = \max_{c\in\{1,\ldots,C\}} \log(\tau_c + 1)$;
- $k_1 \geq 0$: activation strength (how fast the gene is up-regulated);
- $k_2 \geq 0$: repression strength (how fast the gene is down-regulated);
- $t_0 \in [0, 1]$: change time (where the gene reaches the maximum expected expression).

For a valley-shaped gene, the scGTM is the same except that we replace (2.2.2) by

$$\log(\tau_c + 1) = \begin{cases} b - \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) & \text{if } t_c \leq t_0 \\ b - \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}, \tag{2.2.4}$$

where b indicates the baseline (maximum) log-transformed
(expected expression + 1) of the valley-shaped gene. The interpretation of the four key parameters of the valley-shaped scGTM becomes

- $\mu_{\text{mag}} \in [0, b]$: magnitude of the valley, i.e., $b - \mu_{\text{mag}} = \min_{c\in\{1,\ldots,C\}} \log(\tau_c + 1)$;
- $k_1 \geq 0$: activation strength (how fast the gene is up-regulated);
- $k_2 \geq 0$: repression strength (how fast the gene is down-regulated);
- $t_0 \in [0, 1]$: change time (where the gene reaches the minimum expected expression).

Compared to the hill-shaped scGTM, the valley-shaped scGTM has an additional baseline parameter b that needs to be estimated. For simplicity, we estimate b by a plug-in estimator $\hat{b} = \max_{c\in\{1,\ldots,C\}} \log(y_c + 1)$, where $y_1, \ldots, y_C$ are the observed counts of a valley-shaped gene. For the common parameters of the hill- and valley-shaped scGTMs in Section 2.2.1, we next discuss how PSO can provide constrained likelihood estimates for these parameters.

2.2.1 Constrained MLE and the PSO Algorithm

To fit the scGTM to a gene, we first need to ascertain whether the gene is hill- or valley-shaped: we call the gene valley-shaped only if its expression count y_c is negatively correlated with $t_c \in [0, 0.5]$ and positively correlated with $t_c \in [0.5, 1]$; otherwise, we consider the gene hill-shaped. Next, based on the trend shape, we estimate the scGTM parameters. For a hill-shaped gene, we estimate the scGTM parameters $\Theta = (\mu_{\text{mag}}, k_1, k_2, t_0, \phi, \alpha, \beta)^\mathsf{T}$ from the observed expression counts $\boldsymbol{y} = (y_1, \ldots, y_C)^\mathsf{T}$ and cell pseudotimes $\boldsymbol{t} = (t_1, \ldots, t_C)^\mathsf{T}$ using the constrained maximum likelihood method, which respects each parameter's range and ensures the estimation stability. Let $\log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t})$ be the log likelihood function and the

optimization problem is:

$$\begin{aligned} & \max_{\Theta} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) \\ \text{such that } & \min_{c\in\{1,\ldots,C\}} \log(y_c+1) \le \mu_{\text{mag}} \le \max_{c\in\{1,\ldots,C\}} \log(y_c+1)\,, \\ & k_1, k_2 \ge 0\,, \\ & \min_{c\in\{1,\ldots,C\}} t_c \le t_0 \le \max_{c\in\{1,\ldots,C\}} t_c\,, \\ & \phi \in \mathbb{Z}_+\,, \end{aligned} \tag{2.2.5}$$

where

$$\begin{aligned} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) &= \log\left[\prod_{c=1}^{C} \mathbb{P}(Y_c = y_c \mid t_c)\right] \\ &= \sum_{c=1}^{C} \log\Big[(1-p_c) f(y_c|t_c) + p_c\, \mathbb{I}(y_c = 0)\Big] \end{aligned} \tag{2.2.6}$$

and

$$f(y_c|t_c) = \frac{\tau_c^{y_c}}{y_c!} \frac{\Gamma(\phi + y_c)}{\Gamma(\phi)(\phi+\tau_c)^{y_c}} \frac{1}{\left(1+\frac{\tau_c}{\phi}\right)^{\phi}}\,,$$

which can be further specified as a function of Θ based on (2.2.2) and (2.2.3).

For a valley-shaped gene, the constrained MLE problem for estimating parameters in the scGTM is similar and we omit the discussion for space consideration.

There are two difficulties in the optimization problem (2.2.5). First, the likelihood function (2.2.6) is neither convex nor concave. Second, the constraint is linear in μ_{mag}, k_1, k_2, and t_0 but ϕ is a positive integer-valued variable. Hence, conventional optimization algorithms such as P-IRLS in GAM (Wood, 2011, 2017) and L-BFGS in switchDE (Van Loan and Golub, 1996; Campbell and Yau, 2017) are difficult to apply in this case. Metaheuristics is a class of assumptions-free general purpose optimization algorithms that is widely and increasingly used to tackle challenging and high-dimensional optimization problems in the quantitative sciences (Whitacre, 2011a,b; Yang, 2017). PSO is an exemplary and popular

member of this class and has been shown to be effective to solve various types of optimization problems and (Korani and Mouhoub, 2021) provides a recent review of such alogirthms and applications across various disciplines.

PSO first generates a swarm of candidate solutions (known as particles) to the optimization problem (2.2.5). At each iteration, particles change their positions within the constraints, and the algorithm finds the best solution among all particle trajectories. We summarize the vanilla PSO algorithm Bratton and Kennedy (2007) for the constrained MLE of the scGTM in Algorithm 1, and provide further details of PSO in the Supplementary Information. The below algorithm applies PSO to solve our optimization problem (2.2.5).

2.2.2 Approximate Confidence Intervals of the Four Key Parameters in the scGTM

The estimated parameters $\widehat{\Theta} = (\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0, \hat{\phi}, \hat{\alpha}, \hat{\beta})^{\mathsf{T}}$ are next used to construct approximate confidence intervals for μ_{mag}, k_1, k_2, and t_0 using the maximum likelihood theory. Specifically, we calculate the plug-in asymptotic covariance matrix of $(\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0)^{\mathsf{T}}$ as the inverse of the partial Fisher information matrix of the four parameters evaluated at $(\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0)^{\mathsf{T}}$ (detailed derivation in the Supplementary Information). Then we use the diagonal elements of this matrix as the plug-in asymptotic variances of $\hat{\mu}_{\text{mag}}$, $\hat{k}_1$, $\hat{k}_2$, and $\hat{t}_0$, and denote them by $\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})$, $\widehat{\text{Var}}(\hat{k}_1)$, $\widehat{\text{Var}}(\hat{k}_2)$, and $\widehat{\text{Var}}(\hat{t}_0)$, respectively. We then obtain a 95% approximate confidence interval for each of the parameters: $[\hat{\mu}_{\text{mag}}^{\text{lb}}, \hat{\mu}_{\text{mag}}^{\text{ub}}]$, $[\hat{k}_1^{\text{lb}}, \hat{k}_1^{\text{ub}}]$, $[\hat{k}_2^{\text{lb}}, \hat{k}_2^{\text{ub}}]$, and $[\hat{t}_0^{\text{lb}}, \hat{t}_0^{\text{ub}}]$, where

$$\hat{\mu}_{\text{mag}}^{\text{lb}} = \max\left(0,\ \hat{\mu}_{\text{mag}} - 1.96\sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})}\right), \qquad \hat{\mu}_{\text{mag}}^{\text{ub}} = \hat{\mu}_{\text{mag}} + 1.96\sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})},$$

$$\hat{k}_1^{\text{lb}} = \max\left(0,\ \hat{k}_1 - 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_1)}\right), \qquad \hat{k}_1^{\text{ub}} = \hat{k}_1 + 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_1)},$$

$$\hat{k}_2^{\text{lb}} = \max\left(0,\ \hat{k}_2 - 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_2)}\right), \qquad \hat{k}_2^{\text{ub}} = \hat{k}_2 + 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_2)},$$

$$\hat{t}_0^{\text{lb}} = \max\left(0,\ \hat{t}_0 - 1.96\sqrt{\widehat{\text{Var}}(\hat{t}_0)}\right), \qquad \hat{t}_0^{\text{ub}} = \min\left(\hat{t}_0 + 1.96\sqrt{\widehat{\text{Var}}(\hat{t}_0)},\ 1\right).$$

Algorithm 1 PSO for the constrained MLE for the scGTM

Input data: a gene's expression counts and cell pseudotime values

$\boldsymbol{y}$: a $C \times 1$ gene expression count vector;

$\boldsymbol{t}$: a $C \times 1$ cell pseudotime vector;

Input parameters:

F: count distribution: Poisson, NB, ZIP, or ZINB;

H: number of iterations in PSO; set to $H = 100$ by default;

w, c_1, and c_2: hyperparameters of PSO; set to $w = 0.9$, $c_1 = 1.2$, and $c_2 = 0.3$ by default;

Algorithm:

1. Randomly initialize Θ with B particles: $\Theta_1^0, \Theta_2^0, \ldots, \Theta_B^0$;
2. Randomly initialize velocity vectors for the B particles: $\boldsymbol{v}_1^0, \boldsymbol{v}_2^0, \cdots, \boldsymbol{v}_B^0$;
3. For $h = 0$ to H:

(i) Update the best solution of each particle i

$$\widehat{\Theta}_i^h = \underset{\Theta \in \mathcal{A}_i^h}{\arg\max} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) \,,$$

where $\mathcal{A}_i^h = \{\Theta_i^k : k = 0, \cdots, h\}$, $i = 1, \ldots, B$;

(ii) Update the global best solution

$$\widehat{\Theta}^h = \underset{\Theta \in \cup_{i=1}^B \mathcal{A}_i^h}{\arg\max} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) \,;$$

(iii) Update velocity of each particle i

$$\boldsymbol{v}_i^{h+1} = w\boldsymbol{v}_i^h + c_1 r_{i1}^h \left(\widehat{\Theta}_i^h - \Theta_i^h\right) + c_2 r_{i2}^h \left(\widehat{\Theta}^h - \Theta_i^h\right) \,,$$

where r_{i1}^h and r_{i2}^h are independently generated from Unif$(0, 1)$, $i = 1 \ldots, B$;

(iv) Update each particle i

$$\Theta_i^{h+1} = \Theta_i^h + \boldsymbol{v}_i^h \,, \quad i = 1, \ldots, B \,;$$

4. Set $\widehat{\Theta} = \widehat{\Theta}^H$;
5. Calculate 95% approximate confidence intervals of key parameters based on $\widehat{\Theta}$ (Section 2.2.2).

Output:

$-\log L(\widehat{\Theta} \mid \boldsymbol{y}, \boldsymbol{t})$: fitted negative log likelihood value;

$\widehat{\Theta} = (\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0, \hat{\phi}, \hat{\alpha}, \hat{\beta})^\intercal$: estimated parameters;

$[\hat{\mu}_{\text{mag}}^{\text{lb}}, \hat{\mu}_{\text{mag}}^{\text{ub}}]$, $[\hat{k}_1^{\text{lb}}, \hat{k}_1^{\text{ub}}]$, $[\hat{k}_2^{\text{lb}}, \hat{k}_2^{\text{ub}}]$, and $[\hat{t}_0^{\text{lb}}, \hat{t}_0^{\text{ub}}]$: 95% approximate confidence intervals.

2.3 Applications of scGTM

2.3.1 scGTM Outperforms GAM, GLM, LOESS, switchDE, and ImpulseDE2 in Capturing Informative and Interpretable Trends

As an example, we use the *MAOA* gene in the WANG dataset (Wang et al., 2020b) (Supplementary Table S1) to compare the fitted trends of the scGTM, GAM, GLM, LOESS, switchDE, and ImpulseDE2. In the original study, the gene was reported to have a hill-shaped trend. Our comparison results have several interesting observations. First, we show that the scGTM provides more informative and interpretable gene expression trends than the GAM and GLM do when the count outcome comes from the Poisson, ZIP, NB, and ZINB distributions. Fig. 2.2a shows that the scGTM robustly captures the hill-shaped trends for the four distributions and consistently estimates the change time around 0.75, which is where the *MAOA* gene reaches its expected maximum expression. While the GAM also estimates the maximum expression around 0.75, its estimated trends are much more complex. This is likely due to possible overfitting (despite the use of penalization) and consequently, more difficult to interpret them than the scGTM trends (Fig. 2.2b). Unlike the scGTM and GAM, the GLM only allows for capturing monotone trends, making it unable to detect the possible existence of expression change time (Fig. 2.2c). Second, we compare the scGTM with the two existing methods, switchDE and ImpulseDE2, that use models with direct relevance to gene expression dynamics. Although switchDE estimates the activation time around 0.75, similar to the scGTM's estimate change time, it cannot capture the downward expression trend as the cell pseudotime approaches 1.00 due to its monotone nature (Fig. 2.2d). ImpulseDE2 can theoretically capture a hill-shaped trend, but it only fits a monotone increasing trend for the *MAOA* gene (Fig. 2.2e). A likely reason is that the method was designed for time-course bulk RNA-seq data. Third, we compare the scGTM with the LOESS method commonly used for exploratory data analysis. While LOESS outputs a reasonable, though less smooth trend (Fig. 2.2f), it is not probability-based and thus does not have a likelihood. Hence, LOESS does not allow likelihood-based model selection, a functionality of the scGTM. To

summarize, the scGTM fits outperform those from GAM, GLM, LOESS, switchDE and ImpulseDE2 by providing a more informative and interpretable trend with less concern on model overfitting.

In addition to the *MAOA* gene, Wang et al. (2020b) reported 19 other exemplary genes that define menstrual cycle phases and exhibit hill-shaped expression trends along the cell pseudotime. Supplementary Figs. S1-S19 compare the various model fits for the 19 genes and and we observe that the scGTM consistently provides more informative, interpretable trends than the other models.

Besides visually inspecting the fitted expression trends, we compare the AIC values of the scGTM, GAM, and GLM used with the four count distributions fitted to the aforementioned 20 genes. Note that a lower AIC value indicates a model's better goodness-of-fit with the model complexity penalized. Supplementary Fig. S20 shows that the scGTM has comparable or even lower AIC values than the GAM's AIC values, confirming that the scGTM fits well to data despite its much simpler model than GAM's. Based on Fig. 2.2 and Supplementary Figs. S1–S20, we use the scGTM with the Poisson distribution in the following applications for its goodness-of-fit and model simplicity. This choice is consistent with previous research on modeling sequencing data (Silverman et al., 2020) and other count data (Warton, 2005; Campbell, 2021).

2.3.2 scGTM Recapitulates Gene Expression Trends of Endometrial Transformation in the Human Menstrual Cycle

The WANG dataset contains 20 exemplar genes that exhibit temporal expression trends in unciliated epithelia cells in the human menstrual cycle Wang et al. (2020b). The original study also ordered the 20 genes by the estimated pseudotime at which they achieved the maximum expression (Fig. 2.3a; genes ordered from top to bottom), and it was found that the ordering agreed well with the menstrual cycle phases (Fig. 2.3a; the top bar indicates the phases). Comparing the fitted expression trends of the 20 genes by the scGTM, switchDE, and ImpulseDE2, we observe that only the scGTM trends agree well with the data (Fig. 2.3).

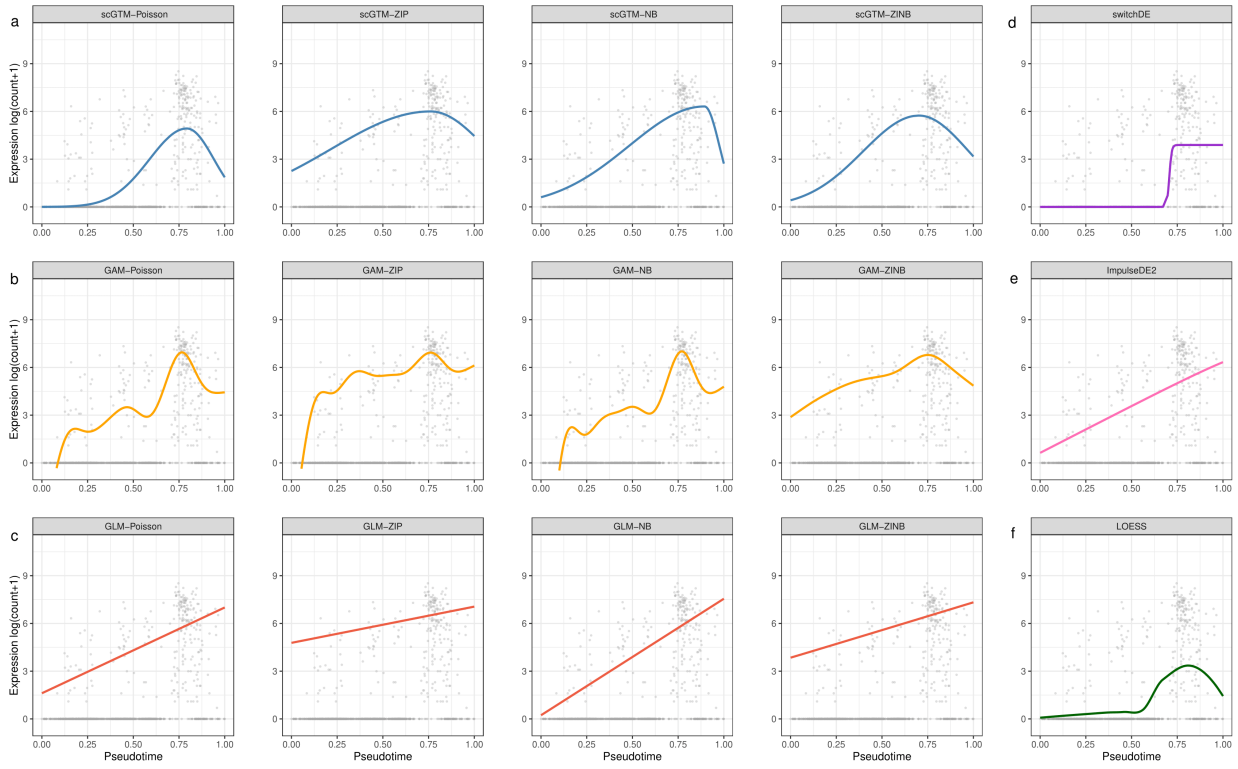


Figure 2.2: Comparison of the scGTM with GAM, GLM, LOESS, switchDE, and ImpulseDE2 for fitting the expression trend of gene *MAOA* in the WANG dataset (Wang et al., 2020b) (Supplementary Table S1). In the first four columns, the three rows correspond to (a) scGTM, (b) GAM, and (c) GLM. From left to right, the first four columns correspond to Poisson, ZIP, NB, and ZINB as the count distribution used in the scGTM, GAM, and GLM. The fifth column corresponds to (d) switchDE, (e) ImpulseDE2 and (f) LOESS. Each panel shows the same scatterplot of gene *MAOA*'s log-transformed expression counts vs. cell pseudotime values, as well as a model's fitted trend. With all four count distributions, the scGTM robustly captures the gene expression trend and estimates the change time around 0.75. In contrast, GLM, switchDE and ImpulseDE2 only describe the trend as increasing; GAM overfits the data and does not output trends as interpretable as the scGTM does; LOESS outputs a reasonable trend, but it does not allow likelihood-based model selection like the scGTM.

Additionally, we evaluate the 20 genes' estimated change times (i.e., t_0) by the scGTM and their estimated activation times by the switchDE. Although the change times and estimation times are both expected to correlate well with the gene ordering in the original study, only the change times estimated by the scGTM fulfills this expectation (Fig. 2.3b–c). Compared with the scGTM, switchDE miscalculates the activation times for many hill-shaped genes whose maximum expression occurs in the middle of the cycle; this is likely due to the fact that switchDE can only capture monotone trends (Fig. 2.3c). Similarly, ImpulseDE2 cannot well capture the trends of those hill-shaped genes (Fig. 2.3d). Unlike switchDE and ImpulseDE2, the scGTM estimates the change times reasonably for almost all genes. For instance, the *GPX3* gene has an estimated change time at 0.88, consistent with its role as a secretory middle/late phase marker gene Wang et al. (2020b).

Besides the 20 exemplar genes, we apply the scGTM, switchDE, and ImpulseDE2 to fit the expression trends of all 1,382 menstrual cycle genes reported in Wang et al. (2020b). Supplementary Fig. S28 shows that the scGTM still outperforms switchDE and ImpulseDE2 for capturing these genes' expression trends. In summary, the scGTM provides useful summaries for gene expression trends in the human menstrual cycle.

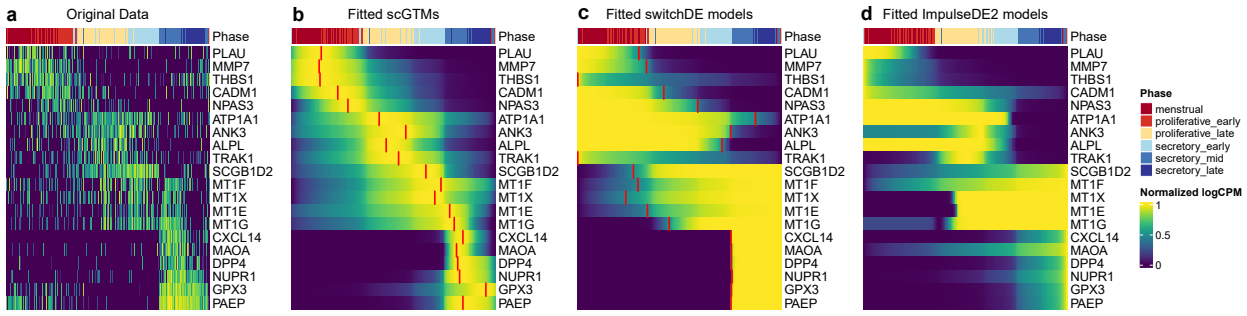


Figure 2.3: Fitted expression trends by the scGTM, switchDE, and ImpulseDE2 for 20 exemplar genes in the WANG dataset (Wang et al., 2020b) (Supplementary Table S1). All panels are ordered by cell pseudotime values from 0 (left) to 1 (right). The top color bars show the endometrial phases defined in the original study. (a) The original expression values along pseudotime. (b) The fitted trends of the scGTM, with the red segments highlighting the estimated change times t_0. (c) The fitted trends of switchDE, with the red segments highlighting the estimated activation times. (d) The fitted trends of ImpulseDE2.

2.3.3 scGTM Identifies Informative Gene Expression Trends after Immune Cell Stimulation

As the second real data application, we use the scGTM to categorize gene expression trends in mouse dendritic cells (DCs) after stimulation with lipopolysaccharide (LPS, a component of gram-negative bacteria) Shalek et al. (2014). First, we apply the likelihood ratio tests to screen the genes that the scGTM fits significantly better than the null Poisson model (in which τ_c and p_c in (2.2.1) do not depend on cell pseudotime t_c). Assuming that the likelihood ratio statistic of every gene follows χ^2_3 as the null distribution, we retain 2405 genes whose Benjamini-Hochberg (BH) adjusted p-values ≤ 0.01.

Second, we use the scGTM's confidence levels of the three parameters t_0, k_1, and k_2 to categorize the 2405 genes into three types: (1) *hill-shaped & mostly increasing genes*: $t_0^{\text{lb}} > 0.5 + 0.1$ (change time occurs at late pseudotime) and $k_1^{\text{lb}} > 1$ (strong activation strength), (2) *hill-shaped & mostly decreasing genes*: $t_0^{\text{ub}} < 0.5 - 0.1$ (change time occurs at early pseudotime) and $k_2^{\text{lb}} > 1$ (strong repression strength), and (3) *valley-shaped genes*. To demonstrate that this categorization is biologically meaningful, we perform gene ontology (GO) analysis on the three gene types and compare the enriched GO terms. Fig. 2.4a shows that the three gene types are enriched with largely unique GO terms, verifying their functional differences. Notably, the hill-shaped & mostly increasing genes are related to immune response processes, showing consistency between their expression trends (activation after the LPS stimulation) and functions (immune response). Further, we visualize 5 illustrative genes from each gene type (Fig. 2.4b) and observe that the scGTM's fitted trends agree well with the data. In conclusion, the scGTM can help users discern genes with specific trends by its trend-informative parameters.

Besides the above three real data applications, we conduct a simulation study (section 2.4) to verify the robustness of the scGTM to gene expression trends not generated from the scGTM assumptions. The simulation results also show that, beyond good interpretability, the scGTM is flexible enough to fit various trends to a similar extent as the GAM does (Supplementary Information S3). Moreover, we use a bootstrap analysis to show that the

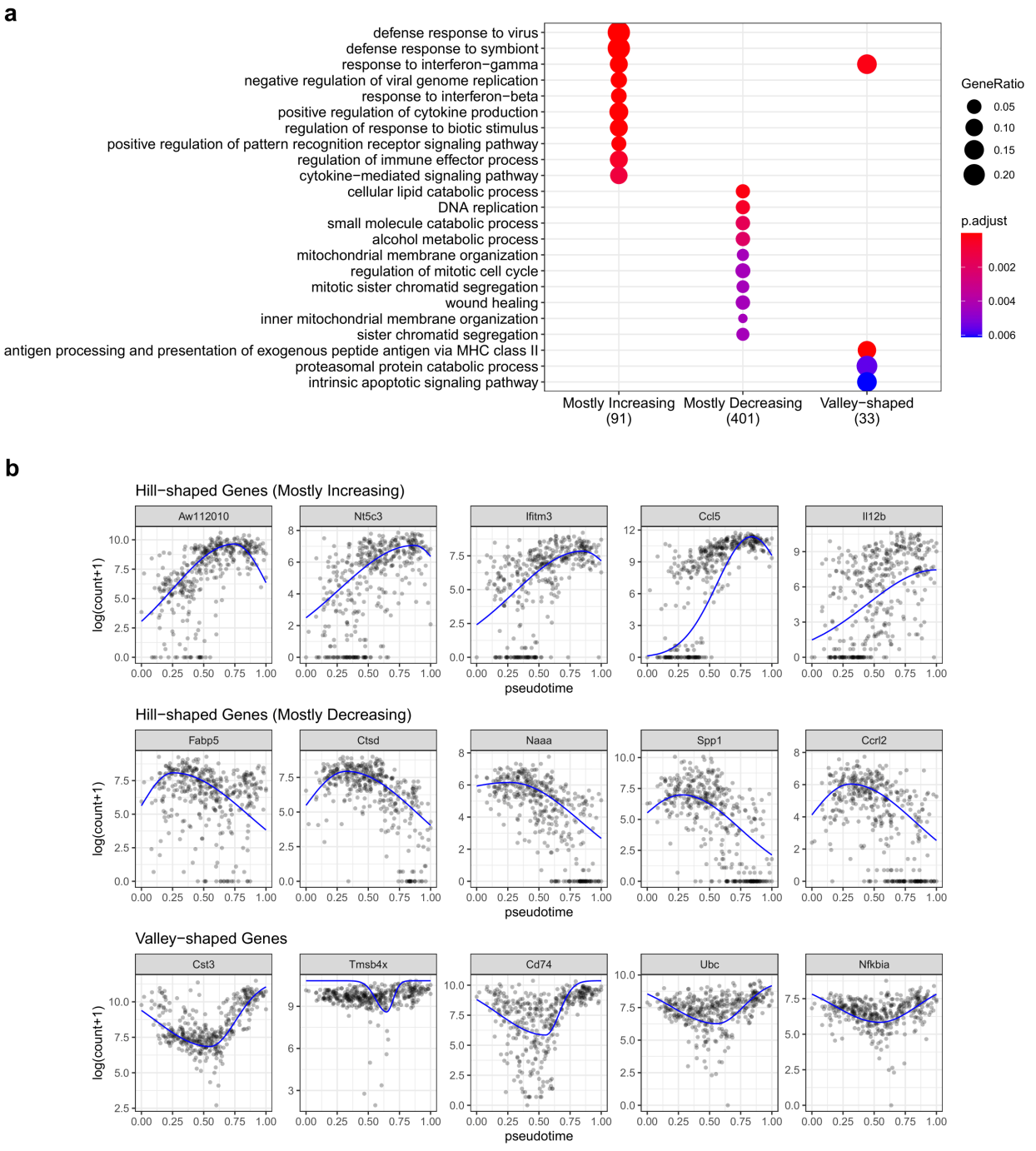


Figure 2.4: Three types of gene expression trends characterized by the scGTM parameters in the LPS dataset (Supplementary Table S1). (a) GO enrichment analysis of the three gene types. The top enriched GO terms are different among the three gene types. Notably, the hill-shaped & mostly increasing genes (1st column) are functionally related to immune responses. (b) Visualization of example genes in the three types. The scatter plots show gene expression data; the trends estimated by the scGTM (blue curves) well match the data.

fitted scGTM trend has a smaller variance than the fitted GAM trend does (Supplementary Information S3), at the cost of a larger bias.

2.4 Robustness and Sensitivity Analysis of scGTM

2.4.1 scGTM outperforms GAM and GLM in balancing goodness-of-fit and model complexity

We compare scGTM with GLM and GAM in terms of relative AIC. The relative AIC is calculated as follows: we first compute the AIC of all models for a particular gene and then divide all AIC values by the minimum AIC value so that the minimum value becomes 1; we call the resulting values the relative AIC values and write $\text{AIC}^i_{\text{rel}} = \text{AIC}^i / \min_j \text{AIC}^j$, where i is the model index. In Fig. 2.5, we plot the boxplots of relative AIC values of different models on the 20 exemplar genes in the WANG dataset (Wang et al., 2020b). From left to right, the four panels correspond to Poisson, ZIP, NB and ZINB respectively. Except for ZINB, scGTM has the top performance with other three distributions in terms of relative AIC. By the definition of AIC and relative AIC, the result suggests that scGTM is relatively robust to the choice of count distribution and does not have an overfitting problem.

2.4.2 Benchmarking scGTM against GAM by simulation and bootstrapping

We design a comprehensive simulation study to compare the scGTM with GAM. The simulation settings are summarized in Table 2.1. The design can be summarized in two aspects. The first aspect is the function of $\log(\tau_c + 1)$: the scGTM function, the quadratic function, and the logistic function; the latter two functions are used to show the robustness of the scGTM to trends that follow other functions. The second aspect is the trend shape: hill, valley, increasing, and decreasing. Together, we have eight function + shape settings, and we generate ten genes under each setting. Given the function of $\log(\tau_c + 1)$, we generate the gene expression count Y_c from a negative binomial distribution with mean τ_c.

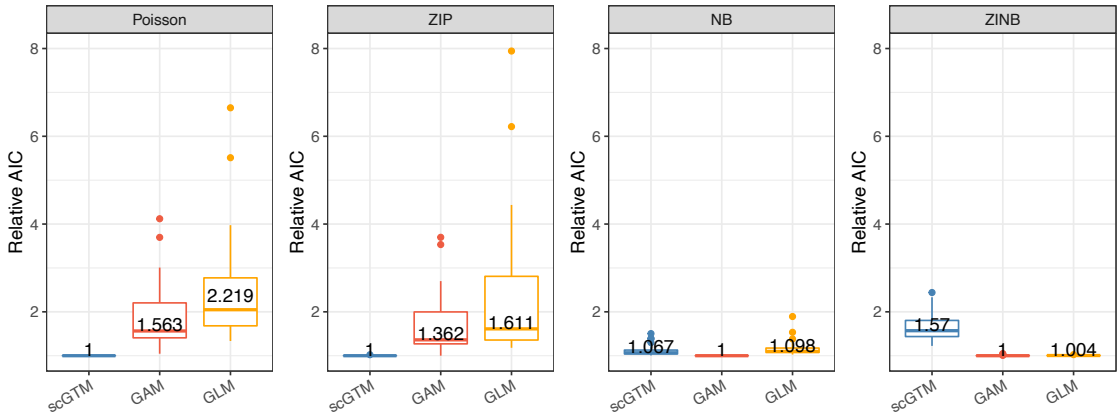


Figure 2.5: AIC comparison (balance of goodness-of-fit and model complexity) of scGTM with GAM and GLM on the WANG dataset (Wang et al., 2020b) (Supplementary Table S1). From left to right, the four panels correspond to the three models with the count distribution as Poisson, ZIP, NB, and ZINB, respectively. In each panel, from left to right, scGTM, GAM, and GLM are shown as blue, red, and orange boxplots, respectively; each boxplot shows the distribution of a model's relative AIC values across genes. A lower relative AIC value indicates better balance of goodness-of-fit and model complexity. With Poisson, ZIP, and NB as the count distribution (the left three panels), scGTM outperforms GAM and GLM.

Table 2.1: Overview of eight simulation settings.

Function	Shape	Formula	Key parameter range(s)
scGTM	hill	$f(t_c) = \begin{cases} \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) \text{ if } t_c \le t_0 \\ \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) \text{ if } t_c > t_0 \end{cases}$	$0.4 < t_0 < 0.6$
scGTM	valley	$f(t_c) = b - \begin{cases} \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) \text{ if } t_c \le t_0 \\ \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) \text{ if } t_c > t_0 \end{cases}$	$0.4 < t_0 < 0.6$
scGTM	increasing	$f(t_c) = \begin{cases} \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) \text{ if } t_c \le t_0 \\ \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) \text{ if } t_c > t_0 \end{cases}$	$1 < t_0 < 1.2$
scGTM	decreasing	$f(t_c) = \begin{cases} \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) \text{ if } t_c \le t_0 \\ \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) \text{ if } t_c > t_0 \end{cases}$	$-0.2 < t_0 < 0$
quadratic	hill	$f(t_c) = b - \mu(t_c - t_0)^2$	$0.4 < t_0 < 0.6;\ \mu > 0$
quadratic	valley	$f(t_c) = b + \mu(t_c - t_0)^2$	$0.4 < t_0 < 0.6;\ \mu > 0$
logistic	increasing	$f(t_c) = \frac{\mu}{1+\exp(-k(t_c - t_0))}$	$0.4 < t_0 < 0.6;\ \mu > 0$
logistic	decreasing	$f(t_c) = \frac{-\mu}{1+\exp(-k(t_c - t_0))}$	$0.4 < t_0 < 0.6;\ \mu > 0$

We first check if the scGTM can correctly recapitulate trend shapes. After fitting both hill- and valley-shaped scGTMs to a gene, we choose the model that has the smaller AIC value. Given the chosen model, we decide if a trend is monotone by checking if the confidence interval of k_1 or k_2 (Section 2.3) contains 0. For example, if a hill-shaped scGTM is chosen and the confidence interval of k_1 contains 0, we consider the trend as monotone decreasing; if a valley-shaped scGTM is chosen and the confidence interval of k_1 contains 0, we consider the trend as monotone increasing. Based on this decision process, the fitted scGTMs can perfectly distinguish between hill- and valley-shaped trends, and they have 97.5% accuracy for distinguishing increasing and decreasing trends. Given that a half of the genes are not simulated from the scGTM assumptions, these results demonstrate the robustness of scGTM.

We next check if the goodness-of-fit of the scGTM is comparable to that of GAM, which is designed to have great flexibility. For every gene g, we calculate the root mean square error e_g between the true trend f and the fitted trend $\hat{f}$:

$$e_g = \sqrt{\frac{\sum_{c=1}^{C}\left[f(t_c) - \hat{f}(t_c)\right]^2}{C}}.$$

A smaller e_g means a better fit. We calculate the average of G simulated genes' e_g's and denote it as $\bar{e} = \frac{1}{G}\sum_{g=1}^{G} e_g$. For all 80 simulated genes (10 genes per setting $\times$ 8 settings), the scGTM performs similarly to GAM ($\bar{e}_{\text{scGTM}} = 0.080$; $\bar{e}_{\text{GAM}} = 0.077$). For the 40 genes simulated from the scGTM assumptions, the scGTM expectedly fits better than GAM does ($\bar{e}'_{\text{scGTM}} = 0.058$; $\bar{e}'_{\text{GAM}} = 0.075$). Fig. 2.6 shows one example gene per simulation setting, including the gene's true trend and the fitted trends by the scGTM and GAM. In particular, the scGTM fits increasing and decreasing trends better than GAM does.

Moreover, we use a bootstrap analysis to show that the fitted scGTM trend has a smaller variance than the fitted GAM trend does, at the cost of a larger bias. Fig. 2.7 shows the fitted scGTM and GAM trends on ten bootstrap samples of the *MAOA* gene in the WANG dataset Wang et al. (2020b), and the scGTM trends are more stable across the bootstrap samples. This is more evident in Fig. 2.8, where the ten fitted trends for each model are

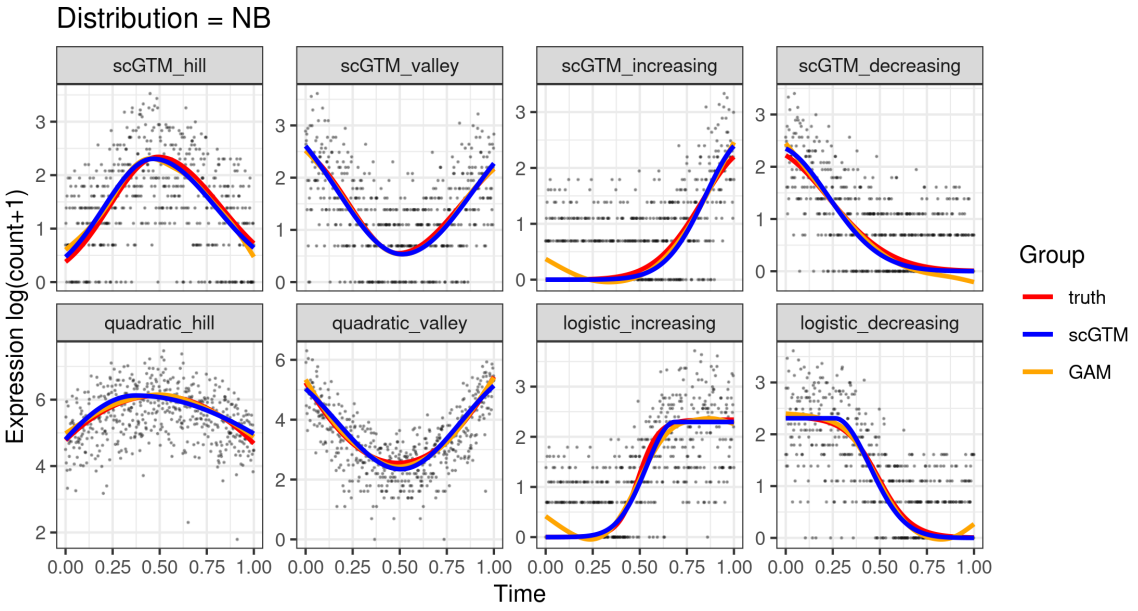


Figure 2.6: Comparison of scGTM with GAM for one example gene under each simulation setting.

overlaying with their mean trend; the root mean square error (between the fitted trends and the mean trend; calculated on 1000 evenly spaced pseudotime values in $[0, 1]$) is 0.399 for the scGTM and 2.799 for GAM.

Further, note that we already used the built-in penalization in the *mgcv* package to reduce the overfitting of GAM when we fit it (the *gam* function). Specifically, there is a smoothing parameter λ to control the wiggliness of the fitted GAM trend, and λ is estimated during the GAM fitting.

2.4.3 scGTM is robust to pseudotime uncertainty

Unlike the observed (true) physical time, the pseudotime is inferred from data and thus intrinsically uncertain. The effects of pseudotime uncertainty on hypothesis testing (i.e., if a gene's expression changes with time) has been discussed in the PseudotimeDE method Song and Li (2021). However, the focus of this work is to proposed the scGTM for interpreting a trend, instead of testing whether a trend is different from a horizontal line, i.e., the focus of PseudotimeDE. Although scGTM does not directly account for the pseudotime uncertainty,

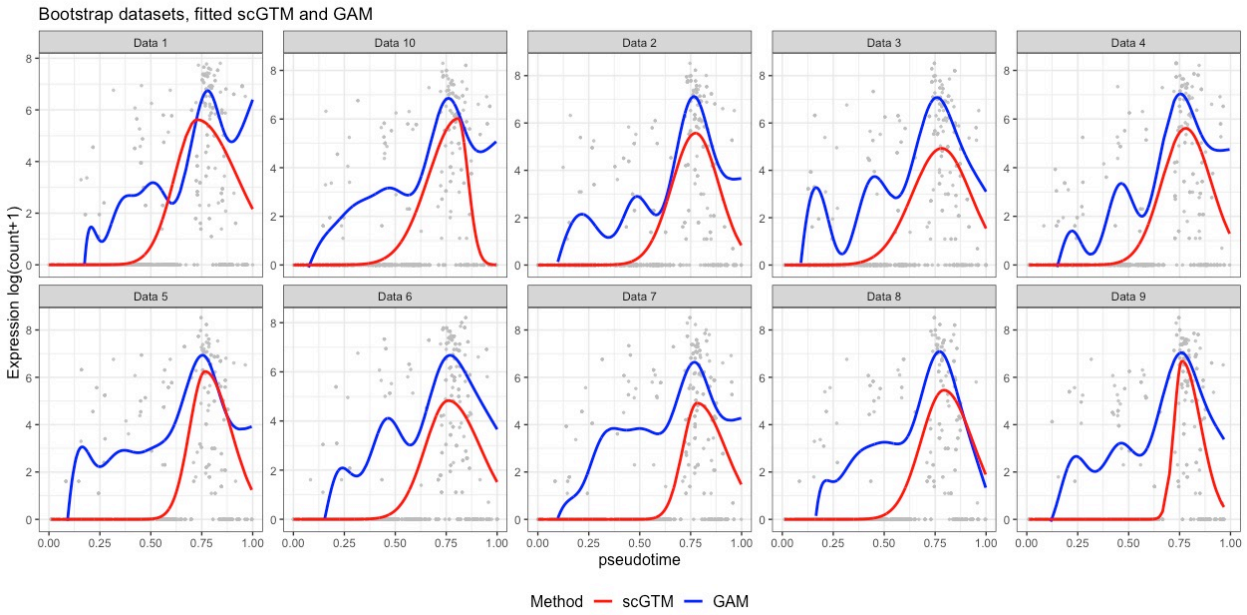


Figure 2.7: Fitted trends of scGTM and GAM on 10 bootstrap samples of the *MAOA* gene in the WANG dataset Wang et al. (2020b).

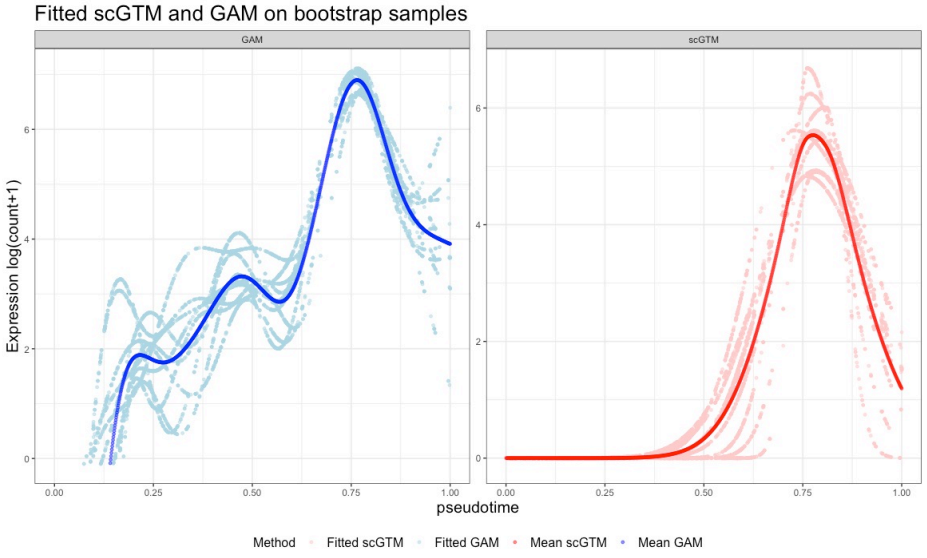


Figure 2.8: The fitted trends of scGTM and GAM on 10 bootstrap samples (light colored scatters) and the mean trends (dark colored curves) of the *MAOA* gene in the WANG dataset Wang et al. (2020b).

we use simulation to show the robustness of scGTM to pseudotime uncertainty in a simple setting.

We first generate the true time t_c and the true trend $f(t_c)$. To introduce uncertainty to t_c, we add normal random noise to obtain the pseudotime $t'_c = t_c + e$, where $e \sim N(0, 0.1^2)$. Fig. 2.9 shows that the scGTM fitted trends are still close to the true trend and correctly capture the trend shape.

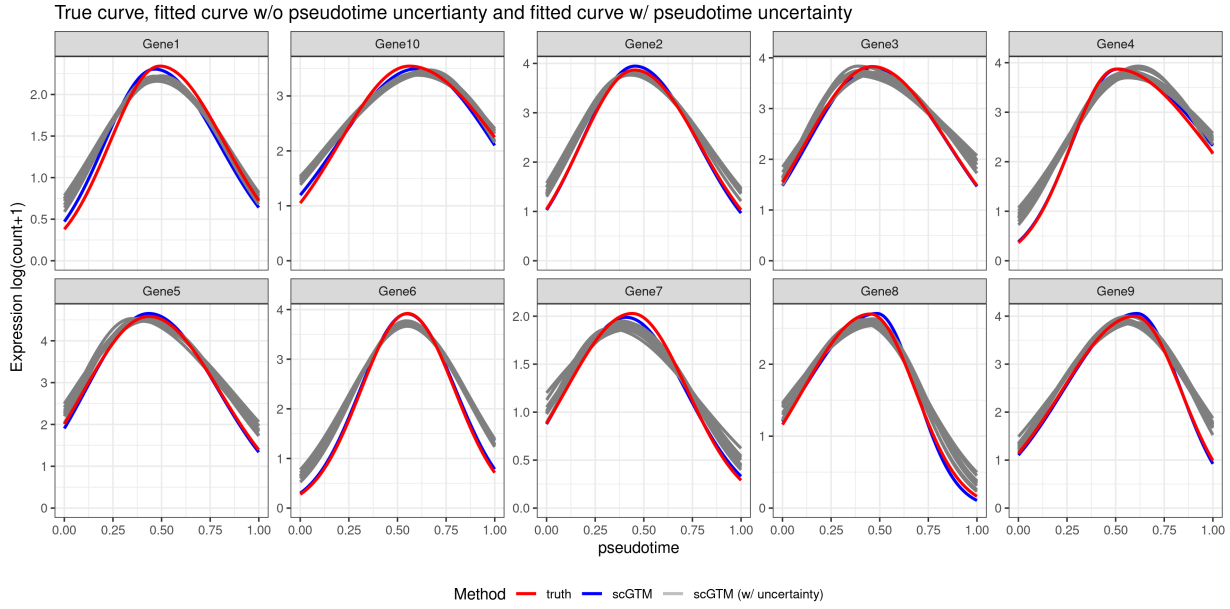


Figure 2.9: The effect of pseudotime uncertainty on scGTM fitting.

2.4.4 scGTM extension can capture more complicated gene trends

Currently the scGTM is designed for detecting simple and easy-to-interpret gene expression trends including monotone, hill-shaped and valley-shaped trends. The reason is that most genes of biological interests are observed to follow one of these simple trends. We deem this reasonable because each pseudotime trajectory is expected to indicate a directional change process, such as development and immune response, along which important genes usually have no more than one hill or valley.

Meanwhile, in practice some genes may exhibit more complicated patterns. Accordingly, the scGTM is extendable by assuming a more complicated mean function, whose estimation can still be achieved by the PSO algorithm (whose major advantage is its flexibility). To demonstrate this functionality of the scGTM, we conduct a simulation study where we use the sine function to generate one gene's true expression trend along the pseudotime. With its mean function set as as the sine function, the scGTM accurately estimates the gene trend (Fig. 2.10).

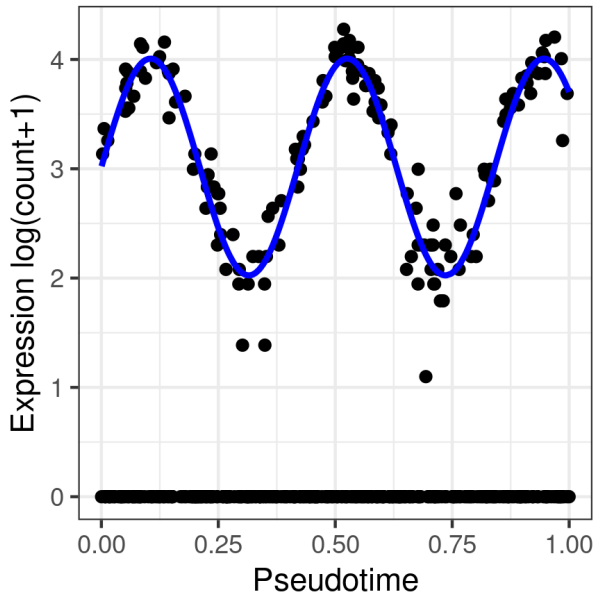


Figure 2.10: scGTM extension correctly captures the sine gene trend.

2.4.5 Applying scGTM on 1,382 human menstrual cycle genes

We choose the 20 genes because they were analyzed and provided with biological interpretations in Wang et al. (2020b). The complete dataset contains 22,036 genes, most of which are not related to human menstrual cycle and thus do not exhibit notable expression trends. To further show the performance of scGTM, we focused on the 1,382 menstrual cycle genes reported in Wang et al. (2020b), and we applied the scGTM to fit these genes' expression trends. The results are still satisfying.

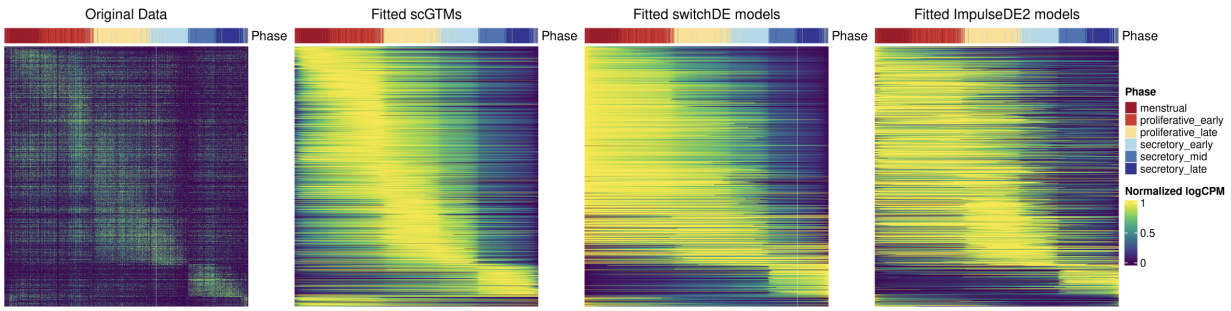
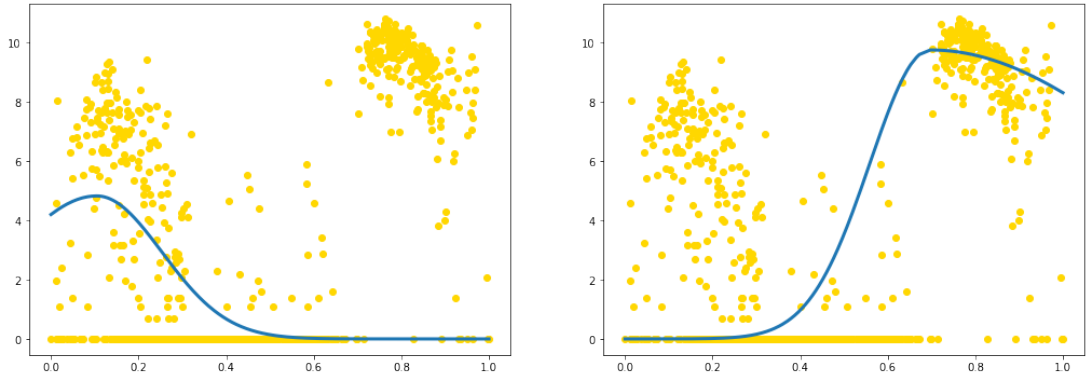


Figure 2.11: Fitted expression trends by scGTM, switchDE, and ImpulseDE2 for 1,382 menstrual cycle related genes in Wang et al. (2020b).

2.4.6 Stagnation: Premature Convergence Issues

The premature convergence issue arises while running the PSO algorithm and in the following we take the *PAEP* gene in WANG Wang et al. (2020b) for illustration. We use Poisson marginal and swarmsize 30. The left panel of Fig. 2.12 shows the result of *PAEP* returned by PSO after 100 iterations while the right panel shows the result after 1000 iterations. It is apparent that the left panel underfits the data and it undergoes a "premature convergence issue" commonly arised in the application of PSO (Bratton and Kennedy, 2007; Larsen et al., 2016).



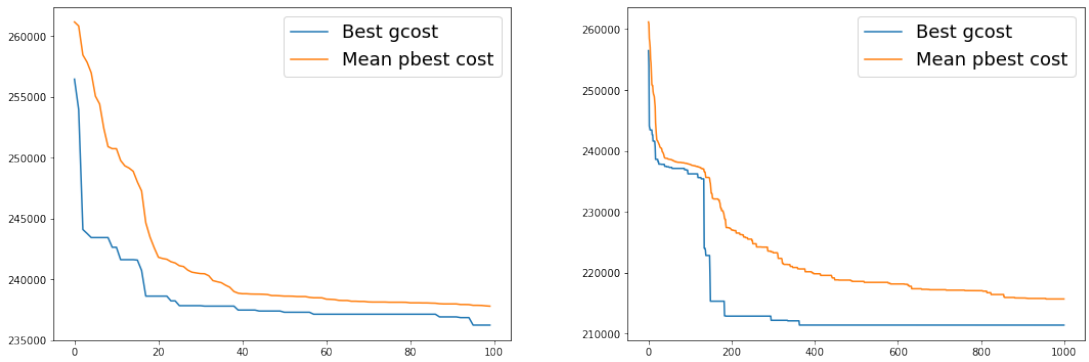
(a) Premature with 100 iterations

(b) Convergence with 1000 iterations

Figure 2.12: Premature convergence of *PAEP* gene

The two panels in Fig. 2.13 further show the cost history of the premature issue of PSO for fitting the *PAEP* gene. The x-axis is the number of iterations and the y-axis is the global

best negative log-likelihood value. There is a premature around 200 iterations and after that there is a big drop, corresponding to the transformation of left hill to right hill in Fig. 2.12.



(a) Cost history with 100 iterations

(b) Cost history with 1000 iterations

Figure 2.13: Premature convergence of *PAEP* gene

To handle the premature convergence issues in PSO and other metaheuristics, scholars have proposed many variants and remedies. For example, local best (lbest) PSO is an alternative (Bratton and Kennedy, 2007). However, it is of great interest to explore that when we have multiple humps or hills, how many iterations we need to prevent PSO from premature convergence. In the next chapter, we further illustrate how metaheuristics can be helpful in a high dimensional data analysis problem.

CHAPTER 3

Metaheuristics in Action: Further Estimation Problems in Statistics

3.1 Preamble

Metaheuristics has been used to find estimates for model parameters and there is work that showed they can outperform those from statistical packages or find them when the latter fail to do so. For example, Cheng et al. (2015) showed that PSO can find more optimal $L1$-estimates for some models than those in statistical packages. In what is to follow, we expand the applications of metaheuristics by highlighting its role in parameter estimation and model optimization across diverse biostatistical contexts. Through a series of examples, we demonstrate the CSO-MA can find more optimal maximum likelihood estimates and also able to find them when some statistical packages cannot. Our applications include finding maximum likelihood estimates for models in bioinformatics and in education studies, and proportional hazard model for analysis in COVID-19 patients in Ethiopia, and a variable selection problem with parameter tuning in ecology.

3.2 Single-cell Generalized Trend Model (scGTM)

In the previous chapter as well as in Cui et al. (2022), we proposed a model called scGTM to study relationship between pseudotime (Trapnell et al., 2014) and gene expression data. The model assumes that the gene expression has a 'hill' trend along the pseudotime and can be modeled using a set of interpretable parameters. Below is a brief description of the model and shows CSO-MA outperforms PSO algorithm for all but one gene in terms of finding the

optimal value of the negative loglikelihood function; details in Cui et al. (2022).

For a hill-shaped gene, the scGTM parameters are $\Theta = (\mu_{\text{mag}}, k_1, k_2, t_0, \phi, \alpha, \beta)^T$ and they are estimated from from the observed expression counts $\boldsymbol{y} = (y_1, \ldots, y_C)^T$ and cell pseudotimes $\boldsymbol{t} = (t_1, \ldots, t_C)^T$ using the constrained maximum likelihood method. Here C is the number of cells and the interpretations of the parameters in the model are given in Section 2.1 of Cui et al. (2022). If $\log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t})$ is the log likelihood function, the optimization problem is:

$$\max_{\Theta} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) \tag{3.2.1}$$

such that

$$\begin{aligned} &\min_{c\in\{1,\ldots,C\}} \log(y_c+1) \le \mu_{\text{mag}} \le \max_{c\in\{1,\ldots,C\}} \log(y_c+1)\,, \\ &k_1, k_2 \ge 0\,, \quad \min_{c\in\{1,\ldots,C\}} t_c \le t_0 \le \max_{c\in\{1,\ldots,C\}} t_c\,, \ \phi \in \mathbb{Z}_+\,, \end{aligned} \tag{3.2.2}$$

where

$$\begin{aligned} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) &= \log\left[\prod_{c=1}^{C} \mathbf{P}(Y_c = y_c \mid t_c)\right] \\ &= \sum_{c=1}^{C} \log\Big[(1-p_c) f(y_c|t_c) + p_c\, \mathbb{I}(y_c = 0)\Big] \end{aligned} \tag{3.2.3}$$

and

$$\begin{aligned} f(y_c|t_c) &= \frac{\tau_c^{y_c}}{y_c!} \frac{\Gamma(\phi + y_c)}{\Gamma(\phi)(\phi+\tau_c)^{y_c}} \frac{1}{\left(1+\frac{\tau_c}{\phi}\right)^{\phi}}\,, \\ \log(\tau_c+1) &= \begin{cases} b + \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) & \text{if } t_c \le t_0 \\ b + \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}, \\ \log\left(\frac{p_c}{1-p_c}\right) &= \alpha \log(\tau_c+1) + \beta\,, \end{aligned}$$

which are all functions of Θ. There are two difficulties in the optimization problem (3.2.1). First, the likelihood function (3.2.3) is neither convex nor concave. Second, the constraint is linear in μ_{mag}, k_1, k_2, and t_0 but ϕ is a positive integer-valued variable. Hence, conventional optimization algorithms, like P-IRLS in GAM (Wood, 2011, 2017) and L-BFGS in switchDE Campbell and Yau (2017) are unlikely able to work well. The authors proposed PSO to solve for the constrained MLEs and a Python package is available online. We now apply CSO-MA to the same problem and compare results from the Python package. In addition, we compared CSO-MA's performance with results from two recently proposed metaheuristic algorithms: the prairie dog optimization algorithm (PDO) proposed by Ezugwu et al. (2022) and the Rutta and Kutta optimization (RUN) algorithm proposed by Ahmadianfar et al. (2021). Table 3.1 displays the negative log likelihood function values found by CSO-MA and PSO for the 20 exemplary genes in Wang et al. (2020b) after 1000 function evaluations of equation (3.2.3) for the two algorithms and it shows that CSO-MA outperformed PSO and PDO in all but three of the 20 genes. The Wilcoxon test of CSO-MA against the other two algorithms produced p-values less than 0.001 (0.00077 for PSO and 0.00026 for PDO), suggesting that CSO-MA indeed outperformed PSO and PDO in this example.

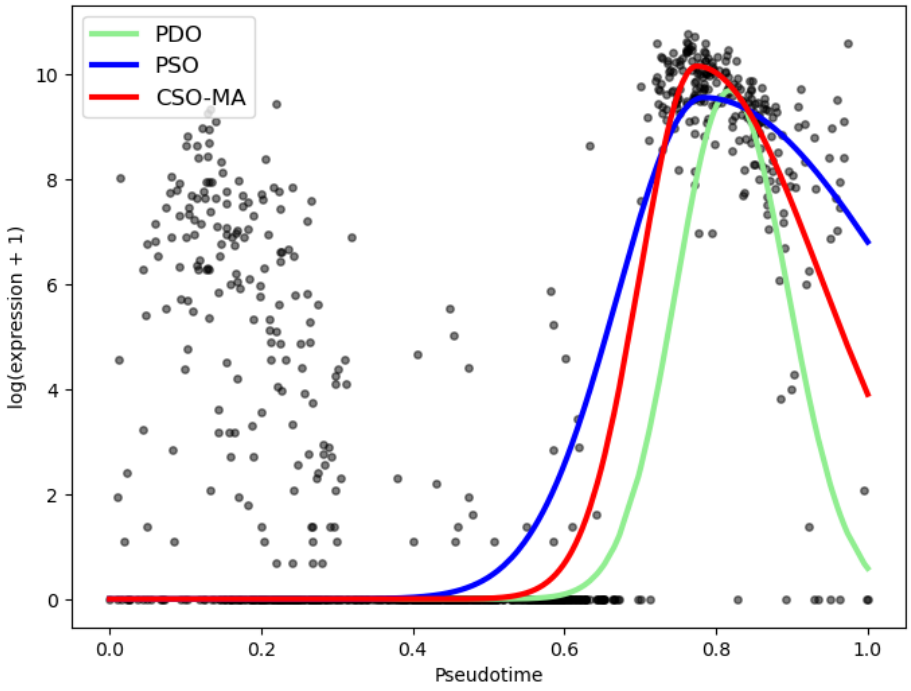
Gene	CSO-MA	PSO	PDO	Gene	CSO-MA	PSO	PDO
PLAU	**1.1115**	1.1291	1.1177	MMP7	**1.6562**	1.6573	1.6963
THBS1	**1.7491**	1.7498	1.7569	CADM1	**0.9903**	0.9907	1.0316
NPAS3	**0.4519**	0.4598	0.4885	ATP1A1	**1.0570**	1.0571	1.1407
ANK3	**1.0473**	1.0501	1.1171	ALPL	**0.6232**	0.6235	0.6315
TRAK1	0.7759	**0.7758**	0.7785	SCGB1D2	2.0608	**2.0501**	2.0952
MT1F	**0.7851**	0.7907	0.8637	MT1X	**0.8060**	0.8065	0.9026
MT1E	**0.6580**	0.6597	0.6735	MT1G	**1.1025**	1.1414	1.1290
CXCL14	0.7939	0.8512	**0.7244**	MAOA	**0.8094**	0.8820	0.8161
DPP4	**0.5503**	0.5535	.5528	NUPR1	**0.7307**	0.7854	0.7739
GPX3	**1.7413**	1.7881	1.7904	PAEP	**2.1034**	2.3693	2.2036

Table 3.1: Optimized negative log likelihood (NLL) values (multiplied by 10^5) obtained by CSO-MA, PSO and PDO after 1000 function evaluations. Lowest NLL values among the three algorithms are in bold for each gene and overall results suggest that CSO-MA outperforms PSO and PDO in almost all cases.

Figure 3.1 displays the fitted PAEP gene given by CSO-MA, PSO and PDO. We observe that CSO-MA captures the "fast decreasing trend" when $t \geq 0.8$ better than PSO does, and

it reaches the higher peak than PDO does. Figures for other genes also show a consistent pattern.

Figure 3.1: Comparison of CSO-MA, PDO and PSO results for the fitted scGTM with gene PAEP.



3.3 Estimation for a Rasch Model

The Rasch model is one of the most widely used item response models in education and psychology research (Embretson and Reise, 2013). Estimating the parameters in the Rasch and other item response models can be challenging and there is continuing interest to estimate them using different methods and studying the various computational issues. For example, (Linacre, 2022; Robitzsch, 2021) reported that there are at least 27 R packages indexed with the word "Rasch" and 11 packages capable of estimating parameters and analysis for the Rasch model.

The expectation-maximization (EM) algorithms is a common method for parameter estimation in statistics (Dempster et al., 1977; Baker and Kim, 2004; Liu et al., 2018).

The Bock-Aitkin algorithm is a variant of the EM algorithm and is one of the most popular algorithms for estimating parameters in the Rasch models (Bock and Aitkin, 1981). Because the Rasch model also has many extensions with applications in agriculture, health care studies and in research in marketing (Mendes et al., 2020; Bezruczko, 2005; Bechtel, 1985), this subsection compares, for the first time, how metaheuristic algorithms perform relative to the Bock-Aitkin's method.

We give a brief review of the Rasch model before we compare the estimation results given by CSO-MA, Bock-Aitkin's (in the R package ltm) and two other metaheuristic algorithms CA and PSO in terms of the likelihood values. In a Rasch model, we work with $N \times I$ binary item response data where 1 indicates correct and 0 indicates incorrect responses. The data come from a cognitive assessment (e.g., math or reading) that includes I test items. A group of N students gave their responses to the I items, and their binary answers to each of the N items were scored and analyzed (Embretson and Reise, 2013). The Rasch model is given by:

$$\text{logit}\big(\mathbf{P}\big(Y_{ji}=1|\theta_j\big)\big)=\theta_j-\beta_i,\ \ \theta_j\sim N(0,\sigma^2). \tag{3.3.1}$$

The item parameter β_i represents the difficulty of item i and parameter θ_j represents the ability of person j. We assume that $\theta_j \sim N(0,\sigma^2)$. This model is called the one-parameter model because it considers one type of item characteristic (difficulty). Let $p_{ji}=\mathbf{P}\big(Y_{ji}=1|\theta_j\big)$ and write the marginal likelihood function for model (3.3.1) as

$$L(\Theta)=\prod_{j=1}^{N}\int\prod_{i=1}^{I}p_{ji}^{Y_{ji}}(1-p_{ji})^{1-Y_{ji}}\pi(\theta)d\theta, \tag{3.3.2}$$

where $\Theta=\big(\beta_1,\cdots,\beta_I,\sigma^2\big)^T$ and $\pi(\theta)$ is the prior of θ.

Metaheuristics has been shown that it can provide superior performance over statistical methods. For instance, Wang and Huang (2014) tackled the challenge of deriving the maximum likelihood estimates for parameters in a mixture of two Weibull distributions with complete and multiple censored data. Their simulation outcomes indicated that the

Particle Swarm Optimization (PSO) frequently outperformed the quasi-Newton method and the EM algorithm in terms of bias and root mean square errors.

In this study, we present similar results and show that the nature-inspired metaheuristic algorithm Mutation Algorithm (CSO-MA) can also give more precise maximum likelihood estimates compared to three of its competitors: PSO, the Bock-Aitkin's method, and the Cat Swarm Algorithm (CA). PSO is legendary and an exemplary nature-inspired swarm based algorithm and CA was introduced by Chu and Tsai (2007), and its effectiveness as an optimizer for a single objective function was demonstrated in Bahrami et al. (2018), where they showed its superior competitive edge against several contemporary top-performing algorithms.

We employed the "Verbal Aggression" data set the R Archive (Bates, 2010) and let NLL denote the minimized value of the negative log-likelihood function. Table 3.2 displays the NLLs from the 4 algorithms, where a swarm size of 30 was used for the 3 metaheuristic algorithms. The hyper-parameter for CSO-MA, was set to $\phi = 0.3$, and the hyper-parameters for PSO and CA were set to the default values in the R package *metaheuristicOpt* (Riza et al., 2018). Evidently, CSO-MA has the smallest NLL value and is the winner. The estimated NNL values from CSO-MA, PSO, and Bock-Aitkin are similar, but that from CA is not, suggesting that CA appears less reliable since its estimated NLLs (gold points and lines on the left panel do not come close to the others.

Algorithm	CSO-MA	Bock-Aitkin	PSO	CA
NLL	4038.77	4072.93	4041.23	4780.49

Table 3.2: Negative log likelihood values from the four algorithms with CSO-MA outperforming the other three algorithms.

Figure 3.2 presents a two-panel visualization. The upper panel illustrates the estimated parameters derived from the four algorithms: CSO-MA, Bock-Aitkin, PSO, and CA. Here, the x-axis represents all 24 parameters (encompassing 23 items in addition to the variance parameter) in the model, while the y-axis depicts their estimated values. The lower panel delineates the progression trajectories of the negative log-likelihood functions of the four algorithms, spanning about 100 function evaluations. The left panel shows that except for the

CA algorithm, Bock-Aitkin, PSO and CSO-MA give similar parameter estimates; the right panel shows that Bock-Aitkin converges fastest in terms of number of function evaluations while PSO is the slowest. However, CSO-MA has the smallest negative log-likelihood value, or equivalently, the largest log-likelihood value.

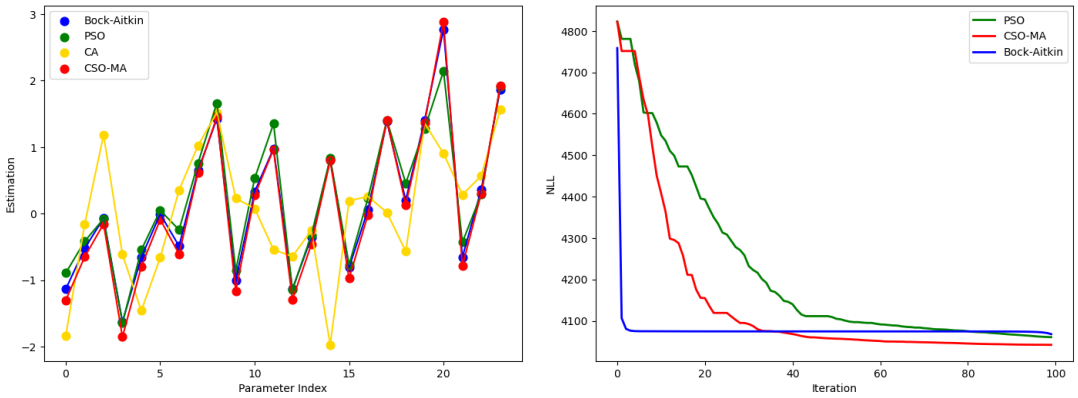


Figure 3.2: The left panel shows estimated parameters from the four algorithms: CSO-MA, Bock-Aitkin, PSO and CA. The x-axis refers to all 24 parameters (23 items plus the variance parameter) in the model and the y-axis refers to the estimated parameter values. The right panel shows the trajectories of the negative log likelihood functions of the four algorithms as they evaluate the negative log-likelihood functions about 100 times.

3.4 M-estimation for Cox Regression in a Markov Renewal Model

In this subsection, we show CSO-MA can solve estimating equations and produce M-estimates for model parameters, that are sometimes more efficient than those from statistical packages. Askin et al. (2017) correctly noted that metaheuristics is rarely used to solve estimating equations in the statistical community.

In a survival study, the experience of a patient may be modelled as a process with finite states (Meira-Machado et al., 2009) and modelling is based on transition probabilities among different states. We take bone marrow transplantation (BMT) as an example. BMT is a primary treatment for leukemia but has major complications, notably Graft-Versus-Host Disease (GVHD), where transplanted marrow's immune cells react against the recipient's cells in two forms: Acute (AGVHD) and Chronic (CGVHD). The main treatment failure is death in remission, often seen in patients with AGVHD or both GVHD types, occurring

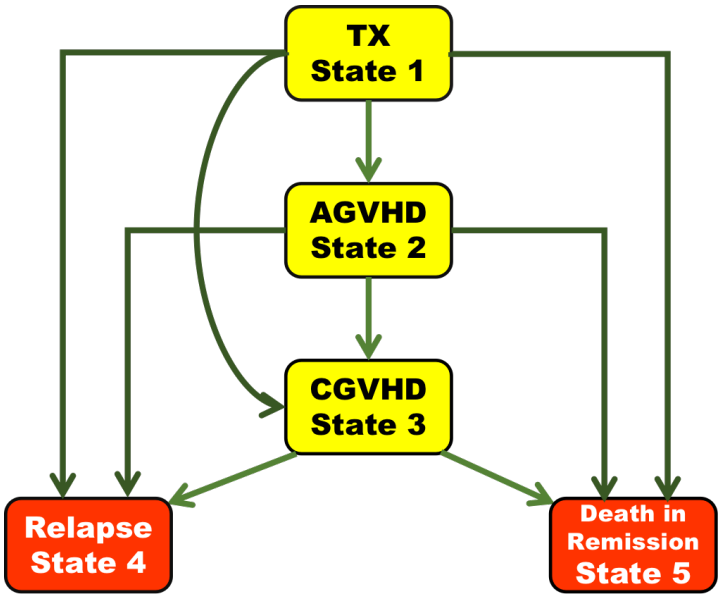


Figure 3.3: A five-state Markov renewal model for BMT failure. Reproduced from Dabrowska et al. (1994). TX = Transplant, AGVHD = Acute Graft-Versus-Host Disease, CGVHD = Chronic Graft-Versus-Host Disease, Relapse = Relapse of leukemia, Death in Remission = Death of a patient who is in remission from leukemia..

unpredictably before relapse. The term "death in remission" in the context of leukemia refers to the death of a patient who is in remission from leukemia. This means the patient has achieved remission, where there are no detectable leukemia cells in the body, but they died from other causes that are not directly related to the active progression of leukemia. However, both AGVHD and CGVHD reduce leukemia relapse risks. Hence, there's a five-state model: transplant (TX), AGVHD, and CGVHD are temporary states, while relapse and death in remission are absorbing states (Dabrowska et al., 1994). Figure 3.3 shows the possible transitions among different states (i.e., TX, AGVHD, CGVHD, Relapse and Death).

To model such a process in a mathematically rigorous way, we assume observations on each individual form a Markov renewal process with a finite state, say $\{1, 2, \cdots, r\}$ (Dabrowska, 2012). That is, we observe a process $(X, T) = \{(X_n, T_n) : n \geq 0\}$ where (for simplicity, we do not consider censoring in this subsection), and $0 = T_0 < T_1 < T_2 < \cdots$ are calendar times of entrances into the states $X_0, X_1, \cdots, X_n \in \{1, 2, \cdots, r\}$. In the BMT example, $r = 5$ and X_n takes values in {TX, AGVHD, CGVHD, Relapse, Death in Remission} and $W_i = T_n - T_{n-1}$ represents the sojourn time staying in the state X_n. We also

observe a covariate matrix $\mathbf{Z} = \{\mathbf{Z}_{ij} : i, j = 1, 2, \cdots, r\}$ where each $\mathbf{Z}_{ij}$ itself is a vector. In practice, we assume that the sojourn time W_n given $X_{n-1} = i$ and $\mathbf{Z}$ has survival probability (Jacod, 1975)

$$\mathbf{P}(W_n > x | X_{n-1} = i, \mathbf{Z}) = \exp\left(- \sum_{k=1, k\neq i}^{r} A_{0,ik}(x) e^{\beta^T Z_{ik}} \right)$$

and the transition probability is $(i \neq j)$

$$\mathbf{P}(X_n = j | X_{n-1} = i, W_n) = \frac{\alpha_{0,ij}(W_n) e^{\beta^T Z_{ij}}}{\sum_{k\neq i} \alpha_{0,ik}(W_n) e^{\beta^T Z_{ik}}},$$

where β is the parameter of interest, $A_{0,ik}(x) = \int_0^x \alpha_{0,ik}(s) ds$ is the baseline cumulative hazard from state i to state k and $\alpha_{0,ik}(x)$ is the hazard function from state i to state k (Cox, 1972). Suppose we observe M iid individuals and suppose the risk process for an individual is given by $Y_i(x) = \sum_{n\geq 1} \mathbb{I}(W_n \geq x, X_{n-1} = i)$. For a fixed x, $Y_i(x)$ counts the number of visits to state i with sojourn time more than x for a particular individual. In the five-state model in Figure 3.3, since we cannot revisit the states that we have already exited, $Y_i(x)$ is a binary variable. Then from Dabrowska et al. (1994); Cook and Lawless (2007); Andersen et al. (2012), the estimating equation for β is

$$\mathbf{U}(\beta) = \sum_{m=1}^{M} \sum_{i\neq j}^{r} \int_0^\infty \left[\mathbf{Z}_{ijm} - \frac{S_{ij}^{(1)}(x, \beta)}{S_{ij}^{(0)}(x, \beta)} \right] dN_{ijm}(x). \tag{3.4.1}$$

Here $N_{ijm}(x) = \sum_{n\geq 1} \mathbb{I}(T_n \leq x, X_n = j, X_{n-1} = i)$, $S_{ij}^{(0)}(x, \beta) = \frac{1}{M} \sum_{m=1}^{M} Y_{im}(x) e^{\beta^T Z_{ijm}}$ and $S_{ij}^{(1)}(x, \beta)$ is the first partial derivative of $S_{ij}^{(0)}$ with respect to β. The M-estimates of β are obtained by solving $\mathbf{U}(\beta) = 0$. To apply CSO-MA to obtain the estimates, we turn the problem of solving $\mathbf{U}(\beta) = 0$ into a minimization problem as follows:

$$\widehat{\beta} = \arg\min_{\beta} \|\mathbf{U}(\beta)\|_p \tag{3.4.2}$$

where $p \in [1, \infty]$ is a user-selected constant. If the solution exists for $\mathbf{U}(\beta) = 0$, then we

have $\min\|\mathbf{U}(\beta)\|_p = 0$ for any $p \geq 1$. Using metaheuristics to creatively solve the system of nonlinear equations Li et al. (2015b); Pant et al. (2019), results from our simulation study suggest that the choice of p does not affect the convergence speed of CSO-MA nor the estimated parameters.

For simulation, we set $p = 2$ and assume $r = 3$, $A_{0,ij}(x) = 0.5x$ for all $i \neq j$, the true parameter vector $\beta = (0.901, 0.759, 0.348)^T$ and elements of the covariance matrix $\mathbf{Z}$ are random uniform variates from $[-1, 1]$. In total, we generated $M = 100$ individuals and the left panel of Figure 3.4 shows one of the realizations. The swarm size for CSO-MA was 20 and we ran it for 100 function evaluations and the right panel of Figure 3.4 shows the convergence of CSO-MA. The estimated parameter is $\widehat{\beta} = (0.908, 0.753, 0.329)^T$, which is close to the true value. The observed vector of biases $(0.007, 0.006, 0.017)^T$ is likely due to both the optimization algorithm and the method of partial likelihood itself. The first issue can be reduced by trying using different initialized values of CSO-MA and the second issue may be solved by having a larger sample size so that consistency of the estimators is guaranteed theoretically. For space consideration, we omit additional simulation results that support the effectiveness of CSO-MA for estimating the true parameters correctly.

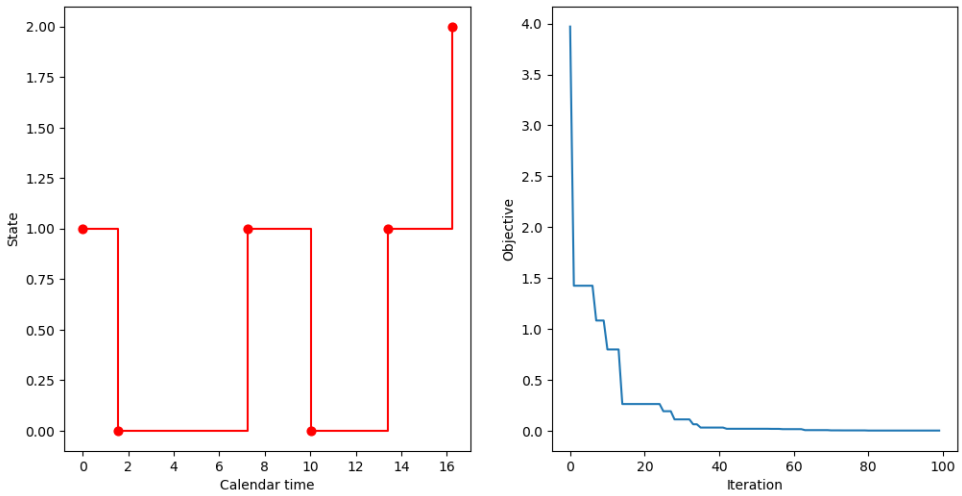


Figure 3.4: Application of CSO-MA to find M-estimates for a Cox regression in a Markov renewal model. The left panel is one of the realizations of 100 individuals; the red dots represent the jump times and the transitions for the pair (X_n, T_n). The right panel shows the convergence trajectory of CSO-MA.

To further investigate the scalability of CSO-MA and compare it with other algorithms,

we perform another simulation study where the state space of X_i consists of two, i.e., $\{1, 2\}$ and 2 is an absorbing state. Consequently, the Markov renewal model is equivalent to a two-state Markov model or a Cox proportional hazards model (Cox, 1972), the sample size is $M = 10,000$ and the β parameter has is the 100×1 vector with all entries equal to 1. The elements of the covariance matrix $\mathbf{Z}$ are again generated uniformly from $[-1, 1]$ to mimic the high-dimensional scenario in statistical applications (Tibshirani, 2009). The simulation is performed on the Matlab 2023a platform. Instead of minimizing the norm of $\mathbf{U}(\beta)$, we minimize the negative partial log-likelihood (NLL) value (Dabrowska et al., 1994). We compare CSO-MA with PDO and Runge Kutta optimization (abbreviated as RUN, it is another recently proposed metaheuristics (Ahmadianfar et al., 2021)) in terms of their optimum values, stability and running time. The results are given in Table 3.3. We run each algorithm 30 times to get reasonable statistical results and the number of function evaluation is set to 1000, the swarm size for each algorithm is set to 30. The results suggest that RUN performs the best in terms of NLL and its stability; The CSO-MA has the best performance in terms of average elapsed time and PDO is the slowest among the three algorithms.

Algorithm	CSO-MA	PDO	RUN
NLL	1868.51 (248.61)	1825.20 (2.79)	1636.39 (4.34)
Elapsed time	250.10s (10.16s)	472.09s (5.59s)	270.40s (2.41s)

Table 3.3: Negative log likelihood values from the three algorithms with CSO-MA outperforming the other two recently proposed algorithms.

3.5 Proportional Hazard Analysis of COVID-19 patients in Ethiopia

COVID-19 is a global public health problem causing high mortality worldwide. This subsection used a proportional hazard model to assess time to death using predictors of mortality among patients hospitalized for COVID-19 in the Arsi zone treatment center in Ethiopia. The data was from medical records of laboratory-confirmed COVID-19 cases hospitalized at Bokoji Hospital (Kaso et al., 2022). The primary goal was to identify potential risk factors of the death caused by COVID-19 among 422 patients (Kaso et al., 2022) and the data is available in the appendix of their paper. The purpose of this subsection is to further

demonstrate the usefulness and flexibility of CSO-MA to obtain M-estimates in the model with the following risk factors taken from Kaso et al. (2022): malignancy, chronic kidney disease, comorbidity, AIDS/HIV, ICU admission, and intranasal oxygen use. All are binary risk factors. The time-to-event outcome is admission to death (in days)

Let X_m be the observed survival time and δ_m be the censoring indicator (1 refers to death and 0 refers to censored) of patient m. Define $Y_m(t) = \mathbb{I}(X_m \geq t)$ and $N_m(t) = \mathbb{I}(X_m \leq t, \delta_m = 1)$ be the risk and counting process respectively (Andersen et al., 2012). Then the score equation for the proportional hazard model is

$$\mathbf{U}(\beta) = \sum_{m=1}^{M} \int_0^{\tau} \left[\mathbf{Z}_m - \frac{S^{(1)}(x,\beta)}{S^{(0)}(x,\beta)} \right] dN_m(x) \tag{3.5.1}$$

and

$$S^{(0)}(x,\beta) = \sum_{m=1}^{M} Y_m(x) e^{\beta^T \mathbf{Z}_m}, \;\; S^{(1)}(x,\beta) = \sum_{m=1}^{M} Y_m(x) \mathbf{Z}_m e^{\beta^T \mathbf{Z}_m}.$$

As in the previous subsection, we apply CSO-MA and solve for the minimizer of (3.5.1) by

$$\widehat{\beta} = \arg\min_{\beta} \|\mathbf{U}(\beta)\|_2^2. \tag{3.5.2}$$

The standard deviation of $\widehat{\beta}$ can be estimated using observed Fisher information matrix (Klein and Moeschberger, 2003; Elashoff et al., 2016). In the following, we compare the objective function values in (3.5.2) returned by CSO-MA with that given by standard package "coxph" in R and we use Breslow's method for handling ties (Klein and Moeschberger, 2003; Elashoff et al., 2016).

Table 3.4 displays results for the case $M = 422$, where $\widehat{\beta}$ is the estimate, $\|\mathbf{U}\|_2$ is defined in Equations (3.5.1) and (3.5.2) and $LPL(\beta)$ refers to the log partial likelihood function (Cox, 1975; Elashoff et al., 2016) defined by

$$LPL(\beta) = \sum_{m=1}^{M} \int_0^{\tau} \left[\beta^T \mathbf{Z}_m - \log S^{(0)}(x,\beta) \right] dN_m(x).$$

We observe from the table that although the data set is not complicated, both CSO-MA and "coxph" produce similar results (up to 3 decimals), suggesting that CSO-MA is reliable for solving estimating equations

	CSO-MA			**coxph**		
Risk factors	$\widehat{\beta}$	$\|\mathbf{U}\|_2$	$LPL(\beta)$	$\widehat{\beta}$	$\|\mathbf{U}\|_2$	$LPL(\beta)$
Malaginance	1.305	1.256×10^{-8}	-277.75	1.304	1.256×10^{-8}	-277.75
Chronic kidney disease	1.685	3.590×10^{-8}	-277.13	1.685	3.590×10^{-8}	-277.13
Comorbidity	1.669	1.831×10^{-10}	-268.03	1.669	1.831×10^{-10}	-268.03
AIDS/HIV	1.810	7.467×10^{-9}	-268.03	1.810	7.467×10^{-9}	-268.03
ICU admission	2.975	1.798×10^{-7}	-259.83	2.974	1.798×10^{-7}	-259.82
Intranasal oxygen use	2.933	3.150×10^{-8}	-253.76	2.933	3.150×10^{-8}	-253.76

Table 3.4: Comparison of CSO-MA and coxph on the COVID-19 data set.

3.6 Find MLE for Log-binomial Model

The Log-binomial model is sometimes used to estimate the relative risk ratio directly, controlling for confounders when the outcome is not rare (Blizzard et al., 2006). However, one notorious problem using log-binomial model in practice is that the iterative methods provided in statistical software usually fail to find the MLE or the convergence cannot be attained. For instance, Blizzard et al. (2006) implemented a simulation study to examine the successful rate of fitting log-binomial model where 12 data generation designs were proposed. Their results showed that only around 20% of well-designed simulated samples had no problem fitting a log-binomial model. The main reason for the fitting problem is that the optimum may be near the boundary of the constrained space or out of the space, where we require the parameter vector to be admissible. Standard iterative methods in software packages are likely to update steps that lead to estimates outside the constrained space and result in a crash.

A solution for better fitting a log-binomial model was proposed by de Andrade et al. (2018), where the original constrained problem of finding MLE is replaced by a sequence of penalized unconstrained sub-problems whose solutions converge to the solution of the original program. To solve each unconstrained sub-problem, they recommended using derivative-free algorithms such as Nelder-Mead or BFGS. The results show that more than 93% of their

simulations produce the correct parameters for the log-binomial models.

In this subsection, we show the flexibility of CSO-MA and demonstrate that it also an effective method for finding MLEs for challenging models, like the log-binomial model. Because the objective is to maximize the log-likelihood function with subject to the constraint that the parameter vector should be admissible, we make it an unconstrained optimization problem by minimizing the following penalized function:

$$-\text{log-likelihood} + \text{penalty of inadmissibility}.$$

We used all 12 simulated scenarios implemented in Blizzard et al. (2006) and were able to confirm that CSO-MA can correctly and stably find all the MLEs with each one running time less than 1 CPU second.

In the following, we compare BAT with several popular nature-inspired algorithms for finding MLE in log-binomial models using simulated data. The model for log-binomial regression stipulates the outcome $y_i \sim \text{Binomial}(n_i, p_i)$ and

$$\log p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{3.6.1}$$

Since p_i is within range 0 to 1, $\log p_i$ is always non-positive, leading to the linear constraint $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \leq 0$. Hence, the optimization problem is:

$$\begin{aligned} &\min_{p_i} -\sum_{i=1}^{n} \log \mathbf{P}(Y_i = y_i | p_i) \\ &\text{s.t. } \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \leq 0. \end{aligned} \tag{3.6.2}$$

For the simulation study, we have three categories of the log-binomial model described in Williamson (2013), where the maximizer of (3.6.1) is on a finite boundary, inside the parameter space or is in the limit. Table 3.5 contains three data sets for estimating the two parameters β_0 and β_1 in the model with only one covariate X taking on three possible values $\{-1, 0, 1\}$. Fig.3.5 displays the contour plots of the negative log likelihood function

constructed from each data set and we observe that its maximum is at the boundary, in the limit of the parameter space and in its interior, respectively, using data sets from left to right.

	(Y=1)	(Y=0)
(X=-1)	10	8
(X=0)	18	9
(X=1)	5	0

MLE at boundary

	(Y=1)	(Y=0)
(X=-1)	0	17
(X=0)	0	21
(X=1)	0	12

MLE at infinity

	(Y=1)	(Y=0)
(X=-1)	2	2
(X=0)	14	3
(X=1)	2	17

MLE at interior

Table 3.5: Three data sets from Williamson (2013).

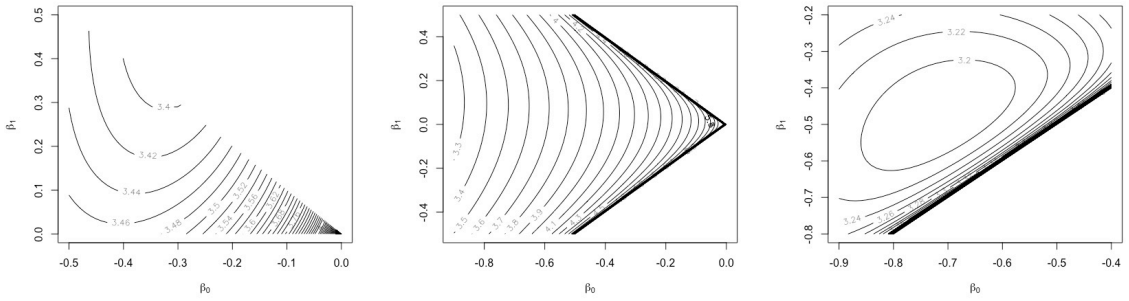


Figure 3.5: Contour plots of the negative log likelihood functions from the three data sets. Left: MLE at boundary; middle: MLE at infinity; right: MLE at interior.

How does BAT perform relative to other metaheuristic algorithms? We mentioned at the onset that there are many other nature-inspired metaheuristic algorithms and it is good practice not to rely on results from a metaheuristic algorithm since such algorithms do not guarantee convergence to the optimum. Accordingly, we additionally implement some of CSO competitors to find the MLEs of parameters in the same log binomial model using each of the three data sets, except that we now restrict values of both β_0 and β_1 in the range $[-10, 10]$. Table 3.6 reports the comparison results, where NLL refers to the negative log likelihood. We run each algorithm 100 times with 100 iterations using their default values and report the estimated quantities, along with their means and the standard deviations.

The results shown in the table support that CSO-MA is an effective algorithm for solving the optimization problem. In particular, we observe that the optimized NLL values from CSO-MA are generally smaller than other NLL values found by the other algorithms and the two

	CSO-MA	ABC	BA	CS	GA	PSO
NLL	29.78 (0.011)	30.28 (0.72)	29.91 (0.17)	34.81 (4.21)	43.50 (12.88)	29.77 (0.0013)
β_0	-0.34 (0.0086)	-0.37 (0.069)	-0.34 (0.019)	-0.59 (0.24)	-0.97 (0.57)	-0.34 (0.0029)
β_1	0.34 (0.0090)	0.29 (0.083)	0.31 (0.035)	0.12 (0.31)	-0.19 (0.58)	0.34 (0.0029)

MLE at boundary

	CSO-MA	ABC	BA	CS	GA	PSO
NLL	2.25 (2×10^{-10})	2.25 (3×10^{-10})	2.25 (1×10^{-6})	2.86 (4×10^{-4})	3.28 (2×10^{-3})	2.25 (0)
β_0	-10 (0)	-10 (0)	-10 (0)	-9.84 (0.14)	-9.74 (0.31)	-10 (0)
β_1	0.17 (4×10^{-4})	0.17 (2×10^{-4})	0.17 (0.025)	0.19 (0.50)	0.22 (0.37)	0.17 (0)

MLE at infinity (NLL $\times10^{-3}$)

	CSO-MA	ABC	BA	CS	GA	PSO
NLL	24.14 (2.44×10^{-4})	24.15 (0.025)	24.17 (0.069)	26.41 (1.98)	27.53 (4.15)	24.14 (6.8×10^{-6})
β_0	-0.70 (0.0024)	-0.71 (0.022)	-0.69 (0.020)	-0.88 (0.26)	-1.11 (0.37)	-0.71 (3.85×10^{-4})
β_1	-0.47 (0.0023)	-0.47 (0.021)	-0.45 (0.033)	-0.42 (0.29)	-0.66 (0.24)	-0.47 (4.06×10^{-4})

MLE at interior

Table 3.6: Minimized negative log likelihood values and estimated parameters. ABC = Artificial Bee Colony Algorithm, BA = Bat Algorithm, CS = Cuckoo Search, GA = Genetic Algorithm, PSO = Particle Swarm Optimization.

parameters are estimated more accurately compared with the larger standard errors from the other algorithms. Genetic algorithm (GA) clearly performs the worst followed by Cuckoo Search (CS) in this particular study, but this does not imply that they are inferior algorithms at all. They may well excel in tackling other optimization problems but these are just intriguing aspects of metaheuristics. The take home message is that one should compare results from a metaheuristic algorithm with several other algorithms and have a higher confidence that the generated results are correct or nearly so when other algorithms also produce similar results.

3.7 Empirical Likelihood and Turnbull's Estimator

The empirical likelihood has a long history in statistics with the advantage of robustness, but it is computationally intensive Lazar (2021) and difficult to compute in practice. In this subsection, we show that CSO-MA can generate competitive results for deriving Turnbull's estimator, a type of empirical likelihood based statistical estimator, hence suggesting its potential for solving other types of empirical likelihood problems. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. survival times for n

subjects, and we only observe the interval censored data:

$$\{(L_i, R_i] : i = 1, 2, \cdots, n\}$$

where $0 \leq L_i < R_i \leq \infty$ are left- and right- endpoint of an observation. That is, we only know that $X_i \in (L_i, R_i]$ instead of their exact values. There are two types of interval censoring:

- The case I interval censoring refers to either $L_i = 0$ or $R_i = \infty$ and it is also called the current status data.
- The case II interval censoring refers to $L_i > 0$ and $R_i < \infty$.

Such censoring mechanism arises commonly in longitudinal studies, and in particular, COVID-19 data analyses (Mingyue et al., 2020; Yin et al., 2021; Tian and Sun, 2022). The counting process approaches do not apply in estimating the survival function of X with interval-censored data, and so we can formulate the likelihood function of our observations by:

$$\mathcal{L}(S) = \prod_{i=1}^{n} \left(S(L_i) - S(R_i)\right),$$

where S is the survival function of X. If $L_i = R_i$ for some i, we replace $S(L_i)$ with its the left limit $S(L_i-)$ so that the likelihood assigns probability mass to $1 - S(\Delta R_i)$. The nonparametric maximum likelihood estimation (NPMLE) is defined as

$$\widehat{S} = \arg\max_{S} \log \mathcal{L}(S).$$

Following Turnbull (1976) and Sun (2006), let $\{s_j\}_{i=0}^{m}$ be the unique ordered elements of $\{0, L_i, R_i : i = 1, \cdots, n\}$, let $\alpha_{ij} = \mathbb{I}(s_j \in (L_i, R_i])$ and let $p_j = S(s_j) - S(s_{j-1})$, for $i = 1, \cdots, n$ and $j = 1, \cdots, m$. If $L_i = R_i$ for some i, i.e., an exact observation, then we replace α_{ij} by $\alpha_{ij} = \mathbb{I}(s_j \in [L_i-, R_i])$ so that $\alpha_{ij} = 1$ if $s_j = L_i$. Hence, the likelihood $\mathcal{L}(S)$ can be written as

$$\mathcal{L}(S) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} p_j \right) \tag{3.7.1}$$

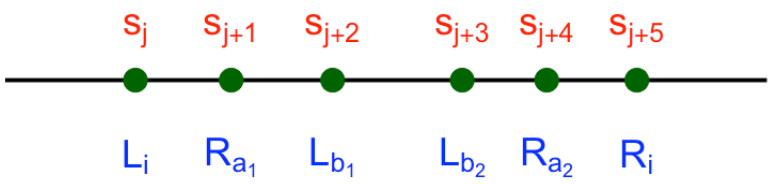
and NPMLE refers to the probability vector $\mathbf{p} = (p_1, \cdots, p_m)^T$. If m is large, then finding $\mathbf{p}$ that

maximizes $\mathcal{L}(S)$ is computationally intractable or inefficient. However, if we know in advance that some (or many) p_j are 0, then the computation will be reduced by a lot. The following lemma shows that it is feasible in practice.

Lemma 1 (Turnbull's intervals)**.** *The p_j can be nonzero only if $s_{j-1} = L_i,\ s_j = R_k$ for some i and some k.*

Proof. Fix i and we consider Figure 3.6 as an example where a_1, a_2, b_1, b_2 are indexes. The i^{th} sum of $\mathcal{L}(S)$ is $\mathcal{L}_i = \sum_{j=1}^{m} \alpha_{ij} p_j$, and it reduces to $\sum_{k=1}^{5} p_{j+k} = S(s_{j+5}) - S(s_j)$. If there is a probability mass for either $(R_{a_1}, L_{b_1}]$, $(L_{b_1}, L_{b_2}]$ or $(R_{a_2}, R_i]$, we can always put the mass to the interval $(L_{b_2}, R_{a_2}]$ from them without changing the value of $\mathcal{L}_i$. On the other hand, the b_2^{th} sum $\mathcal{L}_{b_2}$ will increase by at least the probability assigned to the $(R_{a_1}, L_{b_2}]$.

Figure 3.6: Illustration of Turnbull's intervals.



□

The resulting intervals $(s_{j-1}, s_j]$ for nonzero p_j are referred as the Turnbull's intervals and the resulting estimator $\widehat{\mathbf{p}} = (\widehat{p}_1, \cdots, \widehat{p}_m)^T$ is termed as the Turnbull's estimator.

The "Icens" (Gentleman and Vandal, 2010), "interval" (Fay and Shaw, 2010), "icenReg" (Anderson-Bergman, 2017) are three well-established R packages for handling univariate interval-censored data while "MLEcens" (Groeneboom et al., 2008; Maathuis and Maathuis, 2022) and "intccr" (Park et al., 2019) are designed for bivariate and competing risk interval-censored data, respectively.

In table 3.7, we illustrate the stability of CSO-MA. The lung tumor data is from Hoel and Walburg Jr (1972); Anderson-Bergman (2017), the simulated Weibull data is generated using "simIC-weib" function in "icenReg" with $n = 1000, b1 = 0, b2 = 0, \text{shape} = 0.5,$ and $\text{scale} = 1$ and the simulated Poisson data is generated by $(X, X + Y]$ with sample size 600 where X and Y are independent Poisson variables with rate 30 and 15, respectively. We run CSO-MA and PSO 50

times and each time with 1000 iterations to get reasonable statistical results. The column CSO-MA refers to the average log $\mathcal{L}(S)$ along with its standard deviation returned by CSO-MA. The second column refers to the results given by PSO along with their standard deviation. The third column corresponds to result given by the "EMICM" function in "icens". Though CSO-MA does not output a lower $-\log \mathcal{L}(S)$ than PSO and "EMICM", it has a smaller standard deviation than PSO does, suggesting its robustness among metaheuristic algorithms.

Dataset	CSO-MA	PSO	"EMICM"
Lung tumor data	77.9570 (0.0340)	77.8362 (0.0265)	77.8351 (0)
Simulated Weibull data	806.1604 (0.323)	805.6635 (0.754)	803.1158 (0)
Simulated Poisson data	475.3893 (0.0163)	475.2120 (0.0791)	475.2146 (0)

Table 3.7: Stability of CSO-MA applied to Turnbull's estimator using different datasets.

3.8 Matrix Completion (Missing Data Imputation) in a Two Compartment Model

In this subsection, we apply CSO-MA to a missing data imputation problem in a non-linear Gaussian regression model using simulated data. Missing data arise commonly in engineering, economics, social science and biomedical research. Imputation is one of the most important topics in handling missing data (Little and Rubin, 2019) and EM algorithm Dempster et al. (1977) is a popular choice for imputing multivariate normal data. We briefly describe the problem and the EM algorithm below.

Suppose that $(Y_1, Y_2) \in \mathbb{R}^2$ has a bivariate normal distribution with mean $\mu(\theta) = (\mu_1(x, \theta), \mu_2(x, \theta))$ and a known covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ where θ is a vector of parameters characterizing μ and x is (possibly) a vector of covariates. We observe n realizations $y_i = (y_{i1}, y_{i2}), i = 1, 2, \cdots, n$ and y_{ij} contains missing values for some i and j. Let $Y_{(0)}$ and $Y_{(1)}$ denote the observed and missing parts, respectively. At the $(t+1)^{th}$ iteration, the E step of the algorithm calculates (page 250-251 in Little and Rubin (2019))

$$\mathbf{E}\left(\sum_{i=1} y_{ij} \Big| Y_{(0)}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t+1)}$$

and

$$\mathbf{E}\left(\sum_{i=1} y_{ij}y_{ik}\Big|Y_{(0)},\theta^{(t)}\right)=\sum_{i=1}^{n}\left(y_{ij}^{(t+1)}y_{ik}^{(t+1)}+c_{jki}^{(t+1)}\right)$$

for $j,k=1,2,\cdots,K$ where

$$y_{ij}^{(t+1)}=\begin{cases} y_{ij} & \text{if } y_{ij}\in Y_{(0)} \\ \mathbf{E}\left(y_{ij}\Big|Y_{(0)},\theta^{(t)}\right) & \text{if } y_{ij}\in Y_{(1)} \end{cases}$$

and

$$c_{jki}^{(t+1)}=\begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed.} \\ Cov\left(y_{ij},y_{ik}\Big|Y_{(0)},\theta^{(t)}\right) & \text{if } y_{ij},y_{ik}\in Y_{(1)}. \end{cases}$$

After the E-step, missing values are replaced by the conditional expectation derived above. Next, for the M-step, we maximize the following conditional log-likelihood with respect to θ using CSO-MA:

$$\begin{aligned} &\mathbf{E}\left(l(\theta|Y_{(0)},Y_{(1)})\Big|Y_{(0)},\theta^{(t)}\right) \\ =&-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i^{(t+1)}-\mu(x_i,\theta)\right)^T\Sigma^{-1}\left(\mathbf{y}_i^{(t+1)}-\mu(x_i,\theta)\right)+C \end{aligned} \tag{3.8.1}$$

where $\mathbf{y}_i^{(t+1)}=(y_{i1}^{(t+1)},\ y_{i2}^{(t+1)})$ and C is a constant that is independent of θ.

Wild and Seber (1989) illustrates a two-compartment model with (see also chapter 7 in Fedorov and Leonov (2013))

$$y_{ij}=\mu_j(x_i,\theta)+\epsilon_{ij}, i=1,2,\cdots,n,\ j=1,2,$$

where x refers to time, $(\epsilon_{i1},\epsilon_{i2})\sim_{ind} N(0,\Sigma)$, and

$$\begin{aligned} \mu_1(x,\theta)&=\theta_1e^{-\theta_2x}+(1-\theta_1)e^{-\theta_3x}, \\ \mu_2(x,\theta)&=1-(\theta_1+\theta_4)e^{-\theta_2x}+(\theta_1+\theta_4-1)e^{-\theta_3x}, \end{aligned}$$

and

$$\theta_4=\frac{(\theta_3-\theta_2)\theta_1(1-\theta_1)}{(\theta_3-\theta_2)\theta_1+\theta_2}.$$

Suppose at some time point x, the operator forgot to record either y_{i1},y_{i2} or both and we observe

$Y_{(0)}$ and $Y_{(1)}$ (n observations in total). To make inference about θ, however, we still want to make use of the partially observed data. In this case, we adopt the EM algorithm described above and need to maximize the conditional likelihood (3.8.1).

We analyze a real dataset to illustrate this idea. The dataset comes from Beauchamp and Cornell (1966), see also section 11.2 in Wild and Seber (1989). We randomly mask some of the values to be missing (denote as NA) and present the data in table 3.8.

i	x_i (hours)	y_{i1}	y_{i2}
1	0.33	NA	0.03
2	2	0.84	0.10
3	3	NA	0.14
4	5	0.64	0.21
5	8	0.55	NA
6	12	NA	0.40
7	24	0.27	0.54
8	48	0.12	0.66
9	72	0.06	0.71

Table 3.8: The dataset from Beauchamp and Cornell (1966)

The covariance Σ is taken to be $\begin{pmatrix} 0.075 & -0.06 \\ -0.06 & 0.06 \end{pmatrix}$ (estimated from complete observations) and in the original paper, using full data, the authors estimate the parameters as $\widehat{\theta} = (0.060,\ 0.007,\ 0.093)$. For the EM algorithm, we set the initial θ to be $(0.381,\ 0.021,\ 0.197)$ and set the iteration of CSO-MA to be 200 and $\phi = 0.3$. The whole algorithm alternates between computing expression (3.8.1) and applying CSO-MA to maximize (3.8.1) and we run 10 iterations in total. The imputed results and parameter estimates are given in table 3.9.

We further perform a simulation study (not reported here) with sample size $n = 80$ and 40 missing values in total. The true parameter θ is $(0.4, 0.05, 0.3)$ and the initial value for the EM algorithm is $(0.1, 0.1, 0.1)$. The algorithm terminates after 5 iterations, with the estimated parameter value $\widehat{\theta} = (0.392, 0.056, 0.275)$.

Imputed data			
i	x_i (hours)	y_{i1}	y_{i2}
1	0.33	0.75	0.03
3	3	0.65	0.14
4	5	0.64	0.21
5	8	0.55	0.28
6	12	0.39	0.40
	θ_1	θ_2	θ_3
Estimates	0.394	0.006	1.879

Table 3.9: The imputed dataset and estimated parameters

3.9 High Dimensional D-optimal Design for Generalized Linear Models

In this section, we apply metaheuristics and construct a variety of new optimal designs under a complex situation. In a generalized linear model (GLM) setting, one is usually interested in multiple factors (covariates) and their interactions because a few explanatory factors may not capture the complex structure of the full data adequately. But this will lead to an increasing number of parameters. For example, a 5-factor GLM with second order interactions has 16 parameters in total. Then the optimal design will be at least a $16 \times 6 = 96$ dimensional optimization problem, which is formidable to solve in the pre-computer age. In this subsection, we demonstrate how to apply CSO-MA to find D-optimal designs in such a high dimensional setting and compare the results with the reported optimal designs in Shi et al. (2019).

We give a very brief introduction to GLM below, for a more comprehensive introduction, see Afifi et al. (2011); Dobson and Barnett (2018); McCullagh and Nelder (2019). A GLM (with 5 explanatory covariates $x_1, \cdots, x_5$) assumes that the response y follows an exponential family distribution, i.e.,

$$f(y|x) = h(y) \exp(y\eta - A(\eta))$$

where $A(\eta)$ is the cumulant generating function and η, known as the canonical link, is a linear

function of parameter $\theta^T = (\theta_0. \cdots, \theta_{15})$, i.e.,

$$\eta = \theta_0 + \theta_1 x_1 + \cdots + \theta_5 x_5 + \theta_6 x_1 x_2 + \cdots + \theta_{15} x_4 x_5.$$

For example, the Bernoulli distribution with success rate p corresponds to the canonical link $\eta = \log(\frac{p}{1-p})$, and the resulting model is known as the logistic regression. The Poisson distribution with mean rate λ corresponds to the canonical link $\eta = \exp(\lambda)$, and the resulting model is called the Poisson regression. It is well-known that given a $n \times p$ design matrix $\mathbf{X}$, the Fisher information matrix of the parameter θ is (Dobson and Barnett, 2018)

$$\mathbf{M}(\theta) = \sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^T$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{i5})^T$ is the i^{th} row of the design matrix and $w_i = \frac{\partial^2 A(\eta_i)}{\partial \eta_i^2}, i = 1, 2, \cdots, n$ is the weight associated with $\mathbf{x}_i$. Then a D-optimal design seeks to find a suitable ξ^* with elements

$$\xi^* = \begin{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \end{pmatrix} & \begin{pmatrix} x_{21} \\ x_{22} \\ x_{33} \\ x_{44} \\ x_{55} \end{pmatrix} & \cdots & \begin{pmatrix} x_{n1} \\ x_{n2} \\ x_{n3} \\ x_{n4} \\ x_{n5} \end{pmatrix} \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

such that $\sum_{i=1}^{n} p_i = 0$, $p_i \geq 0$ and

$$\xi^* = \arg\min_{\xi} \log\det \left(\sum_{i=1}^{n} p_i w_i \mathbf{x}_i \mathbf{x}_i^T \right) \tag{3.9.1}$$

Here we note that the number of design points n itself is an unknown parameter that cannot be less than the number of parameters (which is 16 in this case). In practice, we start with a large n and then collapse same design points together later.

To apply CSO-MA solve a high dimensional locally D-optimal design, we use the commands:

We compare the results with the designs found in Table 2 in Shi et al. (2019). Four different models are considered. The first two are logistic regression models with different nominal parameter

values and the next two are Poisson regression models with different nominal parameter values (see Table 1 in Shi et al. (2019)).

Model	CSO-MA	GA	PSO	CSO
Logistic 1	28.88(0.16)	29.54(0.83)	31.05(1.51)	28.80(0.37)
Logistic 2	28.89(0.22)	29.76(1.12)	30.78(1.07)	28.91(0.54)
Poisson 1	-151.67(0.41)	-167.30(1.31)	-163.11(0.92)	-169.04(1.24)
Poisson 2	-100.51(0.32)	-100.14(1.71)	-93.23(1.40)	-100.35(0.64)
Average Runtime	20.1s	95.2s	64.5s	42.3s

Table 3.10: Average criterion values of the locally D-optimal designs found by different algorithms, along with their standard deviations in parentheses.

Table 3.11 shows the D-optimal design found by CSO-MA under the Poisson 2 model. It has only 16 design points, 5 points less than that found in Shi et al. (2019) while still it achieves a high efficiency.

x_1	x_2	x_3	x_4	x_5	p
1.000	-1.000	0.001	1.000	-1.000	0.0629
-1.000	-1.000	-1.000	1.000	-0.505	0.0628
-1.000	1.000	-0.683	-1.000	-1.000	0.0624
-1.000	-0.452	-1.000	1.000	-0.523	0.0625
-1.000	1.000	-1.000	-1.000	-1.000	0.0627
-0.030	-1.000	-1.000	1.000	-0.600	0.0626
-1.000	-0.436	-1.000	1.000	-1.000	0.0627
0.906	-1.000	-1.000	0.222	-1.000	0.0622
-1.000	-1.000	-1.000	1.000	-1.000	0.0622
0.195	-1.000	-1.000	1.000	-1.000	0.0617
-1.000	-1.000	-0.670	1.000	-1.000	0.0627
-1.000	1.000	-1.000	-1.000	0.506	0.0636
-1.000	-0.350	-0.618	1.000	-1.000	0.0627
-1.000	-1.000	-0.611	1.000	-0.419	0.0623
-0.198	1.000	-1.000	-1.000	-1.000	0.0618
-1.000	-1.000	-1.000	0.516	-1.000	0.0622

Table 3.11: A CSO-MA generated 16 point design for Poisson model 2

3.10 A Variable Selection Problem in Ecology

In this subsection, we apply CSOMA to a penalized linear regression problem in ecology. Model selection is essential in much of ecology because ecological systems are often too large and slow-moving for our hypotheses to be tested through manipulative experiments at the relevant temporal and spatial scales (Tredennick et al., 2021).

The data comes from a plateau lake in Yunnan, China, and was collected by a group of researchers at Department of Environmental Engineering, Tsinghua University in 2019. They took the water sample in March (Spring), June (Summer), September (Autumn) and December (Winter). At each time 30 sites were sampled, where 15 sites was from upper water and 15 sites was from bottom water. Due to weather issues at the plateau lake in June, 6 sites were not recorded. Therefore, they collected a water sample of size 114 ($30 \times 4 - 6$). After sampling, they measured key characteristics of water quality, such as the concentration of total Nitrogen (TN), the potential of Hydrogen (pH), etc. The table of measurements of water quality is 114×20 and table 3.12 lists first two samples with 19 measurements.

Sample idx	**Depth**	**Chi-a**	**DO**	**Turbity**	**pH**
1_1D	0.5	34.29	6.1	4.19	9.36
1_1M	0.5	18.36	6.46	15.4	9.47
NH4-N	**NO3-N**	**TN**	**TP**	**TOC**	**TDS**
0.4	0.38	0.96	0.07	22.61	906.7
0.13	0.33	0.96	0.06	22.15	910.4
T	**Ca**	**K**	**Mg**	**Na**	**F**
17.3	7.26	11	68.08	191.38	2.23
16.1	5.02	9.906	68.88	223.35	2.83
CRAP	**16sCRA**				
0.64444	0.361235				
0.0126	0.143714				

Table 3.12: Measurements of water quality.

Cyanobacteria can form dense and sometimes produce algal toxins. The cyanobacteria bloom, which mean the high cyanobacterial density or high proportion of cyanobacteria in phytoplankton, would threaten the aquatic ecosystem function, fisheries and human drinking water safety. Remarkably, the cyanobacterial blooms are increasing in frequency, magnitude and duration glob-

ally (Huisman et al., 2018). The cyanobacteiral bloom is not independant, but is influenced by surrounding environment. To effectively control and prevent the cyanobacterial bloom, one of the most important scientific questions is how other measurements affect CRAP (Cyanobacteria relative abundance in Phytoplankton). High value of CRAP often indicates cyanobacterial bloom. Therefore, if we can control the key factors that are associated with CRAP (or 16sCRA), we can improve environment dramatically.

Linear regression analysis is a default choice for detecting association and outliers. Note that many measurements(covariates) are correlated. For example, NH4-N (Nitrogen in Ammonium) and NO3-N (Nitrogen in Nitrate) are highly correlated with TN (Total concentration of Nitrogen). Thus, in reality, some measurements are more important than others to ecologists. In statistics, Variable selection and penalized regression methods are proposed to address this issue. Hence, we conduct a penalized regression to analyze the data via the CSOMA algorithm.

We denote X as the covariate matrix (i.e., from variable Depth to F in table 3.12) and y as the response vector (variable CRAP in table 3.12). The optimization problem is

$$\min_{\beta} \|y - X\beta\|_2^2 + \rho \left(\sum_{i=1}^{p} P(\beta_j|\lambda, a) \right) \tag{3.10.1}$$

Where ρ is the regularization parameter and

$$P(\beta_j|\lambda, a) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq \lambda \\ \frac{a\lambda|\beta_j| - \beta_j^2 - \lambda^2}{a-1} & \text{if } \lambda < |\beta_j| \leq a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{if } |\beta_j| > a\lambda \end{cases}$$

is a differentiable but non-convex function. First, we standardize each column of X by subtracting its mean and dividing its standard deviation so that each column of X has mean 0 and standard deviation 1. This step is crucial because we want to analyze the relative influence of variables on CRAP and different scales cause confusion. Next, we perform SCAD regression (Fan and Li, 2001) on our data (X, y) for different choices of ρ (see formula (3.10.1)) and optimize it using PSO algorithm. We set 12 different values for ρ, i.e., $10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1, 10, 100$. For each ρ, we record the best particle position found by CSOMA as our estimation for β. The CSO-MA algorithm is initialized with 25 particles and iterates 100 times (i.e., 100

function evaluations). We run the algorithm 50 times for each ρ to analyze the stability of CSO-MA. For illustration purpose, we demonstrate the average and standard deviation of the 50 runs when $\rho = 0.025$ and the results are shown in Table 3.13; further, the average minimum of (3.10.1) when $\rho = 0.025$ is 0.315 with a standard deviation of 0.0009 (the other ρ's have similar standard deviation and minimum values), suggesting the stability of CSO-MA algorithm.

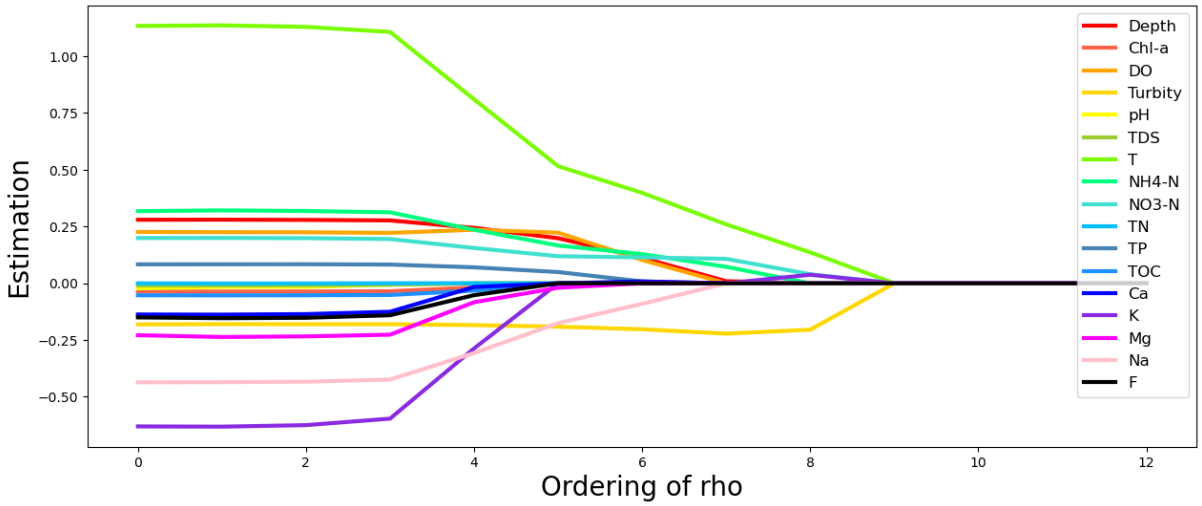
Variable	Average	Standard deviation	Variable	Average	Standard deviation
Depth	0.191	0.012	NO3-N	0.128	0.013
Chl-a	-0.001	0.003	TN	0.000	0.002
DO	0.219	0.015	TP	0.047	0.008
Turbity	-0.195	0.016	TOC	-0.001	0.003
pH	-0.003	0.012	Ca	0.003	0.007
TDS	0.000	0.002	K	-0.002	0.009
T	0.499	0.033	Mg	-0.031	0.027
NH4-N	0.166	0.018	Na	-0.162	0.025
F	-0.016	0.028			

Table 3.13: Average and standard deviation of parameter estimation after 50 times of runs.

Figure 3.7 illustrates the solution path of SCAD using the CSO-MA algorithm. The x-axis represents the scaled ρ values. When ρ decreases from 100, estimation of turbidity (T) deviates from 0 at first. It suggests that turbidity is one of the most important measurements associating with the level of DRAP. One possible reason for such phenomenon is that the turbid water prevents light from penetrating, which in turn indicates a lower amount of the algae carrying out photosynthesis. Further, temperature (T) is another variable deviating from 0 at first. The reason is that the optimum temperature for algae growth is $20 + \ C^{\circ}$ and the lower the temperature, the less active the metabolism of algaeis. In addition, when ρ decreases from 0.05 to 0.01 (x from 7 to 5), parameter estimation for chemical elements, such as K, Mg, Na, all deviates from 0, suggesting that the concentration of chemical elements has slightly different association of CRAP.

This subsection shows CSO-MA can be usefully applied along with SCAD penalized regression to explore association among different components of water quality and how that affect the outcome CRAP. The interpretation of the solution path is in line with scientific common sense.

Figure 3.7: Solution path of SCAD using CSO-MA. Each line represents the trajectory of an estimated coefficient for a predictor variable across the ordered values of the regularization parameter ρ. The y-axis denotes the estimated coefficient values. The x-axis corresponds to the ordinal position of each ρ value in the set $10^{-6}, 10^{-5}, \ldots, 100$, which have been rescaled to 1, 2, ..., for clarity of presentation.



3.11 Parameter Tuning of LASSO Regression

In the previous section, we conduct a penalized regression to select variables. However, the choice of the regularization parameter (or tuning parameter) ρ is selected by-hand. In this section, we focus on another type of penalized regression known as the LASSO regression and discuss how to apply metaheuristics to choose the tuning parameter. In statistical literature, many methods for empirically choosing ρ given a fixed data set have been proposed. These methods can be lumped into three various categories (Homrighausen and McDonald, 2018):

- **GIC-based approach**: Proposed by Schwarz (1978), BIC was originally designed for regression models with unpenalized MLEs, lacking motivation in the penalized likelihood setting. Hence, several modifications of BIC are proposed to address the tuning parameter selection problem (Wang et al., 2009; Wang and Zhu, 2011; Gao et al., 2012; Kwon et al., 2017). Importantly, it has been proved that the extended BIC (EBIC) is selection consistent in both linear models and GLMs (Luo and Chen, 2013). Hui et al. (2015) proposed the extended regularized information criterion (ERIC) for choosing the tuning parameter in adaptive Lasso regression. In addition, Flynn et al. (2013) investigates a variety GIC-based methods with

increasing dimensions.

- **Resampling procedures**: Cross-validation (Allen, 1974) is one of the most common techniques for choosing tuning parameters (Zou et al., 2007). Yu and Feng (2014) proposed a modified version of cross-validation (MCV) which corrects the bias introduced by the LASSO regression. Hall et al. (2009) proposed an m-out-of-n bootstrap algorithm, pointing out that standard bootstrap methods would fail for the LASSO regression. Later, Chatterjee and Lahiri (2011) proposed a modified bootstrap algorithm for LASSO. Based on their theoretical and empirical results, they suggested choosing the tuning parameter that minimizes the bootstrapped approximation to the mean-squared error of the LASSO estimator.

- **Reformulations of the LASSO**: Several alternatives to the original LASSO formulation has been proposed. Some theoretical developments show that the consistency of LASSO estimator requires the knowledge of the variance parameter in the linear model Bickel et al. (2009). Based on this, Sun and Zhang (2012) proposed the scaled LASSO which jointly estimates the regression coefficients and the variance parameter in a sparse linear regression model; Chichignoud et al. (2016) proposed a novel adaptive validation method for tuning parameter selection for LASSO. Belloni and Chernozhukov (2011); Belloni et al. (2011) proposed the square-root lasso, which is also a variant of LASSO that avoids calibrating the tuning parameter with respect to the noise parameter. To further alleviate the problem of parameter tuning, Lederer and Müller (2015) proposed the TREX estimator which standards for tuning-free (T) regression (R) that adapts to the entire (E) design matrix (X). Wang et al. (2020a) proposed the rank LASSO which can be easily simulated and automatically adapts to both the unknown random error distribution and the structure of the design matrix. Other sub-splitting and reformulations are also proposed and are well-summarized in Wu and Wang (2020).

In this subsection, we show that PSO generates reasonable results for the optimal tuning parameter compared with standard statistical packages. Using the same notations in Section 3.10, we can write

the optimization problem of the tuning parameter selection in LASSO regression as

$$\widehat{\lambda} = \arg\min_{\lambda} \mathbb{E}\left(y_{new} - x_{new}^T \widehat{\beta}_\lambda\right)^2 \tag{3.11.1}$$
$$\widehat{\beta}_\lambda = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where x_{new} and y_{new} are new test points and the expectation is taken over the joint distribution of (x_{new}, y_{new}). In practice, the true joint distribution is never known to us, so we have to estimate the expectation. One general approach is to split our data into training and testing datasets, i.e., if X is an $n \times p$ matrix and y is an $n \times 1$ vector, then we split them into (X_{train}, y_{train}) and (X_{test}, y_{test}) so that $X = X_{train} \cup X_{test}$ and $y = y_{train} \cup y_{test}$. Then $\widehat{\beta}_\lambda$ is estimated from the training dataset and the expectation is taken w.r.t. to the empirical distribution induced by the test dataset. To further alleviate the uncertainty caused the random split of training and testing datasets, we can split (X, y) into K disjoint datasets and treat each of them as both training and testing datasets. Then we estimate the expectation via

$$\widehat{\mathbb{E}}\left(y_{new} - x_{new}^T \widehat{\beta}_\lambda\right)^2 = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k} \|y_{test,k} - X_{test,k}\widehat{\beta}_{\lambda,k}\|_2^2 \tag{3.11.2}$$

where $(X_{test,k}, y_{test,k})$ is the k^{th} dataset and n_k is its corresponding sample size (usually this number is the same across datasets so that $n_k = n/K$, in our experiments, it is always the case), and $\widehat{\beta}_{\lambda,k}$ is the LASSO estimate of β using the k^{th} training dataset (i.e., the whole dataset without $(X_{test,k}, y_{test,k})$). The choice of K usually depends on the sample size of the dataset; a rule of thumb is to set $K = 10$ when sample size is large and $K = 5$ when the sample size is small. In practice, it also depends on the computational resources available since as K increases, the required computational power increases drastically. We note that Anguita et al. (2012) discusses the choice of K in detail; Marcot and Hanea (2021) presents a comprehensive simulation study on choosing K and they recommend the value $K = 5$ or 10 is sufficient in most applications. We set $K = 10$ in our Table 3.14. In summary, we have transformed the problem of tuning parameter selection in

LASSO into the following form:

$$\min_{\lambda} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\|y_{test,k} - X_{test,k}\widehat{\beta}_{\lambda,k}\|_2^2 \tag{3.11.3}$$

$$s.t.\ \widehat{\beta}_{\lambda,k} = \arg\min_{\beta}\|y_{train,k} - X_{train,k}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{3.11.4}$$

where $y_{train,k}$ is the union of $y_{test,k'}$ and $X_{train,k}$ is the union of $X_{test,k}$ for $k' \neq k$. Mathematically, let $(X_{train,k}, y_{train,k})$ and $(X_{test,k}, y_{test,k})$ be the k^{th} splitted training and testing datasets, respectively. Then we could write

$$y = y_{train,k} \cup y_{test,k} = \bigcup_{k=1}^{K} y_{test,k},\ \text{and}\ X = X_{train,k} \cup X_{test,k} = \bigcup_{k=1}^{K} X_{test,k}$$

for any $k = 1, 2, \cdots, K$. Although the constraint of the above optimization problem is not "restricting β to a specific region", we can still apply metaheuristics in this case using a two-stage approach: fix a λ, we estimate $\widehat{\beta}_{\lambda,k}$ and then vary λ to minimize the objective function. In this scenario, particles are one-dimensional, i.e., particles correspond to the tuning parameter λ. This two-stage approach is applicable to any type of penalized regression, not just specific to the LASSO regression. For instance, we can replace $\|\beta\|_1$ with the SCAD penalty in Section 3.10 or the elastic net penalty (Zou and Hastie, 2005) and the procedure for applying metaheuristics stays the same. The Elastic net penalty has two tuning parameters (λ, α) and is of the following form

$$\lambda\left(\alpha\|\beta\|_1 + \frac{(1-\alpha)}{2}\|\beta\|_2^2\right)$$

where λ, as usual, represents the regularization strength and $\alpha \in [0, 1]$ controls the trade-off between Lasso and Ridge penalty. When $\alpha = 0$, Elastic net reduces to the Ridge penalty and when $\alpha = 1$, Elastic net coincides with the LASSO penalty.

Our proposed strategy for tackling tuning parameters problem in penalized regression is to formulate it into one that PSO can solve directly, as follows:

- Choose a penalty such as LASSO or SCAD of user's own interest; choose K, the number of cross-validation folds.
- Choose an objective function such as mean squared loss (MSE) for regression or the receiver

operating characteristic (ROC) curve for classification. Then given the parameter estimate based on the penalty, tuning parameter(s) and K, we can estimate the expectation of the objective function (e.g., Equation 3.11.2 for regression).

- Run PSO to minimize the estimate of the expectation by finding the best tuning parameter(s).

We apply our proposed strategy to the same ecology data in Section 3.10 and another *Hitters* data (James et al., 2013) with LASSO, Ridge and Elastic net penalties and compare the results with the standard R package *glmnet* (Hastie et al., 2021) in Table 3.14 (using $K = 10$). Because the split of CV is random, so we need to perform multiple rounds (in this case, 30) of estimation to get a stable estimation of the best tuning parameter(s). The values in the braces are standard errors of tuning parameter(s). The *glmnet* package does not provide options for tuning both parameters (λ, α) in Elastic net but users have to specify α on their own so that our proposed strategy is more general. In practice, one can define grids for tuning parameters and then use *glmnet* to fit the models, such functionality is implemented in the *caret* package (Kuhn et al., 2020). We set the hyperparameters of PSO to be $c_1 = 1.2, c_2 = 0.5, w = 0.9$ and number of iteration is 30 (which is more than sufficient to converge). For LASSO and Ridge penalties, the best tuning parameters for LASSO found by PSO and *glmnet* are quite similar but quite different for Ridge. We note that PSO seems to be more robust compared with *glmnet*. We also perform additional simulation and real data studies using different K values, the results are consistent with the Table 3.14.

The Ecology Data		
Penalty	PSO	*glmnet*
LASSO	0.026 (0.001)	0.027 (0.003)
Ridge	0.116 (0.001)	0.409 (0.242)
Elastic net (λ, α)	0.0098 (0.001), 0.0122 (0.024)	NA

The Hitters Data		
Penalty	PSO	*glmnet*
LASSO	0.0059 (0.001)	0.0063 (0.006)
Ridge	0.113 (0.004)	0.092 (0.113)
Elastic net (λ, α)	0.0082 (0.003), 0.1021 (0.203)	NA

Table 3.14: Comparison of PSO-generated tuning parameter and the standard pacakge *glmnet* using K=10.

Under the choice $\lambda = 0.026$, the parameter estimates of the LASSO regression using the ecology

dataset in given in Table 3.15. In the same table, we also compare the estimates with $\lambda = 0.1$, a randomly selected tuning parameter, to illustrate that different tuning parameters result in distinct parameter estimates, thereby leading to varied scientific conclusions. For instance, $\lambda = 0.026$ suggests that K and Mg have negative effect on the CRC while $\lambda = 0.1$ suggests that such effect is negligible.

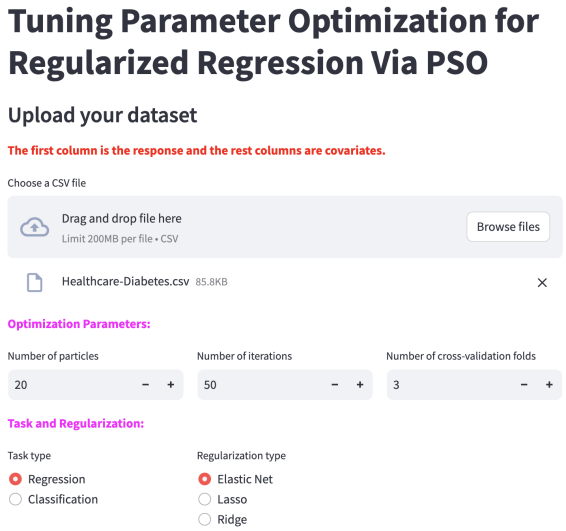
Variable	$\lambda = 0.026$	$\lambda = 0.1$	Variable	$\lambda = 0.026$	$\lambda = 0.1$
Depth	0.191	0.009	NO3-N	0.000	0.000
Chl-a	-0.001	0.000	TN	0.046	0.000
DO	0.213	0.000	TP	0.000	0.000
Turbity	-0.192	-0.220	TOC	0.000	0.000
pH	-0.000	0.000	Ca	0.000	0.000
TDS	0.000	0.000	K	-0.018	0.000
T	0.506	0.260	Mg	-0.168	0.000
NH4-N	0.163	0.073	Na	0.000	0.000
F	0.119	0.108			

Table 3.15: Comparison of LASSO estimates under different λ using the ecology data.

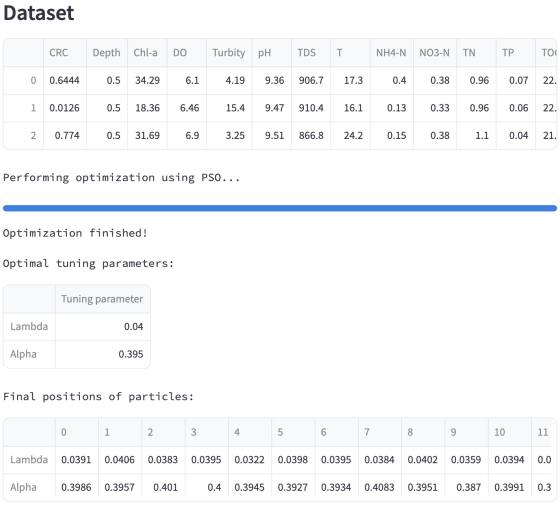
We also created a Python App to illustrate the tuning parameter problem for different penalties via PSO (Figure 3.8) and is publicly available at https://pso-parameter-tuning.streamlit.app/. The App proffers an intuitive interface, enabling users to upload their dataset, with the first column designated as the response variable and the subsequent columns as covariates. Users can tailor the optimization process by selecting the type of regularization, the type of task (regression or classification), the number of particles in the swarm, the number of iterations for convergence, and the folds for cross-validation evaluation (Left panel of Figure 3.8). The PSO algorithm iteratively updates the positions of the particles, representing potential solutions, through the problem space by following the current optimum in a quest to discover the most propitious tuning parameters for the regularized regression model. The best tuning parameters are then printed in the screen together with the final positions of all particles (Middle panel of Figure 3.8). A contour plot further illustrates the swarm's trajectory through the iterations, offering a graphical elucidation of the optimization process and the convergence of the swarm towards the optimal values (Right panel of Figure 3.8).

In conclusion, we have shown that (i) choice of an optimal or appropriate tuning parameter is important; otherwise, one can arrive at different estimate of risks, (ii) using the Ecology and

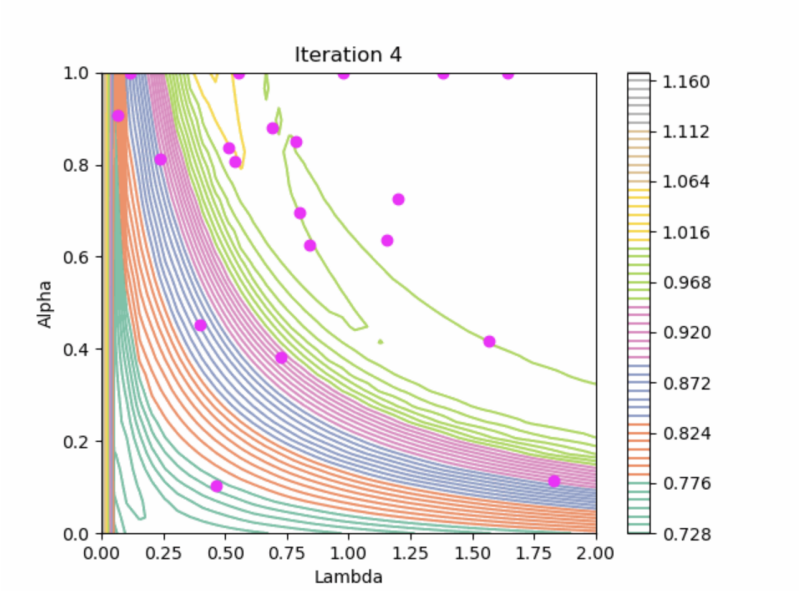
the Hitters example, we demonstrate PSO can provide comparable tuning parameters in terms of standard deviation than current methods of choosing tuning parameters, (iii) the regression estimates can depend on the tuning parameter sensitively and Ridge regression produces a higher tuning parameter compared with LASSO, and (iv) a Python App is available for users who are interested in both parameter tuning topics and applications of metaheuristics (PSO in particular).



(a) Beginning of the App.



(b) Loading data and optimizing.



(c) Visualization of PSO.

Figure 3.8: Illustration of the Python App for tuning parameter optimization.

3.12 A Two-factor Quasi-sequential Design

In this section, we implemented a two-factor quasi-sequential D-optimal designs defined in section 6.3. The fixed points are $(-2,-2),(10,0)$ and $(0,10)$ which correspond to the control group, death level for dose 1 and death level for dose 2, respectively. The pre-specified weight α is set to 0.2 and we restrict the design space to $[-1,8]\times[0,7]$. The model is

$$\begin{aligned} y \sim \mathcal{M}(n,\pi),\ \ \pi &= (\pi_1,\pi_2,\pi_3)^T \\ \log\frac{\pi_1}{\pi_2+\pi_3} &= \eta_1 = \beta_1+\alpha_1 x_1+\alpha_2 x_2 \\ \log\frac{\pi_1+\pi_2}{\pi_3} &= \eta_2 = \beta_2+\alpha_1 x_1+\alpha_2 x_2 \\ \log(\pi_1+\pi_2+\pi_3) &= \eta_3 = 0. \end{aligned}$$

with parameter $\theta=(\beta_1,\beta_2.\alpha_1,\alpha_2)^T$. We put estimation $\widehat{\theta}=(-1.8,3.8,-0.5,-0.4)^T$ to find the quasi-sequential locally D-optimal design. By PSO, the optimal design is

$$\xi_{\text{quasi}-D}=\begin{pmatrix} \begin{pmatrix}-1\\0\end{pmatrix} & \begin{pmatrix}8\\0\end{pmatrix} & \begin{pmatrix}-1\\7\end{pmatrix} & \begin{pmatrix}4\\0\end{pmatrix} & \begin{pmatrix}-2\\-2\end{pmatrix} & \begin{pmatrix}10\\0\end{pmatrix} & \begin{pmatrix}0\\10\end{pmatrix} \\ 0.216 & 0.212 & 0.272 & 0.100 & 0.067 & 0.067 & 0.066 \end{pmatrix} \tag{3.12.1}$$

and the sensitivity surface is shown in figure 3.9, which indicates that the optimal design is indeed a 4-point design in equation 3.12.1. Hence, PSO has successfully derived a two-factor quasi-sequential D-optimal design.

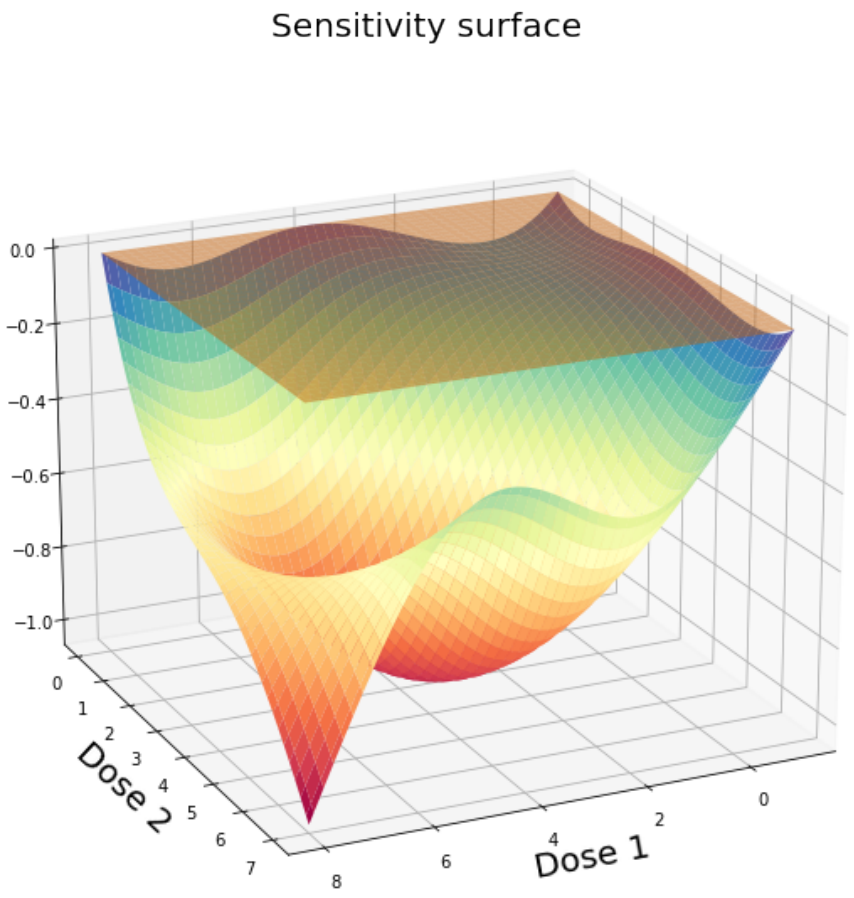


Figure 3.9: Sensitivity surface of the two-factor quasi-sequential D-optimal design.

3.13 A Metric-based Principal Curve Approach for Learning One-dimensional Manifold

Principal curve (Hastie, 1984) is a well-known statistical method oriented in manifold learning using concepts from differential geometry. In this paper, we propose a novel metric-based principal curve (MPC) method that learns one-dimensional manifold of spatial data. Synthetic datasets Real applications using MNIST dataset show that our method can learn the one-dimensional manifold well in terms of the shape. This section is based on Cui and Shao (2024).

3.13.1 A Brief Review on Differential Geometry in Statistics

The application of differential geometry in statistics should be credit to two great Indian statisticians Mahalanobis (2018) (the original paper was published in 1936) and Rao (1945). One of the early pioneer papers of differential geometry in statistics was done by Efron (1975). He defined the concept statistical curvature rigorously for the first time. Following his work, Skovgaard (1984) studied the Riemannian geometry of a family of multivariate normal models in depth. Some other early works include Efron (1978); Atkinson and Mitchell (1981); Amari (1982); Campbell (1985); Barndorff-Nielsen et al. (1986) and Ravishanker et al. (1990). Two excellent monographs in connecting differential geometry and statistics are Murray and Rice (1993) and Amari (2006). In addition, the concept of *tangent space* is borrowed from differential geometry and deeply studied in Bickel et al. (1993). Apart from the conventional work mentioned above, there are also much more interesting and exciting progress going on in bridging these two fields (which are under the fancy names *metric learning* (Izenman, 2008), *topological data analysis* (Rabadán and Blumberg, 2019), *directional statistics* (Mardia et al., 2000), *shape analysis* (Bhattacharya and Bhattacharya, 2012; Dryden and Mardia, 2016) and *functional data analysis* (Wang et al., 2016) , etc.). In most of these works, data are not assumed to be sampled from a regular Euclidean space (say, $\mathbb{R}^d$) but a smooth low-dimensional manifold embedded in $\mathbb{R}^d$ (Do Carmo and Flaherty Francis, 1992; Tapp, 2016). To the best of our knowledge, it was Hastie (1984) who first defined such a manifold in a statistical setting and he named it as the *principal curve*. His idea was later developed in Hastie and Stuetzle (1989) and Tibshirani (1992) and more recently in Ozertem and Erdogmus (2011). We list some of other applications of differential geometry methods in statistics in table 3.16. For

a more comprehensive review, we suggest the article written by Wasserman (2018).

Method / Approach [1]	**Reference**
Principle curves and surfaces	Hastie and Stuetzle (1989)
Kernel principal component analysis (K-PCA)	Schölkopf et al. (1997)
Local linear embedding (LLE)	Roweis and Saul (2000)
Isometric feature mapping (ISOMAP)	Tenenbaum et al. (2000)
Laplacian eigenmap	Belkin and Niyogi (2003)
Hessian eigenmaps	Donoho and Grimes (2003)
Intrinsic dimension estimation (IDE)	Levina and Bickel (2004)
Principal geodesic analysis (PGA)	Fletcher et al. (2004)
Intrinsic statistics on Riemannian manifolds	Pennec (2006)
Nonparametric regression on Riemannian manifolds	Pelletier (2006)
The diffusion maps	Coifman and Lafon (2006)
Shape-space smoothing	Kume et al. (2007)
Mapper algorithm	Singh et al. (2007)
Riemannian K-means	Goh and Vidal (2008); Zhang (2020a)
Locally defined principal curves	Ozertem and Erdogmus (2011)
Computation of Vietoris-Rips filtration	Sheehy (2012)
Probablistic principal geodesic analysis (P-PGA)	Zhang and Fletcher (2013)
Geodesic mixture models (GMM)	Simo-Serra et al. (2017)
Geodesic convolutional neural network (G-CNN)	Masci et al. (2015)
Dirichlet process mixture on spherical manifold	Straub et al. (2015)
Locally adaptive normal distribution (LAND)	Arvanitidis et al. (2016)
Uniform manifold approximation and projection	McInnes et al. (2018)
Statistical inference on Lie groups	Falorsi et al. (2019)
Fréchet regression	Petersen and Müller (2019)
Geometrically enriched latent space approach	Arvanitidis et al. (2020)
Wasserstein regression	Chen et al. (2021); Matabuena et al. (2021)
Pseudotime analysis	Cui et al. (2022)
Invertible kernel PCA (IK-PCA)	Gedon et al. (2023)

Table 3.16: Applications of Differential Geometry in Statistics. [1] This table only includes methods that use concepts or tools from differential geometry. Hence, some other popular manifold learning techniques are not included, such as local discriminant analysis (Hastie and Tibshirani, 1995), random projection (Johnson and Lindenstrauss, 1984), and t-SNE (Van der Maaten and Hinton, 2008).

We note that many methods listed above are published in non-statistical journals. Hence, it is urgent (and also a necessity) for statisticians to develop concepts, tools and methods that can adapt today's societal needs. Further, the *Geomstats* package (Miolane et al., 2020a,b), the *Geoopt* package (Kochurov et al., 2020) and the *Pymanopt* package (Townsend et al., 2016) in Python

and the *umap* package, the *geomorph* package (Adams and Otárola-Castillo, 2013) and the *frechet* package in R (Konopka and Konopka, 2018) provides many useful computational tools for the differential geometry methods mentioned in the table.

The rest of the section is organized as follows. In section 3.13.2, we propose a new approach for learning one-dimensional representation of data and term it as the metric-based principal curve. Simulation studies using synthetic datasets are presented in section 6.5 and real applications using the MNIST dataset is given in section 3.13.4. In appendix 7.2.1, we provide some preliminaries in Riemann geometry.

3.13.2 Metric-based Principal Curve

In this section, we propose a new algorithm for learning a one-dimensional representation of data. We term it the *metric-based principal curve* as it minimizes a metric distance between the raw data and the projected data using smoothing and regression techniques. We give a formal and rigorous introduction below.

Definition 3.13.1 (Principal curve assumption)**.** Let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_p)^T \in \mathbb{R}^p$ be a p-dimensional random vector and $\lambda \in \mathbb{R}$ be a scalar known as the **projection index**. Extending the idea in Tibshirani (1992), the **principal curve assumption** for $(\mathbf{Y}, \lambda)$ is described as follows:

$$Y_j|\lambda \sim_{ind} \mathbb{P}_{Y_j|\lambda}, j = 1, \cdots, p, \tag{3.13.1}$$

$$\mathcal{P}(\lambda) = \mathbb{E}(\mathbf{Y}|\lambda) = \int_{\mathbb{R}^p} y \, \mathbb{P}_{\mathbf{Y}|\lambda}(dy) \tag{3.13.2}$$

where $\mathbb{P}$ represents a general probability measure and the expectation $\mathcal{P}(\lambda)$ is defined as the **principal curve** of $\mathbf{Y}$. Alternatively, using the idea of nonparametric regression, we can also assume that

$$Y_j = f_j(\lambda) + \epsilon_j, \ j = 1, 2, \cdots, p$$

where $f_j(\cdot)$ is a smooth function and ϵ_j is a random error.

Note that the above definition is in the population level, i.e., we have the oracle knowledge $\mathbb{P}_{\mathbf{Y}|\lambda}$. However, in practice, we do not have the access to it and have to estimate either $\mathbb{P}_{\mathbf{Y}|\lambda}$ or f_i from data as well as the projection index λ. There are several projection-expectation (PE)

type algorithms to learn the principal curve $\mathcal{P}(\lambda)$ (Hastie and Stuetzle, 1989; Chang and Ghosh, 1998). However, it is well-known that PE algorithms do not guarantee convergence and different types of definition and algorithm lead to different estimated principal curve (Gerber and Whitaker, 2013). Hence, it is of great interest to develop new algorithms to estimate principal curves. In the following, we propose a metric-based algorithm for learning principal curves and term the estimator **the metric-based principal curve** (MPC).

The idea of MPC starts with a user-specified metric $d(\cdot, \cdot)$ on $\mathbb{R} \times \mathbb{R}$ and a regularization parameter ρ. Suppose we observe the data $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ where each $\mathbf{Y}_i \in \mathbb{R}^p$. The associated projection index $\lambda_1, \cdots, \lambda_n$ is chosen such that the following quantity is minimized.

$$\{\lambda_i\}_{i=1}^n = \arg\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} d(\mathbf{Y}_i, \widehat{\mathbf{Y}}(\lambda_i)) + \rho\ \phi(\{\lambda_i\}_{i=1}^n) \tag{3.13.3}$$

where $\widehat{\mathbf{Y}}(\lambda_i) = (\widehat{f}_1(\lambda_i), \widehat{f}_2(\lambda_i), \cdots, \widehat{f}_p(\lambda_i))^T$ is the fitted value of $\mathbf{Y}_i$ using *under-smooth regression models* $\{f_j\}_{j=1}^p$ and ϕ is a dispersion function characterizing the dispersion of λ's. We provide a list of the smoother f_j, the metric $d(\cdot, \cdot)$, and the dispersion ϕ in table 3.17.

Table 3.17: Some choices of f_j, $d(\cdot, \cdot)$ and $\phi(\cdot)$.

f_j	$d(x, y)$	ϕ *
Smoothing spline	L_d-distance $\|x - y\|_d$	$\sum_{i=1}^{n-1} \lvert\lambda_{(i+1)} - \lambda_{(i)}\rvert$
LOWESS	Mahalanobis distance	$\sum_{i=1}^{n-1} \left(\lambda_{(i+1)} - \lambda_{(i)}\right)^2$
Kernel ridge regression	Chebyshev distance	$\max_i \lvert\lambda_{(i+1)} - \lambda_{(i)}\rvert$
Gaussian process regression	Hellinger distance	Coefficient of variation
Support vector regression	Canberra distance	
Nadaraya-Watson estimator		

* Here $\lambda_{(i)}$ denotes the i^{th} order statistics from $\lambda_1, \cdots, \lambda_n$.

In short, a metric-based principal curve minimizes the mean of distances of all points (feature vectors) projected onto the curve plus a regularization term that penalizes the dispersion of the on-dimensional parameter.

3.13.3 Simulation Studies

In this section, we perform simulation studies based on three synthetic datasets, namely, spiral curve, golden bridge and Arabic numerals seven.

- **Number Seven**

  The generative model is

$$
\begin{aligned}
Y_1 &= t + \epsilon_1 \\
Y_2 &= U_2^X (1+\epsilon_2)^{1-X} \\
Y_3 &= (1+\epsilon_3)^X U_3^{1-X}
\end{aligned}
$$

  where $t \in [0,1]$, $U_2 \sim \mathcal{U}(0,1)$, $U_3 \sim \mathcal{U}(-2,0.7)$, $\epsilon_i \sim_{iid} \mathcal{N}(0,0.1)$ and $X \sim \text{Ber}(0.5)$. For simulation, we take the sample size to be 120 and generate t uniformly spaced within $[0,1]$. For estimation of λ's, we set f_j to be smoothing splines, $d(x,y)$ to be L_2-distance and $\phi(\lambda)$ to be $\sum_{i=1}^{n-1} |\lambda_{(i+1)} - \lambda_{(i)}|$. For prediction, we set f_j to be LOWESS with bandwidth 0.4. The results are shown in the left panel of figure 3.10. In addition, we also fit another MPC for (Y_2, Y_3) only since Y_1 is just a linear function in t. The results are shown in the left panel of figure 3.11.

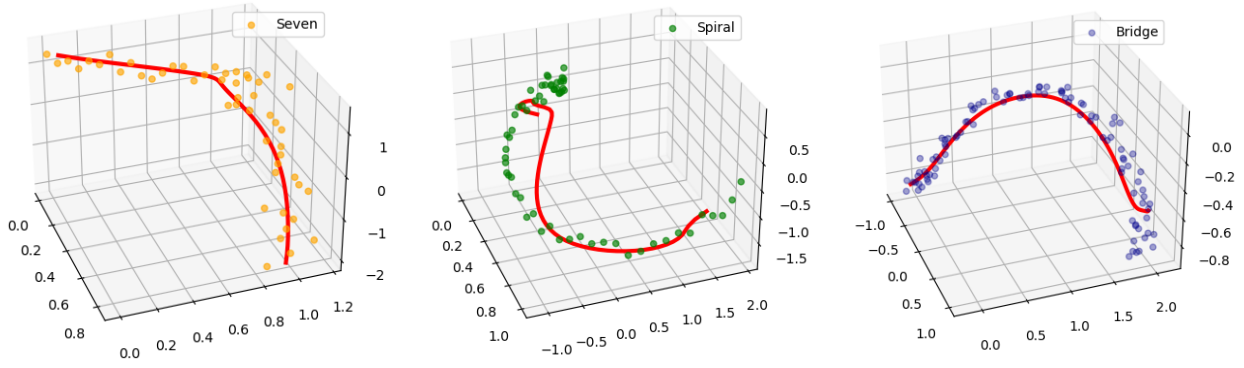


Figure 3.10: Principal curves of seven, spiral and bridge in $\mathbb{R}^3$. Red lines are learned principal curves which represent the trajectory of data manifold.

- **A Spiral Curve**

The generative model is

$$
\begin{aligned}
Y_1 &= t \\
Y_2 &= 2t\cos(6t) + \epsilon_2 \\
Y_3 &= 2t\sin(6t) + \epsilon_3
\end{aligned}
$$

where $\epsilon_i \sim_{iid} \mathcal{N}(0, 0.1)$ and $t \in [0, 1]$. For simulation, we take the sample size to be 120 and generate t uniformly spaced within $[0, 1]$. For estimation of λ's, we set f_j to be LOWESS with bandwidth 5, $d(x, y)$ to be L_2-distance and $\phi(\lambda)$ to be $\sum_{i=1}^{n-1} |\lambda_{(i+1)} - \lambda_{(i)}|$. For prediction, we set f_j to be LOWESS with bandwidth 0.4. The results are shown in the middle panel of figure 3.10. In addition, we also fit another MPC for (Y_2, Y_3) only since Y_1 is just a linear function in t. The results are shown in the middle panel of figure 3.11.

- **The Golden Bridge**

  The generative model is

$$
\begin{aligned}
Y_1 &= t \\
Y_2 &= \sin(2t) + \cos\left(\frac{2}{3}t\right) + \epsilon_2 \\
Y_3 &= -t\sin(2t) + \epsilon_3
\end{aligned}
$$

  where $\epsilon_i \sim_{iid} \mathcal{N}(0, 0.1)$ and $t \in [0, 1]$. For simulation, we take the sample size to be 120 and generate t uniformly spaced within $[0, 1]$. For estimation of λ's, we set f_j to be kernel regression with regularization parameter $\alpha = 10$, $d(x, y)$ to be L_2-distance and $\phi(\lambda)$ to be $\sum_{i=1}^{n-1} |\lambda_{(i+1)} - \lambda_{(i)}|$. For prediction, we set f_j to be smoothing splines. The results are shown in the right panel of figure 3.10. In addition, we also fit another MPC for (Y_2, Y_3) only since Y_1 is just a linear function in t. The results are shown in the right panel of figure 3.11.

3.13.4 Applications to MNIST data

In this section, we apply MPC to the famous MNIST dataset (LeCun et al., 1998). We first sample 150 figures for each handwritten digit from the training set. Each figure can be viewed as a point

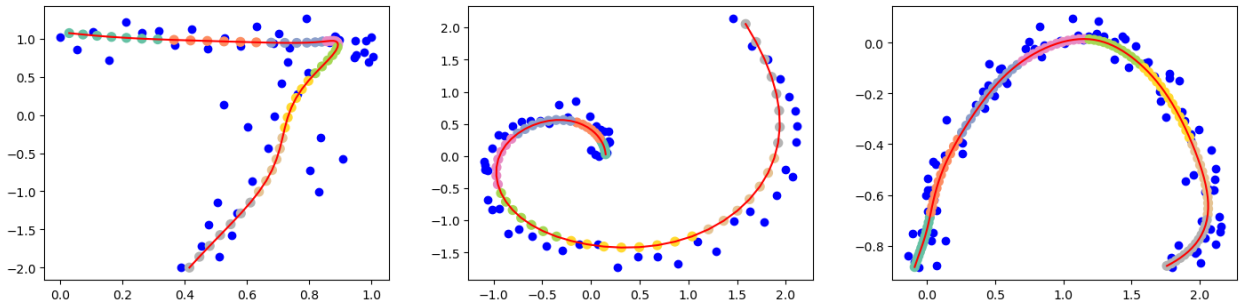


Figure 3.11: Principal curves of seven, spiral and bridge in $\mathbb{R}^2$. Red lines and colorful points are learned principal curves which represent the trajectory of data manifold. Blue points are raw data in $\mathbb{R}^2$.

living in a $28 \times 28 = 784$ dimensional space. Next, we apply uniform manifold approximation and projection (UMAP) algorithm to each digit (150 figures) so that each figure is projected onto $\mathbb{R}^3$. Then we perform an MPC analysis to the projected 3-dimensional data for each digit. We visualize the principal curves of all ten digits from in Figure 3.12 for illustration.

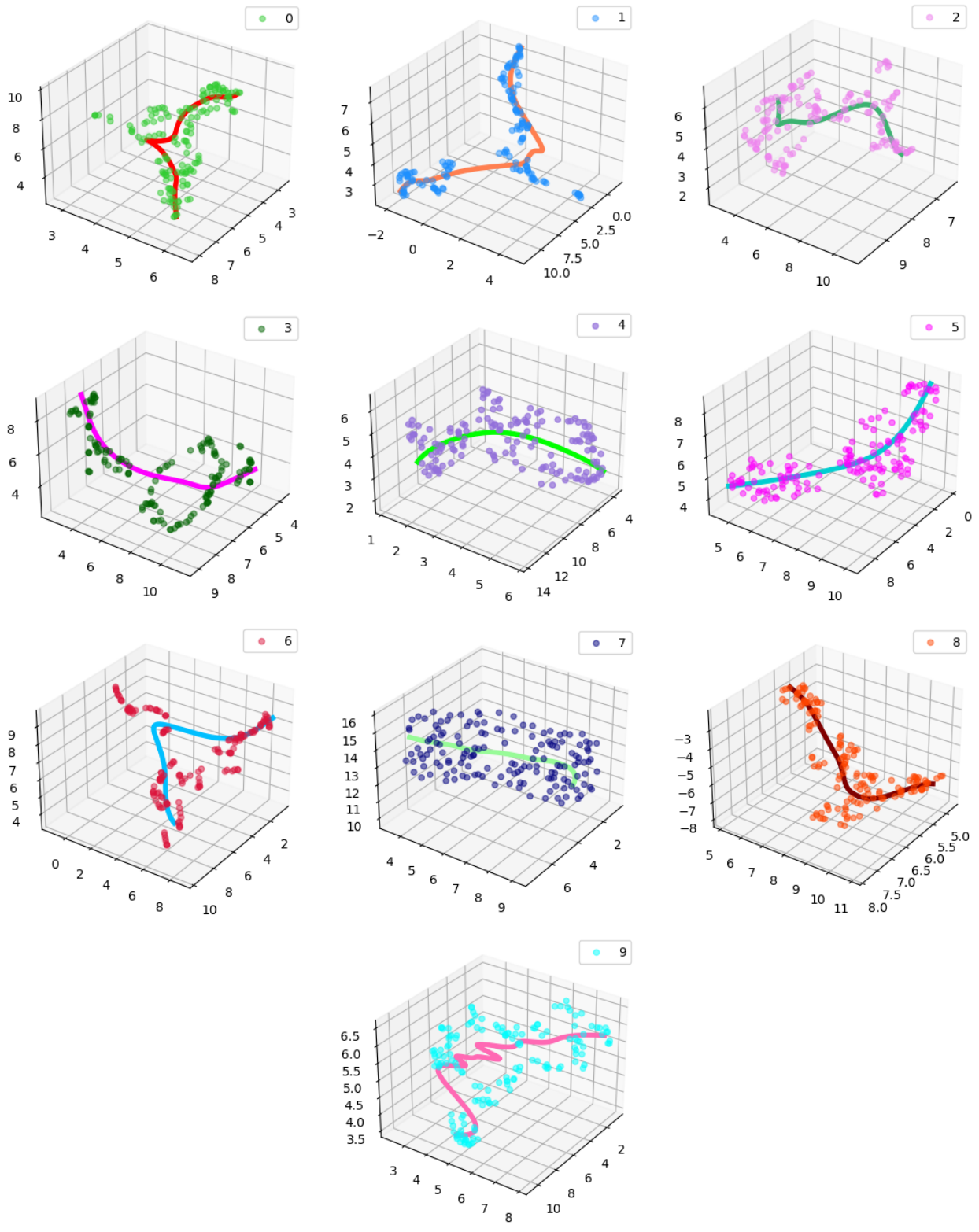


Figure 3.12: Principal curves of MNIST. Blue lines are learned principal curves which represent the trajectory of data manifold.

3.14 Parameter Estimation in Hawkes Process Models

In this section, we demonstrate the potential usage of metaheuristics to Hawkes process models, which has been applied in epidemiology to model the progress of infectious disease such as COVID-19 Rizoiu et al. (2018); Garetto et al. (2021). Hawkes (1971), Hawkes process is a self-exciting point process that can be applied to model the dynamics of disease progression and is proposed in 1971 Hawkes (1971). Different algorithms have been proposed to simulate a Hawkes process, see e.g., Ogata (1981, 1998); Daley et al. (2003); Møller and Rasmussen (2005); Laub et al. (2021). Briefly speaking, a Hawkes process is an extension of the non-homogeneous Poisson process such that the intensity increases as more events happen. Mathematically, let $N(t), t \geq 0$ denote the number of infected patients at time t, then we assume that the intensity of $N(t)$ satisfies Hawkes (1971)

$$\lambda(t) = \nu + \int_0^t g(t-u)dN(u) = \nu + \sum_{t_i<t} g(t-t_i) \tag{3.14.1}$$

where $\nu > 0$ is a constant parameter of a Poisson process, t_i's are jump times of $N(t)$ and g is known as the triggering function. We take it to be exponential, i.e., $g(x) = \alpha \exp(-\beta x)$ where $\beta > \alpha > 0$ are hyper-parameters. One can think of $\lambda(t)$ in the following way: at the beginning, the spread of COVID-19 is at a low speed; as more and more patients are infected, the spread of COVID-19 speeds up (the addition terms $g(t-t_i)$ in $\lambda(t)$). This is indeed the case when the COVID-19 pandemic first breakout Fanelli and Piazza (2020); Bavel et al. (2020). Suppose within a pre-specified time range $[0, T]$, we observe k (infection) events with time points $t_1, t_2, \cdots, t_k$. Then the likelihood of the Hawkess process is Ozaki (1979)

$$L(\nu, \alpha, \beta) = \left(\prod_{i=1}^{k} \lambda(t_i)\right) \exp\left(-\int_0^T \lambda(t)dt\right). \tag{3.14.2}$$

Different algorithms for simulating $N(t)$ and explicitly calculating $\log L(\nu, \alpha, \beta)$ leads to various computational time. In this section, we adopts the algorithm implemented in the R package *hawkes*. It uses Ogata's algorithm Ogata (1981) to simulate both univariate and multivariate Hawkes processes. We compare the BAT algorithm with other metaheuristic algorithms implemented in the R package *metaheuristicOpt* for deriving the MLE of (ν, α, β) Riza et al. (2018). The true parameter vector is $(0.2, 0.5, 0.7)$ and we run each algorithm 100 times with 300 iterations each

to get reasonable statistical results. The tuning parameter are set to the default values in the *metaheuristicOpt* package and the swarm size is set to 30 for all algorithms.

The average and standard errors of negative log-likelihood, estimated ν, α and β and L_2-error are reported in Table 3.18 and Figure 3.13 (the values in brackets are standard error estimates based on 100 runs). Given an estimate $(\widehat{\nu}, \widehat{\alpha}, \widehat{\beta})$, the L_2-error is defined as

$$\frac{1}{3}\left((\widehat{\nu}-\nu)^2+(\widehat{\alpha}-\alpha)^2+(\widehat{\beta}-\beta)^2\right).$$

From the table, we see that PSO performs the best among 5 algorithms in terms of negative log-likelihood values and standard errors. All of PSO standard errors are 0.000 because they are too small to report, i.e., they are smaller than 10^{-10}. In addition to PSO, Harmony search (HS) algorithm has the smallest L_2-error and produces stable parameter estimates. Bat algorithm has an intermediate performance among all 5 algorithms. Finally, Cuckoo search (CS) and Genetic algorithm (GA) do not produce stable results as other 3 algorithms do.

Algorithm	Negative log-likelihood	ν	α	β	L_2-error
BA	2739.851 (9.458)	0.214 (0.026)	0.408 (0.054)	0.572 (0.075)	0.164 (0.087)
CS	2788.553 (29.623)	0.260 (0.086)	0.563 (0.182)	0.859 (0.285)	0.337 (0.201)
GA	2749.475 (12.023)	0.257 (0.031)	0.568 (0.090)	0.833 (0.144)	0.207 (0.112)
HS	2733.447 (1.222)	0.029 (0.006)	0.457 (0.022)	0.648 (0.036)	0.082 (0.021)
PSO	2732.831 (0.000)	0.225 (0.000)	0.442 (0.000)	0.624 (0.000)	0.099 (0.000)

Table 3.18: Average and standard errors of negative log-likelihood, estimated ν, α and β and L_2-error based on various metaheuristic algorithms. BAT = Bat algorithm, CS = Cuckoo search, GA = Genetic algorithm, HS = Harmony search, PSO = Particle swarm optimization.

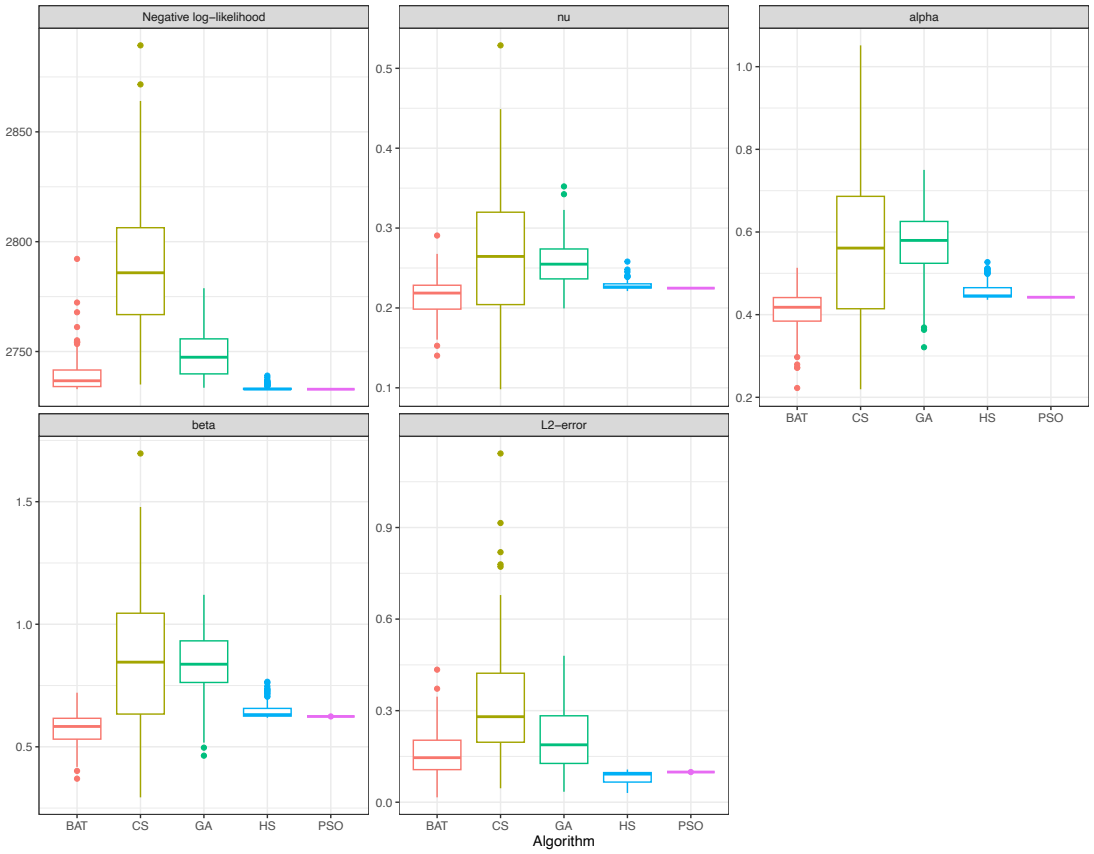


Figure 3.13: Negative log-likelihood, estimated parameters and L_2-error of 100 simulated Hawkes process estimates. BAT = Bat algorithm, CS = Cuckoo search, GA = Genetic algorithm, HS = Harmony search, PSO = Particle swarm optimization.

CHAPTER 4

Instrumental Variable Analysis with Interval-censored and Doubly-censored Outcome

4.1 Preamble

Focusing on the complexities of interval-censored and doubly-censored data, this chapter presents a sophisticated instrumental variable analysis framework. By developing and applying the DPMIV model, we tackle the challenges associated with such data, showcasing the model's efficacy through simulation studies and real-world applications.

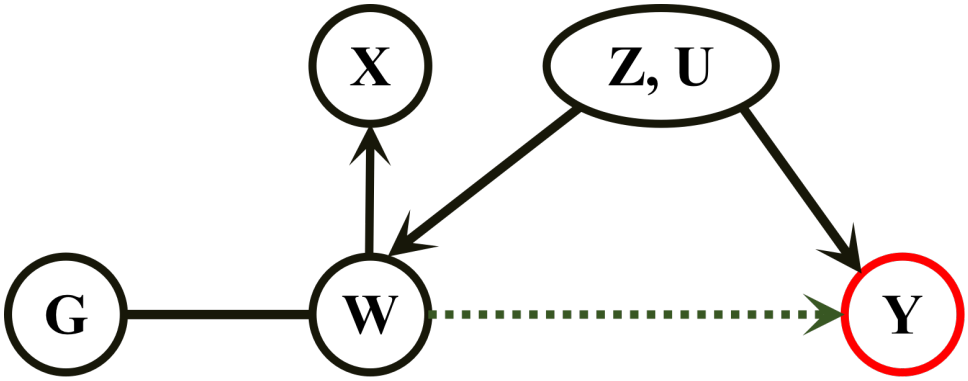
4.2 Introduction to Instrumental Variable Analysis

Estimating the causal effects of covariates on an outcome is a fundamental focus of scientific research. For example, in epidemiology studies, the emphasis often lies on discerning the causal impact of modifiable phenotypes or exposures on disease outcomes, transcending mere associations. However, unlike randomized control trials (RCT), which provide the gold standard for drawing causal inferences, deducing causation from real-world data sources such as electronic health records poses considerable challenges. One of the many hurdles encountered is the presence of unknown or unmeasured confounding factors in observational studies, potentially giving rise to spurious associations between covariates and outcomes (Rubin, 1997; Lin et al., 1998; Greenland et al., 1999; Fewell et al., 2007; Bareinboim et al., 2015; Pearl et al., 2016; VanderWeele et al., 2021). Additionally, measurement errors in the covariates are a common issue in observational studies that can introduce bias into estimates of causal effects (Hernán and Cole, 2009; Yi et al., 2015; Lee and Burstyn, 2016; Sengewald et al., 2019; Shu and Yi, 2019; Yi and Yan, 2021). For example, in Section 4.5, we examine the causal effect of systolic blood pressure (SBP) on the

time to cardiovascular disease (CVD) following the diagnosis of diabetes mellitus (DM) using data from the UK Biobank (UKB) study (Allen et al., 2014). While the UKB dataset includes some common observed covariates, such as age, smoking, and cholesterol, potential confounders that exhibit correlations with both SBP and time-to-CVD, such as annual income and drinking habits, are absent from the dataset (Naimi et al., 2005). Furthermore, it is well-known that SBP measurements are susceptible to errors. As illustrated in Section 4.5, directly regressing time-to-CVD on SBP without accounting for potential unobserved confounders and measurement errors produced a result indicating an unexpected association. The outcome, presented in Table 4.3, suggests a counterintuitive link where higher SBP is associated with longer time-to-CVD, contrary to expectations. In Section 4.4, our simulation studies further highlight that neglecting unobserved confounders and measurement errors can result in significant estimation bias and invalid inference for a causal effect.

Two primary approaches are frequently employed for causal inference: the potential (counterfactual) outcomes framework (Rubin, 1974; Angrist et al., 1996; Imbens and Rubin, 2010, among others) and the graphical causal models (Baiocchi et al., 2014; Burgess et al., 2017, among others). This paper will focus on instrumental variable (IV) analysis under the latter framework, which has been commonly used for mitigating bias arising from unmeasured confounding and measurement errors (Pearl, 2000; Heckman, 2008; Wright et al., 1928; Haavelmo, 1943; Theil, 1958; Goldberger, 1972; Heckman and Robb, 1985; Morgan, 1991; Swanson and Hernán, 2013, among others). IV analysis is also known as the Mendelian Randomization (MR) in epidemiology where genetic markers are used as the instrument (e.g. Gray and Wheatley, 1991; Didelez and Sheehan, 2007; Lawlor et al., 2008; Wehby et al., 2008; VanderWeele et al., 2014; Emdin et al., 2017). The basic structure of Instrumental Variable (IV) analysis can be visually depicted using directed acyclic graphs (DAGs), as illustrated in Figure 4.1, where variables are denoted as nodes, causal relationships are indicated by directed arrows between nodes, and the absence of a direct arrow between two nodes signifies the absence of a direct causal link. In Figure 4.1, Y represents the outcome variable, W is the endogenous covariate that may remain unobserved due to measurement errors, X represents an observed surrogate for the covariate W, Z is a vector encompassing observed confounders, U comprises unobserved confounders, and G serves as an instrument vector. A mathematical representation of this IV analysis structure is provided in Section 2 through equations (4.3.1), (4.3.2), and (4.3.3). Before going further, it is important to note that IV analysis operates

Figure 4.1: Directed acyclic graph of instrumental variable analysis. G is the instrumental variable, W refers to the unobserved endogenous covariate, X refers to the noisy surrogate, Z and U refer to the observed and unobserved confounders, Y is the outcome. β_1 represents the causal effect of W on Y. A line with no arrow indicates association and an arrow indicates a causal relationship in a specific direction.



under three key assumptions:

1. **Independence**: G is independent of both U and the measurement errors in W.

2. **Relatedness**: G is correlated with W.

3. **Exclusion Restriction**: G is independent of Y given W, Z, and U, meaning that any association between G and Y is solely through W.

For continuous outcomes, assuming linear models, classical IV analysis estimates the causal parameter (β_1 in equation (4.3.2)) using a Two-stage Least Squares (TSLS) method. This is achieved by first regressing the observed surrogate X on the instrument G and then regressing the outcome Y on the fitted values of X (Huber (1967); White (1980); Murphy and Topel (1985); Hardin (2002); Hardin and Carroll (2003); Gustafson (2007); Martins and Gabriel (2014) among others). This approach has also been extended to nonlinear models, such as the logistic regression model for binary outcomes, based on the M-estimation method (Amemiya (1985, 1990); Foster (1997) among others). Bayesian IV methods have also been developed for continuous outcomes. For instance, among others, Kleibergen and Van Dijk (1998); Hoogerheide et al. (2007) have discussed parametric Bayesian IV methods based on the assumption that the error terms in a two-stage IV model (4.3.4) and (4.3.5) follow a bivariate normal distribution with conjugate priors. Wang et al. (2023b) combines the linear Bayes method (Hartigan, 1969) with the IV approach using both conjugate and non-conjugate priors. On the other hand, Conley et al. (2008) and Wiesenfarth et al.

(2014) have developed semiparametric Bayesian IV methods using a Gaussian mixture of Dirichlet process (DPM) model (Ferguson, 1983; Lo, 1984; Escobar and West, 1995), and demonstrated by simulations that they are more efficient than their parametric counterparts when the error terms are non-normal. We refer to Baiocchi et al. (2014); Bowden et al. (2021) for some recent surveys of IV analysis methods for causal inference.

In recent years, IV analysis methods have also been developed for time-to-event outcomes (Bijwaard, 2008; Roodman, 2011; Atiyat, 2011; Li and Lu, 2015; Li et al., 2015a; Tchetgen et al., 2015; Kjaersgaard and Parner, 2016; Martinussen et al., 2017, 2019; Martinussen and Vansteelandt, 2020; Lee et al., 2023; Wang et al., 2023a, among others). For instance, one line of research has considered the potential outcome framework to estimate the causal treatment effect and extended the G-estimation method proposed in Robins and Tsiatis (1991) to various time-to-event models with right-censored data including Aalen's additive risk model (Martinussen et al., 2017), a Cox structural model (Martinussen et al., 2019; Wang et al., 2023a), and a competing risks model (Martinussen and Vansteelandt, 2020). More recently, Li and Peng (2023) studied a general class of causal semiparametric transformation models for estimating the complier causal treatment effect with interval-censored data within the potential outcome causal framework. The DAG IV framework depicted in Figure 4.1 has also been studied to address unobserved confounders and/or measurement errors for time-to-event models with right-censored data. For example, Bijwaard (2008) proposed an IV Linear Rank estimator for right-censored time-to-event data, based on a Generalized Accelerated Failure Time (GAFT) model which encompasses the Proportional Hazard model and the Accelerated Failure Time (AFT) model. Kjaersgaard and Parner (2016) proposed a pseudo-observation approach to IV analysis of the survival function, the restricted mean, and the cumulative incidence function for right-censored competing risks data. Li and Lu (2015) developed a parametric Bayesian method for a two-stage IV model (PBIV) assuming bivariate normal errors with right-censored data.

It's important to note that the theoretical underpinnings of most frequentist IV methods for right-censored data primarily rely on the counting process and martingale framework (Andersen and Gill, 1982; Martinussen et al., 2017). However, these frameworks don't easily extend to other censoring schemes, such as interval-censoring. Li and Lu (2015) developed a parametric Bayesian method for a two-stage IV model (PBIV) assuming bivariate normal errors for right-censored data and noted that their approach can potentially be extended to handle more complex censoring

schemes.

This chapter aims to develop an IV analysis tool for estimating the causal effect of an endogenous variable when dealing with unobserved confounders and measurement errors. Our method is particularly tailored for partly interval-censored time-to-event data, where event times are observed exactly for some subjects but left-censored, right-censored, or interval-censored for others (Pan et al., 2020). To the best of our knowledge, this problem has not been previously addressed in the literature. Specifically, we develop a semiparametric Bayesian IV analysis method based on a two-stage Dirichlet process mixture instrumental variable (DPMIV) model for the DAG IV framework illustrated in Figure 4.1. As detailed in Section 2.1, our DPMIV method simultaneously models the first-stage random error term for the exposure variable and the second-stage random error term for the time-to-event outcome using a Gaussian mixture of the Dirichlet process (DPM) model. The DPM model can be broadly understood as a mixture model with an unspecified number of Gaussian components, making it versatile for approximating various error distributions (Ferguson, 1983; Lo, 1984). It relaxes the normal error assumptions and allows the number of mixture components to be determined by the data. It's important to note that our approach can be viewed as a non-trivial extension of the work presented in Conley et al. (2008), transitioning from uncensored data to partly interval-censored data. A fundamental difference in our approach is our use of non-conjugate priors within the DPM model. This choice is pivotal for effectively handling partly interval-censored data, while the previous method relied on the use of conjugate priors tailored for its Markov Chain Monte Carlo (MCMC) sampling algorithm designed for uncensored data. Throughout this paper, we develop an MCMC algorithm tailored for our DPMIV model when applied to partly interval-censored data, and discuss its distinct features in comparison with the approach presented in Conley et al. (2008). For completeness and comparison purposes, we additionally broaden the applicability of the PBIV method as presented in Li and Lu (2015), extending it from right-censored data to partly interval-censored data. We conduct extensive simulations to assess the performance of our DPMIV method and exemplify its practicality and effectiveness through real data applications. Our simulations revealed that compared to the naive method, which ignores unobserved confounders and measurement errors, the proposed DPMIV significantly reduces bias in estimation and substantially improves the coverage probability of the endogenous variable parameter. Moreover, when the errors exhibit a non-normal distribution, the DPMIV approach consistently provides less biased parameter estimates with smaller standard errors while maintaining performance comparable to the parametric

Bayesian approach PBIV in cases where errors follow a bivariate normal distribution. Furthermore, we have developed an R package that facilitates the implementation of both the DPMIV method and the PBIV method for partly interval-censored data. This package is publicly accessible at https://github.com/ElvisCuiHan/PBIV/.

4.3 DPMIV: A Semiparametric Bayesian Instrumental Variable Method for Partly Interval-censored Data

In this section, we present our two-stage Dirichlet process mixture instrumental variable (DPMIV) model. We also outline an MCMC estimation and inference procedure centered around the DPM model, tailored specifically for handling partly interval-censored data. A detailed description of the MCMC algorithm is provided in the Appendix.

We note that our algorithm is more general and also more complicated than Conley et al. (2008) and Wiesenfarth et al. (2014) by putting uniform priors and using random walk M-H algorithm. In other words, it is not only able to handle partly interval-censored data but also continuous and categorical outcome with minor modifications to the likelihood function.

4.3.1 The Model and the Data

Consider the DAG IV framework in Figure 4.1. Let $(Y_i, W_i, X_i, Z_i, U_i, G_i)$ be n independent and identically distributed realizations of (Y, W, X, Z, U, G). Then, assuming linear models, the underlying structure of Figure 4.1 can be represented as follows:

$$W_i = \alpha_0 + \alpha_1{}'G_i + \alpha_2{}'Z_i + \alpha_3{}'U_i + \varepsilon_{1i}, \tag{4.3.1}$$

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2{}'Z_i + \beta_3{}'U_i + \varepsilon_{2i}, \tag{4.3.2}$$

$$X_i = W_i + \varepsilon_{3i}, \quad i = 1, \ldots, n, \tag{4.3.3}$$

where β_1 is the causal effect parameter of interest, and ε_{1i}, ε_{2i}, and ε_{3i} represent independent random errors in models (4.3.1), (4.3.2) and (4.3.3), respectively.

It is easy to see that by substituting the unobserved W_i with $W_i = X_i - \varepsilon_{3i}$ into equations

(4.3.1) and (4.3.2), we obtain the following two-stage linear model:

$$X_i = \alpha_1{}'G_i + \alpha_2{}'Z_i + \xi_{1i}, \tag{4.3.4}$$

$$Y_i = \beta_1 X_i + \beta_2{}'Z_i + \xi_{2i}, \tag{4.3.5}$$

where $\xi_{1i} = \alpha_0 + \alpha_3{}'U_i + \varepsilon_{1i} + \varepsilon_{3i}$ and $\xi_{2i} = \beta_0 + \beta_3{}'U_i + \varepsilon_{2i} - \beta_1\varepsilon_{3i}$ are independent of the instrument G_i, but there is the possibility of them being correlated with X_i and Z_i. This correlation is an important consideration in instrumental variable (IV) analysis and highlights the need for careful modeling and estimation to account for these relationships when estimating the causal effect (β_1).

Our two-stage DPMIV model with time-to-event outcome takes the form of (4.3.4) and (4.3.5) and assumes that the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution with a Dirichlet Process (DP) prior for its mean and variance-covariance parameters:

$$(\xi_{1i}, \xi_{2i})' \sim N_2(\mu_i, \Sigma_i), \tag{4.3.6}$$

$$(\mu_i, \Sigma_i) \sim \text{i.i.d. } H, \tag{4.3.7}$$

$$H \sim \text{DP}(\nu, H_0). \tag{4.3.8}$$

Here $\mu_i = (\mu_{1i}, \mu_{2i})'$, $\Sigma_i = \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix}$, and $DP(\nu, H_0)$ in (4.3.8) is the Dirichlet process (DP) prior with strength parameter ν and base distribution H_0 (Ferguson, 1973).

Assume that instead of observing $(Y_i, W_i, X_i, Z_i, U_i, G_i)$, $i = 1, \ldots, n$, one observes a partly interval-censored data set consisting of n independent and identically distributed observations $(L_i, R_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots n$, where L_i and R_i represent the left and right endpoints of the censoring interval for the outcome variable Y_i, and δ_i is an indicator variable ($\delta_i = 1$ if $Y_i < L_i$ (left-censored); $\delta_i = 2$ if $L_i \leq Y_i \leq R_i$ and $L_i < R_i$ (interval-censored); $\delta_i = 3$ if $Y_i > R_i$ (right-censored); $\delta_i = 4$ if $L_i = Y_i = R_i$ (event)). Our objective is to estimate the causal effect of W_i on Y_i, represented by parameter β_1, based on this partly interval-censored data.

Remark 1: The model (4.3.6)-(4.3.8) for the error $(\xi_{1i}, \xi_{2i})^T$, known as a Dirichlet process mixture (DPM) model (Ferguson, 1983), is a widely used nonparametric Bayesian model. A nice introduction of DP prior and DPM can be found in Ghosal and Van der Vaart (2017). The DPM model can be viewed as a mixture of Gaussians with infinite number of components. Notably, H is

a random discrete distribution that has the same support as H_0, where H_0 is usually a continuous distribution, i.e., $P(H(B) > 0) = 1$ if and only if $H_0(B) > 0$ for any Borel sets B. This discreteness of H randomly clusters different (μ_i, Σ_i) together. The parameters μ_i and Σ_i are the same within one cluster and different across clusters. Note that the marginal distribution of any (μ_i, Σ_i) (by marginalizing out H) is H_0. In other words, given all the parameters and H, the samples $(X_i, Y_i)^T$ are drawn mixtures of normal distributions (i.e., the distribution of (μ, Σ)) and hence are clustered naturally. As a result, the total number of clusters, denoted as k, is random and we denote the cluster indicators as c in later subsections. The posterior distribution of k is determined by both the strength parameter ν and the data. Therefore, the DP prior enables the model to better capture heterogeneity in the error distribution, without using a pre-specified number of clusters. Further, theorem 2 in Eaton (1981) states that a large class of distributions can be represented as a mixture of Gaussian density and an underlying mixing distribution. Similarly, Fejér's theorem states that with a Gaussian kernel, we may approximate any density in L_1 by mixtures (Lo, 1984; Ghosal and Van der Vaart, 2017). Ferguson (1983) pointed out that an infinite mixture of Gaussian densities can approximate any distribution on the real line with any preassigned accuracy in the Lévy metric. These justifies the use of Dirichlet process mixtures. In addition, as Conley et al. (2008) pointed out, putting a prior on ν makes it easier for the data to determine the number of clusters instead of letting users to specify the possible number of clusters in DPM. Hence, in our customized MCMC algorithm, we put a prior on ν so that both small and large number of clusters are possible (Section 4.3.2).

Remark 2: The two-stage model (4.3.4)-(4.3.8) is an extension of the semiparametric IV model proposed in Conley et al. (2008) where we allow Y to be partly interval-censored time-to-event data. Because the likelihood function for censored outcome has a complicated form, conjugate priors (and thus, Gibbs sampler) are not available for $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ (Neal, 2000) and there is no convenience to assume H_0 to be conjugate as in Conley et al. (2008) (we give details of H_0 in Section 4.3.2, for Gibbs sampler, see Equation (3.2) in Neal (2000)). This is another major difference between our algorithm and that in Conley et al. (2008).

Remark 3: Our proposed model can relax the parametric assumption of a specific distribution for the IV model introduced in Li and Lu (2015), and address for potential heterogeneous clustering problems within the context of IV modelling with right censored-data. For completeness, the PBIV assumes that the error $(\xi_{1i}, \xi_{2i})^T$ follows a common bivariate normal distribution and the mean μ

follows a mean 0 normal distribution, ρ follows a uniform distribution on $(-1, 1)$ and σ_1, σ_2 have inverse Gamma distributions, respectively.

Assume that one observes a partly interval-censored data set consisting of n independent and identically distributed observations $(L_i, R_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \ldots n$, where L_i and R_i represent the left and right endpoints of the censoring interval for the outcome variable Y_i, and δ_i is an indicator variable ($\delta_i = 1$ if $Y_i < L_i$ (left-censored); $\delta_i = 2$ if $L_i \leq Y_i \leq R_i$ and $L_i < R_i$ (interval-censored); $\delta_i = 3$ if $Y_i > R_i$ (right-censored); $\delta_i = 4$ if $L_i = Y_i = R_i$ (event)). Our objective is to estimate the causal effect of W_i on Y_i, represented by parameter β_1, based on this partly interval-censored data.

4.3.2 The MCMC Algorithm

Our Bayesian causal inference on β_1 is conducted through its posterior distribution given the data and other parameters. Because an analytical expression for the posterior distribution is not available, we resort to Markov Chain Monte Carlo (MCMC) methods, which are particularly useful in Bayesian statistics (Robert et al., 1999). Notably, due to the non-parametric and discrete nature of the Dirichlet Process (DP) and Dirichlet Process Mixture (DPM), the MCMC algorithm developed by Li and Lu (2015) for the two-stage normal IV model is not applicable in our case, necessitating the development of new algorithms. Various methods exist for drawing posterior samples from a DPM, and both Neal (2000) and Chapter 3 in Müller et al. (2015) provide comprehensive reviews of these methods. In our work, we have developed a customized MCMC procedure to make inference on β_1. In each iteration of the procedure, we sequentially update individual parameters while keeping other parameters fixed at their current states. Below, we outline the key steps of our MCMC algorithm, with a more detailed description provided in Appendix.

Because of the discrete nature of the DP, the DPM model induces a probability on clusters associated with latent $\theta_i = (\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i)^T, i = 1, 2, \cdots, n$ (Antoniak, 1974; Müller et al., 2015). That is, there is a positive probability of having identical values among the θ_i's. Let θ_c, $c = 1, \cdots, k$ be the $k \leq n$ unique values (so that the total number of clusters is k), and $S_j = \{i : \theta_i = \theta_c\}$ be the indices associated with θ_c. Then the multiset $\{S_1, \cdots, S_k\}$ forms a partition of $\{1, 2, \cdots, n\}$ and it is random because θ_i's are random (Müller et al., 2015). For convenience, we represent the clustering by an equivalent set of cluster membership indicators: let

$\vec{C} = \{c_1, \ldots, c_n\}$ be the latent class indicator of a subject, i.e. θ_C consists of all distinct values of θ_i and $\vec{C}$ is a vector of indicators that maps the individuals to the clusters. Note that the numbering of C can be arbitrary. For the two-stage DPMIV model (4.3.4)–(4.3.8), we denote the parameters as $\Theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_C, \vec{C})$. The observed data consists of Data $= (\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$, where $\vec{L} = (L_1, ..., L_n)$, $\vec{R} = (R_1, ..., R_n)$, $\vec{\delta} = (\delta_1, ..., \delta_n)$, $\vec{X} = (X_1, ..., X_n)$, $\vec{Z} = (Z_1, ..., Z_n)$ and $\vec{G} = (G_1, ..., G_n)$. Then the likelihood function is written as

$$\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \quad = \quad P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \cdot P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) \tag{4.3.9}$$

where $P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta)$ is likelihood contributed by the first-stage model (4.3.4) and $P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta)$ is the likelihood based on the second-stage model (4.3.5). We provide details of derivation of both terms in the Appendix.

Given the likelihood function $\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$, our MCMC algorithm draw samples from the following posterior distributions iteratively:

1) $\alpha_1 | \alpha_2, \beta_1, \beta_2, \theta_c, \vec{C}, \nu, \text{Data}$
2) $\alpha_2 | \alpha_1, \beta_1, \beta_2, \theta_c, \vec{C}, \nu, \text{Data}$
3) $\beta_1 | \alpha_1, \alpha_2, \beta_2, \theta_c, \vec{C}, \nu, \text{Data}$
4) $\beta_2 | \alpha_1, \alpha_2, \beta_1, \theta_c, \vec{C}, \nu, \text{Data}$
5) $\vec{C} | \alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c, \nu, \text{Data}$
6) $\theta_c | \alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c, \vec{C}, \text{Data}$
7) $\nu | \alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c, \vec{C}, \text{Data}.$

Draw of $(\boldsymbol{\alpha_1, \alpha_2, \beta_1, \beta_2})$

The algorithm provided by Conley et al. (2008) do not involve α_2 and they break the draw into 2 parts, i.e., α_1 and (β_1, β_2), both with normal priors. Our customized draw of $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ is done by the random walk Metropolis-Hastings (M-H) algorithm. In constrast to potentially correlated normal priors in Conley et al. (2008), independent normal priors are put on each parameters and the proposal distribution is uniform within a certain interval. We note that a suitable length (neither too wide nor too narrow) of the uniform distribution leads to fast convergence of the MCMC algorithm. We set 0.0128 for β_1 and 0.0064 for $\alpha_1, \alpha_2, \beta_2$ in simulation studies, and 0.0584 for β_1 and 0.0128 for $\alpha_1, \alpha_2, \beta_2$ in the UKB example.

Draw of $\vec{C}$ and $\boldsymbol{\theta}_c$

We draw new θ's and update $\vec{C}$ from the base measure H_0, hence it is required to specify H_0. Conley et al. (2008) assumes it is a Normal-Wishart distribution, i.e., $H_0 = \pi(\mu|\Sigma)\pi(\Sigma)$ where $\pi(\mu|\Sigma)$ is a bivariate normal density whose covariance matrix is proportional to Σ and $\pi(\Sigma)$ is a Wishart density. In contrast, since the conjugate prior is not available for censored outcome, we do not assume that the base measure H_0 follows a Normal-Wishart distribution as that in Conley et al. (2008). Instead, we assume independent priors on H_0, i.e., $H_0 = \pi(\mu_1)\pi(\mu_2)\pi(\sigma_1^2)\pi(\sigma_2^2)\pi(\rho)$ where $\pi(\cdot)$ is an abuse of notation for priors. For simulation studies, we set $\pi(\mu_1)$ and $\pi(\mu_2)$ to be normal density with mean 0 and pre-specified large variances (we set it to 10 and it works well), $\pi(\sigma_1^2)$ and $\pi(\sigma_2^2)$ to be inverse-gamma with pre-specified small shape and scale parameters (we set them to 0.1 and 0.001), $\pi(\rho)$ to be uniform within $[-1, 1]$. These correspond to non-informative (or vague) priors (Li and Lu, 2015). For the UKB study in Section 4.5, we use slightly informative priors (here "slightly informative" means we use 5% of the samples as the training data to get posterior distributions of parameters and use them as priors for the the remaining 95% interval-censored data.) and the details are given in the Appendix.

We adopts algorithm 8 in Neal (2000) for non-conjugate priors to update $\vec{C}$ and θ_c while Conley et al. (2008) use the Gibbs sampler in Bush and MacEachern (1996) (see also algorithm 2 in Neal (2000)). We note that according to Neal (2000), posterior samples using the algorithm 8 has the smallest auto-correlation among other MCMC algorithms.

Draw of ν

It is tricky to set the prior and update the posterior for ν as we indicated in **Remark 2**. Given Data and k, the number of distinct values of θ_c, the distribution of ν is independent of Θ (Ghosal and Van der Vaart, 2017). Hence, computation of $\nu|k, n$ (n is the sample size) requires a prior for ν and a marginal expression for $k|\nu, n$. Antoniak (1974) derived the expression for $k|\nu, n$ and Conley et al. (2008) suggested a prior for ν on the discrete grid between $\overline{\nu}$ and $\underline{\nu}$ so that ν can be interpreted as groups of observations:

$$P(\nu) \propto \left(\frac{\overline{\nu} - \nu}{\overline{\nu} - \underline{\nu}}\right)^{\omega} \cdot I(\underline{\nu} < \nu < \overline{\nu}).$$

We note it is also workable for ν to be continuous as in our algorithm (see Appendix). In our simulation study and real data examples, we set $\overline{\nu}$ and $\underline{\nu}$ to be 0.1 and 4.8 so that the modes of k are 1 and 16 for sample size equals to 100, respectively.

By iterating the procedure described above, a sufficiently large amount of MCMC samples can be generated from the posterior distribution. Posterior mean of a parameter can be used as an estimation of the parameter. Credible intervals of the parameters can be constructed by using the empirical quartiles of the simulated samples. Convergence of the MCMC algorithm can be examined visually by graphical methods including trace plots and histograms, and quantitatively by using the Brooks-Gelman-Rubin diagnostics (Brooks and Gelman, 1998). We implemented this method in C programming language, due to its relatively fast process in large number of iterations. Our C program is available online at `https://github.com/ElvisCuiHan/BayesianIVAnalysis`.

4.4 Simulation Studies

We conducted extensive simulations to evaluate the performance of proposed two-stage DPMIV method for partly interval-censored time-to-event data under a variety of scenarios. Additionally, we include two other methods for reference in our simulation analysis: 1) the naive single-stage accelerated failure time (AFT) model for partly interval-censored data (Huang and Wellner, 1997; Anderson-Bergman, 2017), which does not account for unobserved confounders and measurement errors, and 2) the two-stage PBIV method, as described in Appendix, which extends the parametric Bayesian IV method introduced by Li and Lu (2015) from right-censored data to partly interval-censored data.

We simulated data from model (4.3.4)-(4.3.5) with a two-dimensional instrument G_i and a two-dimensional observed confounder U_i, both following a standard bivariate normal distribution $N(0, I_2)$. The regression parameters in equation (4.3.4) were set as $\alpha_1 = (0.5, 0.5)^T$, and $\alpha_2 = (0.5, 0.5)^T$. The regression parameters in equation (4.3.5) were set as $\beta_1 = -1$, and $\beta_2 = (0.8, 0.8)^T$. We considered six scenarios for the bivariate distribution of $(\xi_{1i}, \xi_{2i})^T$:

1. Bivariate normal distribution.
2. Bivariate exponential distribution as described in Equation 18 in Nagao and Kadoya (1971).
3. Mixture of two bivariate normal distributions with different means but the same variance-covariance matrix.
4. Mixture of two bivariate normal distributions with the same mean but different variance-

covariance matrices.

5. Mixture of five bivariate normal distributions, mimicking the distribution of the female cohort estimated from the UK Biobank dataset using the DPMIV method.

6. Mixture of five bivariate normal distributions, mimicking the distribution of the male cohort estimated from the UK Biobank dataset using the DPMIV method.

Detailed specifications for the bivariate distribution of $(\xi_{1i}, \xi_{2i})^T$ under these six simulation scenarios can be found in Table 7.11.

Similar to the simulation settings in Pan et al. (2020), we generate partly interval-censored data as follows. In each simulated dataset, we first set around 25% individuals to have exact event times observed. Next, we assume L_i has an exponential distribution with hazard rate 2 and $R_i - L_i$ has another independent exponential distribution with hazard rate 2. Then left-, interval- and right-censored observations are determined by whether Y_i is less than L_i, within $(L_i, R_i]$ or greater than R_i, resulting in a approximate censoring rate around 20%, 20%, 35% and 25% (left-, interval-, right-censored and event). Finally, we considered different sample sizes $n = 300, 500$ and 1000 under each scenario.

Table 4.2 presents a summary of the simulated bias, standard deviation (SD), and coverage probability (CP) for the causal parameter β_1 using the three aforementioned methods based on 100 Monte Carlo replications. Additionally, the proposed DPMIV method, we also report the average of the estimated number of clusters k.

As observed in Table 4.2, the naive single-stage AFT model estimate generally exhibits substantial bias and unacceptably low coverage probability, which underscores the critical need to address unobserved confounders and measurement errors.

The PBIV estimate demonstrates satisfactory performance in scenario 1 (normal model) and scenarios 2 and 4 when the error distribution is, or can be approximated by, a mixture of 1 or 2 normal components. Nevertheless, it exhibits substantial bias and very low coverage probability in scenarios 5 and 6, where the error distribution involves a mixture of a larger number of normal components (five). These findings underscore the limited robustness of the PBIV method under certain scenarios.

Our proposed DPMIV method consistently delivers robust and stable performance, with mini-

Table 4.1: Specification of the bivariate distribution of $(\varepsilon_{1i}, \varepsilon_{2i})^T$ under six simulation scenarios

Scenario 1		Normal				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	100%	0.5	0.500	0.5	1.000	0.424
Scenario 2		Bivariate exponential				
Component	Proportion	μ_1	σ_1	μ_2	σ_2	ρ
1	100%	\	0.300	\	0.300	0.300
Scenario 3		Normal mixture I				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	50%	0.630	0.300	-0.630	0.300	0.500
2	50%	-0.630	0.300	0.630	0.300	0.500
Scenario 4		Normal mixture I				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	50%	0.000	0.700	0.000	0.700	0.357
2	50%	0.000	0.050	0.000	0.050	0.600
Scenario 5		Normal mixture III				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	72%	1.882	0.015	1.511	1.110	0.107
2	18%	1.783	0.022	-2.370	0.204	-0.081
3	5%	1.260	0.112	1.265	0.226	0.996
4	3%	1.941	0.095	1.128	0.493	0.345
5	2%	1.922	0.052	-0.701	2.347	0.401
Scenario 6		Normal mixture IV				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	50%	4.985	0.015	5.011	0.966	0.076
2	20%	4.585	0.024	4.265	0.177	-0.051
3	10%	4.830	0.103	5.265	0.255	0.878
4	10%	4.983	0.084	5.256	0.633	0.484
5	10%	4.924	0.055	3.880	2.264	0.670

The simulation studies puts six different distributions on the bivariate random error $(\xi_{1i}, \xi_{2i})^T$ in the DPMIV model (4.3.4)-(4.3.5). The first scenario is bivariate normal with mean $(0.5, 0.5)^T$ and covariance matrix $\begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 1 \end{pmatrix}$. The second scenario is a bivariate exponential distribution where the density is given in the Equation 18 in Nagao and Kadoya (1971). For this distribution, we only need to specify the two scale parameters σ_1 and σ_2 and the correlation parameter ρ. The third scenario is a mixture of two bivariate normal distributions with equal proportion, separate means and same covariances, i.e., $0.5 \times N\left(\begin{pmatrix} 0.63 \\ -0.63 \end{pmatrix}, \begin{pmatrix} 0.3 & 0.15 \\ 0.15 & 0.3 \end{pmatrix}\right) + 0.5 \times N\left(\begin{pmatrix} -0.63 \\ 0.63 \end{pmatrix}, \begin{pmatrix} 0.3 & 0.15 \\ 0.15 & 0.3 \end{pmatrix}\right)$. The fourth scenario is also a mixture of two bivariate normal distributions with equal proportion but the same means and different covariances, i.e., the density is $0.5 \times N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.7 & 0.25 \\ 0.25 & 0.7 \end{pmatrix}\right) + 0.5 \times N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.05 & 0.03 \\ 0.03 & 0.05 \end{pmatrix}\right)$. The fifth and sixth scenarios are five-component normal mixtures (with different proportions) that mimics the estimated error distribution in Section 4.5.

mal bias and satisfactory coverage probability across all six scenarios and various sample sizes. In the first scenario, it exhibits similar bias and standard deviation (SD) to PBIV and outperforms in the remaining four scenarios, correctly identifying the number of clusters k as 1. In scenario 2, where there isn't a correct number of clusters, DPMIV estimates k as 2, providing a normal mixture approximation to the bivariate exponential distribution. While it may appear that k is over-estimated in scenarios 3 and 4, it's worth noting that the estimation of random errors reveals two dominant components, with negligible sample sizes in the remaining clusters. As for scenarios 5 and 6, as the sample size increases, DPMIV correctly estimates k as expected.

Lastly, in Figure 4.2, we depict the true and estimated log-density error distribution by DPMIV under different sample sizes. The results align with our expectations, showing that as the sample size increases, DPMIV accurately estimates the random error distribution.

In our comprehensive simulation studies, we have broadened the scope to include a variety of different scenarios focusing particularly on scenarios with low event rates, such as a censoring rate leaving only 5% observable events. We have also explored a smaller effect size where $\beta_1 = -0.363$, mirroring the causal effect magnitude found in the UKB data. Furthermore, we have assessed the performance of our methods across a spectrum of instrument strengths, considering from weak to strong instrumental strengths at 2%, 15%, 35%, and 50% respectively. The outcomes of these additional simulations have been consistent with the findings reported in the main text, reaffirming the robustness of our methods under a wide array of conditions.

Table 4.2: β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.365	0.083	0%	0.027	0.128	100%	0.112	0.187	90%	1.000
		500	0.373	0.081	0%	0.043	0.101	99%	0.063	0.124	94%	1.008
		1000	0.365	0.053	0%	0.010	0.072	98%	0.024	0.081	92%	1.002
2	Exponential	300	0.147	0.061	11%	0.008	0.069	90%	0.037	0.074	99%	1.194
		500	0.158	0.040	1%	0.010	0.053	96%	0.023	0.055	100%	1.239
		1000	0.153	0.029	0%	0.002	0.037	96%	0.013	0.034	94%	1.994
3	Normal Mixture I	300	0.236	0.050	0%	0.012	0.105	88%	0.007	0.077	92%	3.987
		500	0.233	0.032	0%	0.016	0.086	88%	0.007	0.060	98%	3.969
		1000	0.225	0.029	0%	0.026	0.061	89%	0.006	0.041	98%	3.046
4	Normal Mixture II	300	0.236	0.050	0%	0.024	0.078	94%	0.011	0.059	96%	7.330
		500	0.233	0.032	0%	0.025	0.061	97%	0.009	0.042	95%	7.056
		1000	0.225	0.029	0%	0.014	0.043	92%	0.003	0.028	96%	6.373
5	Normal Mixture III	300	0.164	0.195	89%	0.451	0.128	0%	0.017	0.056	95%	4.069
		500	0.106	0.078	92%	0.207	0.086	24%	0.012	0.054	94%	4.603
		1000	0.093	0.058	90%	0.129	0.065	52.5%	0.003	0.037	98%	4.980
6	Normal Mixture IV	300	0.152	0.176	78%	0.649	0.128	0%	0.025	0.113	93%	4.172
		500	0.202	0.167	62%	0.497	0.114	4%	0.022	0.066	90%	4.922
		1000	0.167	0.107	66%	0.350	0.089	3%	0.006	0.045	91%	5.286

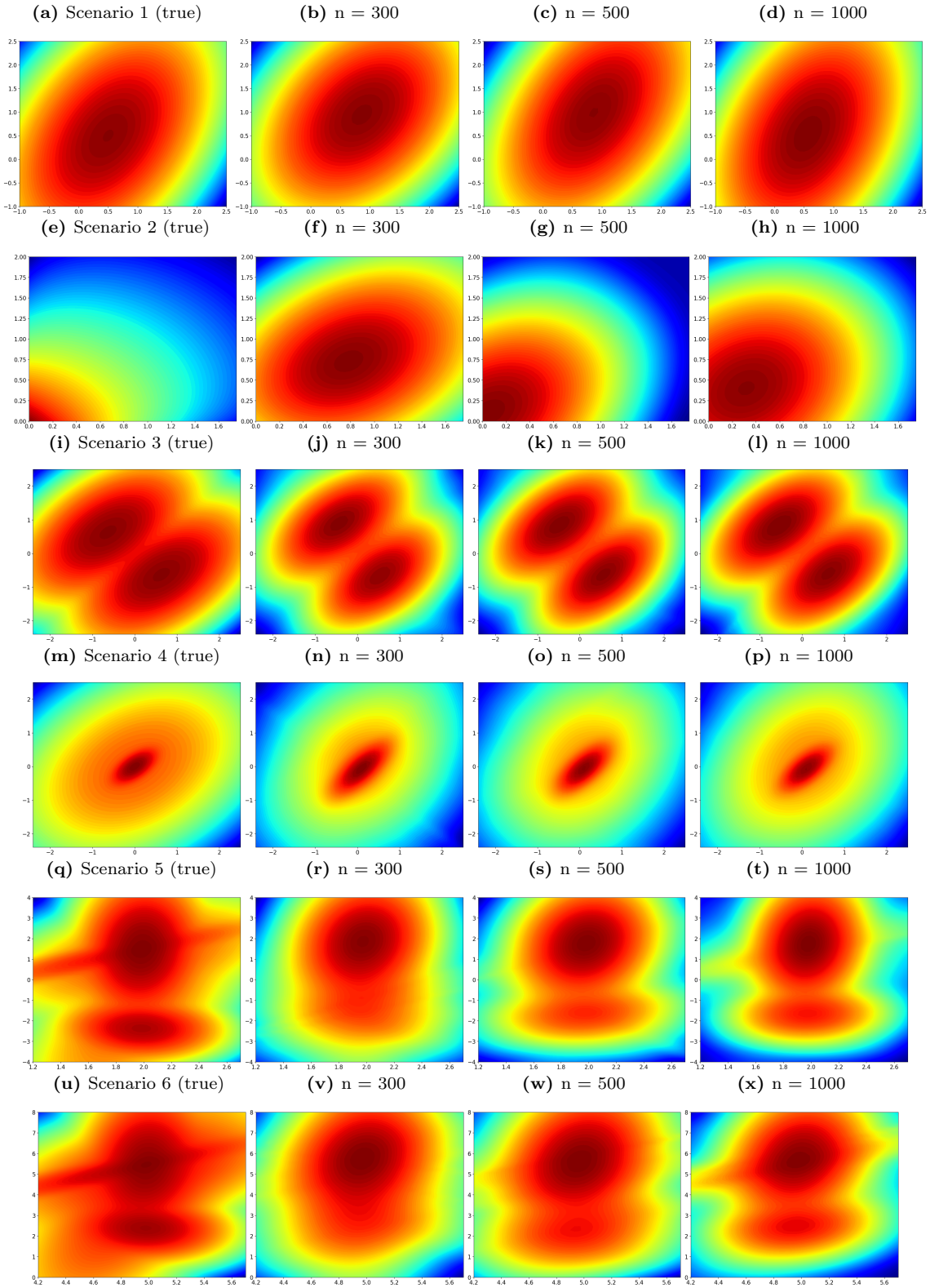
- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

4.5 IV Analysis of Interval-censored UK Biobank Data

The UK Biobank (UKB) cohort comprises 500,000 individuals aged 40 to 69 years at baseline, recruited between 2006 and 2010 at 22 assessment centers across the United Kingdom. Participants were followed up until January 1, 2018, or until their date of death. This extensive resource provides data on genotyping, clinical measurements, assays of biological samples, and self-reported health behavior. As an illustrative example, we will investigate the causal effects of systolic blood pressure (SBP) on the time-to-development of cardiovascular disease (CVD) from the onset of diabetes mellitus (DM) with a focus on White individuals, including 3,141 females and 5,029 males, who developed DM before CVD. Descriptive statistics of baseline characteristics for this subgroup are summarized in Appendix. It is commonly known that CVD is associated with death (Amini et al., 2021), i.e., death is a competing risk for CVD. Hence, we construct a composite event that is either CVD or death and adjust the time-to-event outcome accordingly. A significant challenge arises from time stamp ambiguities regarding the onset of DM in the UKB data, a common issue in many electronic health record (EHR) datasets for various diseases. Consequently, the time-to-development of CVD from the onset of DM is partly interval-censored with 8.9% interval-censored and 91.1% right-censored in the UKB white female cohort and 16.4% interval-censored and 83.6% right-censored in the white male cohort.

For both the male and female cohorts, we applied a DPMIV model (4.3.4)-(4.3.8). In this model, Y_i represents the log-transformed time-to-development of cardiovascular disease (CVD) from the onset of diabetes mellitus (DM). The endogenous covariate of interest, X_i, corresponds to the standardized log-transformed SBP level. The instrumental variables G_i consist of 15 SNPs known to be associated with SBP (refer to Appendix for details on SNP selection). Additionally, the vector Z_i encompasses observed potential confounders, such as age at recruitment, cholesterol levels, body mass index (BMI), smoking status (yes vs. no), and physical activity level measured in metabolic equivalents (MET, range: 2-8 h/week). The priors used in the DPMIV model are informed by a training set consisting of 5% of the total dataset (see Appendix E). The instrumental variable strength (partial R-squared) of G are 0.056 for the female cohort and 0.056 for the male cohort. Subsequently, we employ a log-normal accelerated failure time (AFT) model (4.3.5) for analyzing partly interval-censored data, Anderson-Bergman (2017) with estimated coefficients and standard errors serving as hyperparameters for the second-stage priors on β_1, β_2 and ξ_{i2} in the DPMIV model.

Figure 4.2: True and estimated error distributions of the DPMIV method for simulation studies under different sample sizes.



The specifics of these priors used in both cohorts of the UKB data are detailed in Appendix. For the DPMIV method, we run 6 chains separately with length 1,200,000, 200 thinning and 200,000 burn-in samples. We also run 51 chains with length 3,200,000 and 200,000 burn-in samples and the results are the same. A large thinning value reduces the auto-correlation among posterior samples and a large burn-in value ensures the chain enters the stationary distribution (Robert et al., 1999).

Table 4.3 provides a summary of the estimated causal effect (β_1), its associated standard error, and the 95% credible interval (CI) for both the male and female cohorts using three distinct methods: the DPMIV method, the PBIV method, and a naive single-stage AFT model designed for interval-censored data using the "icenReg" R package (Anderson-Bergman, 2017).

Table 4.3 also reveals that, in the case of the female cohort, the causal effect estimated by the proposed DPMIV method is $\hat{\beta}_1 = -0.363$ (95% CI = (-0.670, -0.092)). This finding indicates that a higher systolic blood pressure (SBP) level is associated with a significantly shorter time-to-cardiovascular disease (CVD) from the onset of diabetes mellitus (DM). To put this into perspective, if SBP increases by 10%, then the expected survival time from DM to CVD will be shortened by a factor of $10\% \times \beta_1 \approx 3.6\%$. Interestingly, this result aligns with recent findings in the literature, as reported by studies such as Chan et al. (2021) and Wan et al. (2021). Moreover, our DPMIV analysis indicates that the error distribution is a mixture of $k = 5$ bivariate normal distributions, with two dominant clusters (the mixing proportions are around 97% and 1.5% for the white female cohort and 95% and 2.5% for the white male cohort). This observation is further substantiated by density contour plots of the estimated error distribution displayed in Figure 4.3.

It is important to highlight that, in the case of the female cohort, the naive single-stage AFT model produced a positive estimated coefficient $\hat{\beta}_1 = 0.262$. This unexpected result suggests that a higher systolic blood pressure (SBP) level is associated with a longer time-to-cardiovascular disease (CVD) from the onset of diabetes mellitus (DM), contrary to what one might intuitively expect. This anomaly can likely be attributed to the omission of significant confounders, such as annual income and drinking habits, as well as potential measurement errors in SBP by the naive single-stage AFT model. These findings underscore the critical need to address unobserved confounders and measurement errors when conducting causal analyses.

It is worth noting that the PBIV method also yielded a positive causal effect estimate for the female cohort, with $\hat{\beta}_1 = 0.589$ (95% CI = (0.245, 0.870)). However, this seemingly unreasonable

result from the PBIV method may be attributed to the highly heterogeneous error distribution characterized by five clusters, as estimated by the DPMIV method. To this end, we recall that our simulation studies in Section 3 (as seen in scenarios 5 and 6 in Table 4.2) suggest that under such conditions, the PBIV method may face significant challenges, which can result in substantially biased causal effect estimates and thus misleading findings.

Finally, Table 4.3 unveils remarkably similar results for the male cohort.

Table 4.3: Comparison of approaches for the analysis of UKB data

Female Cohort (n=3141) Partial R-squared = 0.056			
	Estimated causal effect (β_1)	SE	95% CI
DPMIV with SNPs as instruments	-0.363	0.146	(-0.670, -0.092)
PBIV with SNPs as instruments	0.589	0.181	(0.245, 0.870)
Naive AFT Model	0.262	0.230	(-0.201, 0.703)
Male Cohort (n=5029) Partial R-squared = 0.056			
	Estimated causal effect (β_1)	SE	95% CI
DPMIV with SNPs as instruments	-0.356	0.092	(-0.537, -0.173)
PBIV with SNPs as instruments	0.562	0.150	(0.255, 0.837)
Naive AFT Model	0.288	0.150	(-0.005, 0.581)

DPMIV with SNPs as instruments refers to our proposed method with the selected 15 SNPs. PBIV with SNPs as instruments refers to the extension of parametric Bayesian method proposed in Li and Lu (2015) where the details of the algorithm are given in the appendix. Single-stage AFT model without instruments refers to the interval-censored AFT model implemented in Anderson-Bergman (2017) and we do not include instruments in the model.

4.5.1 Using PSO to Find Maximal Correlation between SNPs and SBP

A key assumption in IV analysis is that the correlation between instruments (SNPs) and the causal effect covariate (SBP) should not be too small. Hence, we apply PSO to select a linear combination of SNPs (15 in total) such that the maximal correlation (Rényi, 1959) between the SNPs and SBP is maximized.

Figure 4.3: Log-density contour plot of random errors (ξ_1, ξ_2) of the Dirichlet process mixture model for the UKB data.

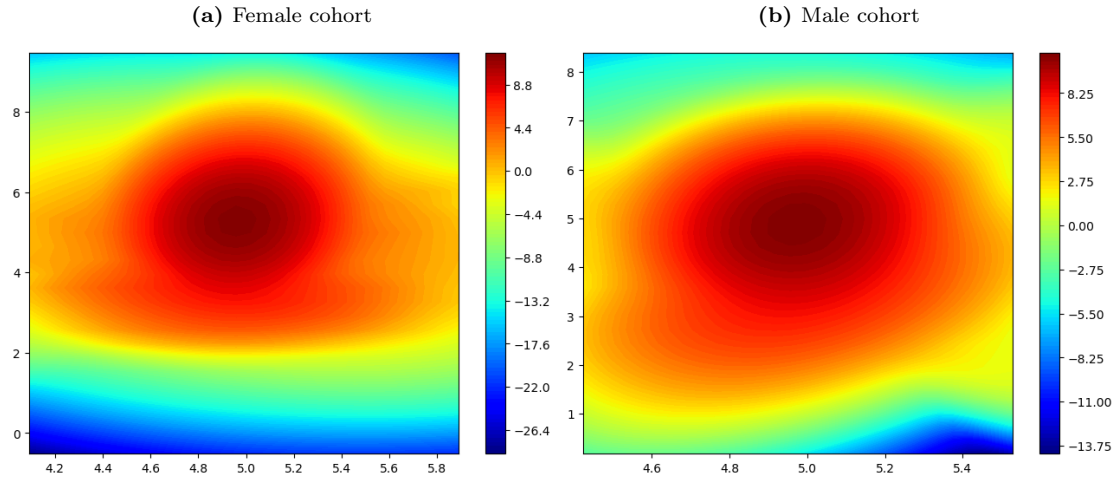
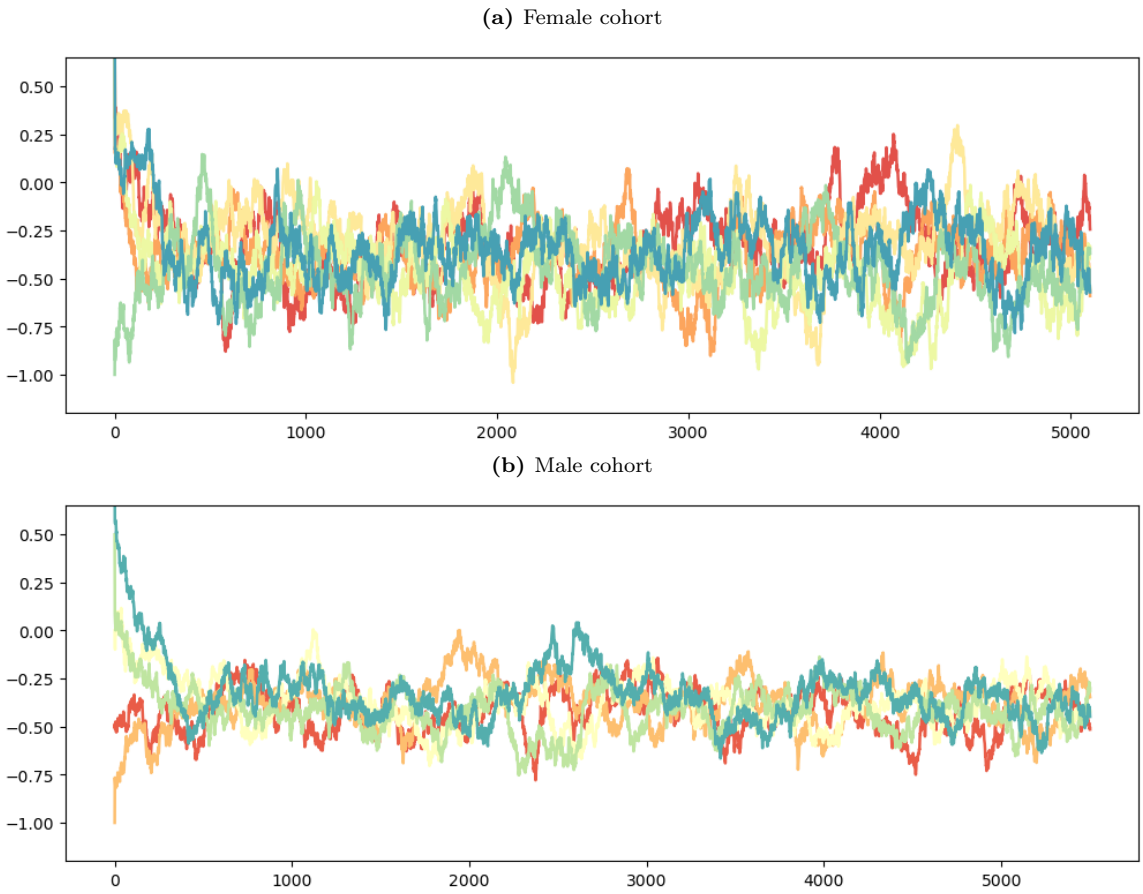


Figure 4.4: Trace plots of causal effect β_1 of the Dirichlet process mixture model for the UKB data.



4.6 Extensions to Doubly Interval-censored Data with Interval-censored Covariates

In this section, we first give a brief review on handling doubly interval-censored data (Sun, 2006) and then extend our previous approach to deal with it.

4.6.1 Introduction to Doubly Interval-censored Data and Problem Formulation

Doubly interval-censored data, also known as doubly censored data in literature, arise in studies where both the time of the originating event (denote as S_i) and the failure event (denote as T_i) are either right- or interval-censored (Kim et al., 1993). The basic structure of doubly interval-censored is

$$\{(U_i, V_i], (L_i, R_i], \mathbf{Z}_i, i = 1, \cdots, n\} \tag{4.6.1}$$

where

- $S_i \in (U_i, V_i]$ is the 1^{st} interval-censored time (time-to-DM in our application).
- $T_i \in (L_i, R_i]$ is the 2^{nd} interval-censored time (time-to-CVD in our application).
- $\mathbf{Z}_i$ is a vector of (possibly time-dependent and even interval-censored) covariates (age, sex, MET, etc.).
- $Y_i = T_i - S_i$ is the time of interest (time from DM to CVD in our application).

Another example of doubly censored data occurs frequently in acquired immune deficiency syndrome (AIDS) cohort studies (Sun et al., 1999) where we are interested in estimating the time from the Type-I human immunodeficiency virus (HIV-1) infection to the diagnosis of AIDS. The HIV-1 infection time is our S_i. It is interval-censored and we only observe $(U_i, V_i]$ in practice because the recruitment of HIV-1 positive patients into the studies and the fact that the infection times of these patients can usually only be determined retrospectively to lie in some intervals. The time to AIDS on-site is our T_i while in practice, we only observe $(L_i, R_i]$ due to the same reason. We note that in this case, if it is censored, then it is right-censored, i.e., $R_i = +\infty$.

4.6.1.1 Brief Review of Methods for Doubly Interval-Censored Data

There is a growing number of statistical literature dealing with doubly censored data in the past few decades, focusing on nonparametric estimation of survival functions, estimation of survival quantities, parametric and semiparametric regression models. We give a brief historical review below.

Given a set of interval-censored data without covaraites, Turnbull (1976) proposed to maximize the empirical likelihood function and used a self-consistency algorithm; his approach was termed as Turnbull's estimator. However, Turnbull's estimator only applies to interval-censored data so extension to doubly censored data is needed. De Gruttola and Lagakos (1989) is one of the first papers that generalizes Turnbull's estimator and his self-consistency algorithm to analyze doubly censored data in AIDS studies. Chang (1990) established the weak convergence result of Turnbull's estimator for doubly censored data. Kim et al. (1993) extended the previous work to a Cox's proportional hazards model with doubly censored data using a marginal likelihood approach. That is, they compute the marginal likelihood of observations (4.6.1) and maximize it with respect to regression coefficients. For doubly censored current status data, an accelerated failure time (AFT) model was proposed by Rabinowitz and Jewell (1996). In Sun et al. (1999) and Sun et al. (2004), the authors proposed a general regression framework when S_i is interval-censored and T_i is right-censored and they applied counting process and martingale theory to establish the corresponding asymptotic results for Cox's proportional hazards model and Aalen's additive risk model. Because of it simplicity and generality, we give an introduction of their approach in the sequel.

Suppose that the hazard of $\widetilde{E}_i$ at time t is of form (Sun, 2006)

$$\lambda_i(t) = \lambda_0(t)\exp(\mathbf{Z}_i^T\beta)$$

given $\mathbf{Z}_i$ where $\lambda_0(t)$ is an unknown baseline hazard and β is the vector of regression coefficients. For estimation of β, define the risk process $Y_i(t|\widetilde{S}_i) = \mathbb{I}(\widetilde{E}_i - \widetilde{S}_i \geq t)$ and the counting process $N_i(t|\widetilde{S}_i) = \mathbb{I}(\widetilde{E}_i - \widetilde{S}_i \leq t, \delta_i = 1)$. Using the conventional notation, let $\mathbf{S} = (\widetilde{S}_1, \cdots, \widetilde{S}_n)$ and

$$S^{(j)}(t;\beta|\mathbf{S}) = \frac{1}{n}Y_i(t|\widetilde{S}_i)\mathbf{Z}_i^j\exp(\mathbf{Z}_i^T\beta),$$

$j = 0, 1$, where $\mathbf{Z}_i^0 = 1$ and $\mathbf{Z}_i^1 = \mathbf{Z}_i$. Conditioned on $\mathbf{S}$, we can estimate β solving the score

equation (Andersen et al., 2012)

$$U_p(\beta|\mathbf{S}) = \int_0^\tau \sum_{i=1}^n \left[Z_i - \frac{S^{(1)}(t;\beta|\mathbf{S})}{S^{(0)}(t;\beta|\mathbf{S})} \right] dN_i(t|\widetilde{S}_i) = 0$$

where τ is the upper bound of $\widetilde{E}_i - \widetilde{S}_i$. Because $\mathbf{S}$ is unknown (or unobservable), it is natural to use the profile likelihood idea and marginalize $\mathbf{S}$. That is, we first estimate the cumulative distribution function of the $\widetilde{S}_i$'s based on interval-censored data $(\widetilde{L}_i, \widetilde{R}_i]$ using Turnbull's estimator. We denote it as $\widehat{H}$. Then we estimate β by the solution, say, $\widehat{\beta}$, to the following marginalized estimating equation

$$U_p(\beta, \widehat{H}) = \left(\prod_{i=1}^n a_i^{-1} \right) \int_{\widetilde{L}_1}^{\widetilde{R}_1} \cdots \int_{\widetilde{L}_n}^{\widetilde{R}_n} U_p(\beta|\mathbf{s}) \prod_{i=1}^n \left[d\widehat{H}(s_i) \right] = 0, \tag{4.6.2}$$

where $a_i = \int_{\widetilde{L}_i}^{\widetilde{R}_i} d\widehat{H}(s_i), i = 1, 2, \cdots, n$. If all $\widetilde{S}_i$'s are exactly observed, then $\widehat{\beta}$ boils down to the maximum partial likelihood estimator.

Sun (2001) also developed a nonparametric test for doubly censored data. More recently, Ji et al. (2012) proposed a quantile regression and asymptotic results, including the uniform consistency and weak convergence, are established. Wong et al. (2023) proposed a sieve maximum likelihood approach to address the infeasibility of maximum likelihood estimator. The methods mentioned above are from a frequentist perspective. Bayesian methods are also develoepd for doubly censored data. Among the early work within a Bayesian paradigm, Komárek and Lesaffre (2006) and Komárek and Lesaffre (2008) are two representative papers and they proposed Bayesian AFT model for doubly censored data. The Cox's proportional hazards model with priors are studied in Yu (2010). Jara et al. (2010) developed a more elaborated model based on Pitman-Yor processes (Pitman and Yor, 1997) and it can handle multivariate doubly censored data. Recently, Zeng et al. (2016) proposed a EM-type algorithm for semiparametric transformation models with interval-censored data with the presence of time-dependent covariates.

4.6.1.2 Brief Review of Methods for Interval-censored Covariates

Besides doubly censored data, interval-censored covariates also puts a great challenge in front of applied statisticians. Goggins et al. (1999) presents an example of interval-censored covariate

arising from the AIDS Clinical Trial Group (ACTG) 181. In the study, the interest is to model the relationship between the status of cytomegalovirus (CMV) and the onset of active CMV end-organ disease. Because only interval-censored data are available for the time-to-CMV shedding so it is an interval-censored covariate in the model (Goggins et al., 1999; Goggins and Finkelstein, 2000; Sun, 2006). According to Morrison et al. (2022), there are three mainstreams of handling interval-censored covariate:

- Midpoint imputation of the interval-censored covariate (Rubin, 1976; Little and Rubin, 2019), i.e., if a covariate $X_i \in (a_i, b_i]$, then we replace it with $\frac{a_i+b_i}{2}$;
- Multiple imputation with uniform distribution of the interval-censored covariate (Konikoff et al., 2013);
- Joint modeling or simultaneuous estimation of the nuisance distribution of the interval-censored covariate and the regression model (e.g., Frailty model) (Hsiao, 1983; Goggins et al., 1999; Gómez et al., 2003; Topp and Gómez, 2004; Du et al., 2021; Morrison et al., 2022; Melis et al., 2023).

For the third mainstream, the approach of Goggins et al. (1999), referred to as the GFZ method, suggested to treat Z_i (the interval-censored covariate) as a latent variable and they developed a Monte Carlo EM algorithm (Levine and Casella, 2001) for parameter estimation. One open problem with the GFZ method is that no asymptotic justification is available yet for the derived parameter estimates. Section 10.3.2 in Sun (2006) provides an extension of the GFZ method to doubly censored data with interval-censored covariate. His method is to integrate out $\mathbf{Z}_i$ in equation (4.6.2) using a NPMLE of the cumulative distribution function of the $\mathbf{Z}_i$'s. The approach of Gómez et al. (2003), referred to as the GEL approach, and the approach of Morrison et al. (2022), referred to as the MLB approach, model the distribution of time-to-event, interval-censored covariate and longitudinal outcomes simultaneously. In addition, an EM estimating procedure is proposed by Morrison et al. (2022).

4.6.2 An Imputation-based Approach Based on Turnbull's Estmator

In this section, we propose a new approach for handling doubly censored data. The idea can be described as follows.

- We first perform Turnbull's estimator on the first interval-censored data.
- We then sample the imputed $\widehat{S}_i$ based on each individual's $(U_i, V_i]$. If $U_i = V_i$, then there is no need to sample but let $S_i = U_i = V_i$ (exact event).
- Based on the imputed $\widehat{S}_i$'s, we define the imputed time of interest $\widehat{Y}_i = T_i - \widehat{S}_i$ and the imputed second interval-censored time $(L_i - \widehat{S}_i, R_i - \widehat{S}_i]$ for $i = 1, 2, \cdots, n$.
- Perform DPMIV analysis based on the vector of covariates $\mathbf{Z}_i$ (including the exposure X_i) and the second interval-censored data.
- Repeat the above four steps M times to create multiple imputed datasets.

One of the advantages of the proposed approach is that it reduces the more complicated doubly interval-censored data to partly interval-censored data by imputing the first interval-censored outcome. The imputation procedure is based on sampling from the conditional probability vector estimated by Turnbull's estimator.

4.6.3 Simulation Studies

Similar to the partly interval-censored case, we conduct simulations to evaluate the performance of proposed two-stage DPMIV method for doubly interval-censored time-to-event data under a variety of scenarios. We compare the proposed imputation-based estimator with two other imputation methods: uniform imputation and midpoint imputation. Additionally, we include two other methods for reference in our simulation analysis: 1) the naive single-stage accelerated failure time (AFT) model for partly interval-censored data (Huang and Wellner, 1997; Anderson-Bergman, 2017), which does not account for unobserved confounders and measurement errors, and 2) the two-stage AFT model, i.e., we first regress the exposure on instruments and then regress the outcome on the fitted exposure.

We simulated data from model (4.3.4)-(4.3.5) with a two-dimensional instrument G_i and a two-dimensional observed confounder U_i, both following a standard bivariate normal distribution $N(0, I_2)$. The regression parameters in equation (4.3.4) were set as $\alpha_1 = (0.5, 0.5)^T$, and $\alpha_2 = (0.5, 0.5)^T$. The regression parameters in equation (4.3.5) were set as $\beta_1 = -1$, and $\beta_2 = (0.8, 0.8)^T$.

Similar to the partly interval-censored case, we considered six scenarios for the bivariate distribution of $(\xi_{1i}, \xi_{2i})^T$:

1. Bivariate normal distribution.
2. Bivariate exponential distribution as described in Equation 18 in Gumbel (1960).
3. Mixture of two bivariate normal distributions with different means but the same variance-covariance matrix.
4. Mixture of two bivariate normal distributions with the same mean but different variance-covariance matrices.
5. Mixture of five bivariate normal distributions, mimicking the distribution of the female cohort estimated from the UK Biobank dataset using the DPMIV method.
6. Mixture of five bivariate normal distributions, mimicking the distribution of the male cohort estimated from the UK Biobank dataset using the DPMIV method.

Detailed specifications for the bivariate distribution of $(\xi_{1i}, \xi_{2i})^T$ under these six simulation scenarios can be found in Table 7.11.

Similarly to the simulation settings in Sun et al. (2004), we generate doubly interval-censored data (U_i, V_i), S_i, (L_i, R_i), T_i as follows. In each simulated dataset, we assume

- $\log S_i \sim 2 \times Beta(5, 5)$;
- $S_i - U_i \sim \text{Uniform}(0, 0.2)$, $V_i - S_i \sim \text{Uniform}(0, 0.2)$;
- $T_i = \exp(Y_i) + S_i$;
- T_i is assumed to be right-censoring only, i.e., either $R_i = L_i$ or $R_i = +\infty$;
- Right-censoring is generated as $C_i \sim \text{Uniform}(V_i, V_i + 2)$, then $L_i = \min(C_i, T_i)$, $R_i = L_i$ if $L_i = T_i$ else $R_i = +\infty$.

A natural question is, how many multiple imputation datasets do we need? We conduct a simulation study to show the effect of the number of imputed datasets on the bias of causal effect estimate in Figure 4.5. The x-axis is the number of imputed datasets and the y-axis represents six different

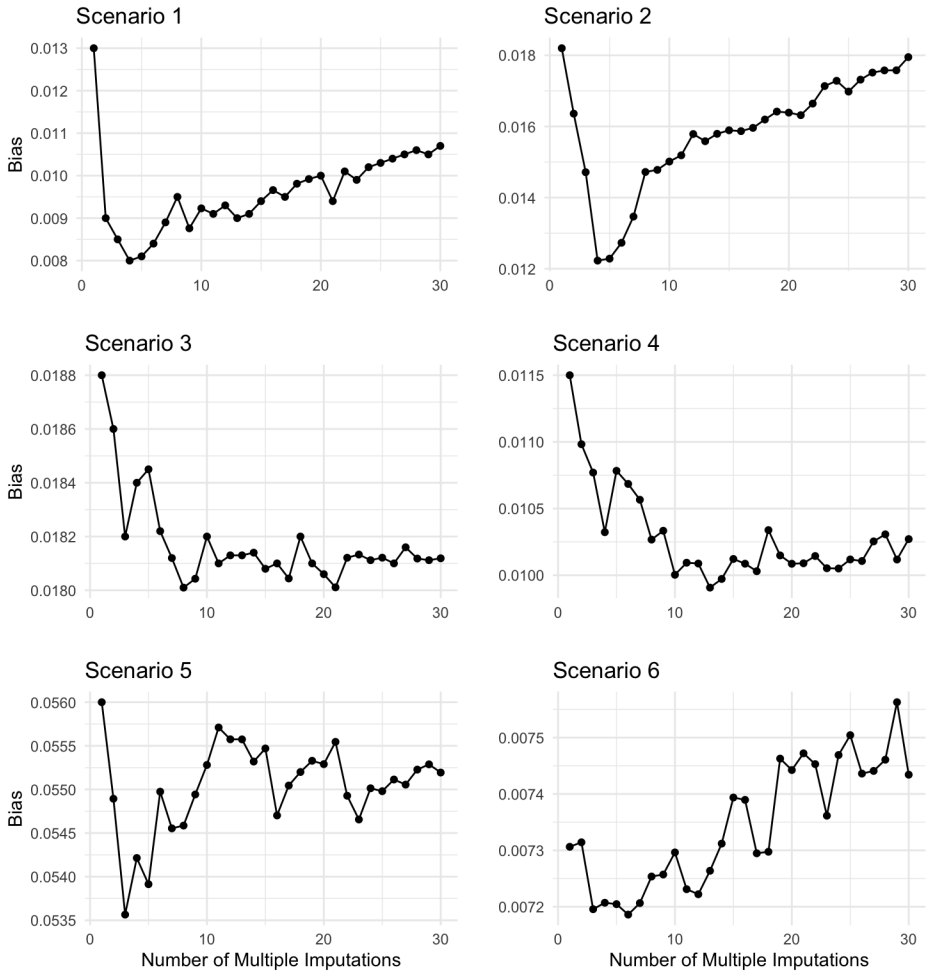


Figure 4.5: Bias of β_1 Using Multiple Imputations. The x-axis represents the number of multiple imputations for sample size 500; the y-axis represents the average bias using Turnbull's estimator for imputation under different simulation scenarios.

simulation scenarios. From the figure, we recommend that $5\tilde{1}0$ imputed datasets would be sufficient to produce reasonable results and in the simulation study, we set it to 5.

Finally, we considered different sample sizes $n = 300, 500$ and 1000 under each scenario. Table 4.4 presents a summary of the simulated bias, standard deviation (SD), and coverage probability (CP) for the causal parameter β_1 using the three aforementioned methods based on 100 Monte Carlo replications. Additionally, we also report the average of the estimated number of clusters k using the DPMIV method.

Table 4.4: β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; Two-stage AFT estimate refers to AFT model with instrumental variables; DPMIV refers to our proposed method.

Scenario	Error Distribution	Imputation	n	Two-stage AFT estimate			Single-stage AFT estimate			DPMIV estimate			
				Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	Midpoint	300	0.098	0.118	90%	0.400	0.073	1%	0.036	0.122	92%	1.453
		Uniform	300	0.098	0.121	91%	0.391	0.075	1%	0.022	0.125	93%	1.690
		Turnbull	300	0.107	0.126	90%	0.371	0.078	1%	0.006	0.127	93%	1.906
		Midpoint	500	0.072	0.090	94%	0.393	0.056	1%	0.038	0.094	95%	1.651
		Uniform	500	0.071	0.092	95%	0.386	0.057	1%	0.031	0.095	94%	1.694
		Turnbull	500	0.072	0.095	96%	0.371	0.059	1%	0.005	0.096	95%	1.456
		Midpoint	1000	0.057	0.064	88%	0.393	0.040	1%	0.044	0.066	89%	1.876
		Uniform	1000	0.054	0.065	91%	0.384	0.041	1%	0.030	0.067	91%	1.968
		Turnbull	1000	0.050	0.066	91%	0.376	0.042	1%	0.023	0.067	99%	2.032
2	Exponential	Midpoint	300	0.154	0.153	93%	0.167	0.042	5%	0.038	0.046	82%	5.755
		Uniform	300	0.132	0.162	93%	0.219	0.049	0%	0.019	0.059	85%	6.285
		Turnbull	300	0.121	0.170	95%	0.277	0.054	1%	0.004	0.058	89%	7.387
		Midpoint	500	0.140	0.117	91%	0.201	0.028	1%	0.045	0.035	76%	7.367
		Uniform	500	0.118	0.124	92%	0.213	0.038	1%	0.027	0.043	89%	8.183
		Turnbull	500	0.099	0.129	96%	0.250	0.041	1%	0.014	0.045	92%	7.595
		Midpoint	1000	0.156	0.082	64%	0.145	0.025	1%	0.028	0.025	79%	3.950
		Uniform	1000	0.127	0.087	77%	0.201	0.027	1%	0.021	0.030	85%	9.346
		Turnbull	1000	0.108	0.028	87%	0.223	0.029	1%	0.010	0.031	94%	8.610
3	Normal Mixture I	Midpoint	300	0.124	0.179	97%	0.253	0.073	2%	0.044	0.109	96%	2.233
		Uniform	300	0.114	0.184	98%	0.280	0.076	1%	0.033	0.118	95%	2.302
		Turnbull	300	0.101	0.191	98%	0.328	0.081	1%	0.003	0.123	96%	2.343
		Midpoint	500	0.130	0.138	92%	0.245	0.056	1%	0.038	0.073	95%	2.617
		Uniform	500	0.114	0.141	95%	0.273	0.059	1%	0.021	0.079	95%	2.731
		Turnbull	500	0.103	0.145	97%	0.307	0.061	1%	0.018	0.081	94%	2.903
		Midpoint	1000	0.140	0.097	75%	0.243	0.041	1%	0.042	0.046	83%	2.760
		Uniform	1000	0.119	0.100	89%	0.275	0.042	1%	0.025	0.051	90%	3.270
		Turnbull	1000	0.109	0.101	91%	0.291	0.043	1%	0.008	0.053	91%	3.516

Table 4.5: β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Continued.

Scenario	Error Distribution	Imputation	n	Two-stage AFT estimate			Single-stage AFT estimate			DPMIV estimate			
				Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
4	Normal Mixture II	Midpoint	300	0.064	0.081	93%	0.228	0.051	2%	0.014	0.065	96%	2.595
		Uniform	300	0.066	0.084	95%	0.217	0.053	2%	0.041	0.066	88%	2.958
		Turnbull	300	0.068	0.088	89%	0.195	0.055	9%	0.016	0.067	90%	3.022
		Midpoint	500	0.046	0.063	92%	0.234	0.039	1%	0.014	0.043	87%	2.730
		Uniform	500	0.047	0.065	92%	0.273	0.040	1%	0.024	0.044	90%	2.915
		Turnbull	500	0.051	0.067	93%	0.207	0.042	2%	0.013	0.044	93%	2.853
		Midpoint	1000	0.033	0.045	93%	0.230	0.028	1%	0.008	0.028	97%	2.711
		Uniform	1000	0.033	0.046	94%	0.217	0.029	1%	0.018	0.029	94%	3.081
		Turnbull	1000	0.037	0.047	93%	0.209	0.030	1%	0.010	0.029	99%	3.087
5	Normal Mixture III	Midpoint	300	0.261	0.139	51%	0.314	0.135	31%	0.128	0.142	77%	2.571
		Uniform	300	0.249	0.145	60%	0.302	0.140	38%	0.093	0.145	83%	2.469
		Turnbull	300	0.230	0.154	68%	0.281	0.149	48%	0.084	0.157	85%	2.612
		Midpoint	500	0.231	0.109	44%	0.296	0.105	19%	0.085	0.101	84%	3.224
		Uniform	500	0.216	0.113	50%	0.281	0.109	24%	0.053	0.101	89%	3.244
		Turnbull	500	0.198	0.119	61%	0.259	0.115	35%	0.054	0.106	91%	3.244
		Midpoint	1000	0.225	0.077	15%	0.284	0.075	1%	0.062	0.063	86%	3.408
		Uniform	1000	0.209	0.081	26%	0.269	0.078	2%	0.019	0.067	91%	3.632
		Turnbull	1000	0.190	0.084	36%	0.252	0.081	10%	0.018	0.067	92%	3.636
6	Normal Mixture IV	Midpoint	300	0.306	0.183	61%	0.422	0.173	17%	0.095	0.129	75%	2.755
		Uniform	300	0.302	0.189	61%	0.417	0.178	19%	0.081	0.136	89%	3.535
		Turnbull	300	0.285	0.199	65%	0.404	0.188	33%	0.054	0.135	87%	3.656
		Midpoint	500	0.316	0.125	29%	0.406	0.135	8%	0.052	0.097	85%	3.225
		Uniform	500	0.306	0.129	32%	0.392	0.139	15%	0.028	0.097	92%	3.204
		Turnbull	500	0.289	0.136	44%	0.373	0.145	20%	0.004	0.100	94%	3.428
		Midpoint	1000	0.294	0.101	10%	0.403	0.096	1%	0.033	0.064	87%	3.367
		Uniform	1000	0.277	0.105	21%	0.389	0.099	2%	0.009	0.065	87%	3.387
		Turnbull	1000	0.262	0.107	33%	0.376	0.101	1%	0.014	0.067	93%	3.406

Table 4.6: DPMIV β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring and different imputation strategies.

Scenario	Error Distribution	n	Midpoint			Right			Uniform			Turnbull		
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	Bias	SD	CP
1	Normal	300	0.035	0.120	93%	0.072	0.139	89%	0.021	0.123	94%	0.007	0.125	94%
		500	0.039	0.092	96%	0.133	0.109	72%	0.030	0.093	95%	0.006	0.094	96%
		1000	0.043	0.064	90%	0.111	0.076	65%	0.029	0.065	92%	0.022	0.065	100%
2	Exponential	300	0.037	0.044	83%	0.067	0.041	58%	0.020	0.057	86%	0.005	0.056	90%
		500	0.044	0.033	77%	0.062	0.029	43%	0.028	0.041	90%	0.015	0.043	93%
		1000	0.029	0.023	80%	0.061	0.019	11%	0.022	0.028	86%	0.011	0.029	95%
3	Normal Mixture I	300	0.043	0.107	97%	0.074	0.142	89%	0.032	0.116	96%	0.002	0.121	97%
		500	0.037	0.071	96%	0.096	0.097	85%	0.020	0.077	96%	0.017	0.079	95%
		1000	0.041	0.044	84%	0.108	0.063	64%	0.024	0.049	91%	0.007	0.051	92%
4	Normal Mixture II	300	0.013	0.063	97%	0.080	0.066	81%	0.040	0.064	89%	0.015	0.065	91%
		500	0.013	0.041	88%	0.074	0.046	64%	0.023	0.042	91%	0.012	0.042	94%
		1000	0.007	0.026	98%	0.057	0.030	51%	0.017	0.027	95%	0.009	0.027	100%
5	Normal Mixture III	300	0.127	0.140	78%	0.091	0.170	96%	0.092	0.143	84%	0.083	0.155	86%
		500	0.084	0.099	85%	0.004	0.131	97%	0.052	0.099	90%	0.053	0.104	92%
		1000	0.061	0.061	87%	0.051	0.081	89%	0.018	0.065	92%	0.017	0.065	93%
6	Normal Mixture IV	300	0.094	0.127	76%	0.082	0.178	94%	0.080	0.134	90%	0.053	0.133	88%
		500	0.051	0.095	86%	0.004	0.129	95%	0.027	0.095	93%	0.003	0.098	95%
		1000	0.032	0.062	88%	0.034	0.079	93%	0.008	0.063	88%	0.013	0.065	94%

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

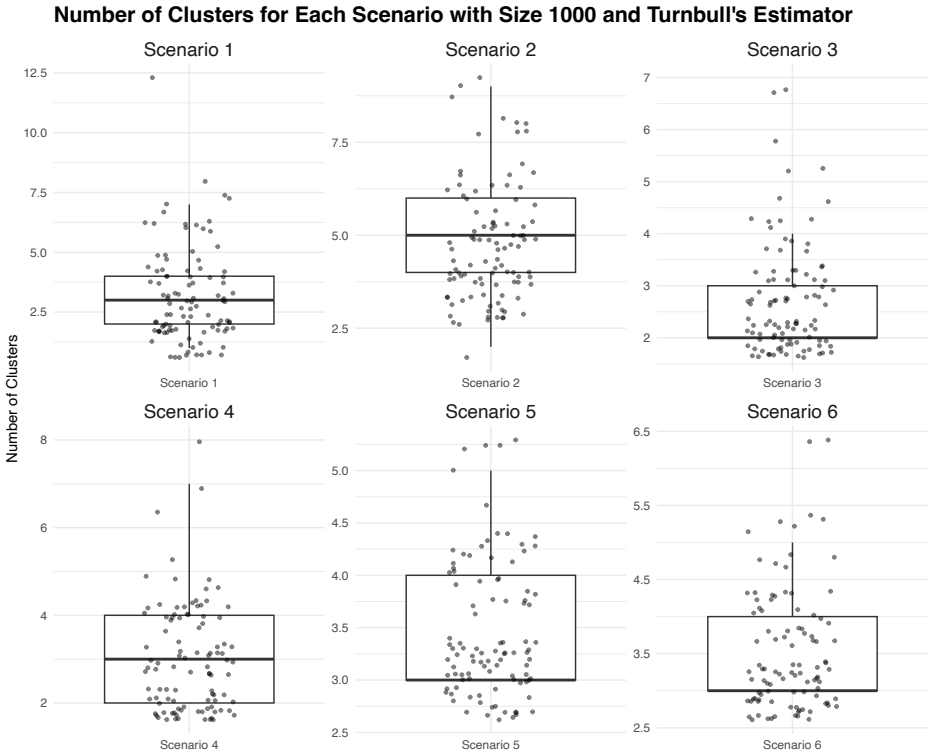


Figure 4.6: Number of Clusters for Each Scenario with Size 1000 and Turnbull's Estimator

Table 4.4 presents a summary of the simulated bias, standard deviation (SD), and coverage probability (CP) for the causal parameter $\beta_1 = -1$ using three different methods: the single-stage AFT estimate, the two-stage AFT estimate, and the proposed DPMIV estimate, based on 100 Monte Carlo replications. Additionally, for the proposed DPMIV method, the table reports the average of the estimated number of clusters k.

As observed in Table 4.4, the naive single-stage AFT model estimate generally exhibits substantial bias and unacceptably low coverage probability across all scenarios and sample sizes. This highlights the critical need to address unobserved confounders and measurement errors, as failing to do so results in biased estimates and poor coverage.

The two-stage AFT estimate performs better than the single-stage estimate, showing reduced bias and improved coverage probabilities. However, its performance varies depending on the scenario and the sample size, indicating that while it addresses some of the issues related to confounders and measurement errors, it does not fully resolve them.

Our proposed DPMIV method consistently delivers robust and stable performance, with min-

imal bias and satisfactory coverage probability across all six scenarios and various sample sizes. In the first scenario, it exhibits similar bias and standard deviation (SD) to the two-stage AFT estimate and outperforms it in terms of coverage probability. The DPMIV method also correctly identifies the number of clusters k as being close to the expected value, indicating its ability to adapt to the underlying data structure.

In Scenario 2, where the error distribution is exponential, the DPMIV method estimates k as 5.755, 6.285, and 7.387 for the different imputation methods with $n = 300$, suggesting a normal mixture approximation to the bivariate exponential distribution. This indicates the method's flexibility in modeling different types of data distributions. In Scenarios 3 and 4, which involve normal mixture distributions, the DPMIV method shows minimal bias and high coverage probability. It tends to estimate k as slightly higher than the true number of clusters, reflecting the presence of additional components in the data that capture the variability in the distribution. For Scenarios 5 and 6, which involve more complex normal mixture distributions, the DPMIV method's performance remains strong as the sample size increases. The method provides accurate estimates of β_1 with good coverage probabilities, and the estimated number of clusters k aligns well with the expected values, indicating its effectiveness in handling complex data structures.

Comparing the three imputation strategies (midpoint, uniform, and Turnbull), we observe notable differences in their performance:

- **Midpoint Imputation**: This strategy tends to show higher bias and lower CP in some scenarios compared to the other methods. While it is simple to implement, its performance is less reliable, especially in scenarios with more complex error distributions.
- **Uniform Imputation**: This strategy generally performs better than midpoint imputation, showing reduced bias and improved CP across most scenarios. It provides a good balance between simplicity and accuracy, making it a reasonable choice for many applications.
- **Turnbull Imputation (Proposed Strategy)**: Turnbull imputation consistently outperforms the other two strategies, particularly in scenarios with complex error distributions. It shows the least bias and highest CP across all scenarios, highlighting its robustness and reliability. The Turnbull method's superior performance makes it the preferred choice for handling mixed censoring in instrumental variable analysis.

Overall, the DPMIV method with Turnbull's estimator imputation strategy demonstrates superior performance in estimating the causal parameter β_1 under different simulation scenarios.

4.6.4 IV Analysis of Doubly Interval-censored UK Biobank Data

The UK Biobank (UKB) cohort comprises 500,000 individuals aged 40 to 69 years at baseline, recruited between 2006 and 2010 at 22 assessment centers across the United Kingdom. Participants were followed up until January 1, 2018, or until their date of death. This extensive resource provides data on genotyping, clinical measurements, assays of biological samples, and self-reported health behavior. As an illustrative example, we will investigate the causal effects of systolic blood pressure (SBP) on the time-to-development of cardiovascular disease (CVD) from the onset of diabetes mellitus (DM) with a focus on White individuals, including 3,141 females and 5,029 males, who developed DM before CVD. Descriptive statistics of baseline characteristics for this subgroup are summarized in Appendix E.

A significant challenge arises from time stamp ambiguities regarding the onset of DM in the UKB data, a common issue in many electronic health record (EHR) datasets for various diseases. The exact DM onset date remains unknown, only known to fall between two visit times. Consequently, the time-to-DM is partly interval-censored with 28.7% interval-censored, 16.2% left-censored, and 55.1% event in the UKB white male cohort and 37.8% interval-censored, 1.4% left-censored, and 60.7% event in the white female cohort. Similarly, the time-to-CVD is partly interval-censored with 85.1% right-censored, and 14.9% event in the UKB white male cohort and 2.6% interval-censored, 87.7% right-censored, and 9.7% event in the white female cohort.

Applying the proposed imputation-based approach, we first imputed five interval-censored time-to-CVD datasets for each cohort and then run DPMIV for each imputed dataset. Figure 4.7 has shown two samples of the imputed NPMLE (the top is from the female cohort and the bottom five is from the male cohort). More samples are provided in the Appendix 7.3.8. The left panels represent conditional survival probabilities from the Turnbull's estimator and the right panels plot the empirical cumulative distribution function (ECDF) based on the left panels; the grey dashed lines are cdf of a uniform distribution. It is clear that the empirical distribution of the UKB study is far from a uniform distribution. Hence, it might not be appropriate to impute the time-to-DM using either midpoint imputation or uniform distribution approaches.

Some of the priors used in the DPMIV model are partly informed by the results from a preliminary two-stage least squares estimation method. To elucidate, we initially conduct an ordinary linear regression of X on G (SNPs) (refer to Equation (4.3.4)), utilizing the estimated coefficients and standard errors as hyperparameters for the first-stage priors on α_1, α_2 and ξ_{i1} in the DPMIV model.

Table 4.7 shows the doubly interval-censored analogue of Table 4.3 where the CIs of DPMIV and PBIV are constructed by combining posterior samples using different datasets and then taking the quantiles. Figure 4.8 shows the doubly interval-censored counterpart of Figure 4.4. The results are consistent with the interval-censored case, suggesting that our proposed method is reliable and stable. Similar density contour plots of the estimated error distribution are displayed in Figure 4.9, providing evidence for the efficacy of our proposed approach.

4.7 Discussion

In this chapter, we have developed DPMIV, a semiparametric Bayesian approach for IV analysis to examine the causal effect of a covariate on a partly interval-censored time-to-event outcome, in the presence of unobserved confounders and/or measurement errors in the covariate. We show by simulations that the proposed method largely reduces bias in estimation and it greatly improves coverage probability of the endogenous parameter, compared to the 'simple method' where the unobserved confounders and measurement errors are ignored and PBIV, the parametric Bayesian approach. The method works well in a variety of settings, provided that the instrumental variable assumptions described early in introduction are satisfied. We also extend the methods to doubly interval-censored data and it produces consistent results compared with the interval-censored data using the UKB data.

We note that one of the limitations of our MCMC algorithm is that it uses Neal's algorithm 8 which is slow and inefficient with a large number of auxiliary parameter m. We put the computational modification as one of our future works. In addition, a first step in IV analysis is to select which IVs are valid under the three assumptions in the introduction. Kang et al. (2016) developed the R package "sisVIVE" to select instrumental variables that are valid under the three IV assumptions and estimate the causal effect simultaneously. It is of interest to develop a Bayesian

method for selecting instrumental variables and estimating the causal effect simultaneously. One alternative is to replace the uniform priors in our algorithm with horseshoe priors (Carvalho et al., 2009) and we put it as our future work.

Table 4.7: Comparison of approaches for the analysis of doubly interval-censored UKB data. For completeness, we extend PBIV to handle doubly interval-censored data in a similar fashion as DPMIV does. The AFT model without instruments assumes a parametric log-normal error. For uniform and Turnbull's imputation strategies, we impute 5 different datasets and for each imputed dataset, we run 5 different chains. For the midpoint imputation strategy, there is only one imputed dataset and we also run 5 different chains.

Female Cohort				
	Imputation	Estimate of β_1	SE	95% CI
DPMIV with SNPs as instruments	Midpoint	-0.280	0.151	(-0.589, 0.005)
DPMIV with SNPs as instruments	Right	-0.104	0.162	(-0.429 , 0.214)
DPMIV with SNPs as instruments	Uniform	-0.271	0.170	(-0.605, 0.054)
DPMIV with SNPs as instruments	Turnbull	-0.351	0.177	(-0.694, -0.006)
PBIV with SNPs as instruments	Midpoint	0.756	0.216	(0.396, 1.334)
PBIV with SNPs as instruments	Right	0.688	0.162	(0.392, 1.026)
PBIV with SNPs as instruments	Uniform	0.875	0.198	(0.483, 1.264)
PBIV with SNPs as instruments	Turnbull	0.609	0.234	(0.113, 0.971)
AFT Model without instruments	Midpoint	0.396	0.195	(0.013,0.780)
AFT Model without instruments	Right	0.358	0.224	(-0.081,0.798)
AFT Model without instruments	Uniform	0.355	0.186	(-0.009,0.720)
AFT Model without instruments	Turnbull	0.447	0.237	(-0.017,0.812)

Male Cohort				
	Imputation	Estimate of β_1	SE	95% CI
DPMIV with SNPs as instruments	Midpoint	-0.329	0.118	(-0.527, -0.114)
DPMIV with SNPs as instruments	Right	-0.269	0.142	(-0.556, -0.008)
DPMIV with SNPs as instruments	Uniform	-0.344	0.128	(-0.614, -0.125)
DPMIV with SNPs as instruments	Turnbull	-0.405	0.144	(-0.672, -0.135)
PBIV with SNPs as instruments	Midpoint	0.603	0.200	(0.134, 0.989)
PBIV with SNPs as instruments	Right	0.452	0.132	(0.206, 0.714)
PBIV with SNPs as instruments	Uniform	0.495	0.197	(0.069 0.802)
PBIV with SNPs as instruments	Turnbull	0.527	0.155	(0.256, 0.870)
AFT Model without instruments	Midpoint	0.495	0.113	(0.274, 0.716)
AFT Model without instruments	Right	0.659	0.134	(0.396, 0.922)
AFT Model without instruments	Uniform	0.511	0.112	(0.292, 0.731)
AFT Model without instruments	Turnbull	0.503	0.129	(0.249, 0.756)

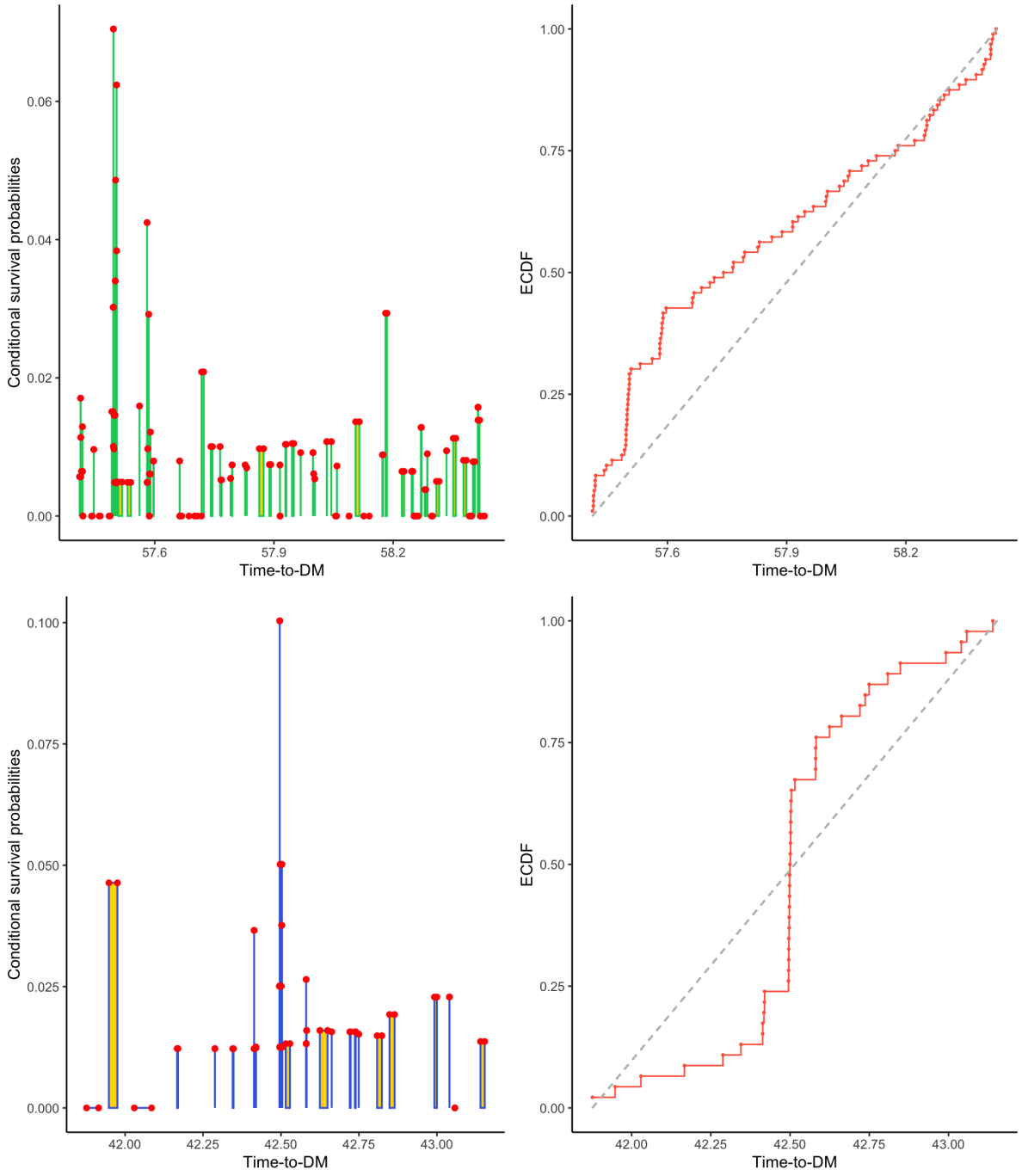
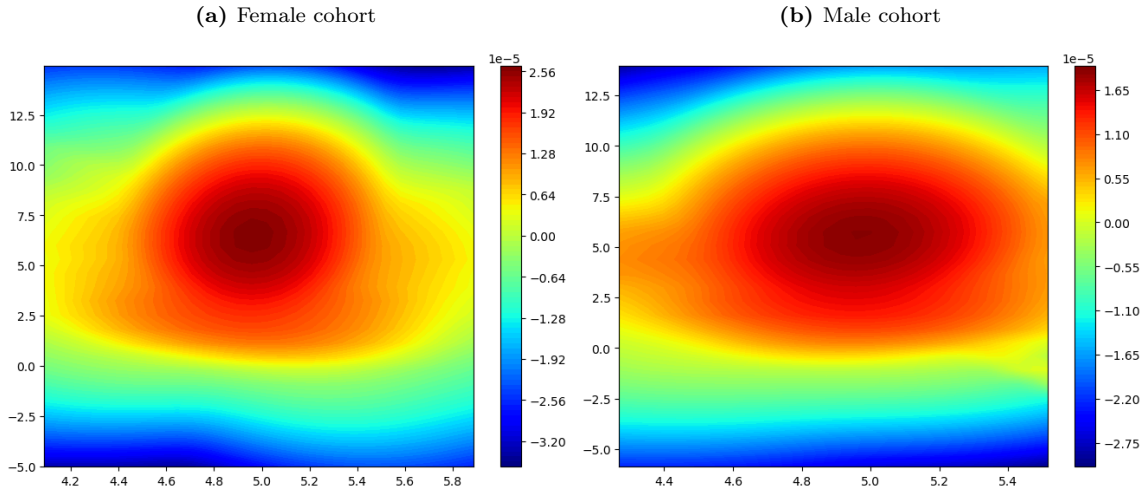


Figure 4.7: Two samples of the imputed NPMLE (Top: female; Bottom: male). The left panels represent conditional survival probabilities from the Turnbull's estimator. A single verticle line means it puts a probability mass at that particular age; a gold rectangle means it puts the probability on the interval. The right panels plot the empirical cumulative distribution function (ECDF) based on the left panels and the grey dashed lines are cdf of a uniform distribution.

Figure 4.8: Trace plots of the Female (top) and the Male (bottom) Cohorts. Each trace comes from an imputed dataset.

Figure 4.9: Log-density contour plot of random errors (ξ_1, ξ_2) of the DPMIV for the doubly interval-censored UKB data using Turnbull's estimator.



CHAPTER 5

Introduction to Optimal Design and Its Applications in Regression Models

5.1 Preamble

In this chapter, we show metaheuristics can be used to provide improved inference for toxicology experiments by finding more efficient designs. In section 5.2, we first illustrate the motivation of optimal design, and review the fundamental theory of optimal approximate designs, including the representation of information matrices and convex optimality criteria. The preliminary work consists of two separate parts. In section 5.3.1, we provide analytical formulae for various types of optimal designs and apply metaheuristics to find global optimal designs based on equivalence theorems. In section 5.4, we apply compound optimality criteria and build an R Shiny app to help toxicologists to design experiments at a lower cost.

5.2 Introduction to Optimal Design

5.2.1 Motivation of Optimal Design

In a dose–response experiment, decisions regarding the dose range, the number of doses, the dose levels, and the number of experimental units at each dose are sometimes made predicated on nebulous criteria. These are design issues that can potentially have a substantial impact on the quality of the statistical inference at the end of the study, yet they are decided in some cases on an ad-hoc basis. Frequently, an equal number of experimental units are assigned at each dose. When the doses are equally spaced, these are called uniform designs in the statistical literature and while they are appealing and intuitive, it has been shown that they can be inefficient, depending on the

goal of the study and the underlying model assumed. For example, Wong and Lachenbruch (1996) showed that performance of such designs can depend sensitively on the choice of the number of doses in a uniform design, the model, and the optimality criteria. Therefore, each aspect in the design of the study must be carefully considered to realize maximum accuracy in the information. Such attention to detail will enhance reproducibility, thus addressing a current issue in animal experimentation (Giles, 2006) and reducing the overall cost of experiments. More specifically, if the current cost for producing a new drug is 10 dollars per dose, then using optimal design theory, one is able to reduce the cost to 5 dollars per dose.

To optimally design an experiment, model assumptions are required to work out the mathematical and statistical details. Invariably, the goal is formulated as an objective function defined on the user-specified dose range (or design interval) that depends on the statistical model and the design. The optimization of the criterion can then be performed among a specific class of designs, for example, among all designs with five doses, or among all designs on a given dose interval. The resulting optimal design is therefore model-based and, as a consequence, can be highly model-dependent, suggesting that choice of a statistical model for the dose–response study is also important.

5.2.2 Basic Concepts of Optimal Approximate Design

Optimal approximate designs in clinical trials can help investigators achieve higher quality results for the given resource constraints (Schwaab et al., 2006; Jóźwiak and Moerbeek, 2013; Sverdlov et al., 2020; Zhou et al., 2021). The creation of this field can be traced back to Smith (1918). From 1950s to 1980s, the field of approximate design has witnessed a booming development (Federov, 1972; Kiefer, 1974; Pázman, 1986; Atkinson et al., 2007; Silvey, 2013) and we give a brief introduction below.

Consider a linear model $\mathbb{E}(y) = \mu = \beta^T f(x)$ where μ is the expectation of y. A k-point design ξ is a $2 \times k$ matrix of the following form

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_{k-1} & x_k \\ p_1 & p_2 & \cdots & p_{k-1} & p_k \end{pmatrix}$$

where x_i's are called design points and p_i's are non-negative weights that sum to 1. In practice, we

usually have a total of n observations and np_i is not an integer in general. Hence, we choose the closest integer of np_i and assign the corresponding dose x_i to these individuals. For this reason, ξ is referred as approximate design in literature. The information matrix associated with design ξ is

$$\mathfrak{M}(\xi) = \int_{\mathcal{X}} f(x)f(x)^T\xi(dx) = \sum_{i=1}^{k} p_i f(x_i)f(x_i)^T$$

where $\xi(dx)$ is the measure induced by p_i's. The optimal design seeks to find a ξ^* that minimizes $\phi(\mathfrak{M}(\xi))$ where $\phi(\cdot)$ is a real-valued function. Commons choices of $\phi(\cdot)$ and their terminologies are given in the following table

Table 5.1: List of Common Choices of $\phi(\cdot)$

Optimality[a]	**Choice of ϕ**	**Remarks**
A	$\mathrm{Tr}(\mathfrak{M}^{-1})$[b]	Sum of variances
c	$\mathrm{Var}(g(\widehat{\beta}))$[c]	Variance of $g(\widehat{\beta})$
D	$\log\det\mathfrak{M}^{-1}$	Log-volume of the ellipse
E	$\min\lambda_i(\mathfrak{M}) = \lambda_{min}$[d]	Length of minor axis
G	$\max_i(\mathfrak{M}^{-1})_{ii}$	Maximum of $\mathrm{Var}(\widehat{\beta})$
I	$\int_{\mathcal{X}} f(x)^T\mathfrak{M}^{-1}f(x)\mu(dx)$	Integrated variance

[a] For example, the first line reads A-optimality.

[b] Tr refers to the trace function of a matrix.

[c] g is a function of β, $\widehat{\beta}$ is the MLE of β and the variance of $g(\widehat{\beta})$ can be derived using Delta method, i.e., $\widehat{\mathrm{Var}}(g(\widehat{\beta})) = \nabla g(\widehat{\beta})^T\mathfrak{M}^{-1}\nabla g(\widehat{\beta})$ and ∇ refers to the gradient operator.

[d] λ's refer to eigenvalues of $\mathfrak{M}$.

Further, if we extend the linear model framework to generalized linear models, then the information matrix depends on parameters (see section 5.3.1). In this case, one usually plug-in plausible parameter values and then calculate the optimal design ξ^*. We call it ϕ-optimal design, where ϕ refers to D-, A-, c-, E-, etc. In practice, researchers would like to consider different criteria simultaneously, leading to the so-called compound criteria. For a comprehensive review of different optimality criteria, see Chapter 10 of Atkinson et al. (2007) or the review paper by Fedorov (2010).

To verify that the resulting design is globally optimal, that is, optimal among all possible designs, one needs to apply the equivalence theorem (Kiefer, 1974) and plot the sensitivity functions to check. Different optimality criteria corresponds to different types of equivalence theorems and sensitivity functions, hence for brevity, we state them only if necessary. In addition, Chen et al. (2022) provides a comprehensive review on the application of PSO in optimal approximate design.

5.2.3 The General Equivalence Theorem

The celebrated general equivalence theorem is not a single theorem but a series of theorems that specify the ϕ-optimality of a design. It was first proved by Kiefer (1959) and Kiefer and Wolfowitz (1960) showing the equivalence of D- and G-optimal designs in classical Gaussian linear models. Later, Fedorov (1971) extended it to multivariate linear regression models and White (1973) generalized it to non-linear models. The well-known paper by Kiefer (1974) proposed new optimality criteria and provided a general class of equivalence theorems. In 1980, S.D. Silvey collected the general theory of both linear and non-linear cases in his monograph (Silvey (2013)). Making a step further, Cook and Wong (1994) and Clyde and Chaloner (1996) provided theorems on the equivalence of compound designs and constraind designs. Non-trial extensions to multivariate non-linear case were also studied by a number of authors, for example, see the appendix in Zocchi and Atkinson (1999). Chapter 9 and 10 in Atkinson et al. (2007) collects many of the theorems in the linear case while Fedorov and Leonov (2013) mainly deals with non-linear models.

Let $\mathcal{X}$ be a given compact subset of Euclidean k-space, to be known as a design space. Suppose the response vector $\mathbf{y}$ is related to $\mathbf{x} \in \mathcal{M}$ via generalized linear models. Let $\widetilde{s}(x)$ be the quasi-score vector defined as the square root of the Fisher information matrix based on a single observation (See Section 6.4.2.4 for an example). In the linear regression case, $\widetilde{s}(x)$ is simply x itself.

Let $\mathcal{P}$ be the class of probability measures on the Borel sets of $\mathcal{X}$. Any $\xi \in \mathcal{P}$ will be called a design measure. Let $M(\xi)$ be the Fisher information matrix associated with the design measure ξ. That is, suppose for each $\mathbf{x} \in \mathcal{X}$, the response vector $\mathbf{y}$ follows an exponential family distribution

$$M(\xi) = \int_{\mathcal{X}} \widetilde{s}(\mathbf{x})\widetilde{s}(\mathbf{x})^T \xi(d\mathbf{x})$$

Further define $\mathcal{M} = \{M(\xi) : \xi \in \mathcal{P}\}$. Sometimes we drop the dependence on ξ and simply write

M or $M_i \in \mathcal{M}$.

Definition 5.2.1 (Gâteaux Derivative)**.** Let ϕ be a real-valued function defined on the $k \times k$ symmetric matrices and bounded above on $\mathcal{M}$. We allow ϕ to take the value $-\infty$ on $\mathcal{M}$. The Gâteaux Derivative of ϕ at M_1 in the direction of M_2 is:

$$G_\phi(M_1, M_2) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \{\phi(M_1 + \epsilon M_2) - \phi(M_1)\}. \tag{5.2.1}$$

An important property of G_ϕ is:

$$G_\phi(M_1, \sum_i a_i M_i) = \sum_i a_i G_\phi(M_1, M_i),$$

for all real a_i. For details, see page 74 in Silvey (2013) and page 241 in Rockafellar (1970). This fact is the key to our subsequent derivations.

Definition 5.2.2 (Fréchet Derivative)**.** The Fréchet derivative of ϕ at M_1 in the direction of M_2 is:

$$F_\phi(M_1, M_2) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} [\phi((1-\epsilon)M_1 + \epsilon M_2)]. \tag{5.2.2}$$

Lemma 2 (Basic Properties of Fréchet Derivative (Silvey, 2013))**.** *Some basic properties of Fréchet dericative are*

1. *Since $\mathcal{M}$ is convex, $\phi((1-\epsilon)M_1 + \epsilon M_2)$ is automatically defined.*
2. *Concavity of ϕ implies that*
$$\frac{1}{\epsilon} [\phi((1-\epsilon)M_1 + \epsilon M_2)]$$
is a non-increasing function of $\epsilon \in (0, 1]$.
3. *By letting $\epsilon = 1$ and concavity of ϕ implies that*
$$F_\phi(M_1, M_2) \geq \phi(M_2) - \phi(M_1).$$

4. *By definition,* $F_\phi(M_1, M_2) = G_\phi(M_1, M_2 - M_1)$. *Hence, if* ϕ *is differentiable and* $\sum_i a_i = 1$,

$$F_\phi(M_1, \sum_i a_i M_i) = \sum_i a_i F_\phi(M_1, M_i).$$

5. *If* $\widetilde{M}$ *is a random matrix,* ϕ *is differentiable, we have*

$$\mathbb{E}F_\phi(M_1, \widetilde{M}) = F_\phi(M_1, \mathbb{E}\widetilde{M}).$$

6. *Suppose* $\mathbf{s}$ *is a random* k*-vector with distribution* ξ *and* $M(\xi) = \mathbb{E}(\mathbf{s}\mathbf{s}^T)$. *If* ϕ *is differentiable at* $M(\xi)$,

$$\mathbb{E}F_\phi(M(\xi), \mathbf{s}\mathbf{s}^T) = F_\phi(M(\xi), \mathbb{E}(\mathbf{s}\mathbf{s}^T)) = F_\phi(M(\xi), M(\xi)) = 0.$$

These properties enable us to construct ϕ-optimal design measures in the next theorem.

Theorem 5.2.1 (The General Equivalence Theorem (Kiefer, 1974; Silvey, 2013))**.** We assume that $M(\xi)$ is of form $M(\xi) = \sum_{i=1}^{l} p_i \mathbf{s}(x)\mathbf{s}(x)^T$ and $\sum_i^l p_i = 1$ are nonnegative design weights. Note this is indeed the case in generalized linear models. If ϕ is concave on $\mathcal{M}$ and differentiable at $\mathcal{M}(\xi^*)$, then the following are equivalent:

(1) ξ^* is ϕ-optimal;

(2) The Fréchet derivative

$$F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) \leq 0 \tag{5.2.3}$$

for all $x \in \mathcal{X}$ and the maximum of F_ϕ is 0 which occurs when x is at the points of support of ξ^*;

(3) The design ξ^* satisfies

$$\max_{x\in\mathcal{X}} F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) = \min_\xi \max_{x\in\mathcal{X}} F_\phi(M(\xi), \mathbf{s}(x)\mathbf{s}(x)^T). \tag{5.2.4}$$

Proof. We prove that (2) $\Rightarrow$ (1), (1) $\Rightarrow$ (2), (1) $\Rightarrow$ (3) and (3) $\Rightarrow$ (1).

- (2) $\Rightarrow$ (1): Since any $M(\xi)$ is of form $M(\xi) = \sum_{i=1}^{l} p_i \mathbf{s}(x)\mathbf{s}(x)^T$, we have that

$$F_\phi(M(\xi^*), M(\xi)) = \sum_i p_i F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) \leq 0$$

by property 4 in 2. By property 3 in 2, we have

$$\phi(M(\xi)) - \phi(M(\xi^*)) \leq F_\phi(M(\xi^*), M(\xi)).$$

Hence, $M(\xi^*)$ is ϕ-optimal.

- (1) $\Rightarrow$ (2): By (1) we have

$$\phi((1-\epsilon)M(\xi^*) + \epsilon M(\xi)) - \phi(M(\xi^*)) \leq 0$$

for any ξ and $\epsilon \in [0, 1]$. Note that $(1-\epsilon)M(\xi^*) + \epsilon M(\xi) = M((1-\epsilon)\xi^* + \epsilon\xi)$ so that $(1-\epsilon)M(\xi^*) + \epsilon M(\xi)$ is within the domain $\mathcal{M}$. Dividing both sides by ϵ and letting it goes down to 0 gives the desired result. To show the maximum is attained at the support points of ξ^*, by property 6 in 2, we have $\mathbb{E}\left(F_\phi(M(\xi^*), \mathbf{s}(x^*)\mathbf{s}(x^*)^T)\right) = 0$ for any design ξ^* and x^* with distribution induced by ξ^*. If ξ^* is discrete with finite support $x_1, \cdots, x_n$, then we must have $F_\phi(M(\xi^*), \mathbf{s}(x_i)\mathbf{s}(x_i)^T) = 0, i = 1, \cdots, n$.

- (1) $\Rightarrow$ (3): Let x^* be a random vector with distribution induced by ξ^*. By property 6 in 2, we have $\mathbb{E}\left(F_\phi(M(\xi^*), \mathbf{s}(x^*)\mathbf{s}(x^*)^T)\right) = 0$. Hence,

$$\max_{x \in \mathcal{X}} F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) \geq 0.$$

By (2), we have $F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) \leq 0$ and it follows that

$$\min_\xi \max_{x \in \mathcal{X}} F_\phi(M(\xi), \mathbf{s}(x)\mathbf{s}(x)^T) = \max_{x \in \mathcal{X}} F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) = 0.$$

The minimum is attained when $\xi = \xi^*$ if such a ξ^* exists.

- (3) $\Rightarrow$ (1): If ξ^+ satisfies the condition. Suppose a ϕ-optimal design exists and denote it as

ξ^*, then we know

$$\max_{x\in\mathcal{X}} F_\phi(M(\xi^*), \mathbf{s}(x)\mathbf{s}(x)^T) = 0 = \min_{\xi}\max_{x\in\mathcal{X}} F_\phi(M(\xi), \mathbf{s}(x)\mathbf{s}(x)^T).$$

But this suggests that

$$\max_{x\in\mathcal{X}} F_\phi(M(\xi^+), \mathbf{s}(x)\mathbf{s}(x)^T) = 0,$$

which implies ξ^+ is also ϕ-optimal.

□

There are several classes of optimal designs lay outside the general Fisher information-based framework. One is known as optimal discriminating designs. $T-$ and $KL-$optimal designs fall into this category. Another type of optimal designs is known as asymptotic optimal designs. It is, indeed, based on functional of matrices. However, the matrices here are no longer Fisher information but asymptotic variance matrices of the corresponding estimators. These estimators can be constructed from likelihood functions, estimating equations or M-/Z-estimators in general (Dette and Trampisch, 2012).

5.3 Binary Regression Models

5.3.1 Two-parameter Binary Regression

Binary endpoints are common in clinical trials, to name a few: beetle mortality and embryogenic anthers in toxicology studies; tumor progression status in cancer studies; low-density lipo-protein (LDL) cholesterol levels (desirable vs undesirable). If we can determine the dose level most efficiently under some criteria, then we would be able to reduce the cost of experiments significantly. Some preliminary work on D-optimal design for binary regression is given in Haines et al. (2007); Atkinson et al. (2007); Kabera and Haines (2012). However, there lack a detailed and unified framework for binary regression with different types of link functions under D-optimality. This project is trying to fill in the gap. Once the gap is being filled, practitioners can choose their own favorite link functions in practice to construct an D-optimal design quickly. Further, not many user-friendly packages have been developed for practitioners, so we develop a Python Package for

clinicians and biostatisticians to use.

In the following, we first give a review on binary regression and then derive the key equation for D-optimal design. Finally, we illustrate the use of the equation using different examples.

Assume that for $i = 1, \cdots, n$, the response y_i is a binary outcome with covariate $x_i \in \mathbb{R}^d$. The y_i's are independently distributed and the density is

$$p(y|x,\pi) = p(y|\eta(x,\beta)) = \exp\left(y \ln \frac{\pi}{1-\pi} + \ln(1-\pi)\right)$$

where β is a p-dimensional parameter of interest and $\pi = F(\eta) = \int_{-\infty}^{\eta} f(s)ds$. Let $\mathcal{X}$ be the design space and the design ξ be

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_{n-1} & x_n \\ p_1 & p_2 & \cdots & p_{n-1} & p_n \end{pmatrix}$$

where for all i, $x_i \in \mathcal{X}$, $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. Let $g(x,\beta) = \mathbb{E}\left(-\frac{\partial^2 \log p(y|x,\pi)}{\partial\beta\partial\beta^T}\right)$ be the Fisher information associated with a single point x, then

$$g(x,\beta) = \omega x x^T$$

where $\omega = F'(\eta)^2/(\pi(1-\pi)) = f(\eta)^2/(\pi(1-\pi))$. The information matrix associated with the design ξ is

$$\begin{aligned} \mathfrak{M}(\xi) &= \int_{\mathcal{X}} g(x,\beta)\xi(dx) \\ &= \sum_{i=1}^n p_i \omega_i x_i x_i^T, \end{aligned}$$

where

$$\omega_i = \frac{f(\eta)^2}{(\pi(1-\pi))},\ \sum_{i=1}^n p_i = 1,\ p_i \geq 0.$$

Here are some examples of commonly used models in practice.

- (Logit) The most famous model is the logistic regression with logit link.

$$\begin{aligned} f(\eta_i) &= \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}, \\ \eta_i &= \ln \frac{\pi_i}{1-\pi_i}, \\ \pi_i &= \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \\ \mathfrak{M}(\xi) &= \sum_{i=1}^{n} \frac{p_i \exp(\eta_i)}{(1+\exp(\eta_i))^2} x_i x_i^T. \end{aligned}$$

- (Probit) Prior to the presense of logit link, one uses the probit link.

$$\begin{aligned} f(\eta_i) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\eta_i^2\right), \\ \eta_i &= \Phi^{-1}(\pi_i), \\ \pi_i &= \Phi(\eta_i), \\ \mathfrak{M}(\xi) &= \sum_{i=1}^{n} \frac{p_i \exp(-\eta_i^2)}{2\pi\Phi(-\eta_i)\Phi(\eta_i)} x_i x_i^T. \end{aligned}$$

  where $\Phi(\cdot)$ is the cumulative function of standard normal.

- (Laplace) If we want the rate of decay is faster than student t but slower than probit, then Laplace density is an alternative:

$$\begin{aligned} f(\eta_i) &= \frac{1}{2} \exp\left(-|\eta_i|\right), \\ \eta_i &= F^{-1}(\pi_i), \\ F(\eta_i) &= \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(\eta_i)\left(1-\exp(-|\eta_i|)\right), \\ \mathfrak{M}(\xi) &= \sum_{i=1}^{n} \frac{p_i \exp(-2|\eta_i|)}{4F(\eta_i)S(\eta_i)} x_i x_i^T \end{aligned}$$

  where $\mathrm{sgn}(\cdot)$ is the sign function and $S(\cdot) = 1 - F(\cdot)$.

5.3.2 Some Optimal Design Results on Binary Regression Models

A D-optimal design seeks to find a design ξ^* such that $\det \mathfrak{M}(\xi)$ is maximized. It is well known that if we know the D-optimal design for a k-parameter model is supported at k-points, then all points are equally weighted (Pázman, 1986; Wong, 2021). In the following, we always assume $k = 2$, then

$$\begin{aligned}
\det \mathfrak{M}(\xi) &= \frac{1}{4} \det \begin{pmatrix} \omega_1 + \omega_2 & \omega_1 x_1 + \omega_2 x_2 \\ \omega_2 x_2 + \omega_1 x_1 & \omega_1 x_1^2 + \omega_2 x_2^2 \end{pmatrix} \\
&\propto (\omega_1 + \omega_2)\left(\omega_1 x_1^2 + \omega_2 x_2^2\right) - (\omega_1 x_1 + \omega_2 x_2)^2 \\
&= \left[\omega_1^2 x_1^2 + \omega_1\omega_2(x_1^2 + x_2^2) + \omega_2^2 x_2^2\right] - \\
&\quad \left[\omega_1^2 x_1^2 + 2\omega_1\omega_2 x_1 x_2 + \omega_2^2 x_2^2\right] \\
&= \omega_1\omega_2\left(x_1^2 - 2x_1x_2 + x_2^2\right) \\
&\propto \frac{f(\eta_1)^2 f(\eta_2)^2}{F(\eta_1)S(\eta_1)F(\eta_2)S(\eta_2)}(x_1 - x_2)^2
\end{aligned} \tag{5.3.1}$$

where for $i = 1, 2$, $\eta_i = \beta_0 + \beta_1 x_i, F(\eta_i) = \int_{-\infty}^{\eta_i} f(s)ds, S(\eta_i) = 1 - F(\eta_i)$. Now it is natural to consider, if we are given a maximizer (η_1^*, η_2^*) of Formula 5.3.2, is it unique? The short answer is no unless η_1^* and η_2^* is symmetric around $2a$ and we provide a lemma below.

Lemma 3. *If $f(s)$ is symmetric, i.e., $f(a + s) = f(a - s)$ for some a, then for any design*

$$\xi^* = \begin{pmatrix} \frac{\eta_1^* - \beta_0}{\beta_1} & \frac{\eta_2^* - \beta_0}{\beta_1} \\ 0.5 & 0.5 \end{pmatrix}$$

we have

$$\det \mathfrak{M}(\xi^*) = \det \mathfrak{M}(\xi')$$

where

$$\xi' = \begin{pmatrix} \frac{2a - \eta_2^* - \beta_0}{\beta_1} & \frac{2a - \eta_1^* - \beta_0}{\beta_1} \\ 0.5 & 0.5 \end{pmatrix}$$

Proof. For $i = 1, 2$, let $x_i^* = \frac{\eta_1^* - \beta_0}{\beta_1}$ and $x_i' = \frac{2a - \eta_2^* - \beta_0}{\beta_1}$, then

$$(x_1^* - x_2^*)^2 = \left(\frac{\eta_1^* - \eta_2^*}{\beta_1}\right)^2 = (x_1' - x_2')^2$$

Next, let $\eta_1' = 2a - \eta_2^*$ and $\eta_2' = 2a - \eta_1^*$, we have

$$f(\eta_1^*) = f(a + (\eta_1^* - a)) = f(a - (\eta_1^* - a)) = f(2a - \eta_1^*) = f(\eta_2')$$

and similarly, $f(\eta_2^*) = f(\eta_1')$.

Finally, we have

$$\begin{aligned} F(\eta_1^*) &= F(a + (\eta_1^* - a)) \\ &= S(\eta_2') \end{aligned}$$

and $S(\eta_1^*) = F(\eta_2')$, $F(\eta_2^*) = S(\eta_1')$, $S(\eta_2^*) = F(\eta_1')$. □

WLOG, we may assume that $f(s) = f(-s)$. As an example, one such f is logistic density $f(s) = e^{-s}/(1 + e^{-s})^2$. Then we have

$$\det \mathfrak{M}(\xi) = \frac{f(\eta_1)^2 f(\eta_2)^2}{F(\eta_1)S(\eta_1)F(\eta_2)S(\eta_2)} \left(\frac{\eta_1 - \eta_2}{\beta_1}\right)^2$$

and taking the logarithm of $\det \mathfrak{M}(\xi)$ and setting the derivative w.r.t. η_1 equal to zero gives

$$\frac{2f'(\eta_1)}{f(\eta_1)} - \frac{f(\eta_1)}{F(\eta_1)} + \frac{f(\eta_1)}{S(\eta_1)} + \frac{2}{\eta_1 - \eta_2} = 0 \tag{5.3.2}$$

We denote the above **the key equation** and call it the **WC** equation where WC stands for Wong and Cui. For a symmetric two-point design, we let $\eta_1 = -\eta_2$, then

$$\det \mathfrak{M}(\xi) = \frac{f(\eta_1)^4}{F(\eta_1)^2 S(\eta_1)^2} \left(\frac{2\eta_1}{\beta_1}\right)^2$$

and taking the logarithm of $\det \mathfrak{M}(\xi)$ and setting the derivative w.r.t. η_1 equal to zero gives

$$\frac{2f'(\eta_1)}{f(\eta_1)} - \frac{f(\eta_1)}{F(\eta_1)} + \frac{f(\eta_1)}{S(\eta_1)} + \frac{1}{\eta_1} = 0 \tag{5.3.3}$$

The resulting solutions provide the design points of the D-optimal and G-optimal designs among all 2-point designs. To verify that it is optimal among all possible designs, we need to calculate the sensitivity function based on the equivalence theorem (Atkinson et al., 2007; Wong, 2021).

Theorem 5.3.1 (Equivalence theorem (Atkinson et al., 2007; Wong, 2021))**.** Let $\mathfrak{M}(\xi)$ be the information matrix associated with design ξ, then the following are equivalent ($\dim(\mathcal{X}) = d$),

1. The design ξ^* is D-optimal, i.e., $\xi^* = \arg\min_\xi \det \mathfrak{M}(\xi)$.
2. The inequality $\psi(\mathbf{x}, \beta) = w(\mathbf{x})(\mathbf{x}^T \mathfrak{M}(\xi^*)^{-1}\mathbf{x}) - d \leq 0$ holds for all $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ where $w(\mathbf{x})$ is a weight depending on the link $\eta(\mathbf{x})$ and ψ is called the sensitivity function.

In the following, we apply the key equation to 3 examples and verify the results using PSO. In short, we write η for $\beta_0 + \beta_1 x$. We omit the derivation of the explicit form of information matrices since it is straightforward.

Example 5.3.1 (Logit)**.** For this problem, $f(\eta) = \exp(\eta)/(1 + \exp(\eta))^2$ and $w = 1/((1+\exp\eta)(1-\exp\eta))$. Plug-in all necessary elements, the key equation is

$$2 - \frac{4\exp\eta}{1+\exp\eta} + \frac{2}{\eta} = 0$$

Solving it numerically, we obtain $\eta_1 = +1.5434$ and $\eta_2 = -1.5434$.

The Figure 5.1 demonstrates the sensitivity functions of two locally D-optimal designs with logit link and specified parameter values.

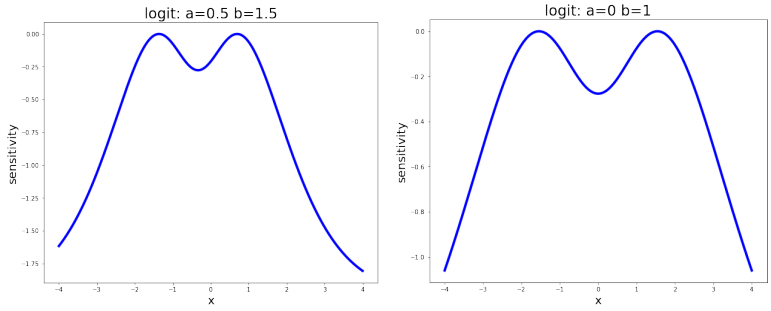


Figure 5.1: Sensitivity function for logit link.

Example 5.3.2 (Probit)**.** For this problem, $f(\eta) = \exp\left(-\frac{1}{2}\eta^2\right)/\sqrt{2\pi}$ and $\Phi(\eta) = \int_\infty^\eta f(s)ds$ and $w = \exp(-\eta^2)/(2\pi\Phi(\eta)(1-\Phi(\eta)))$. Plug-in all necessary elements, the key equation is

$$\frac{2}{\eta} - 4\eta - 2f(\eta)\left(\frac{1}{\Phi(\eta)} - \frac{1}{1-\Phi(\eta)}\right) = 0$$

Solving it numerically, we obtain $\eta_1 = +1.1381$ and $\eta_2 = -1.1381$.

The two panels of Figure 5.2 demonstrates the sensitivity functions of two locally D-optimal designs with probit link and specified parameter values.

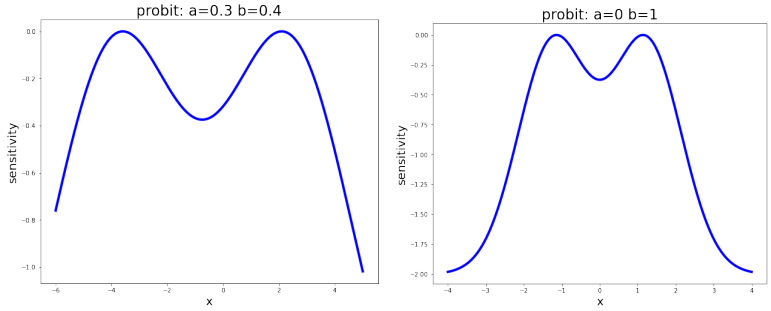


Figure 5.2: Sensitivity function for probit link.

Example 5.3.3 (Laplace)**.** For this problem, we have $f(\eta) = \exp(-|\eta|)/2$ and $w = 1/(2\exp(|\eta|) - 1)$. Plug-in all necessary elements, the key equation is

$$-4\ \mathrm{sgn}(\eta_1) - \frac{2\exp(-|\eta_1|)}{1 + \mathrm{sgn}(\eta_1)(1 - \exp(-|\eta_1|))} + \frac{2\exp(-|\eta_1|)}{1 - \mathrm{sgn}(\eta_1)(1 - \exp(-|\eta_1|))} + \frac{2}{\eta_1} = 0$$

Solving it numerically, we obtain $\eta_1 = +0.7680$ and $\eta_2 = -0.7680$.

However, the left panel of Figure 5.3 has shown that $(+0.7680, -0.7680)$ is NOT a locally D-optimal design. According to Federov's algorithm (Atkinson et al., 2007), it suggests that we need to add a design point at 0. This is empirically verified by Particle Swarm Optimization (PSO) in section 1.3 using the Python package "pyswarms" (Miranda, 2018). The right panel of Figure 5.3 has shown the sensitivity function of the three point design generated by PSO.

5.3.3 Applications

In this section, we apply the developed theory to a dataset which comes from toxicology studies using sea urchins (the data is provided in Collins et al. (2022)). There are two endpoints (failure types): EDA/D and Radial:Ab and we use the second endpoint for illustration. The concentration level for the second endpoint is within 0 to 450 μM, and we re-scale it to $[0, 0.45]$ by dividing 1000. We run two binary regression models using logit and complementary log-log (Cox regression) link

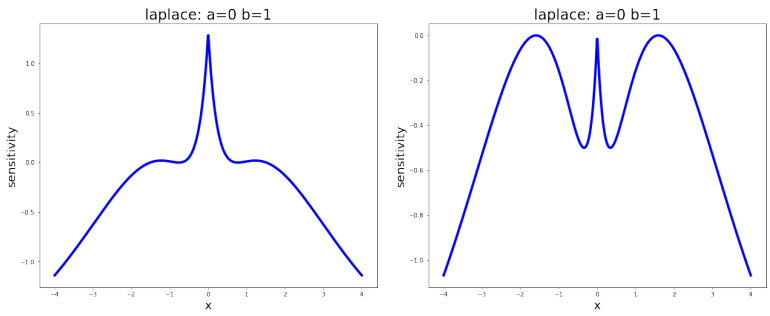


Figure 5.3: Sensitivity function for Laplace link.

functions respectively. The results are generated by 'gtsummary' package in R Sjoberg et al. (2021) and given in Table 5.2.

Table 5.2: Binary regression with two different link functions using sea urchin data

Logit link			
Characteristic	Estimation	95% CI	p-value
β_0	-4.5	(-4.7,-4.4)	< 0.001
β_1	20	(19,21)	< 0.001
Cox regression			
Characteristic	Estimation	95% CI	p-value
β_0	-3.7	(-3.8, -3.6)	< 0.001
β_1	14	(13, 14)	< 0.001

The fitted dose-response curve (in this case, concentration-response curve) is given in Figure 5.4: the orange and dodgerblue curves correspond to Cox regression and logit link respectively. The dots represent the true observations in Collins et al. (2022) with concentration level greater than 450 removed. Hence, by the WC equation 5.3.2 for two-parameter binary regression, the D-optimal designs are

$$\xi_{\text{logit}} = \begin{pmatrix} 0.1478 & 0.3022 \\ 0.5 & 0.5 \end{pmatrix}$$

$$\xi_{\text{Cox}} = \begin{pmatrix} 0.1687 & 0.3343 \\ 0.5 & 0.5 \end{pmatrix}$$

and the sensitivity functions are given in Figure 5.5 (left panel: logit link; right panel: Cox regression). Multiplying by 1000, the resulting D-optimal designs at the original scale are

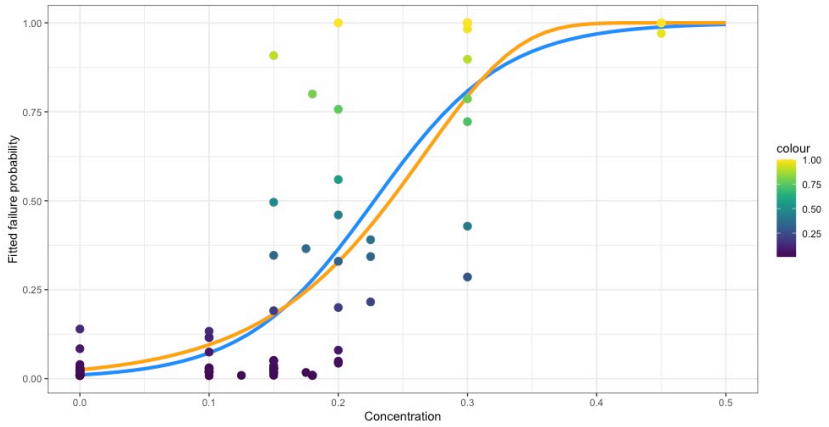


Figure 5.4: The fitted concentration-response curves.

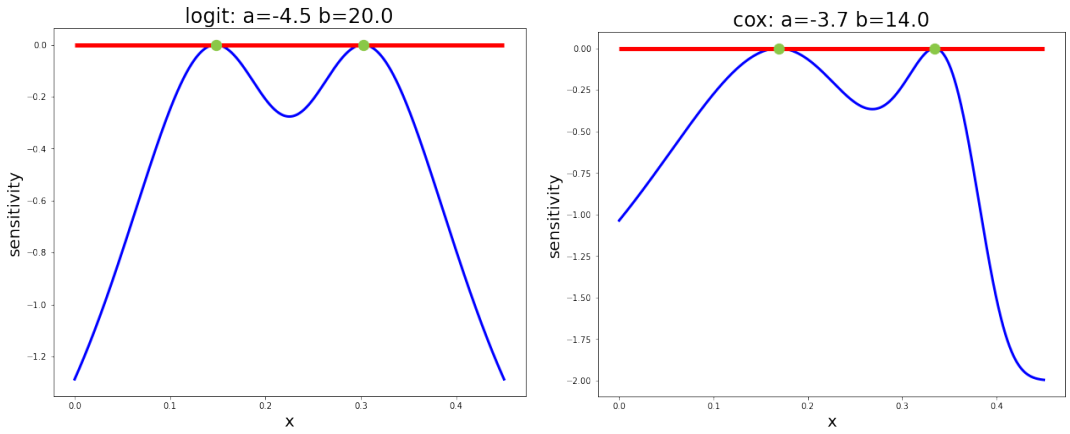


Figure 5.5: Sensitivity functions for the sea urchin data.

$$\xi^*_{\text{logit}} = \begin{pmatrix} 147.8 & 302.2 \\ 0.5 & 0.5 \end{pmatrix} \tag{5.3.4}$$

$$\xi^*_{\text{Cox}} = \begin{pmatrix} 168.7 & 334.3 \\ 0.5 & 0.5 \end{pmatrix} \tag{5.3.5}$$

Comparing them with the original design given in Collins et al. (2022):

$$\xi^*_{\text{original}} = \begin{pmatrix} 0 & 100 & 125 & 150 & 175 & 180 & 200 & 225 & 300 & 450 \\ 0.254 & 0.148 & 0.0129 & 0.169 & 0.0263 & 0.0338 & 0.128 & 0.0370 & 0.155 & 0.0360 \end{pmatrix} \tag{5.3.6}$$

we find that the D-optimal design reduces the number of required concentration levels significantly.

5.4 Applications of Beta Regression Models to Toxicity Studies

In toxicity studies, one is interested in designing a dose–response experiment, and more specifically, a concentration–response experiment (FDA, 2003). As we have indicated in the motivation, existing methods are always based on uniform design or some nebulous criteria, leading to inefficient designs in practice. Hence, the goal of this project is to develop a more quantitative approach to designing a dose–response experiment.

We have proposed a model-based approach to determine the dose–response relationship using sea urchins. It can provide the most accurate statistical inference for the underlying parameters of interest, which may be estimating one or more model parameters or pre-specified functions of the model parameters, such as lethal dose, at maximal efficiency (Collins et al., 2022).

Our model assumes that the response follows a Beta distribution whose density is $f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$. By assuming $\alpha = \exp(\alpha_1 + \alpha_2 x)$ and $\beta = \exp(\beta_1 + \beta_2 x)$ where x is the actual dose range (Wu et al., 2005), the mean response rate of the Beta distribution is

$$\frac{1}{1+\exp\{(\beta_1-\alpha_1)+(\beta_2-\alpha_2)x\}} \tag{5.4.1}$$

where $\Theta^T = (\alpha_1, \alpha_2, \beta_1, \beta_2)$ are unknown model parameters that controls the shape of the response curve. We apply formula 5.4.1 to construct a dual-objective optimal design and estimate a specific LD_p (the dose concentration expected to result in p% of the urchins succumbing) and Θ as accurately as possible. To be more specific, we seek a design ξ to maximize the dual efficiency

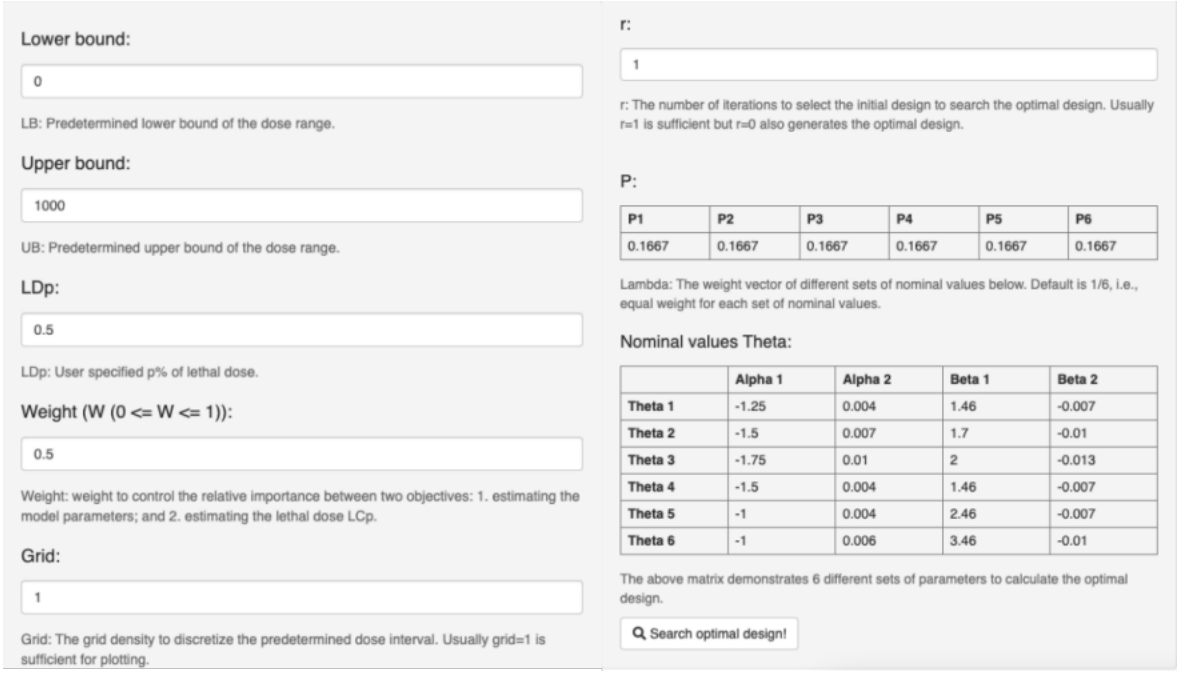
$$e_{\text{dual}} = W \times \log e_D + (1-W) \times \log e_c \tag{5.4.2}$$

where W is a value within $[0,1]$ and e_{dual}, e_D and e_c are *dual-* D- and c- efficiencies respectively (Atkinson et al., 2007). Details for rescaling the two criteria into efficiencies and how the weight W is properly chosen are discussed in Cook and Wong (1994) and theorem 5.4.1. In general, smaller values of W imply less emphasis on the D-optimality criterion. Additional examples with R codes for related models are given in Hyun et al. (2018).

Since the locally optimal design depends on unknown parameters Θ, we provide one set of nominal values of Θ based on maximum likelihood estimation (MLE) from concentration–response data for the sea urchin study with various trimethoprim concentrations (Conc) (Collins et al., 2022). To investigate if the optimal designs were robust to misspecified nominal parameter values, five additional sets of parameter values were selected that provide reasonable approximations to response curves (Collins et al., 2022). Both e_D and e_c are given in our publication.

We also implemented an app based on R Shiny (Wickham, 2021) and practitioners can find it available online at `https://elviscuihan.shinyapps.io/Dc_optimal_design/`. As Figure 5.6 has shown, the app allows users to choose their own hyper-parameters as well as nominal values and find the locally optimal design by clicking the "Search optimal design" button.

Figure 5.6: Illustration of the R Shiny App



In practice, we use the following procedure to proceed.

1. Estimate parameters of the Beta model using pilot data.

2. Based on the estimated parameters, we further specify more groups of parameters, each with a weight.

3. Use estimated and specified parameters to compute the Fisher information matrix of a design ξ.

4. Compute the asymptotic variance of the quantities of interest (LD_p, MTD, etc.) by delta-method.

5. Construct an expression of the dual optimal criterion for the design ξ with a specified weight W.

6. Derive the optimal design ξ^* using either gradient-based or meta-heuristic algorithms.

7. Use the equivalence theorem to check the resulting design ξ^* is globally optimal.

5.4.1 The Cook-Wong Theorem

Finally, we apply Cook-Wong theorem to choose W so that one of the two efficiencies can achieve a certain pre-specified level (say, D-efficiency at 90%), and we provide a modified version (convexity is not required) and a modified proof is given below.

Theorem 5.4.1 (Cook-Wong, 1994)**.** Suppose ϕ_1 and ϕ_2 are two non-positive finite continuous functionals and $\lambda \in (0, 1)$ and let ξ_λ be the solution to the compound optimization problem

$$\begin{aligned} \xi_\lambda &= \arg\max_{\xi} \phi(\xi|\lambda) \\ &= \arg\max_{\xi} \left(\lambda\phi_1(\xi) + (1-\lambda)\phi_2(\xi)\right) \end{aligned} \tag{5.4.3}$$

and let ξ_c be the solution to the constrained optimization problem

$$\begin{aligned} &\max\ \phi_2(\xi) \\ &s.t.\ \phi_1(\xi) \geq c \end{aligned} \tag{5.4.4}$$

where WLOG $c \in [-\infty, 0]$. Then 5.4.3 implies 5.4.4 for some suitable c and 5.4.4 implies 5.4.3 for some suitable λ if $\phi_1(\xi_\lambda)$ is left-continuous at $\lambda = 1$ and right-continuous at $\lambda = 0$.

Proof. ($\Rightarrow$): Define $c_\lambda = \phi_1(\xi_\lambda)$. Suppose there $\exists$ a ξ^* such that $\phi_2(\xi_\lambda) \leq \phi_2(\xi^*)$ and $\phi_1(\xi^*) \geq c_\lambda$,

then

$$\begin{aligned}\phi(\xi_\lambda|\lambda) &= \lambda\phi_1(\xi_\lambda) + (1-\lambda)\phi_2(\xi_\lambda) \\ &= \lambda \times c_\lambda + (1-\lambda)\phi_2(\xi_\lambda) \\ &\leq \lambda\phi_1(\xi^*) + (1-\lambda)\phi_2(\xi^*) \\ &= \phi(\xi^*|\lambda) \\ &\leq \phi(\xi_\lambda|\lambda)\end{aligned}$$

where the last inequality is due to the compound optimality of ξ_λ. Hence, we conclude ξ_λ solves the constrained optimization problem for $c = c_\lambda$.

($\Leftarrow$): Let $c^* = \phi_1(\xi_c)$, then the result follows from definition if either $c^* = \phi_1(\xi_{\lambda=1})$ or $c^* = \phi_1(\xi_{\lambda=0})$. Hence WLOG we assume

$$\phi_1(\xi_{\lambda=0}) < c^* < \phi_1(\xi_{\lambda=1})$$

By the left- and right-continuity of $\phi_1(\xi_\lambda)$ and lemma 3.1 in Pshenichnyi (2020) (page 71), $\exists$ a λ^* such that $c^* = \phi_1(\xi_{\lambda^*})$. The compound optimality of ξ^* implies $\phi_2(\xi_{\lambda^*}) \geq \phi_2(\xi_{\lambda_c})$ while the constrained optimality implies $\phi_2(\xi_{\lambda^*}) \leq \phi_2(\xi_{\lambda_c})$. In conclusion, we have

$$\phi(\xi_{\lambda^*}|\lambda^*) = \phi(\xi_c|c),$$

which means ξ_c is also a solution to the compound optimization problem. □

We have reviewed how to construct an optimal design based on different criteria. The optimization is complex in general, numerical methods need to be applied to find the optimal design. In the next chapter, we demonstrate how metaheuristics can be used in a novel approach to design toxicology experiments sequentially.

CHAPTER 6

Failure of Optimal Design Theory? A Case Study in Toxicology Using Sequential Robust Optimal Design Framework

6.1 Motivation and Introduction

6.1.1 Importance of the Chapter

As statisticians, we should always ask "what others (say, toxicologists) need", instead of saying "what we have in our beautiful theory". **This work represents a groundbreaking collaboration between statisticians and toxicologists, addressing the urgent need to reduce costs in toxicology experiments using sea urchins.** Unlike previous studies in sequential optimal design, which are largely theoretical or simulation-based, our research directly engages with practical challenges faced by toxicologists. **Surprisingly, our findings reveal that parameter estimates derived from optimal designs often fail to outperform those obtained from conventional uniform designs in practice.**

This apparent discrepancy arises not from the failure of optimal design theory but from a multitude of real-world factors. These include (1) practical constraints, such as adherence to Food and Drug Administration (FDA) guidelines and budgetary limitations (U.S. Department of Health and Human Services, 2015; Buckley et al., 2020); (2) the influence of genetic shifts in experimental organisms, which complicate the transferability of theoretical designs across datasets; and (3) the inherent challenges of implementing complex designs in dynamic experimental settings. These factors highlight the critical gap between theoretical advancements and practical applications in toxicological studies.

To bridge this gap, we propose several remedies:

1. **For statisticians**: Design frameworks must incorporate robust features like control groups and endpoints (e.g., lethal doses) to align with biological and practical considerations.

2. **For toxicologists**: Greater control over genetic shifts in experimental organisms can improve the consistency and applicability of optimal designs.

3. **For researchers and practitioners**: Developing accessible, user-friendly tools is essential to implement theoretical designs effectively in real-world toxicology.

While the future of optimal design in practice remains uncertain, our work highlights the need for continuous adaptation and collaboration. By addressing these challenges, we aim to strengthen the connection between theoretical statistics and experimental toxicology, paving the way for more impactful research and applications.

6.1.2 Introduction

In this era of rapid advancement in statistical methods and computational technology, the time has come to revisit our approach to toxicology experiments. The once-dominant uniform design, which takes an equal number of observations at each of the equally spaced doses was conceived in an age before the dawn of computational simulations, has limitations, and it is susceptible to empiricism and unable to incorporate improved statistical methodology from the field of optimal designs (Tse-Tung, 2014). With at our disposal now, the future calls for multiple parallel iterative approaches, where experimentation moves in concert with technology. This shift is not a choice, but an inevitability. The march of progress is spiral—driven by constant iterations, failures, and self-renewal (Marx, 2000).

The field of experimental toxicology is a fascinating and dynamic arena within the scientific landscape. Unlike many traditional disciplines where experiments are conducted under controlled, predictable conditions, toxicology embraces the inherent unpredictability of biological responses. In fields such as chemistry or physics, experiments often follow meticulously designed trajectories, with every variable tightly controlled to minimize surprises. Toxicologists, on the other hand, explore the intricate interactions between substances and biological systems, frequently uncovering unexpected

adverse effects. This unpredictability is precisely what makes toxicology both challenging and captivating—every experiment has the potential to reveal something new and unforeseen about the way our bodies interact with the world around us (Schoen, 1996; Costa et al., 2010).

In contemporary toxicology laboratories, researchers frequently adopt a stepwise experimental approach, where the outcomes of each testing phase guide subsequent experiments (Collins et al., 2022). This method, while rooted in decades of practical wisdom, often lacks the efficiency and precision that a more structured experimental design could provide. Enter the world of optimal experimental design—a transformative approach that could revolutionize toxicology. Imagine a process where every experiment is not just an isolated trial, but a part of a strategic, interconnected sequence that builds upon the last. With optimal design, toxicologists can use the data they gather to refine their approach in real-time, saving time, reducing costs, and accelerating the discovery process with precision and accuracy (de la Calle-Arroyo et al., 2023).

Historically, toxicologists have relied on straightforward experimental designs, such as evenly spaced dose levels or uniform designs (Casarett et al., 2008; Fedorov and Leonov, 2013). While simple to implement, these approaches often lead to inefficiencies—extended timelines, unnecessary costs, and suboptimal use of data. Moreover, the complexity of toxicological systems, including dose-response nonlinearity, low-dose effects, and phenomena like hormesis (Calabrese and Baldwin, 2003), necessitates experimental strategies that adapt to emerging insights. This demands a move away from static, empirical methods to a more dynamic, data-driven approach that integrates modern statistical tools (Dragalin et al., 2008a; Gertsch and Wong, 2024).

The secret to designing these efficient experiments lies in their flexibility and precision (Holland-Letz and Kopp-Schneider, 2015). By clearly defining the goals of a toxicology study—whether it's identifying a toxic threshold, studying low-dose effects, or detecting complex phenomena like hormesis—researchers can leverage statistical models to guide their experimental designs. This approach maximizes the accuracy of estimates while streamlining the experimental process. Even though optimal design theory wasn't initially created with sequential experiments in mind, it offers a rich toolbox that can be adapted to meet the evolving needs of toxicology (Cui et al., 2024b).

In this project, we present an innovative framework that marries the precision of optimal design theory with the adaptive, step-by-step nature of toxicological testing (Koutra et al., 2021). Our method uses data from each experimental phase to fine-tune subsequent steps, ensuring that

each iteration propels researchers closer to their objectives. By applying optimal design criteria, toxicologists can select dose levels that maximize statistical precision, all while reducing both costs and labor. Here's a glimpse into our step-by-step approach:

- **Initial Experiment Design**: Based on their expertise, the toxicologist selects an initial set of doses and conducts the first round of experiments.
- **Modeling the Data**: After analyzing the initial data, the toxicologist identifies the best-fitting statistical model, which will guide the design of the next batch of experiments.
- **Optimal Dose Selection**: Using the fitted model, optimal design criteria can help choose the next set of doses. For example, a D-optimal design focuses on estimating model parameters with maximum accuracy, while a G-optimal design minimizes the variance across the entire dose-response surface. Alternatively, an integrated criterion can emphasize different parts of the response curve, depending on the specific goals of the study. These designs often require fewer dose levels than traditional approaches, making the next round of experiments more efficient.
- **Evaluating and Adjusting**: After the second batch of experiments, the data are analyzed again. The model is either confirmed through graphical or statistical checks (e.g., a lack-of-fit test) or adjusted to reflect new insights. If the model remains consistent with earlier findings, the efficiency of the current design is compared to that of the optimal design.
- **Refining the Design**: If the current design is found to be less efficient, additional dose levels may be introduced to enhance precision. The process is repeated as necessary.

The following sections demonstrate how our method can be applied to real-world toxicology experiments, showcasing significant improvements in efficiency and cost-effectiveness. By embracing this streamlined, data-driven approach, toxicologists can transform the way they approach experimental design, moving from initial hypotheses to conclusive findings with greater speed and precision. The rest of the chapter is organized as follows: Section 6.2 provides a comprehensive review of existing sequential optimal design methodologies and highlights their limitations in toxicological applications. Section 6.3 introduces our proposed quasi-sequential robust optimal design framework, detailing its theoretical foundation and implementation. Section 6.4 applies this

framework to real-world toxicology experiments, showcasing its effectiveness in improving dose-response modeling and experimental efficiency. Section 6.5 provides simulation studies based on bivariate probit model. Finally, Section 5 discusses the implications of our findings, potential limitations, and directions for future research.

6.2 Literature Review: Sequential Optimal Design and Applications

Sequential optimal design has been a cornerstone in experimental research, allowing for iterative refinement of experimental parameters based on accumulated data. This approach has found applications across various fields, including toxicology, clinical trials, and epidemiology.

Theoretical Foundations: The concept of sequential design traces its roots to early works in optimal experimental design, such as those by Silvey (2013) and Fedorov (1971). Sequential designs aim to enhance efficiency by adapting experimental conditions dynamically. Park and Faraway (1998) introduced a nonparametric sequential approach for estimating response curves, emphasizing the advantage of iterative adaptation in achieving greater precision with fewer samples. Similarly, Hughes-Oliver and Rosenberger (2000) utilized compound D-optimality in adaptive group testing to estimate rare trait prevalences efficiently, underscoring the method's applicability to nonlinear and heteroscedastic models. Lane (2020) developed a theoretical framework that utilizes observed Fisher information as a priori.

Applications in Dose-Response Studies: One prominent application of sequential optimal design is in dose-response studies. Dragalin et al. (2008b) proposed a two-stage design for dose-finding in clinical trials, incorporating both efficacy and safety considerations through a bivariate probit model. This methodology demonstrated how sequential designs could address ethical and practical constraints in clinical research by optimizing dose allocations iteratively. Stacey (2007) extended this idea using a Bayesian framework, showcasing how adaptive designs could improve the precision of dose-response models while reducing the risk to participants.

Real-World Implementations: Real-world implementations of sequential designs have highlighted their versatility. For instance, Wright and Bailer (2006) discussed sequential methods in toxicology, where adaptive designs facilitated the efficient identification of toxic thresholds. In

biomedical research, Qiu and Wong (2023) employed metaheuristic algorithms like particle swarm optimization to develop optimal designs for continuation-ratio models, addressing complex dose-response relationships in early-phase clinical trials. Similar algorithms are booming in optimal design and other fields of statistics (Cui et al., 2024a).

Challenges and Advancements: Despite their advantages, sequential designs face challenges, including dependence on prior information and computational complexity. Recent advancements aim to mitigate these issues. de la Calle-Arroyo et al. (2023) introduced D-augmentation techniques to enhance design flexibility, enabling better model discrimination and adequacy checks. Park and Faraway (1998) highlighted the importance of robust algorithms that adapt to uncertain initial conditions, ensuring reliability in applications with limited prior data.

6.3 Methodology

In this section, we detail the proposed sequential robust optimal design framework, which builds on existing methodologies such as Wang et al. (2013). We also outline the mathematical framework, the sequential design process across two stages, and discuss an augmented design approach to account for practical requirements. Finally, we demonstrate the applicability of this framework to proportional odds models commonly used in toxicological studies.

6.3.1 Notations

The sequential robust optimal design aims to iteratively refine experimental designs across multiple stages to achieve greater efficiency and precision. The core idea is to leverage information from earlier stages of the experiment to optimize the design of subsequent stages, ensuring robustness to model uncertainties.

Let $[x_L, x_U]$ denote the design space, where x_L and x_U represent the lower and upper bounds of the experimental range. An initial proportion of samples, α, is allocated to Stage I, while the remaining $(1-\alpha)$ samples are reserved for Stage II. The optimality criterion, such as D-optimality, c-optimality, or dual-optimality, guides the design process by optimizing the Fisher information matrix across stages.

6.3.2 Proposed Sequential Robust Optimal Design Scheme

In the following, we propose a sequential robust optimal design scheme which is a modification of Wang et al. (2013). We provide a comparison between two methods in Table 6.1.

- **Inputs**:
  - $[x_L, x_U]$, the design space;
  - α, the proportion of samples performed at Stage I (so $(1-\alpha)$ is the proportion of samples performed at Stage II).
  - Optimality criterion (e.g., D-optimality, c-optimality or dual-optimality).
- **Stage I**:
  - Perform the initial design which could be evenly spaced on the original or log-scale or be designed using toxicologist's expert knowledge;
  - Collect a total of N_1 samples;
  - Repeat the procedure K times.
  - Choose a regression model that fits the first set of data well.
- **Stage II**:
  - Utilize the information from Stage I to perform Stage II design for the remaining $(1-\alpha)$ experiments:
    - Based on the K sets of data and the regression model that we have chosen, we estimate K sets of nominal values.
    - Based on the optimality criteria we have chosen, we compute the robust optimal design that optimizes the criterion function where the Fisher information matrix is the combination of the first stage design and the to-be-determined second stage design. It is robust in the sense that we utilizes K different sets of nominal values.
  - Perform the statistical modeling and inference using the obtained Stage II design to get the fitted model.

- **Outputs**: The fitted models and the quantity that we are interested in (e.g., parameter estimates, ED50, oral radialization 50, etc.).

6.3.3 Augmented Optimal Design

To address practical considerations in toxicological studies, such as the inclusion of control groups, we propose an augmented design approach. The augmented design incorporates additional dose levels to account for specific experimental requirements.

6.3.3.0.1 Incorporating a Control Group The design can include a zero-dose level to estimate baseline responses, as follows:

$$\xi = \begin{pmatrix} 0 & x_1 & \cdots & x_n \\ p & (1-p)p_1 & \cdots & (1-p)p_n \end{pmatrix},$$

where p is the weight assigned to the control group ($x = 0$) and p_i are the weights for other dose levels. Following Gollapudi et al. (2013), a common choice is $p = 1/(n+1)$.

6.3.3.0.2 Incorporating High Dose Levels To capture endpoints such as lethal effects, the design can include an additional high-dose level x^*:

$$\xi = \begin{pmatrix} 0 & x_1 & \cdots & x_n & x^* \\ \alpha_1 & (1-\alpha)p_1 & \cdots & (1-\alpha)p_n & \alpha_2 \end{pmatrix},$$

where $\alpha_1 + \alpha_2 = \alpha$ and x^* ensures the inclusion of high-dose observations.

6.3.4 Application to Proportional Odds Models

The proposed framework is particularly suitable for proportional odds models, widely used in toxicological studies to analyze ordinal responses. These models relate the cumulative probability of a response to dose levels through a logistic function:

$$\log \frac{P(Y \leq k \mid x)}{P(Y > k \mid x)} = \beta_0 + \beta_1 x,$$

where Y denotes the response category, x is the dose level, and β_0, β_1 are model parameters.

6.3.4.0.1 Stage I Design for Proportional Odds Models

In Stage I, dose levels are selected to ensure sufficient coverage of the response range, allowing for accurate parameter estimation. Data from this stage provide initial estimates of β_0 and β_1.

6.3.4.0.2 Stage II Design for Proportional Odds Models

Stage II incorporates the estimated parameters into the optimal design criteria. The Fisher information matrix for the proportional odds model is used to determine dose levels that maximize parameter precision while maintaining robustness to uncertainty.

6.3.4.0.3 Outputs

The final model provides parameter estimates and key toxicological endpoints, such as the effective dose (ED50) and thresholds for higher response categories.

Table 6.1: Comparison of Wang et al. (2013) and the Proposed Scheme

Aspect	Wang et al. (2013)	Proposed Scheme
Framework	Two-stage design for dose-response modeling.	Sequential robust design adaptable to various contexts.
Stage I Design	Evenly spaced doses.	Flexible: uniform, log-scale, or expert-driven.
Stage II Design	Bootstrap-based optimization.	Robust criteria using Fisher information.
Variance Heterogeneity	Explicitly modeled.	Similar but more flexible.
Efficiency	High computational cost due to bootstrapping.	Faster with analytical optimization.
Practicality	Limited to basic designs.	Supports control groups and extreme doses (augmented design).

6.3.5 Optimizer

The Particle Swarm Optimizer (PSO) is the main optimizer used in this chapter and it is widely available online (Miranda, 2018; Riza et al., 2019).

6.4 Case Study: Toxicology Experiments

In this section, we explore the application of our proposed quasi-sequential robust optimal design framework to real-world toxicology experiments using sea urchin embryos. Sea urchin embryos are an important biological model for understanding dose-response relationships due to their sensitivity to environmental toxins and their relevance to broader ecological and human health concerns. These experiments aim to identify critical dose levels, such as the effective dose for 50% response (ED50) and lethal dose for 50% mortality (LD50), which are fundamental metrics for assessing toxicity (Calabrese and Baldwin, 2003; Gollapudi et al., 2013).

The challenges of these studies arise from high biological variability due to genetic heterogeneity and environmental influences. Conventional static experimental designs, such as uniform or evenly spaced dose levels, often fail to capture complex dose-response relationships or adapt to emerging patterns in the data. These limitations result in inefficiencies, including increased costs, wasted resources, and reduced precision in parameter estimates (Ritz et al., 2015; Hartung and Rovida, 2009).

Our proposed framework, leveraging Particle Swarm Optimization (PSO) and robust design principles, addresses these issues by iteratively refining the experimental design across two stages. This approach balances efficiency and flexibility, enabling the exploration of critical dose-response regions while adapting to uncertainties in biological systems. In this case study, we demonstrate the framework's efficacy using experimental data from sea urchin toxicology studies.

Section 6.4.1 details the experimental data collected and its structure, highlighting the variability across different timelines and genetic groups. Section 6.4.2.1 discusses the modeling approaches and diagnostic measures used to evaluate the data. Section 6.4.2.2 and 6.4.2.3 present the optimal designs generated using our framework, comparing them with conventional methods. Finally, Section 6.4.2.4 provides a theoretical condition that ensures the generated designs are indeed optimal.

6.4.1 Description of Four Datasets

6.4.1.1 The First Dataset

The first dataset was collected by Dr. Collins between June 23rd and August 19th, 2022, over nine days of experiments. The dose levels and design weights were determined solely by the toxicologist's expertise. Table 6.2 illustrates the typical data structure, capturing essential experimental variables such as dose levels, embryo response categories, and durations.

date	dose	duration	observed	normal	radial	0 spicules	dead/delayed
2022.6.23	0	1-24h	108	107	0	0	1
...	...	...	...	...	...	...	...

Table 6.2: A typical sea urchin data structure.

- Date: the date of the row being recorded.
- Dose: dose level with unit mg/cm^3, the lowest dose level is 0.
- Duration: the total experimental time of the sea urchin embryo; 1-24h means that embryo has spent 24 hours in the solution.
- Observed: total number of observed sea urchin embryos.
- Normal: number of observed normal embryos.
- Radial: Number of embryo that has radialization.
- 0 spicules and dead/delayed: Together with other unrecorded abnormal status, these categories record the number of observed abnormal sea urchin embryos.

6.4.1.2 The Second, Third and Fourth Datasets

The second, third, and fourth datasets were collected based on the framework developed in Section 6.3, utilizing optimal design theory and PSO. These datasets represent the second-stage designs in our study. Specifically, the second dataset was collected in December 2022, while the third and fourth datasets were gathered between January 2024 and April 2024. Detailed descriptions of the

design processes for these datasets, along with their associated statistical analyses (including the first dataset), are provided in Section **??**.

6.4.2 Sequential Robust Optimal Designs

6.4.2.1 Analysis and Model Selection Based on the First Dataset

Each experiment records the number of embryos across four outcome categories: normal, radialization, 0 spicules, and dead/delayed. The experimental design points (dose levels) and corresponding weights varied significantly across the nine days, as summarized in Table 6.3. We fitted multiple

Date	Design points	Design weights	Eff_D	Eff_c
06/23	[0, 1, 5, 10, 30, 100]	[0.16, 0.17, 0.17, 0.19, 0.17, 0.16]	0.448	0.640
07/07	[0, 100, 300, 1000]	[0.25, 0.25, 0.25, 0.25]	0.755	0.199
07/14	[0, 1, 3, 10, 30, 100, 300, 3000, 10000]	[0.12, 0.12, 0.1, 0.11, 0.11, 0.11, 0.11, 0.10, 0.12]	0.900	0.430
07/19	[0, 3, 10, 30, 100, 300, 10000, 30000]	[0.13, 0.12, 0.12, 0.14, 0.11, 0.12, 0.15, 0.11]	0.869	0.425
07/21	[0, 0.3, 3, 30, 300, 3000, 30000]	[0.14, 0.16, 0.13, 0.13, 0.13, 0.17, 0.14]	0.830	0.317
07/26	[0, 1, 10, 30, 100, 300, 10000, 20000, 30000]	[0.11, 0.13, 0.10, 0.12, 0.11, 0.11, 0.11, 0.11, 0.11]	0.843	0.356
07/28	[0, 0.3, 1, 30, 3000, 10000, 20000, 30000]	[0.13, 0.14, 0.12, 0.12, 0.13, 0.14, 0.11, 0.12]	0.692	0.141
08/11	[0, 0.3, 1, 10, 100, 3000, 10000, 20000]	[0.11, 0.12, 0.14, 0.13, 0.13, 0.12, 0.13, 0.12]	0.814	0.305
08/19	[0, 0.3, 1, 30, 300, 3000, 10000, 20000]	[0.11, 0.13, 0.11, 0.12, 0.12, 0.14, 0.16, 0.11]	0.836	0.249

Table 6.3: D- and c-efficiencies of designs used in the first dataset.

regression models to analyze the data, including ordinal regression (proportional odds), continuation ratio, and adjacent categories logit models. Model selection was guided by AIC and BIC values (Table 6.4). Based on the AIC and BIC results, we choose to use proportional odds model with logit link for fitting the 9 day data separately. The fitted proportional odds model with logit link using the whole first dataset is given in Figure 6.1 while the fitted models using 9 day data separately is given in Figure 6.2 and estimated parameters are given in Table 6.5. Later we utilize the estimated

Model	AIC	BIC
Cumulative logit model	7212.557	7221.376
Proportional odds model with Cauchit link	7973.305	7994.327
Proportional odds model with logit link	6747.137	6768.160
Adjacent categories logit model	7720.505	7727.119
Continuation-ratio logit model	7656.651	7663.265

Table 6.4: Comparison of trinomial models

parameters to construct the so-called robust designs (Section 6.4.2.3) (Collins et al., 2022).

The proportional odds model with logit link was selected for its superior performance. Figure 6.2 illustrates the fitted models for each experimental day, highlighting variations in dose-response relationships due to biological variability.

Although the data is not independent across 9 days, the resulting dose-response curve can still provide invaluable information when we have different designs. Hence, the resulting second stage design stage is expected to be more robust against the design that we only use one set of nominal values (i.e., the one that we fit the 9-day data all altogether). Taking the first panel in Figure 6.2 as an example, the dose levels are low compared with others, reflecting potential genetic variants in sea urchins across different timelines (since the second stage design is performed at a different time compared with the first one).

Date	β_1	β_2	α
06/23	2.328	9.845	-1.562
07/07	2.077	10.686	-1.303
07/14	2.157	9.342	-1.019
07/19	2.516	9.127	-1.086
07/21	2.186	8.029	-0.960
07/26	2.380	8.359	-1.040
07/28	2.442	8.331	-1.037
08/11	2.449	8.121	-1.021
08/19	2.506	7.800	-0.979

Table 6.5: 9 sets of nominal values based on the first dataset.

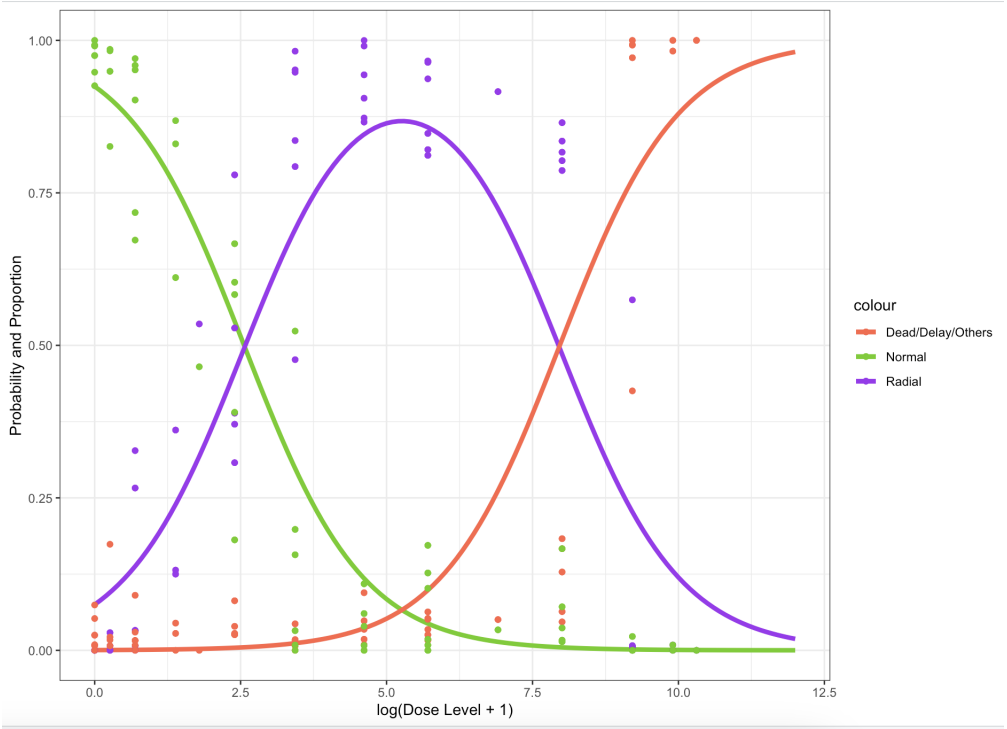


Figure 6.1: The fitted proportional odds model with logit link using the whole first dataset.

6.4.2.2 The Second Dataset: Locally Optimal Design

We construct a locally optimal design to compare it with our proposed two-stage robust design. To construct locally optimal design, we need to first fit the whole dataset and then use the nominal values (i.e., fitted parameter estimates) to compute an optimal design numerically. To estimate the parameters $\theta = (\beta_1, \beta_2, \alpha)$ as accurate as possible, we consider minimizing the volume of the confidence set:

$$S = \left\{\theta : (\widehat{\theta} - \theta)^T M^{-1} (\widehat{\theta} - \theta) \leq \chi^2_{df,0.95}\right\}$$

where $\widehat{\theta}$ is the estimator of θ, M is the asymptotic variance of $\widehat{\theta}$ and $\xi^2_{df,0.95}$ is the 95%-quantile of Chi-square distribution with degrees of freedom df. It can be shown that the volume of S is proportional to the determinant of M^{-1}, which is the D-optimality in literature.

In addition, RD50 is another important quantity of interest (personal communication with Dr. Collins). Hence, minimizing the asymptotic variance of an estimator of RD50 is another goal of the design. The asymptotic variance of RD50 has a neat form and we call it the c-optimality. Then a dual-optimality criteria can be constructed as a convex combination of D- and c-optimality.

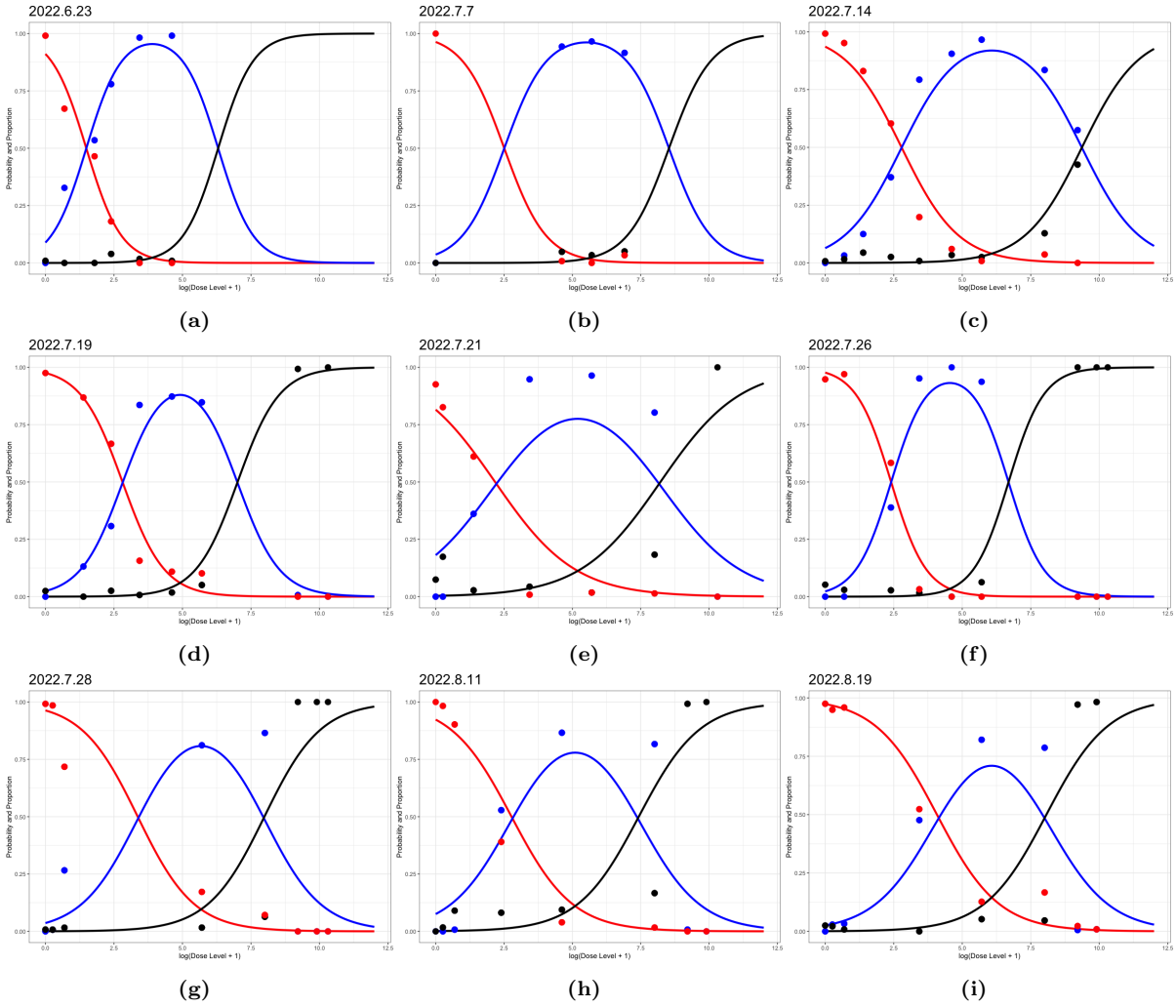


Figure 6.2: Daily fitted proportional odds models with logit link based on the first dataset. Red represents the observed and predicted proportion of normal embryos; blue represents the observed and predicted proportion of radial embryos; black represents the observed and predicted proportion of dead/delayed embryos.

The dual-optimal design is then defined as

$$\xi_{dual} = \arg\min_{\xi \in \Xi} \Phi_{dual}(M) = \arg\min_{\xi \in \Xi} \left(\lambda \frac{\Phi_D(M)}{3} - (1-\lambda) \log \Phi_c(M) \right) \tag{6.4.1}$$

where $\Phi_D(M) = \log\det M$ and $\Phi_c(M) = \left(\nabla_\theta x|_{\theta=\widehat{\theta}}\right)^T M(\xi)^{-1} \left(\nabla_\theta x|_{\theta=\widehat{\theta}}\right)$ and $\lambda \in [0,1]$ is a weight to trade-off between D- and c-optimality. The optimization problem 6.4.1 is constrained because the design space is $[0, +\infty)$ and the design weights are all within $[0, 1]$ with summation 1. Hence, we apply Particle Swarm Optimization (PSO) (Miranda, 2018) to solve it with the special choice $\lambda = 0.5$, indicating an equal importance of both criteria. Using the estimated parameters of the first dataset in table 6.7, the resulting locally dual-optimal design with equal weights and additional zero dose level is given in the table 6.6 which forms the basis of the second dataset. The sensitivity function is shown in figure 6.3. Note that the table shows the raw scale of dose level while for implementation, we use log(Dose level + 1) so that the design space is the same as the original space $[0, +\infty)$ but the numerical issues can be lightened.

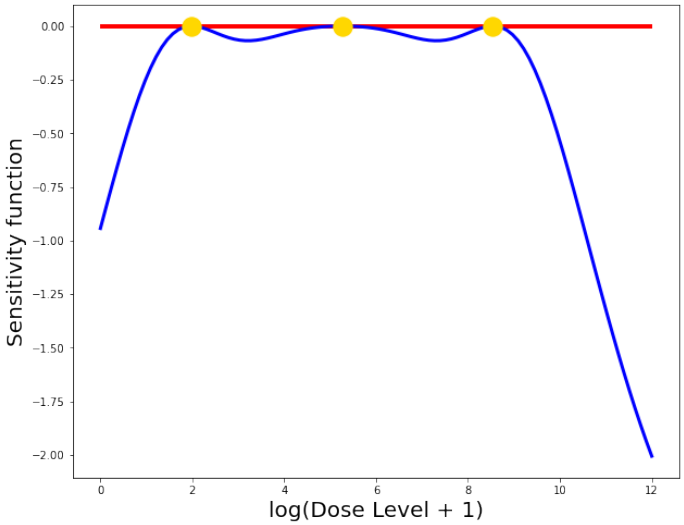


Figure 6.3: Sensitivity function of the dual-optimal design. The x-axis represents the log-transformed dose levels, while the y-axis shows the sensitivity function values. The blue curve represents the sensitivity function across these dose levels, while the red line at the top of the plot highlights the zero-sensitivity baseline.

The estimated parameters and their standard errors based on the second dataset (also called the December data) is given in the right column of Table 6.7 and the fitted dose-response curve is given in Figure 6.4a. It is obvious that the curve does not fit the data well and the corresponding standard

Date	Design points	Design weights	Eff_D	Eff_c
ξ_D	[5.77, 161.4, 4391.52]	[0.33, 0.34, 0.33]	1.000	0.372
ξ_c	[12.01]	[1.00]	Singular	1.000
ξ_{dual}	[9.08, 59.7, 4290]	[0.65, 0.175, 0.175]	0.798	0.748
12/03	[0, 9.05, 59.7, 4290]	[0.169, 0.521, 0.169, 0.169]	0.767	0.688
12/09	[0, 9.08, 59.7, 4290]	[0.169, 0.521, 0.169, 0.169]	0.767	0.688
12/10	[0, 9.08, 59.7, 4290]	[0.169, 0.521, 0.169, 0.169]	0.767	0.688

Table 6.6: D- and c-efficiencies of the December data.

errors are not small enough as expected (though we are using dual-optimality here). Hence, it is not enough to just use optimal design alone, and we are expecting that the two-stage design may perform better. To investigate the gap between the reality and theory in depth, we conjecture that there could a genetic shift in the dataset (see also Figure 6.2), meaning that the tolerance of sea urchin changes in the genetic composition of populations over time, often in response to environmental pressures such as temperature, pH, or salinity. In sea urchins, like many marine organisms, these shifts can influence key developmental processes in embryos and can impact the resilience of populations to climate change and ocean acidification (Jorde and Wooding, 2004).

	The first dataset (June-August)	The second dataset (December)
Total observations	8163	2070
β_1	2.506 (0.055)	0.593 (0.081)
β_2	7.800 (0.134)	6.106 (0.231)
α	0.979 (0.016)	0.719 (0.030)
$RD50$	2.580 (0.089)	0.780 (0.149)

Table 6.7: Ordinal regression models based on two datasets.

6.4.2.3 The Third and Fourth Datasets: Two-stage Robust Optimal Design

The name "robust" comes from the fact that we use different sets of nominal values to construct the optimal design (Zang et al., 2005; Collins et al., 2022).The optimal design is constructed as follows:

$$\xi_{opt} = \arg\max_{\xi} \sum_{i=1}^{K} \Phi(\xi, \widehat{\theta}_i)$$

where each $\Phi(\xi, \theta_i)$ is a criterion function that uses the nominal value θ_i. In the case of D-optimality, $\Phi(\xi, \theta_i)$ is the log determinant of Fisher information matrix while in the case of dual-optimality for estimating parameters and oral radialization 50%, $\Phi(\xi, \theta_i)$ is the convex combination of log D-efficiency and log c-efficiency (Hyun and Wong, 2015).

We construct the two-stage robust optimal designs based on the first dataset. To achieve this, we need to determine how many additional doses we want add to the first dataset. In our real application, we set this as the same number of total observations as the first dataset ($\alpha = 0.5$). Further, we determine to put a fixed proportion to zero dose level (the control group) and 10,000 dose level (the endpoint to make sure all sea urchins are dead). The resulting two-stage robust D-optimal design is given in Table 6.8 while the two-stage robust dual-optimal design (with equal weights) is given in Table 6.9.

Design points	0	14	55	683	4808	10000
Design weights	0.225	0.145	0.112	0.151	0.142	0.225

Table 6.8: Two-stage robust D-optimal design based on the first dataset.

Design points	0	5	25	989	3727	10000
Design weights	0.215	0.200	0.305	0.033	0.032	0.215

Table 6.9: Two-stage robust dual-optimal design with equal weights based on the first dataset.

Based the above suggested designs, we collect additional toxicological data and we have the following four datasets:

- The first dataset from June-August;
- The second dataset in December which is based on the first dataset and is dual-optimal (not robust);
- The third dataset which is based on the two-stage robust D-optimal design;
- The fourth dataset which is based on the two-stage robust dual-optimal design.

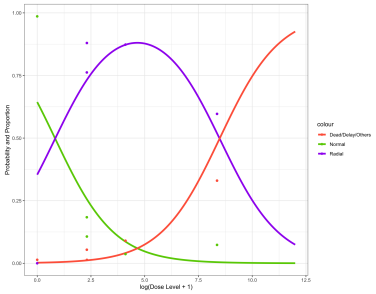
The results of fitted ordinal regression models for the third and fourth datasets are shown in Table 6.10 and Figure 6.4. Table 6.10 provides fitted parameters from ordinal regression models

with a logit link function using two robust design strategies: the two-stage robust D-optimal and the two-stage robust dual-optimal designs. Here, we note that the total number of observations is slightly higher for the dual-optimal design (927 compared to 889 for the D-optimal design), as the actual number of observations may fluctuate a little in contrast to theoretical calculations. Parameter estimates, such as β_1, β_2, and α, as well as the RD50, are provided along with standard errors. The smaller standard errors in the dual-optimal design indicate potentially greater precision for this design. Interestingly, the standard errors of the dual-optimal design are generally lower than the D-optimal designs, which contradicts with the intuition and the original intention of both designs. We further explore this issue in Table 6.11. In addition, the RD50 estimate is slightly higher under the dual-optimal design (2.955 compared to 2.650), which could imply differences in dose-response sensitivity captured by each design.

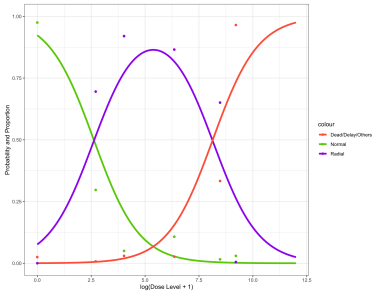
Table 6.11 provides a detailed look at the $D-$ and $c-$ optimality values and their efficiencies under different design settings. The efficiencies are computed in a relative manner, i.e., it is the actual efficiency divided by the largest efficiency among 4 designs and sample sizes (total number of observations) are scaled so results from different designs are comparable. Here $\Theta_i, i = 1, 2, 3, 4$ correspond to the parameters that we estimated from four different designs. The Conventional design refers to the original dataset designed by Dr. Michael Collins; the Dual, Robust-D and Robust Dual designs refer to the locally dual-optimal, robust-D and robust-dual designs based on Θ_1 and daily fitted models (see Figure 6.5) respectively. Interestingly and sarcastically, the Robust D design never performs the best in terms of D-optimality among all scenarios (Θ's). One most plausible explanation is still due to the genetic shift of the dataset, making it awkward for the new dataset it has collected. Apart from that, all designs perform as expected and in some scenarios the Dual design performs better than the Robust Dual design in terms of both D- and c-optimality, suggesting that there is always a gap between theory and practice.

	Two-stage robust D-optimal design	Two-stage robust Dual-optimal design
Total observations	889	927
β_1	2.694 (0.181)	2.065 (0.127)
β_2	7.449 (0.430)	5.313 (0.298)
α	0.906 (0.043)	0.787 (0.040)
$RD50$	2.650 (0.154)	2.955 (0.123)

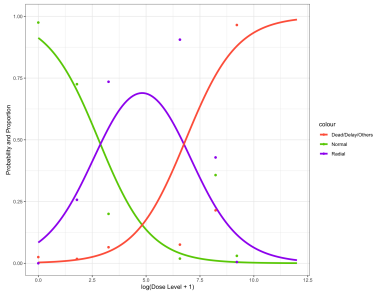
Table 6.10: Fitted ordinal regression models with logit link based on two-stage robust designs.



(a) Fitted proportional odds model based on the December data (the second dataset).



(b) Fitted dose-response curve based on the two-stage robust D-optimal design (the third dataset).



(c) Fitted dose-response curve based on the two-stage robust Dual-optimal design (the fourth dataset).

Figure 6.4: Dose-response curves fitted using various datasets and design strategies, illustrating the probability (or proportion) of different outcomes across dose levels on a logarithmic scale. The curves represent three outcome categories: "Dead/Delayed/Others" (red), "Normal" (green), and "Radial" (purple). Plot (a) shows the dose-response curve using a proportional odds model based on the December data, representing a conventional approach. Plot (b) uses the two-stage robust D-optimal design, adjusting the dose-response curve to improve model robustness. Plot (c) applies the two-stage robust dual-optimal design, further refining the dose-response prediction, particularly for the "Radial" and "Dead/Delayed/Others" categories, and demonstrating the enhanced adaptability of the dual-optimal approach. These variations highlight the effects of different optimization criteria on dose-response predictions across varying outcome probabilities.

Estimates	Datasets	1 (Conventional)	2 (Dual)	3 (Robust D)	4 (Robust Dual)
Θ_1	D-optimality	5.918	6.293	6.118	6.417
	D-efficiency	1.000	0.882	0.935	0.847
	c-optimality	11.990	6.305	17.103	10.191
	c-efficiency	0.526	1.000	0.369	0.619
Θ_2	D-optimality	4.997	5.556	5.039	5.163
	D-efficiency	1.000	0.830	0.986	0.946
	c-optimality	17.961	17.160	18.202	13.753
	c-efficiency	0.766	0.801	0.756	1.000
Θ_3	D-optimality	5.783	6.159	6.025	6.386
	D-efficiency	1.000	0.882	0.923	0.818
	c-optimality	12.582	7.048	18.508	12.106
	c-efficiency	0.560	1.000	0.381	0.582
Θ_4	D-optimality	4.891	5.063	5.202	5.353
	D-efficiency	1.000	0.944	0.902	0.857
	c-optimality	14.891	8.943	19.921	12.787
	c-efficiency	0.601	1.000	0.449	0.699

Table 6.11: Comparison among four datasets. $\Theta_1, \Theta_2, \Theta_3, \Theta_4$ stand for estimated parameters $(\beta_1, \beta_2, \alpha)$ from the four different datasets.

6.4.2.4 Equivalence Theorems

In this subsubsection, we derive the equivalence theorems for the two-stage robust dual-optimal design with an additional 0 dose level under the ordinal regression setting.

We start with a more general case which is the multivariate logistic regression. The multivariate

logistic regression is an extension of the usual logistic regression for categorical outcomes with or without intrinsic ordering (personal communication with Dr. Weng Kee Wong). Suppose $\mathbf{y} \sim \mathcal{M}(n, \pi)$ and we assume

$$\begin{aligned}
&\eta = X\beta,\ \pi = \mu(\theta) \\
&\eta = g(\boldsymbol{\pi}) = \psi^{-1}(\boldsymbol{\pi}) = C^T \log(L\boldsymbol{\pi}) \\
&\theta = r(\eta) = \log \boldsymbol{\pi} = \log g^{-1}(\eta)
\end{aligned}$$

where X is the design matrix and C and L are matrices of contrasts and marginal indicators (Glonek and McCullagh, 1995). Different choices of C and L lead to different models such as proportional odds model, adjacent categories logit model and continuation ratio logit model. Further, the nominal logistic regression can also be written in this form (Glonek and McCullagh, 1995). Most importantly, by introducing L we ensure the mapping $\boldsymbol{\pi} \to \eta$ is of full rank so that we can write $\theta = \log \boldsymbol{\pi}$. Thus, we implicitly put the linear constraint $\mathbf{1}^T \exp(\theta) = 1$ so that in this case $A(\theta) = 0$.

For convenience, we note:

- y, $\boldsymbol{\pi}$, θ and η are all $k \times 1$ vectors;
- β is a $p \times 1$ vector;
- X is a $k \times p$ matrix of full row rank;
- L and C are $l \times k$ matrices both of full column rank.

The Fisher information of β is given by

$$\begin{aligned}
\mathbf{M}(\beta) &= X^T \left(C^T D^{-1} L\right)^{-T} \mathbf{V}^{-1} \left(C^T D^{-1} L\right)^{-1} X \\
&= \widetilde{s}(X)\widetilde{s}(X)^T
\end{aligned} \tag{6.4.2}$$

where $D = \mathrm{Diag}(L\boldsymbol{\pi})$ and

$$\begin{aligned}
\widetilde{s}(X) &= X^T \left(C^T D^{-1} L\right)^{-1} \mathbf{V}^{-1/2}, \\
\mathbf{V}^{-1/2} &= \mathrm{Diag}(1/\sqrt{\boldsymbol{\pi}}).
\end{aligned}$$

The $\widetilde{s}(X)$ representation is particularly useful we are deriving equivalence theorems and sensitivity functions in optimal design theory. More properties of multivariate logistic regression and its extensions are given in Glonek and McCullagh (1995); Lang (1996); Bu et al. (2020).

Example 6.4.1 (Ordinal Trinomial Regression (Zocchi and Atkinson, 1999))**.** Following the previous example, the 3-dimensional response vector $y = (y_1, y_2, y_3)^T \sim \mathcal{M}(1, \pi)$ has an intrinsic ordering, i.e., y_1, y_2 and y_3 correspond to normal, radialization and dead, respectively. We model y using a proportional odds model with common slope:

$$\begin{aligned}
\eta_1 &= \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) = \beta_1 + \alpha x \\
\eta_2 &= \log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right) = \beta_2 + \alpha x \\
\eta_3 &= \log(\pi_1 + \pi_2 + \pi_3) = 0
\end{aligned}$$

In this case, the parameter θ and the design matrix X are

$$\theta = \begin{pmatrix} \beta_1 & \beta_2 & \alpha & 0 \end{pmatrix}^T, \qquad X = \begin{pmatrix} 1 & 0 & x & 0 \\ 0 & 1 & x & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and the choices of L and C^T are

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad C^T = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

In a usual sequential design setting, we add a point x to the existing design ξ so that the corresponding design efficiency (say, D-efficiency) is improved. However, as we mentioned in Section 6.3.3, this is not always what toxicologists want in practice. In practice, we always want a "control group" under any experimental design. In other words, there should always a 0 dose level

group. In this case, the design becomes

$$\xi = \begin{pmatrix} 0 & x_1 & \cdots & x_n \\ \alpha & (1-\alpha)p_1 & \cdots & (1-\alpha)p_n \end{pmatrix}$$

where $p_i, i = 1, \cdots, n$ are nonnegative design weights and $\alpha \in [0, 1]$ is another weight assigned to the control group $x = 0$. In the following, we derive the expression of Fisher information when we add a new design point to the original design.

Example 6.4.2 (Adding a new design point)**.** In example 6.4.1, the Fisher information matrix is of form

$$M(\beta) = \widetilde{s}(X)\widetilde{s}(X)^T$$

where $\widetilde{s}(X)$ is a $p \times k$ matrix. Let $M_1 = \sum_{i=1}^n p_i \widetilde{s}(X_i)\widetilde{s}(X_i)^T$ be the information matrix associated with ξ_1 and p_i are nonnegative design weights. Also let $M_0 = \widetilde{s}(X_0)\widetilde{s}(X_0)$ be the information matrix at point 0. If we want to add 0 to the design ξ_1 with weight $\alpha \in (0, 1)$, then by Sherman-Woodbury-Morrison, the corresponding D-optimality criteria becomes

$$\det\left((1-\alpha)M_1 + \alpha M_0\right) = (1-\alpha)^p \det(M_1) \det\left(I + \frac{\alpha}{1-\alpha}\widetilde{s}(X_0)^T M_1^{-1}\widetilde{s}(X_0)\right). \tag{6.4.3}$$

Further, the c-optimality for estimating $g(\theta)$ becomes

$$\begin{aligned} &\nabla g^T \left((1-\alpha)M_1 + \alpha M_0\right)^{-1} \nabla g = \\ &\frac{\nabla g^T M_1^{-1} \nabla g}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} \nabla g^T M_1^{-1} \widetilde{s}(X_0) \left(I + \frac{\alpha}{1-\alpha}\widetilde{s}(X_0)^T M_1^{-1}\widetilde{s}(X_0)\right)^{-1} \widetilde{s}(X_0)^T M_1^{-1} \nabla g \end{aligned} \tag{6.4.4}$$

where ∇g is the gradient of $g(\theta)$ at θ.

Following example 6.4.1, the Fisher information matrix of β associated with ξ is $M(\xi) = \alpha s_0 s_0^T + (1-\alpha)M_1$ where $s_0 = \widetilde{s}(X)|_{x=0}$ and M_1 is the Fisher information matrix associated with the design with point 0 removed. Let us consider D-optimality, i.e., $\Phi_D(M) = \log\det M$, then by example 6.4.2, we have

$$\Phi_D(M) = \log\det(M_1) + p\log(1-\alpha) + \log\det\left(I + \frac{\alpha}{1-\alpha} s_0^T M_1^{-1} s_0\right). \tag{6.4.5}$$

Suppose α is fixed so that M_1 is the only free variable in the expression and we write $\Phi_D(M_1)$ instead of $\Phi_D(M)$. By standard matrix calculus, the Fréchet derivative of $\Phi_D(M)$ at M_1 in the direction of M_2 is

$$F_{\Phi_D}(M_1, M_2) = (1-\alpha)\mathrm{Tr}\left(M^{-1}(M_2 - M_1)\right) \tag{6.4.6}$$

$$= \mathrm{Tr}\left(M_1^{-1}M_2 - I\right) - \mathrm{Tr}\left(\left(\frac{1-\alpha}{\alpha}I + s_0^T M_1^{-1} s_0\right)^{-1}\left(s_0^T(M_1^{-1}M_2M_1^{-1} - M_1^{-1})s_0\right)\right). \tag{6.4.7}$$

Now the general equivalence theorem applies using the fact that $\Phi_D(M_1)$ is concave in M_1. For c-optimality, we need to use the implicit function theorem (Rudin, 1976) to compute the asymptotic variance of, say, radialization 50% estimate.

Example 6.4.3 (Radialization 50%)**.** In this example, we are interested in finding a specific dose level of a substance that leads to a 50% reaction rate (referred to as RD50), meaning it causes a measurable response in 50% of the population. Using a trinomial regression model, which is a statistical method for analyzing outcomes that can take on three possible values, we aim to calculate this dose level. However, there isn't a simple formula to solve for this dose directly based on the model parameters. Instead, we use mathematical techniques to approximate the effect of small changes in the model parameters on our estimate of the RD50 dose. This process allows us to calculate an "asymptotic variance," or a measure of how accurate and stable our estimated RD50 is likely to be with large sample sizes. Additionally, we look at a related measure, LD50, which indicates a 50% lethal dose, and compute the variability for a ratio of these two metrics. This approach is valuable because it helps us assess the reliability of the dose estimates used in toxicological studies.

Suppose we have the trinomial regression model (Example 6.4.1) with parameters $\theta^T = (\alpha, \beta_1, \beta_2, b)$ and dose level x (with the implicit constraint b=0). Suppose the interest is in estimating radialization 50% (RD50), i.e., the dose level x such that $\pi_2 = 0.5$. An equation for solving x is

$$\frac{1}{1+\exp(-\eta_2)} - \frac{1}{1+\exp(-\eta_1)} - 0.5 = 0.$$

There is no closed form solution of x in terms of θ. However, it is more important for us to derive the partial derivatives $\partial x/\partial\theta^T$ so that Δ-method can be applied to approximate the asymptotic

variance of estimated radialization 50%. Denote the left hand side of the equation as $f(x,\theta)$, then by the implicit function theorem, we have

$$\begin{aligned} A &= f'(x,\theta) \\ &= \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial \alpha} & \frac{\partial f}{\partial \beta_1} & \frac{\partial f}{\partial \beta_2} & \frac{\partial f}{\partial b} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\alpha \exp(-\eta_2)}{(1+\exp(-\eta_2))^2} - \frac{\alpha \exp(-\eta_1)}{(1+\exp(-\eta_1))^2} \\ \frac{x \exp(-\eta_2)}{(1+\exp(-\eta_2))^2} - \frac{x \exp(-\eta_1)}{(1+\exp(-\eta_1))^2} \\ -\frac{\exp(-\eta_1)}{(1+\exp(-\eta_1))^2} \\ \frac{\exp(-\eta_2)}{(1+\exp(-\eta_2))^2} \\ 0 \end{pmatrix}^T \end{aligned}$$

so that the desired $\nabla_\theta x$ can be computed numerically by plug-in estimates of θ, i.e.,

$$\widehat{\nabla_\theta x}^T = \left(\frac{\partial f/\partial \alpha}{\partial f/\partial x}, \frac{\partial f/\partial \beta_1}{\partial f/\partial x}, \frac{\partial f/\partial \beta_2}{\partial f/\partial x}, \frac{\partial f/\partial b}{\partial f/\partial x} \right).$$

The asymptotic variance of radialization 50% is thus

$$\mathbb{V}(\widehat{x}) \approx \left(\nabla_\theta x|_{\theta=\widehat{\theta}} \right)^T M(\xi)^{-1} \left(\nabla_\theta x|_{\theta=\widehat{\theta}} \right)$$

where $M(\xi)$ is the Fisher information matrix of θ associated with design ξ evaluated at $\widehat{\theta}$. Further, if we are also interested in lethal dose 50% (LD50) , and we want to estimate the ratio (suggested by Dr. Collins) $r = \frac{LD50}{RD50}$, then the asymptotic variance of $\widehat{r}$ can be approximated as

$$\mathbb{V}(\widehat{r}) = \begin{pmatrix} -\frac{\widehat{x}_{LD50}}{\widehat{x}^2_{RD50}} \\ \frac{1}{\widehat{x}_{RD50}} \end{pmatrix}^T \widehat{Cov} \begin{pmatrix} \widehat{x}_{LD50} \\ \widehat{x}_{RD50} \end{pmatrix} \begin{pmatrix} -\frac{\widehat{x}_{LD50}}{\widehat{x}^2_{RD50}} \\ \frac{1}{\widehat{x}_{RD50}} \end{pmatrix}$$

where the 2×2 covariance matrix of $\widehat{x}_{LD50}$ and $\widehat{x}_{RD50}$ is computed by, again, the Δ-method.

Plug-in $M(\xi) = \alpha s_0 s_0^T + (1-\alpha) M_1$ where $s_0 = \widetilde{s}(X)|_{x=0}$ and M_1 is the Fisher information matrix associated with the design with point 0 removed. Let us consider c-optimality, i.e., $\Phi_c(M) =$

$\log \mathbb{V}(\widehat{x})$, then by example 6.4.2, we have

$$\Phi_c(M) = \log \left[\frac{c^T M_1^{-1} c}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} c^T M_1^{-1} \widetilde{s}(X_0) \left(I + \frac{\alpha}{1-\alpha} \widetilde{s}(X_0)^T M_1^{-1} \widetilde{s}(X_0) \right)^{-1} \widetilde{s}(X_0)^T M_1^{-1} c \right] \tag{6.4.8}$$

where $c = \nabla_\theta x|_{\theta=\widehat{\theta}}$ is the gradient computed using the implicit function theorem. The Fréchet derivative of Φ_c at M in the direction of M_2 is

$$F_{\Phi_c}(M, M_2) = -\mathrm{Tr}\left(cc^T M^{-1} \left(M_2 - M \right) M^{-1} \right) \tag{6.4.9}$$

$$= c^T M^{-1} c - c^T M^{-1} \widetilde{s}(x) \widetilde{s}(x)^T M^{-1} c. \tag{6.4.10}$$

where M is defined above and provided $M_2 = \widetilde{s}(x)\widetilde{s}(x)^T$. Now we have all the necessary elements to derive the equivalence theorem for the two-stage robust dual-optimal design. A robust dual optimal design optimizes the following criterion

$$\Phi_{\text{robust-dual}}(M) = \frac{1}{K} \sum_{i=1}^{K} \left(\frac{\lambda_i}{p} \Phi_D(M|\widehat{\theta}_i) + (1-\lambda_i) \Phi_c(M|\widehat{\theta}_i) \right) \tag{6.4.11}$$

where K is the number of sets of nominal values, p is the dimension of θ_i λ_i is the weight for $D-$ and $c-$optimality using the i^{th} nominal values and $\widehat{\theta}_i$ represents the i^{th} nominal values.

Theorem 6.4.1 (Equivalence theorem for augmented two-stage robust dual-optimal design)**.** Suppose at first stage, we perform K sets of experiments and derive K sets of nominal values (denoted as $\widehat{\theta}_1, \cdots, \widehat{\theta}_K$). Given specific weights $\lambda_1, \cdots, \lambda_K$ and the proportion of added zero dose level $\alpha_i, i = 1, \cdots, K$, suppose we are trying optimize the robust dual-optimal criterion 6.4.11. Then a design ξ is robust dual-optimal if and only if for any x, the Fisher information associated with ξ (denoted as M), satisfies

$$\frac{1}{K} \sum_{i=1}^{K} \left(\frac{\lambda_i}{p} F_{\Phi_D}(M, \widetilde{s}(x)\widetilde{s}(x)^T|\widehat{\theta}_i) + (1-\lambda_i) F_{\Phi_c}(M, \widetilde{s}(x)\widetilde{s}(x)^T|\widehat{\theta}_i) \right) \leq 0 \tag{6.4.12}$$

where $F_{\Phi_D}(M, \widetilde{s}(x)\widetilde{s}(x)^T|\widehat{\theta}_i)$ is given by 6.4.6 and $F_{\Phi_c}(M, \widetilde{s}(x)\widetilde{s}(x)^T|\widehat{\theta}_i)$ is given by 6.4.9.

According to Theorem 6.4.1, the researcher could apply a dual-criterion approach where weights are assigned to both D- and c-optimality objectives. The design that satisfies the conditions laid

out in Theorem 3.4 would optimize this balance, ensuring robustness by minimizing variability across different possible parameter values for the drug's effect. By using a two-stage approach, the researcher can refine the design after an initial set of experiments, focusing on the most informative dose levels identified in the first stage.

Further, Theorem 6.4.1 guides the researcher in setting up the dual-optimal design that not only optimizes parameter estimation (D-optimality) but also ensures accurate dose-response relationship modeling (c-optimality) across various experimental conditions. This robustness is crucial in toxicology, where small inaccuracies in dose-response estimates can lead to significant errors in understanding the drug's safety and efficacy.

An Application We apply Theorem 6.4.1 to the two-stage robust design scheme developed in Section 6.3. Here we consider $\alpha = 0.225$, i.e., 22.5% of the second-stage observations will be put as 0 dose level and 22.5% of the second-stage observations will be put as 10,000 dose level. Note that 0 dose serves as the control group while $10,000$ dose level is to ensure that all sea urchins are dead, which is an important endpoint in toxicology. The sensitivity plot of the sequential D-optimal design is given in Figure 6.5 and it verifies the global optimality of our implemented design despite of a little numerical fluctuation. We include more sensitivity plots under different optimality criterion and different α-values in the Appendix **??**.

6.5 Simulation Study: Sequential Optimal Designs for Bivariate Probit Model

In this section, we develop two-stage optimal designs for bivariate probit model described in Dragalin et al. (2008b). Such a model has numerous applications in the context of Phase I or II studies, exploiting dose-finding that accounts for both efficacy and safety.

6.5.1 The Model and the Fisher Information Matrix

Recently, various approaches have been suggested for dose escalation studies based on observations of both undesirable events and evidence of therapeutic benefit. bivariate probit model has been applied to model such a response and we give a brief introduction below; for a more comprehensive

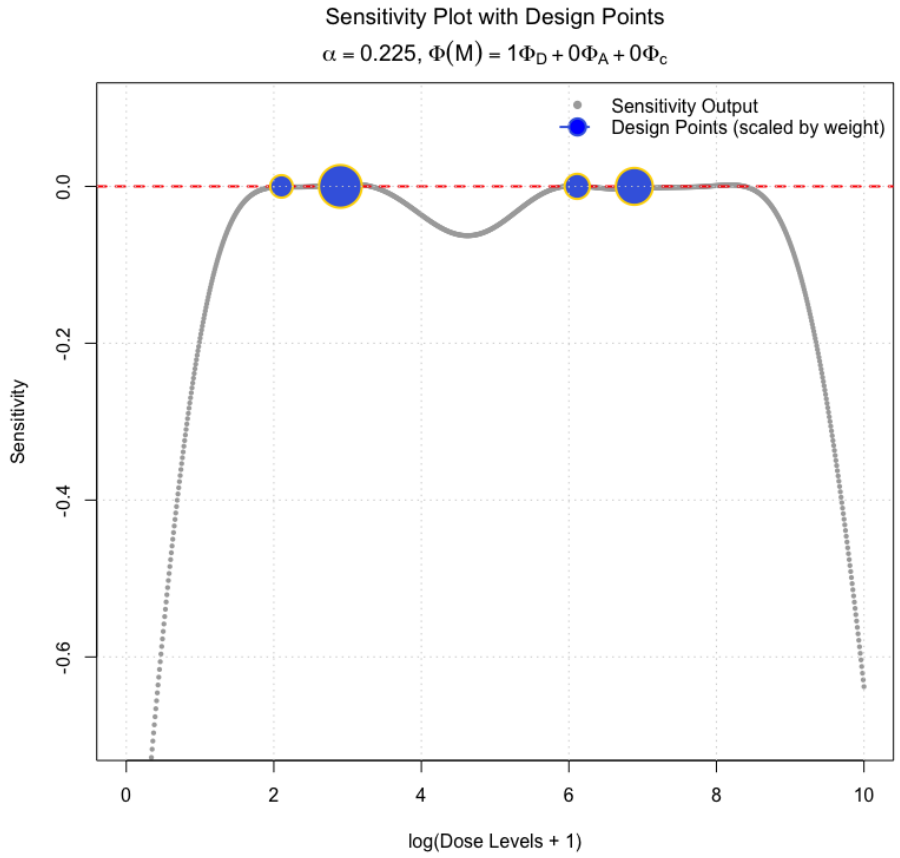


Figure 6.5: Sensitivity function of the sequential robust D-optimal design. The x-axis represents the log-transformed dose levels, while the y-axis shows the sensitivity function values. The blue curve represents the sensitivity function across these dose levels, while the red line at the top of the plot highlights the zero-sensitivity baseline.

review, see McCullagh and Nelder (2019).

Let $Y \in \{0,1\}$ represent efficacy response and $Z \in \{0,1\}$ represent toxicity response, where 1 indicates occurrence and 0 indicates non-occurrence. In our example, the efficacy response is 'no VTE' and the toxicity response is 'bleeding'. A possible dose is denoted by x. The probabilities of different response combinations are defined as:

$$p_{yz}(x) = \Pr(Y = y, Z = z \mid x), \quad y, z = 0, 1$$

The probit model for correlated responses (see Chib and Greenberg (1998); Bekele and Thall (2006)) is given by:

$$p_{11}(x,\theta) = F(\theta_1^T f_1(x), \theta_2^T f_2(x), \rho) = \int_{-\infty}^{\theta_1^T f_1(x)} \int_{-\infty}^{\theta_2^T f_2(x)} \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{v}^T\Sigma^{-1}\mathbf{v}\right) dv_1 dv_2$$

Here, $\theta = (\theta_1^T, \theta_2^T)$, and the variance–covariance matrix is:

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

The matrix Σ is assumed known, though similar results can be derived if ρ is unknown. The functions $f_1(x)$ and $f_2(x)$ include relevant covariates, such as $f_1(x) = f_2(x) = (1, x)^T$ for modeling a single drug effect, or $f_1(x) = f_2(x) = (1, x_1, x_2, x_1x_2)^T$ for modeling drug combinations with interaction effects (Ashford and Sowden, 1970). Additional covariates, such as age, weight, or dosage frequency, can also be incorporated into f_1 and f_2. The marginal distributions for the probabilities of efficacy $p_{1.}(x,\theta)$ and toxicity $p_{.1}(x,\theta)$ are:

$$p_{1.}(x,\theta) = F(\theta_1^T f_1(x)) \quad \text{and} \quad p_{.1}(x,\theta) = F(\theta_2^T f_2(x))$$

where

$$F(v) = \int_{-\infty}^{v} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

The other probabilities are derived as follows:

$$p_{10}(x,\theta) = p_{1.}(x,\theta) - p_{11}(x,\theta)$$

$$p_{01}(x,\theta) = p_{\cdot 1}(x,\theta) - p_{11}(x,\theta)$$

$$p_{00}(x,\theta) = 1 - p_{1\cdot}(x,\theta) - p_{\cdot 1}(x,\theta) + p_{11}(x,\theta)$$

For clarity, we sometimes omit arguments like x and θ when the context is clear.

6.5.1.1 Information Matrix for Bivariate Probit Model

The normalized and unnormalized Fisher information matrices are derived in detail in Dragalin et al. (2008b) and we present their results here. For a more general procedure of deriving Fisher information for optimal designs, see Atkinson et al. (2014).

Given a design

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_{n-1} & x_n \\ p_1 & p_2 & \cdots & p_{n-1} & p_n \end{pmatrix},$$

then the (normalized) Fisher information matrix under the bivariate probit model is

$$M(\xi,\theta) = \sum_{i=1}^{n} p_i \mu(x_i,\theta)$$

where $\mu(x_i,\theta)$ is the elemental information for a single observation at dose x:

$$\mu(x,\theta) = C_1 C_2 (P - pp^T)^{-1} C_2^T C_1^T$$

with

$$C_1 = \begin{pmatrix} \psi(\theta_1^T f_1) f_1 & 0 \\ 0 & \psi(\theta_2^T f_2) f_2 \end{pmatrix}, \quad C_2 = \begin{pmatrix} F(u_1) & 1 - F(u_1) & -F(u_1) \\ F(u_2) & -F(u_2) & 1 - F(u_2) \end{pmatrix}$$

$$u_1 = \frac{\theta_2^T f_2 - \rho \theta_1^T f_1}{\sqrt{1-\rho^2}}, \quad u_2 = \frac{\theta_1^T f_1 - \rho \theta_2^T f_2}{\sqrt{1-\rho^2}}$$

$$P = \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{10} & 0 \\ 0 & 0 & p_{01} \end{pmatrix} \quad \text{and} \quad p = (p_{11}, p_{10}, p_{01})^T$$

where $\psi(u) = \partial F(u)/\partial u$ denotes the probability density function of the standard normal distribu-

tion.

6.5.2 *D*-optimality and *L*-optimality

If the interest is purely in estimating the parameters $\theta = (\theta_1^T, \theta_2^T)$, then we can maximize the determinant of the total normalized information matrix $M(\xi, \theta)$. The resulting design is referred to as the D-optimal design. In practice, we may have specific quantities that are of great interest. For example, we may specify targeted probabilities $(p_{1.}^*, p_{.1}^*)$ in advance, representing the desired efficacy and toxicity levels. Suppose we want the response probabilities $(p_{1.}(x,\theta), p_{.1}(x,\theta))$ are closest to the targeted probabilities in the following sense:

$$d(\theta) = \min_x \left\{ w[p_{1.}(x,\theta) - p_{1.}^*]^2 + (1-w)[p_{.1}(x,\theta) - p_{.1}^*]^2 \right\} \tag{6.5.1}$$

where w is weight between 0 and 1. In other words, we are going to solve for the dose level X^* such that the minimum in the above is obtained. Then the L-optimality is defined as

$$\Psi(M(\xi,\theta)) = L^T(\theta) M^{-1}(\xi,\theta) L(\theta), \;\; L(\theta) = \frac{\partial X^*}{\partial \theta}. \tag{6.5.2}$$

It can also be viewed as A-criterion with $A = L(\theta)L^T(\theta)$ or c-optimality since $L(\theta)$ is a vector. Statistically speaking, $\Psi(M(\xi,\theta))$ represents the asymptotic variance of the estimator X^*. For locally optimal designs, we replace θ with its estimate $\widehat{\theta}$. As pointed out in Dragalin et al. (2008b), it is complicated to work with $p_{1.}$ and $p_{1.}$ directly; so we replace them with their quantiles, i.e., $\phi_1(x,\theta) = F^{-1}(p_{1.}(x))$, $\phi_2(x,\theta) = F^{-1}(p_{.1}(x))$, $\phi_1^* = F^{-1}(p_{1.}^*)$ and $\phi_2^* = F^{-1}(p_{.1}^*)$. Dragalin et al. (2008b) shows that the minimizer in this case is

$$X^* = \frac{(1-w)\theta_{22}(\phi_2^* - \theta_{21}) + w\theta_{12}(\phi_1^* - \theta_{11})}{w\theta_{12}^2 + (1-w)\theta_{22}^*}$$

and its partial derivative w.r.t. θ can be derived easily.

6.5.3 Two Extensions

6.5.3.1 Extension with Penalty

Dragalin et al. (2008b) suggests to introduce penalty into $\Psi(M(\xi,\theta))$ for clinical trial considerations, see also Haines et al. (2003); Dragalin and Fedorov (2006); Dragalin et al. (2008a). Specifically, they define the total normalized penalty

$$\Phi(\xi,\theta) = \int_{\mathcal{X}} \phi(x,\theta)\xi(dx)$$

where ϕ is a user-specified penalty function. For example, to control the costs associated with undesirable events like lack of efficacy or the occurrence of toxicity, then one can define

$$\phi(x,\theta) = p_{1\cdot}^{-C_E}(1-p_{\cdot 1})^{-C_T} \tag{6.5.3}$$

where C_E and C_T are positive constants that quantify the relative importance of penalties for lack of efficacy and occurrence of toxicity, respectively. We use ϕ defined in Equation 6.5.3 in the simulation studies (Section 6.5.4). The resulting optimal design is

$$\xi^* = \arg\min_{\xi\in\Xi} \Psi\left(\frac{M(\xi,\theta)}{\Phi(\xi,\theta)}\right).$$

Note that the special choice $C_E = C_T = 0$ corresponds to the un-penalized case.

6.5.3.2 Extension to Two-stage Designs

In practice, estimating θ often relies on pilot data, leading to a two-stage design approach. This method is similar to the one outlined in Section 6.3.2. Specifically, we begin by fixing a proportion $\alpha \in [0,1]$ of the total pre-specified observations to be collected as pilot data. From this pilot data, we then obtain an estimate $\widehat{\theta}$ using maximum likelihood estimation or other suitable methods. Based on $\widehat{\theta}$, a locally optimal design is constructed according to a chosen criterion, such as D-optimality or L-optimality. The remaining $(1-\alpha)$ proportion of the observations is subsequently collected following this optimized design.

6.5.4 Simulation Studies

In this subsection, we apply PSO to solve for both the D- and L-optimal designs in bivariate probit model and compare the results with those reported in Dragalin et al. (2008b) where they used the first-order algorithm (Atkinson et al., 2007; Fedorov and Hackl, 2012). Although the sensitivity plots fail to confirm the optimality of designs, PSO-generated designs beat the reported designs under most scenarios in terms of criterion values.

Dragalin et al. (2008b) suggests a nominal value $\widehat{\theta} = (-0.9, 1.9, -3.98, 3)$ obtained from a five-dose evenly spaced uniform design (such design is common when we do not have any prior knowledge):

$$\xi_u = \begin{pmatrix} 0.2 & 0.5 & 0.8 & 1.1 & 1.4 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}.$$

Based on the estimated $\widehat{\theta}$, they also constructed and reported the D-optimal and L-optimal designs using first-order algorithms. For comparison, we take $\rho = 0.5, w = 0.5$ and $C_E = C_T = 1$, consistent with the approach outlined in the chapter. The results are presented in Table 6.12 where we compare PSO-generated designs with the three benchmark designs reported in Dragalin et al. (2008b). As an illustration, consider the value 8.332 in row one, column one of the table. This value represents the criterion for the PSO-generated D-optimal one-stage design without penalty. Specifically, it corresponds to the one-stage D-optimality criterion without any penalties applied (i.e., $C_E = C_T = 1$). This design is contrasted with the two-stage extension discussed in Section 6.5.3.2. Highlighted values in bold represent the best performance within each category, indicating the most efficient design strategy for minimizing D-optimality and L-optimality under specific conditions.

The PSO approach used here is effective in finding designs that minimize specific criteria, making it an adaptive and efficient method to optimize both D-optimality and L-optimality. Key insights from this table include the advantages of PSO-generated designs, the impact of using one-stage versus two-stage designs, and the influence of applying penalties.

Firstly, PSO-generated designs consistently outperform the benchmark designs in both one-stage and two-stage configurations, indicating that PSO effectively identifies optimal dose levels and weight distributions to improve statistical efficiency. Two-stage designs generally yield better

performance than one-stage designs, as seen in lower values for both D-optimal and L-optimal criteria (e.g., 8.332 for one-stage D-optimality without penalty versus 8.413 for two-stage). This performance improvement with two-stage designs demonstrates the added flexibility and refinement they offer. Furthermore, applying penalties (i.e., adjusting C_E and C_T parameters) slightly increases the values of D-optimal and L-optimal criteria, highlighting a trade-off between strict optimality and practical constraints. Nevertheless, even with penalties, PSO-generated designs remain more efficient than the benchmark designs, underscoring the robustness of the PSO approach in real-world scenarios where constraints often exist.

The superior performance of PSO-generated designs is evident across various design criteria. For example, PSO-generated designs achieve the lowest values in each category, such as 8.332 for one-stage D-optimality without penalty and 0.434 for L-optimality, confirming that they are highly efficient solutions. This consistency suggests that PSO is highly effective regardless of configuration—whether in one-stage, two-stage, penalized, or non-penalized setups. The PSO approach's adaptability allows researchers to fine-tune experimental designs to meet specific needs, such as balancing dose-response optimization or addressing therapeutic study requirements. In practical terms, the ability of PSO-generated designs to minimize D-optimality and L-optimality with or without penalties positions it as a versatile tool for improving experimental design efficiency, especially in fields like clinical or pharmaceutical studies where optimal dose selection is crucial. These findings highlight PSO's potential to enhance the quality of experimental data, reduce resource usage, and support more accurate conclusions in research.

6.5.5 Python Streamlit App

Additionally, we have developed a Python Streamlit application, which is accessible online at https://optimaldesignbivariateprobit.streamlit.app/.

6.6 Discussion

In this chapter, we have designed a new sequential design scheme for toxicologists to design their experiments more efficiently and more robust in estimating dose-response relationships as well as important endpoints they are interested in. We provide equivalence theorems for checking the

optimality of the design in terms of estimating the RD50 and volume of the confidence ellipsoid. Further, based on the four datasets that Dr. Collins has collected, we empirically show that

- The zero-dose level is crucial in either estimating endpoint or dose-response curves while the extremely high dose level is necessary when we want to estimate the dose-response curve. In conventional statistical literature, the added practical dose levels are always neglected by statisticians;
- The two-stage designs are robust in the way they are constructed. For instance, the two-stage robust D-optimal has relatively high D-efficiencies under different sets of estimated parameters.

Table 6.12: Comparison of D- and L-optimality Across One-stage and Two-stage Designs

Designs	No Penalty				With Penalty			
	One-stage		Two-stage		One-stage		Two-stage	
	D	**L**	**D**	**L**	**D**	**L**	**D**	**L**
PSO-1 D-optimal [w/op]	**8.332**	1.025	8.413	0.868	33.166	499.573	33.167	423.126
PSO-1 L-optimal [w/op]	13.820	**0.434**	9.222	0.521	43.702	762.282	39.104	914.753
PSO-2 D-optimal [w/op]	8.389	1.167	**8.346**	0.921	37.448	1667.525	37.405	1317.194
PSO-2 L-optimal [w/op]	16.245	0.648	9.392	**0.506**	40.978	314.485	34.125	245.512
PSO-1 D-optimal [wp]	10.763	3.623	8.538	1.100	**15.290**	11.233	13.064	3.410
PSO-1 D-optimal [wp]	9.576	1.555	8.449	0.807	16.710	**9.253**	15.583	4.806
PSO-2 D-optimal [wp]	16.730	89.155	8.867	1.581	19.544	180.210	**11.133**	3.196
PSO-2 L-optimal [wp]	23.636	2994.772	8.785	1.607	25.984	5385.837	11.682	**2.890**
Reported D-optimal	8.654	0.680	8.551	0.682	31.193	190.555	31.090	191.143
Reported L-optimal	14.297	0.464	9.309	0.512	40.293	308.995	35.305	340.631
Uniform design	8.605	0.741	8.605	0.740	38.792	1403.310	38.792	1403.316

Note: wp = with penalty, w/op = without penalty

CHAPTER 7

Supplementary Materials

7.1 Supplementary Information for Chapter 2

7.1.1 Fitted trends of 19 genes in the WANG dataset

In this section, we present the other 19 exemplar genes (in addition to *MAOA*) in the WANG dataset Wang et al. (2020b) and their fitted trends by the scGTM, GAM, GLM, LOESS, switchDE, and ImpulseDE2. The interpretation of each figure is the same as the figure in the main text.

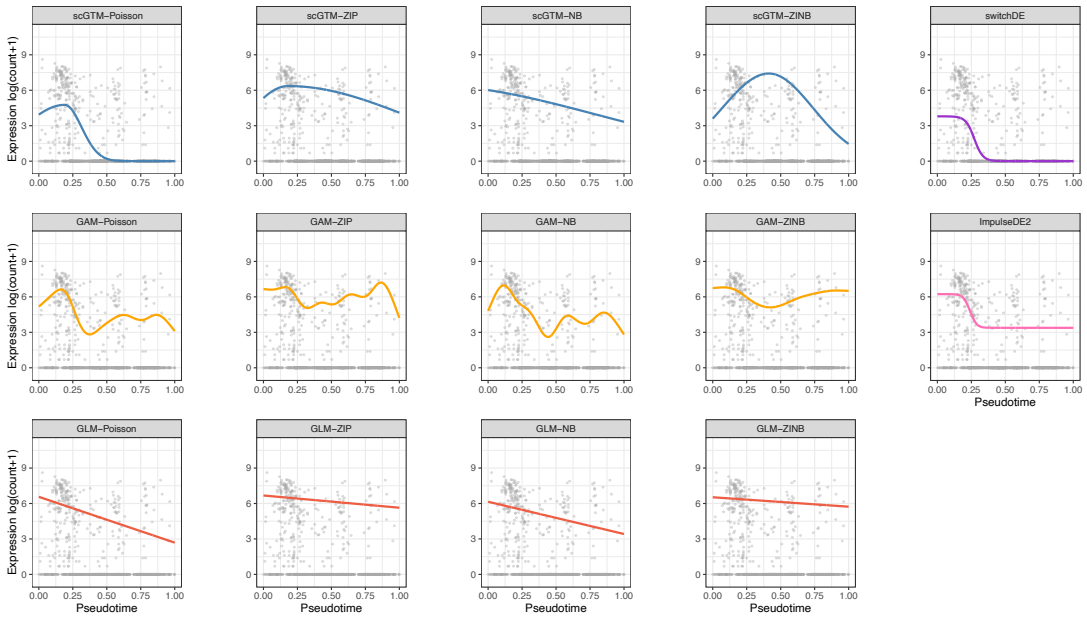


Figure 7.1: *P*LAU

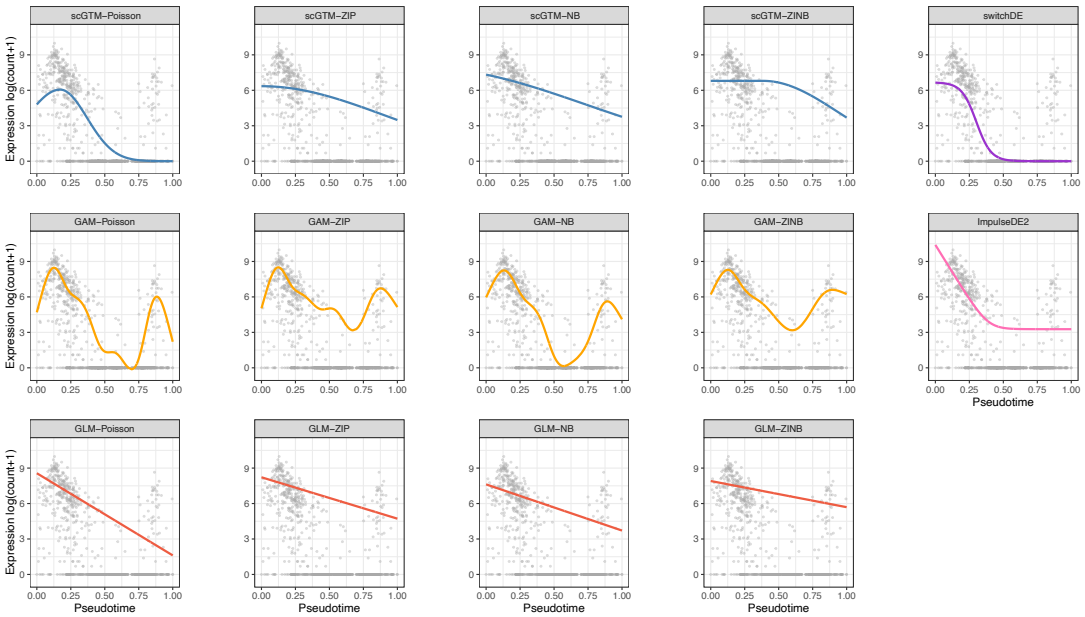


Figure 7.2: *M*MP7

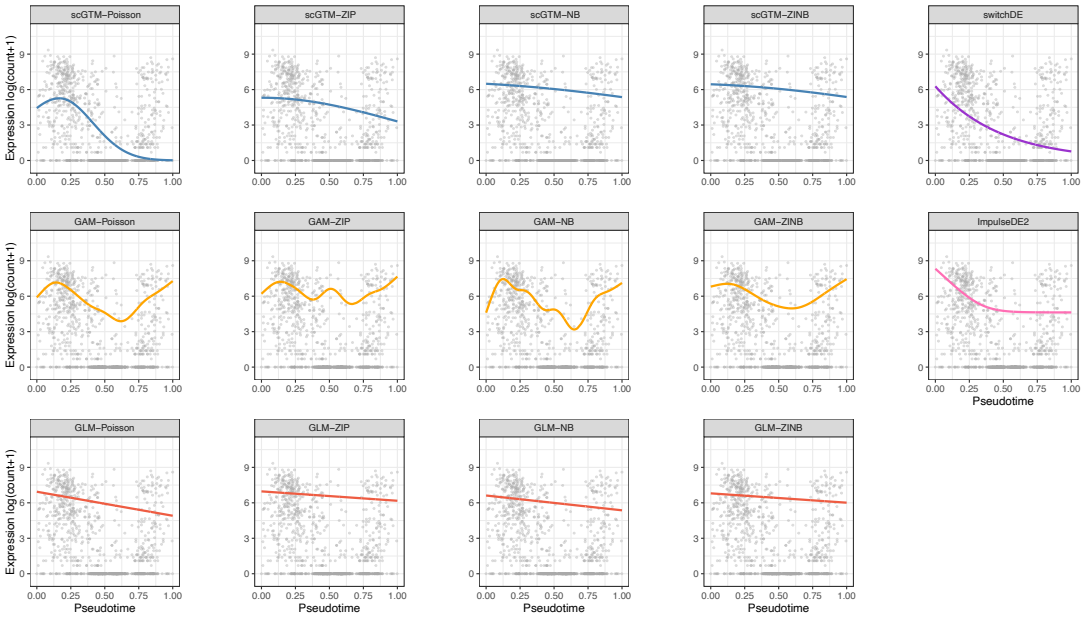


Figure 7.3: *T*HBS1

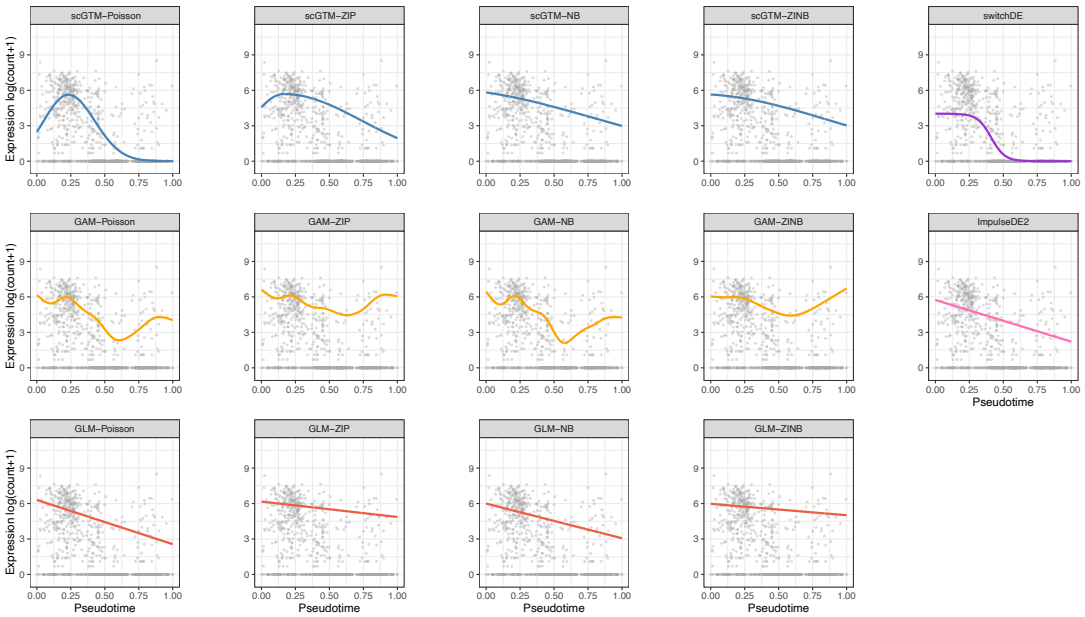


Figure 7.4: *C*ADM1

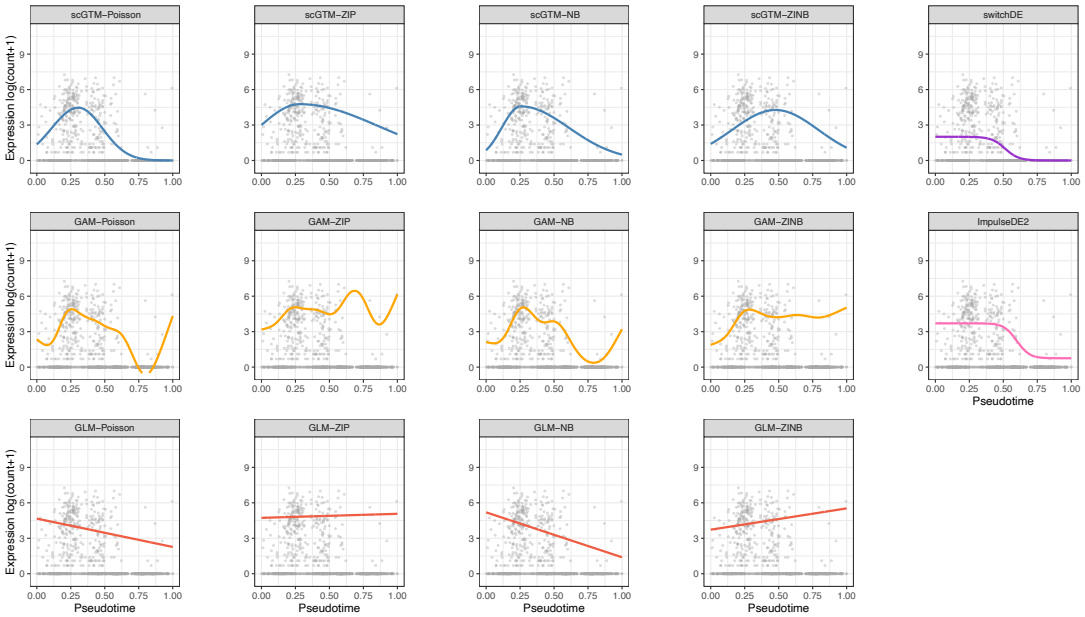


Figure 7.5: *N*PAS3

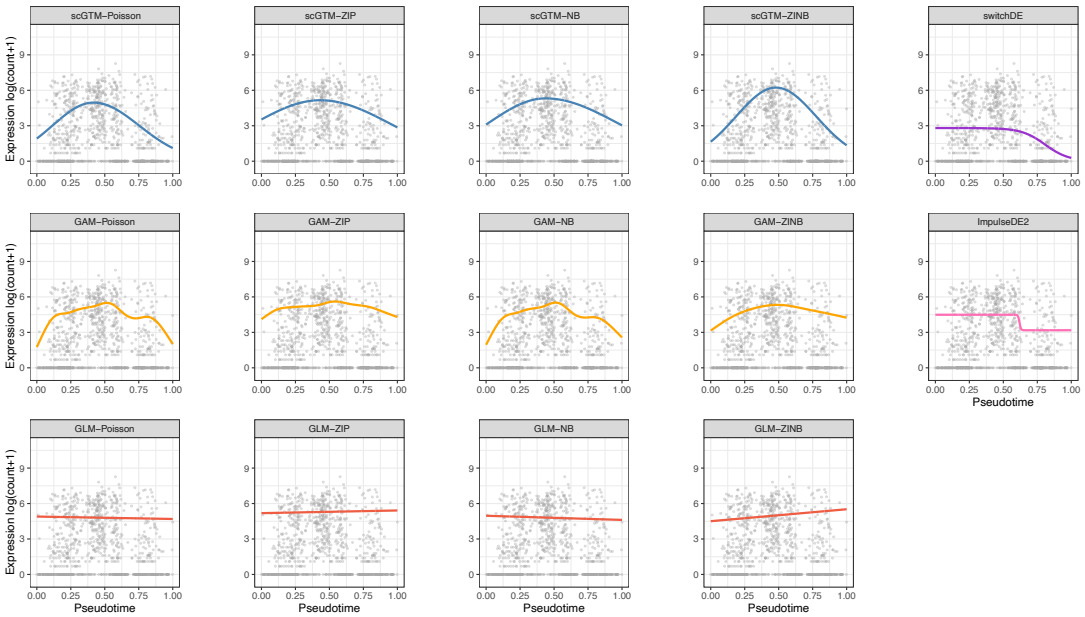


Figure 7.6: *A*TP1A1

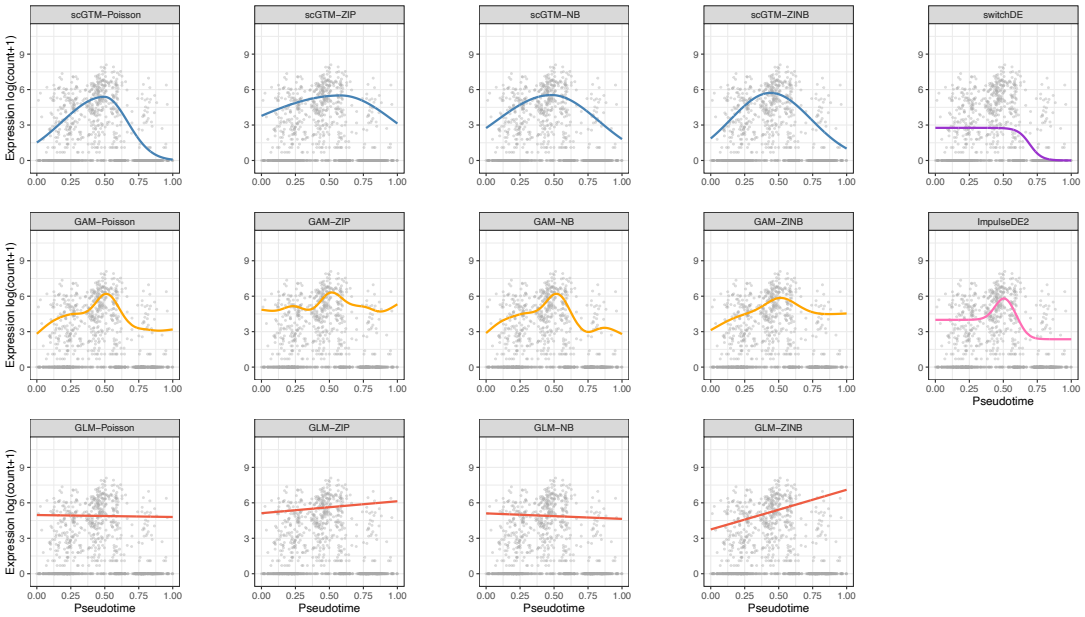


Figure 7.7: *A*NK3

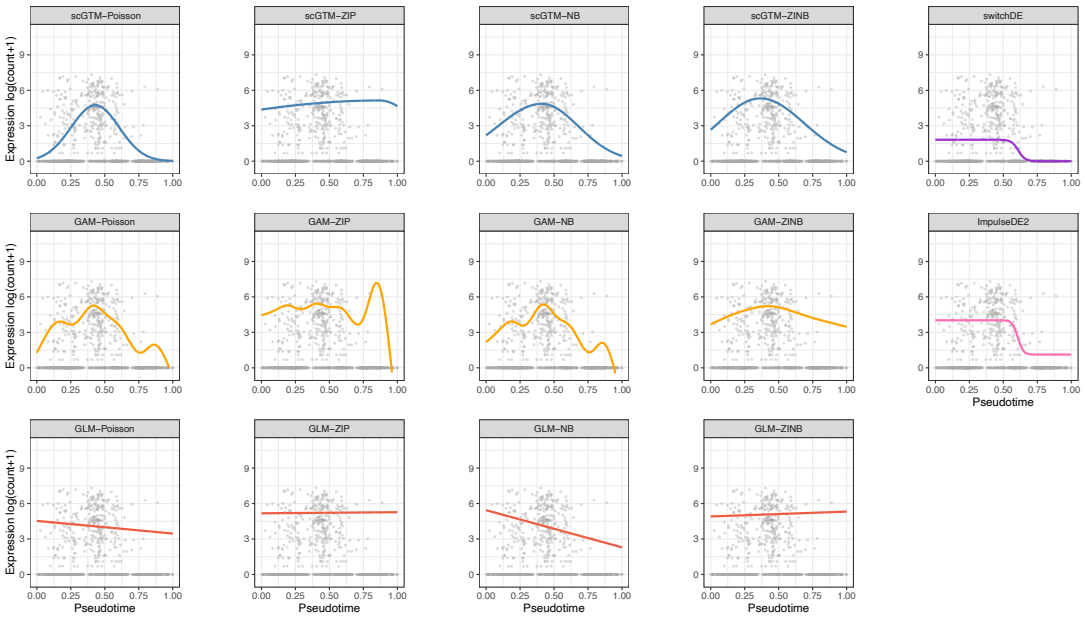


Figure 7.8: *A*LPL

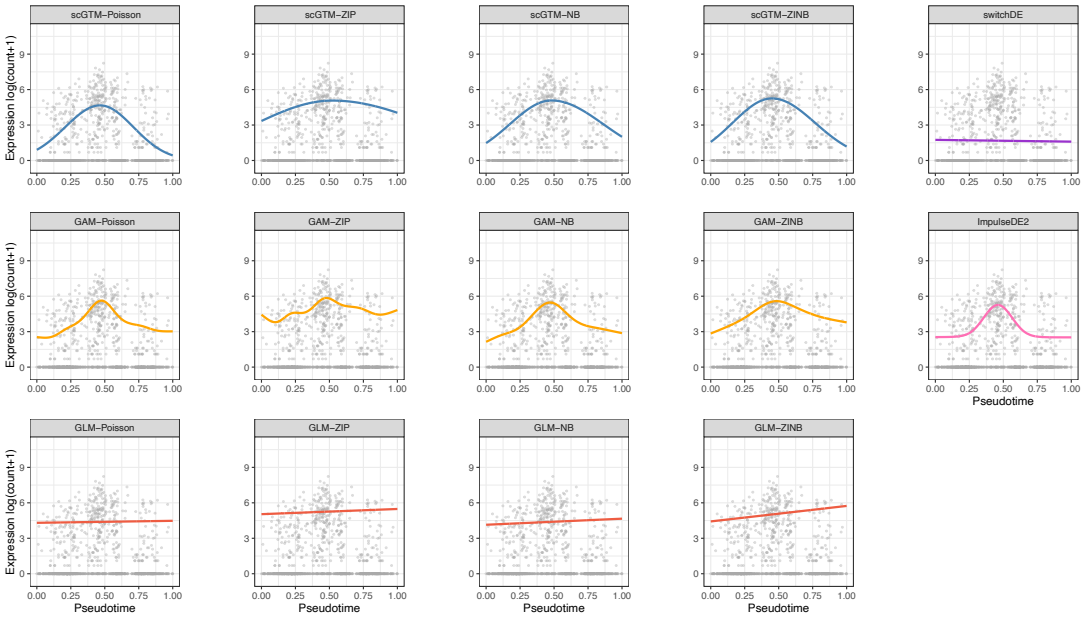


Figure 7.9: *T*RAK1

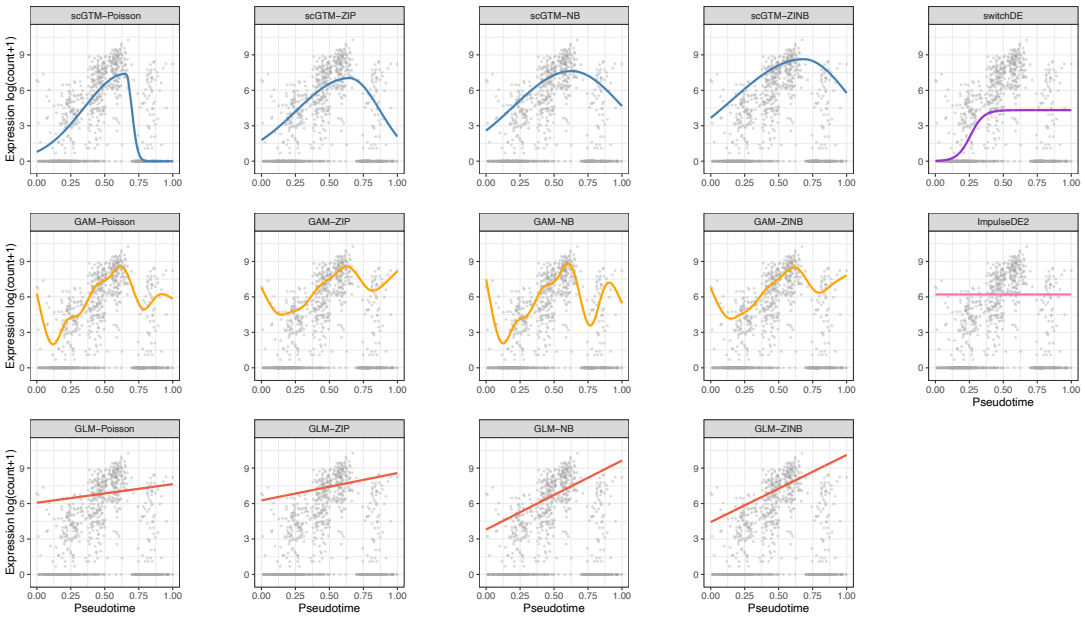


Figure 7.10: *S*CGB1D2

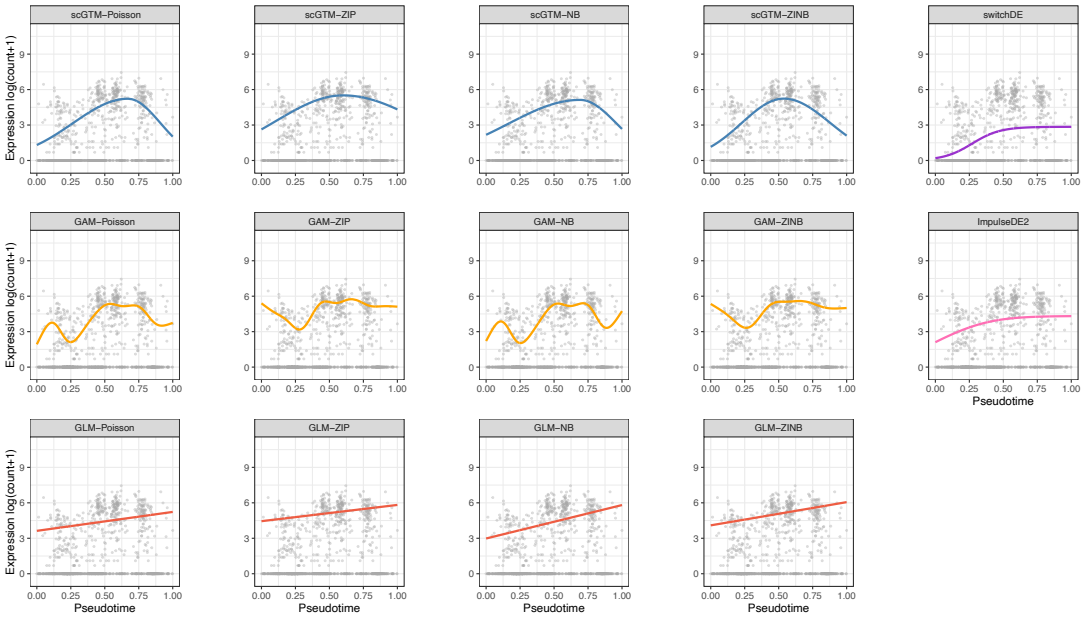


Figure 7.11: *M*T1F

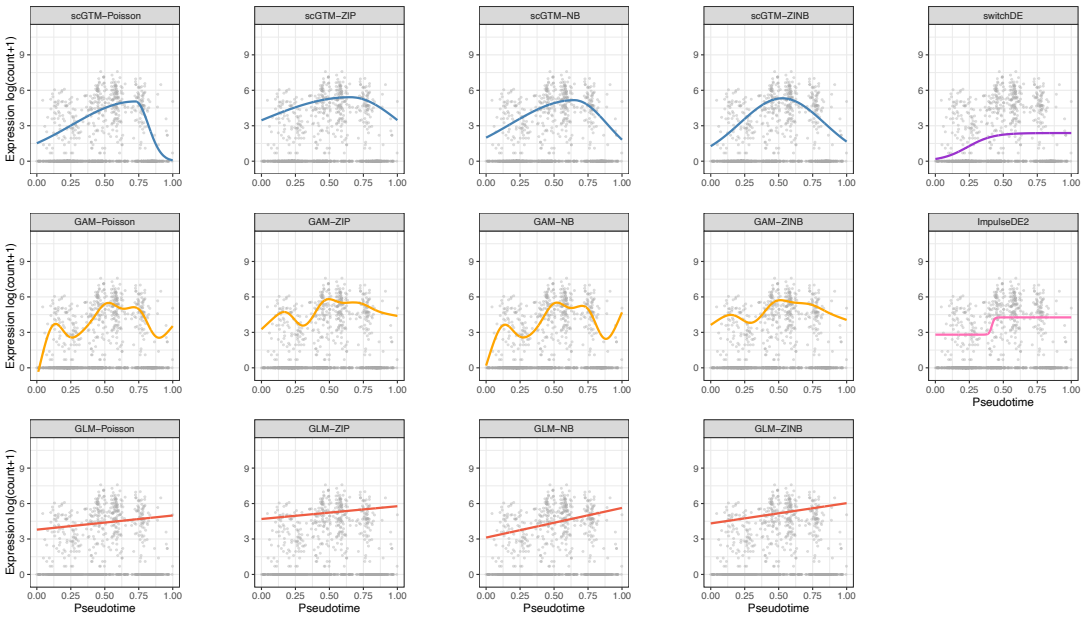


Figure 7.12: MT1X

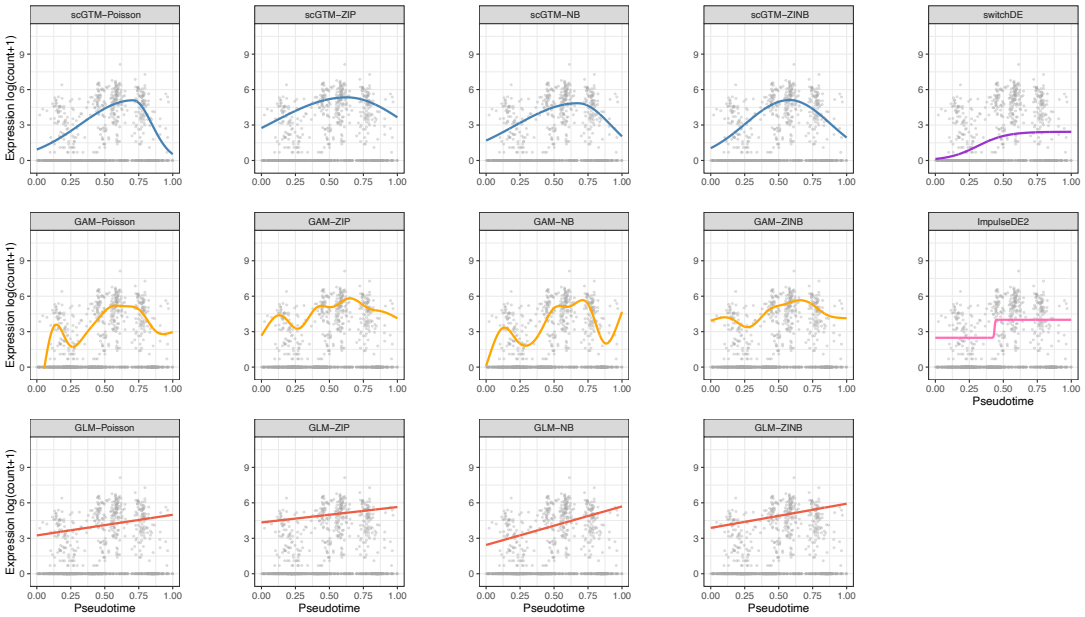


Figure 7.13: MT1E

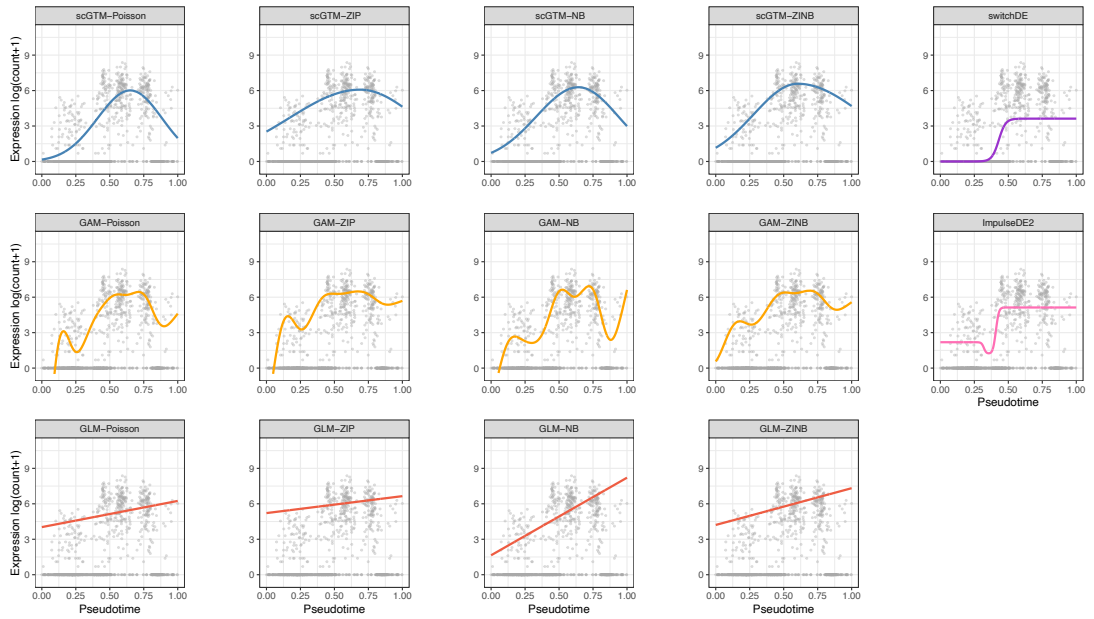


Figure 7.14: *M*T1G

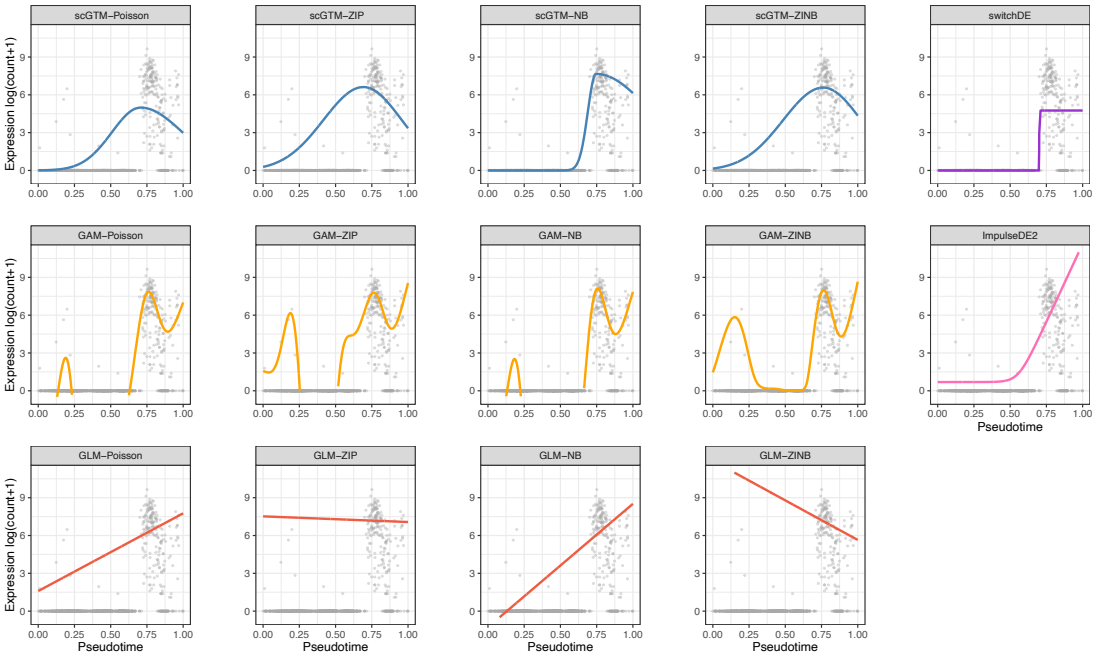


Figure 7.15: *C*XCL14

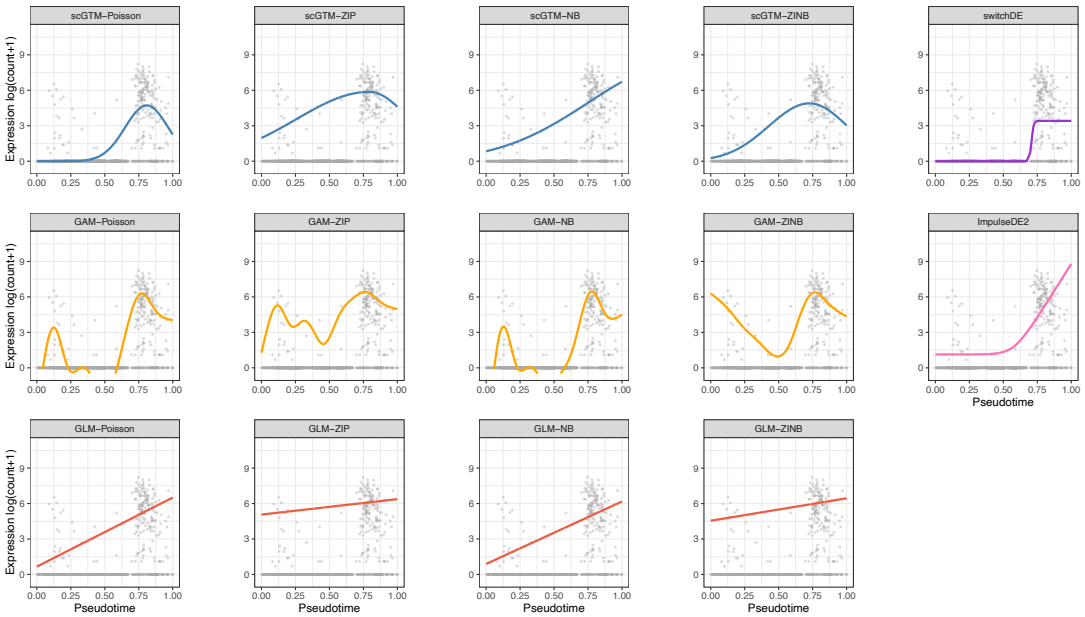


Figure 7.16: *D*PP4

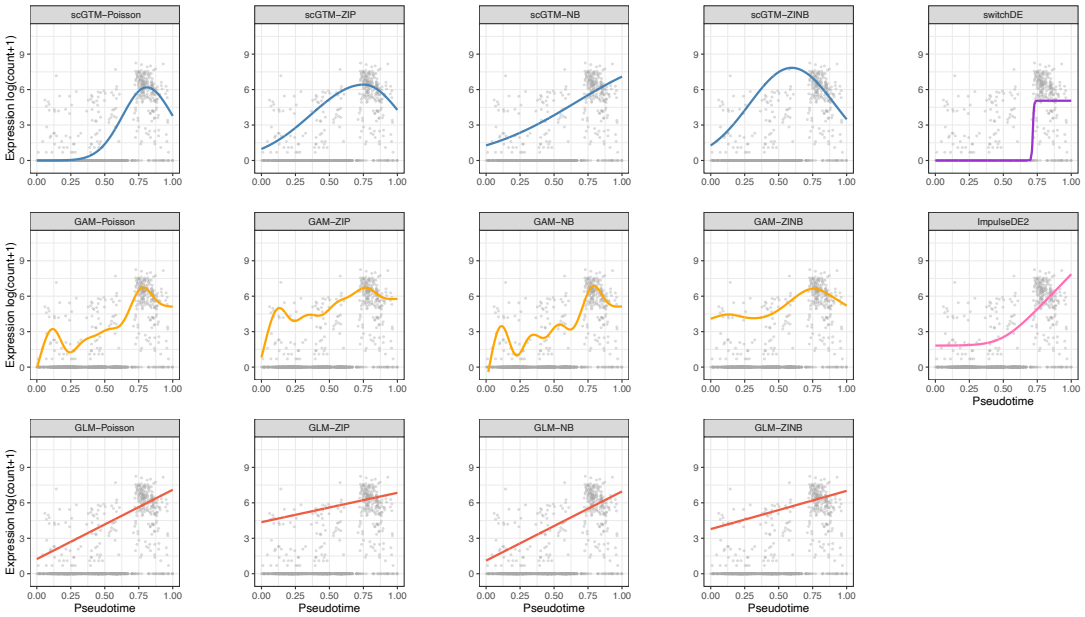


Figure 7.17: *N*UPR1

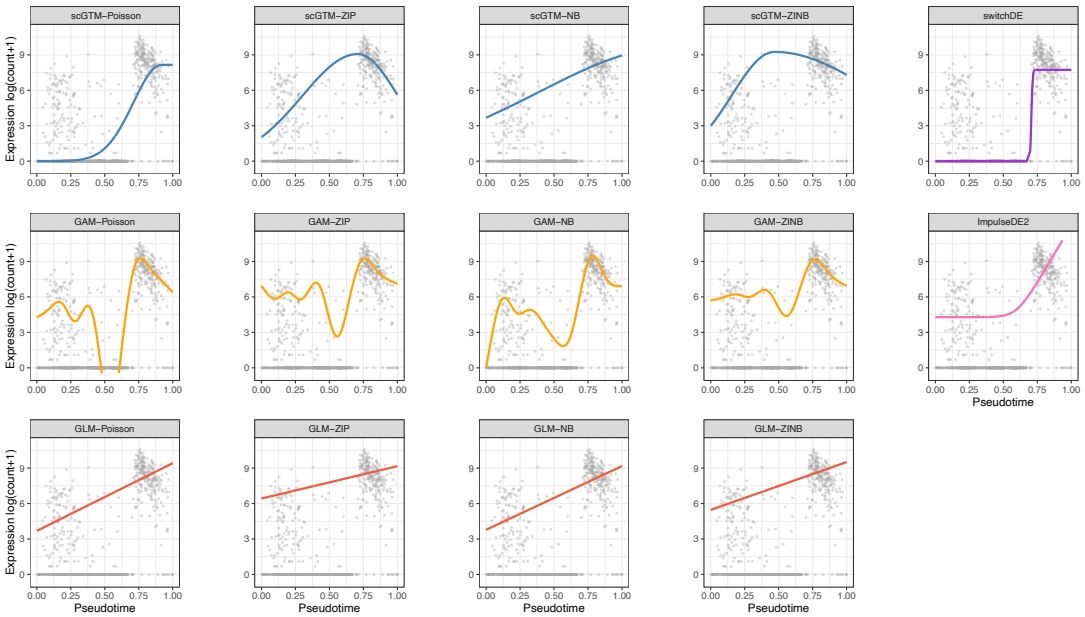


Figure 7.18: *GPX3*

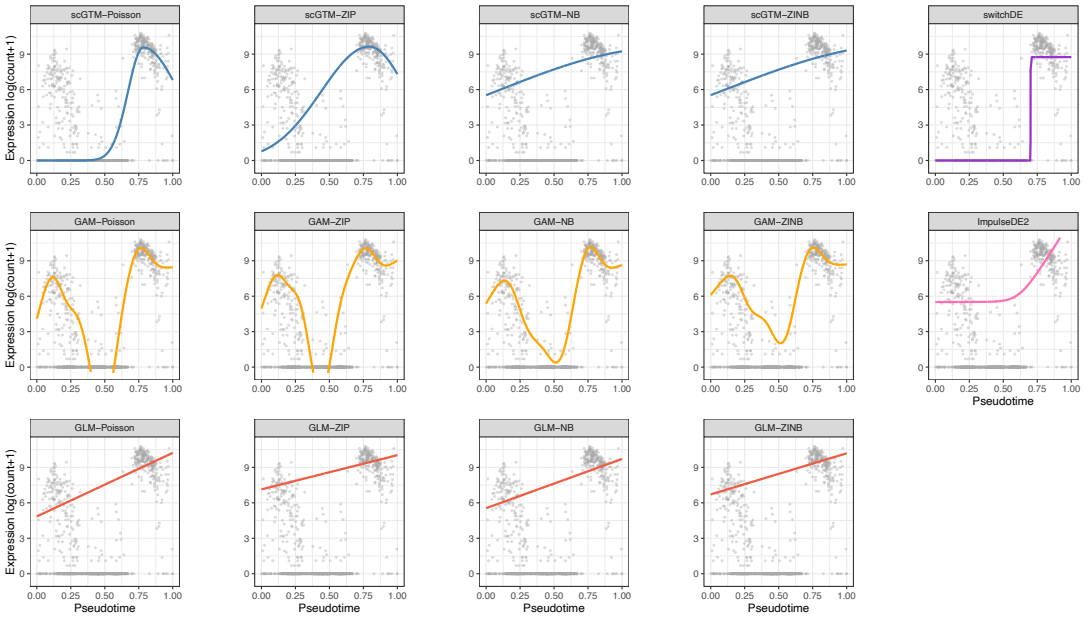


Figure 7.19: *PAEP*

7.1.2 Derivation of Fisher Information for Confidence Interval Construction

Below we derive the Fisher information for the key parameters

$$\Theta^* = (\mu_{\text{mag}}, k_1, k_2, t_0)^\mathsf{T}$$

of the scGTM with Poisson distribution (for demonstration purposes; used for the results in Section 2.3.3). Recall the scGTM for a hill-shaped gene is

$$\begin{aligned} Y_c &\sim \text{Poisson}(\tau_c)\,, \\ \log(\tau_c + 1) &= \begin{cases} \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right) & \text{if } t_c \le t_0 \\ \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}. \end{aligned} \tag{7.1.1}$$

The Fisher information for τ_c alone is $\mathcal{I}_{\text{Poi}}(\tau_c) = 1/\tau_c$, $c = 1, \ldots, C$, and every τ_c is related to Θ^* via (7.1.1). Then by the chain rule, the Fisher information for Θ^* is

$$\begin{aligned} \mathcal{I}_{\text{Poi}}(\Theta^*) &= \sum_{\{c:\ t_c \le t_0\}} \left(1 + \frac{1}{\exp(f_1) - 1}\right) \boldsymbol{x}_c \boldsymbol{x}_c^\mathsf{T} + \sum_{\{c:\ t_c > t_0\}} \left(1 + \frac{1}{\exp(f_2) - 1}\right) \boldsymbol{x}_c \boldsymbol{x}_c^\mathsf{T}\,, \\ \text{where} \quad \boldsymbol{x}_c &= \begin{cases} (f_1/\mu_{\text{mag}}, \quad (t_c - t_0)^2 f_1, \quad 0, \quad 2k_1(t_c - t_0) f_1)^\mathsf{T} & \text{if } t_c \le t_0 \\ (f_2/\mu_{\text{mag}}, \quad 0, \quad (t_c - t_0)^2 f_2, \quad 2k_2(t_c - t_0) f_2)^\mathsf{T} & \text{if } t_c > t_0 \end{cases}, \\ f_1 &= \mu_{\text{mag}} \exp\left(-k_1(t_c - t_0)^2\right), \\ f_2 &= \mu_{\text{mag}} \exp\left(-k_2(t_c - t_0)^2\right). \end{aligned} \tag{7.1.2}$$

Then the estimated asymptotic covariance of $\widehat{\Theta}^*$ is $\widehat{\mathcal{I}}_{\text{Poi}}^{-1}(\widehat{\Theta}^*)$.

7.1.3 Datasets, R packages, and R functions used in this paper

Table 7.1: Overview of datasets used.

Dataset	Sequencing protocol	Gene #	Cell #	Description	Ref
LPS	Fluidigm c1	4018	390	mouse bone-marrow-derived dendritic cells after stimulation with LPS	Shalek et al. (2014)
WANG	Fluidigm C1	22036	984	human unciliated epithelia cells during the menstrual cycle	Wang et al. (2020)
GYRUS	10x Genomics Chromium	2291	678	mouse developing dentate gyrus	Hochgerneret al.(2018)

Table 7.2: Overview of R packages and functions used for fitting GLMs and GAMs.

Model	R package	R function[1]	Parameter `family`[2]
GLM-Poisson	`stats`	`glm()`	`poisson()`
GLM-ZIP	`mgcv`	`gam()`	`ziP()`
GLM-NB	`mgcv`	`gam()`	`negbin()`
GLM-ZINB	`zigam`	`zinbgam()`	
GAM-Poisson	`mgcv`	`gam()`	`poisson()`
GAM-ZIP	`mgcv`	`gam()`	`ziP()`
GAM-NB	`mgcv`	`gam()`	`negbin()`
GAM-ZINB	`zigam`	`zinbgam()`	

[1] The R function to call in the R package.

[2] The `family` parameter (i.e., distribution) to specify in the R function. For example, `ziP()` refers to the ZIP distribution and can be specified in the `gam()` function in the `mgcv` package. There is no such parameter in the `zinbgam()` function in the `zigam` package.

7.1.4 Additional detail of analysis in the paper

7.1.4.1 Pseudotime inference

For the LPS and GYRUS datasets, we use the R package `slingshot` (version 2.0.0) to infer cell pseudotime. We use the top 2 principal components on the $\log(\text{count} + 1)$ matrix as the input of `slingshot`. For the WANG dataset, the pseudotime is provided by the authors of the original study Wang et al. (2020b).

7.1.4.2 GO analysis

We use the R package `clusterProfiler` (4.0.5) to perform GO analysis in Section 3.3. We set the p-value cutoff and q-value cutoff as 0.01 and 0.05, respectively. We set the ontology type as "BP (Biological Process)". We use the function `clusterProfiler::simplify` to further reduce the redundancy in GO terms.

7.1.4.3 Visualization

Most figures are made with the R package `ggplot2` (version 3.3.5). Figure 5 is generated by the R package `ComplexHeatmap` (version 2.9.3).

7.2 Supplementary Information for Chapter 3

7.2.1 Some Preliminaries in Riemann Geometry

In this section, we provide some preliminaries on differential geometry for more details, we refer to Do Carmo and Flaherty Francis (1992) and Tapp (2016). Suppose all points $x \in \mathbb{R}^d$ are column vectors and $\mathcal{S}_{++}^d$ represents the set of symmetric $d \times d$ positive definite matrices. The manifold is denoted $\mathcal{M}$, and its tagent space at $x \in \mathcal{M}$ is denoted $\mathcal{T}_x\mathcal{M}$.

Definition 7.2.1 (Riemannian manifold)**.** A Riemannian manifold is a smooth manifold $\mathcal{M}$, equipped with a positive definite Riemannian metric $M(x)$ $\forall x \in \mathcal{M}$, which is a smoothly varying inner product $\langle v, u \rangle_x = v^T M(x) u$ in the tangent space $\mathcal{T}_x\mathcal{M}$.

Definition 7.2.2 (Geodesic)**.** Let $\mathcal{M}$ be a Riemannian manifold. A geodesic curve $\gamma : [0, 1] \to \mathcal{M}$ is a length-minimizing smmooth curve connecting two given points $x, y \in \mathcal{M}$, i.e.,

$$\gamma(t) = \arg\min_c L(t, c, c') \tag{7.2.1}$$

$$L(t, c, c') = \int_0^1 \sqrt{c'(t)^T M\left(c(t)\right) c'(t)} dt \tag{7.2.2}$$

$$\gamma(0) = x \text{ and } \gamma(1) = y, \tag{7.2.3}$$

where L is a functional of t, c and c', $c'(t) \in \mathcal{T}_{c(t)}\mathcal{M}$ is the velocity of the curve c at t and M is the Riemannian metric tensor. Formula 7.2.3 is referred as *boundary value conditions*.

Theorem 7.2.1 (Euler-Lagrange equation for geodesic (Hauberg et al., 2012))**.** At minima of $L(t, c, c')$, the Euler-Lagrange equation must hold, i.e.,

$$\frac{\partial L}{\partial \gamma} = \frac{d}{dt} \frac{\partial L}{\partial \gamma'}.$$

Hence, geodesic curves embedded in $\mathbb{R}^d$ satisfy the following system of second-order ordinal differential equations (ODE):

$$M(\gamma(t))\gamma''(t) = -\frac{1}{2} \left(\frac{\partial \text{vec}[M(\gamma(t))]}{\partial \gamma(t)} \right)^T \left(\gamma'(t) \otimes \gamma'(t) \right), \tag{7.2.4}$$

where $\otimes$ denotes the Kronecker product and $\text{vec}(\cdot)$ stacks the columns of a matrix into a vector.

7.3 Supplementary Information for Chapter 4

7.3.1 Likelihood Derivation for the Semiparametric Bayesian Approach with Arbitrary Censoring

Recall that our model is

$$X_i = \alpha_1{}'G_i + \alpha_2{}'Z_i + \xi_{1i} \tag{7.3.1}$$

$$Y_i = \beta_1 X_i + \beta_2{}'Z_i + \xi_{2i} \tag{7.3.2}$$

where the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution with Dirichlet process (DP) prior:

$$(\xi_{1i}, \xi_{2i})' \sim N_2(\mu_i, \Sigma_i) \tag{7.3.3}$$

$$(\mu_i, \Sigma_i) \sim \text{i.i.d. } H \tag{7.3.4}$$

$$H \sim \text{DP}(\nu, H_0) \tag{7.3.5}$$

where DP refers to the Dirichlet process.

Recall that $\vec{C} = \{c_1, \ldots, c_n\}$ is the latent class (or "cluster") indicator of a subject, and $\theta_C = \{\theta_c : c \in c_1, \ldots, c_n\}$, where $\theta_c = \{\mu_{1c}, \mu_{2c}, \sigma^2_{1c}, \sigma^2_{2c}, \rho_c\}$, i.e. θ_C consists of all distinct values of $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma^2_{1i}, \sigma^2_{2i}, \rho_i\}$ and $\vec{C}$ is a vector of indicators that maps the individuals to the clusters. Note that the numbering of C can be arbitrary. We denote the total number of clusters as k. For the two-stage IV model (4.3.4)–(4.3.8), we denote the parameters as $\Theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_C, \vec{C})$. The observed data consists of $(\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$, where $\vec{L} = (L_1, ..., L_n)$, $\vec{R} = (R_1, ..., R_n)$, $\vec{\delta} = (\delta_1, ..., \delta_n)$, $\vec{X} = (X_1, ..., X_n)$, $\vec{Z} = (Z_1, ..., Z_n)$ and $\vec{G} = (G_1, ..., G_n)$. Due to censoring of the event times $\vec{Y}$, the likelihood function cannot be derived based on the bivariate distribution given by (4.3.6) directly. We construct the likelihood function by using the marginal likelihood of the first-stage model (4.3.4) and the conditional likelihood of the second-stage model (4.3.5). The likelihood function is (details

are given below):

$$\begin{aligned}\mathcal{L}(\Theta \mid \vec{L},\vec{R},\vec{\delta},\vec{X},\vec{Z},\vec{G}) &= P(\vec{X},\vec{Z},\vec{G} \mid \Theta)\cdot P(\vec{L},\vec{R},\vec{\delta} \mid \vec{X},\vec{Z},\vec{G},\Theta) \\ &= \prod_{i=1}^{n} f_{1i}(X_i,Z_i,G_i)\cdot[1-S_i(L_i \mid X_i,Z_i,G_i)]^{I\{\delta_i=1\}} \\ &\quad \cdot[S_i(L_i \mid X_i,Z_i,G_i)-S_i(R_i \mid X_i,Z_i,G_i)]^{I\{\delta_i=2\}} \\ &\quad \cdot S_i(R_i \mid X_i,Z_i,G_i)^{I\{\delta_i=3\}}\cdot f_{2i}(L_i \mid X_i,Z_i,G_i)^{I\{\delta_i=4\}}\end{aligned} \tag{7.3.6}$$

where

$$\begin{aligned} f_{1i}(X,Z,G) &= \phi\left(\frac{X-\mu_{1i}-\alpha_1{}'G-\alpha_2{}'Z}{\sqrt{\sigma_{1i}^2}}\right) \\ f_{2i}(Y \mid X,Z,G) &= \phi\left(\frac{Y-\mu_{2i}-\beta_1 X-\beta_2{}'Z-\frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X-\mu_{1i}-\alpha_1{}'G-\alpha_2{}'Z)}{\sqrt{(1-\rho_i{}^2)\sigma_{2i}^2}}\right) \\ S_i(Y \mid X,Z,G) &= 1-\Phi\left(\frac{Y-\mu_{2i}-\beta_1 X-\beta_2{}'Z-\frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X-\mu_{1i}-\alpha_1{}'G-\alpha_2{}'Z)}{\sqrt{(1-\rho_i{}^2)\sigma_{2i}^2}}\right)\end{aligned}$$

$i=1,\ldots n$. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. The detailed derivation of the likelihood is given in the appendix. Note that functions $f_{1i}(\cdot)$, $f_{2i}(\cdot)$ and $S_i(\cdot)$ are specifically for subject i, since the subjects have different distribution parameters given by the DP prior.

For time-to-event data subject to arbitrary-censoring, the likelihood of $(\vec{L},\vec{R},\vec{\delta})$ is:

$$\mathcal{L}=\prod_{i=1}^{n}(1-S(L_i))^{I\{\delta_i=1\}}\,(S(L_i)-S(R_i))^{I\{\delta_i=2\}}\,S(R_i)^{I\{\delta_i=3\}}f(L_i)^{I\{\delta_i=4\}} \tag{7.3.7}$$

where $S(y)=Pr(Y>y)$ is the survival distribution function and $f(y)=-\frac{d}{dy}S(y)$ is the survival time density function.

Based on the Two-stage IV model (4.3.4) and (4.3.5), the likelihood function of observing

$(\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ can be written as:

$$\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \quad = \quad P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \cdot P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) \tag{7.3.8}$$

where the first part is the marginal likelihood of the first-stage model (4.3.4), and the second part is the conditional likelihood of the second-stage model (4.3.5).

For the first part: From the bivariate normality assumption of ξ_1 and ξ_2 given by (4.3.6), the marginal distribution of ξ_{1i} is:

$$\xi_{1i} \sim N(\mu_{1i}, \sigma_{1i}^2)$$

which gives the marginal density function for the first-stage model (4.3.4):

$$f_{1i}(X, Z, G) \quad = \quad \phi\left(\frac{X - \mu_{1i} - \alpha_1{}'G - \alpha_2{}'Z}{\sqrt{\sigma_{1i}^2}}\right)$$

$i = 1, \ldots n$. Therefore, the likelihood of observing $\vec{X}$, $\vec{Z}$ and $\vec{G}$ is:

$$P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \quad = \quad \prod_{i=1}^{n} f_{1i}(X_i, Z_i, G_i) \tag{7.3.9}$$

For the second part: From the bivariate normality assumption of ξ_1 and ξ_2 given by (4.3.6), the conditional distribution of ξ_{2i} given ξ_{1i} is:

$$\xi_{2i} \mid \xi_{1i} \sim N\left(\mu_{2i} + \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(\xi_{1i} - \mu_{1i}),\ (1 - \rho_i^2)\sigma_{2i}^2\right)$$

$i = 1, \ldots n$. Since $\xi_{1i} = X_i - \alpha_1{}'G_i - \alpha_2{}'Z_i$ from the first-stage model (4.3.4), the conditional distribution becomes:

$$\xi_{2i} \mid X_i, Z_i, G_i \sim N\left(\mu_{2i} + \frac{\sigma_{2i}}{\sigma_{1i}}\rho_i(X_i - \mu_{1i} - \alpha_1{}'G_i - \alpha_2{}'Z_i),\ \ (1 - \rho_i^2)\sigma_{2i}^2\right)$$

Therefore, given $\vec{X}$, $\vec{Z}$, $\vec{G}$ and Θ, the second-stage model (4.3.5) has conditional survival function

$$\begin{aligned} S_i(T \mid X, Z, G) &= P(Y > T \mid X, Z, G) \\ &= P(\beta_1 X + \beta_2{}' Z + \xi_2 > T) \\ &= P(\xi_2 > T - \beta_1 X - \beta_2{}' Z) \\ &= 1 - \Phi\left(\frac{T - \mu_{2i} - \beta_1 X - \beta_2{}' Z - \frac{\sigma_{2i}}{\sigma_{1i}} \rho_i (X - \mu_{1i} - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho_i^2)\sigma_{2i}^2}} \right) \end{aligned}$$

and conditional density function

$$\begin{aligned} f_{2i}(T \mid X, Z, G) &= -\frac{\partial}{\partial t} S_i(T \mid X, Z, G) \\ &= \phi\left(\frac{T - \mu_{2i} - \beta_1 X - \beta_2{}' Z - \frac{\sigma_{2i}}{\sigma_{1i}} \rho_i (X - \mu_{1i} - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho_i^2)\sigma_{2i}^2}} \right) \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. From (7.3.7), we have

$$\begin{aligned} P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) = \prod_{i=1}^{n} & [1 - S_i(L_i \mid X_i, Z_i, G_i)]^{I\{\delta_i = 1\}} \\ & \cdot [S_i(L_i \mid X_i, Z_i, G_i) - S_i(R_i \mid X_i, Z_i, G_i)]^{I\{\delta_i = 2\}} \\ & \cdot S_i(R_i \mid X_i, Z_i, G_i)^{I\{\delta_i = 3\}} \cdot f_{2i}(L_i \mid X_i, Z_i, G_i)^{I\{\delta_i = 4\}} \end{aligned} \tag{7.3.10}$$

From (7.3.8), (7.3.9) and (7.3.10), we have the joint likelihood function (7.3.6).

7.3.2 Details on Pre-processing the UKB Data

7.3.2.1 Definition of the Outcome

Recall that a total of approximately 500,000 participants were included in this study and a total of approximately 26,000 participants (5.3%) had prevalent diabetes at the start of the study (age of diabetes diagnosis was recorded from self-reported data and where missing supplemented using Hospital Episode Statistics (HES) data). The goal is to quantify the time from diabetes diagnosis

to complications. Below we use CVD complication as an example. For each individual, we output (L_i, R_i), where $L_i \leq R_i$, such that $T_i \in (L_i, R_i]$. For the notations used for interval-censored data, see for example Sun (2006).

We start from the diabetes cohort curated before with $n \approx 26k$ (both T1D and T2D). For each individual in this cohort, DM diagnosis is from UKB assessment, admission data, or primary care data, along with the first known evidence date $T_{DM} = \min(\text{self\_DM}, \text{diagnosis\_DM})$. We determine the time-to-diabetes diagnosis (or time-to-DM), denoted as S_i, was subject to interval censoring, indicating that it falls within the interval $(U_i, V_i]$. Here, U_i represents the left endpoint of the interval, corresponding to the last recorded visit time before a negative diabetes diagnosis. Conversely, V_i signifies the right endpoint of the interval, denoting the first recorded visit time when a positive diabetes diagnosis was made. For determining U_i, the following variables played a crucial role:

- `last_a1c_lt48_pre_DMEHR`: This variable recorded the date of the last non-diabetic HbA1c level (<48mmol/mol) before the occurrence of the first diabetic event.
- `date_last_visit_negDMEHR`: It indicated the date of the last electronic health record (EHR) visit (either to a hospital or primary care provider) before the first diabetic event.
- `last_a1c_lt48_pre_higha1c`: This variable marked the date of the last non-diabetic HbA1c level (<48mmol/mol) recorded prior to the first occurrence of a high HbA1c level.

On the other hand, for defining V_i, the study relied on the following variables:

- `first_dm_hosp`: This variable indicated the date when diabetes was first recorded in the hospital records.
- `first_a1c_gt48`: It represented the date when the first diabetic HbA1c level (>48 mmol/mol) was recorded.
- `first_dm_EHR`: The date of the first recorded diagnosis of diabetes in the EHR, which was defined as the minimum value between `first_dm_hosp` and `first_dm_pc` (primary care).

For each individual in this cohort, we determine whether CVD complications occurred or not from UKB assessment(s), admission records, and primary care data. If occurred, we determine the date, T_{CVD}, and the censoring mechanism is determined below.

- If the DM first diagnosis date is unknown (DM diagnosis date before T_{DM}) and CVD complication has not occurred yet ($T_{CVD} = \infty$), then the time-to-event is lower bounded by ($L_i = \max$(last UKB assessment time of this individual, last admission record date, last primary care record date) - T_{DM}). Then the time-to-event is type I right-censored ($R_i = \infty$).

- If the DM first diagnosis date is known and CVD complication has not occurred yet ($T_{CVD} = \infty$), then the time-to-event is lower bounded by

$$L_i = \max(T_{Umax}, T_{Amax}, T_{PCmax}) - T_{DM}$$

  where the definitions of T_{Umax}, T_{Amax} and T_{DM} are given below. Then the time-to-event is type II right-censored ($R_i = \infty$).

- If the DM first diagnosis date is unknown (DM diagnosis date before T_{DM}) and CVD complication has already occurred, then the time-to-event is lower bounded by ($L_i = T_{CVD} - T_{DM}$) and upper bounded by ($R_i = T_{CVD} -$ birth date). In this case, the time-to-event is interval-censored.

- Finally, if the DM first diagnosis date is known (DM diagnosis date = T_{DM}) and CVD complication has already occurred, then the time-to-event is exactly observed ($L_i = R_i = T_{CVD} -$ DM diagnosis date).

To sum up, we have determined the following quantities:

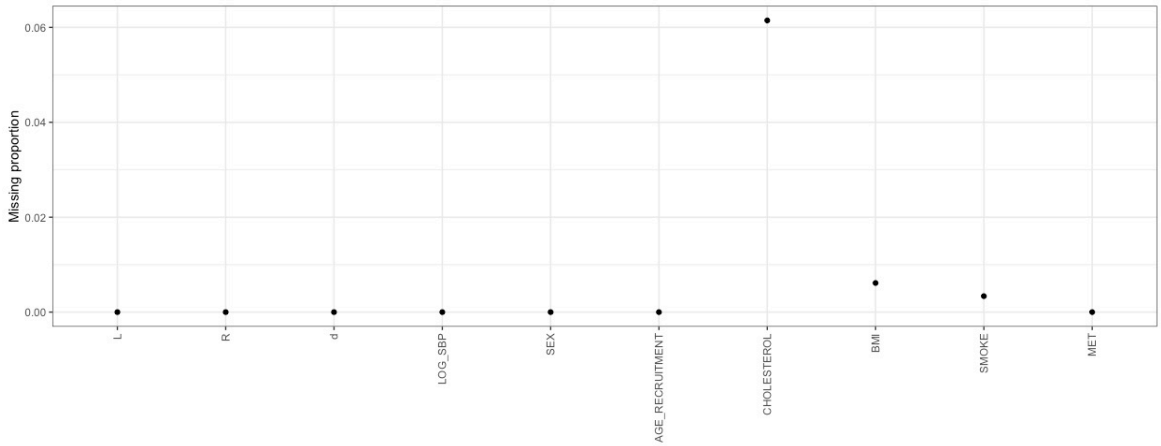
- T_{CVD}: date of CVD complications first occurrence.
- T_{DM}: date at diabetes first occurrence, i.e., min(self_DM , diagnosis_DM)
- T_{Umax}: date at the last UKB assessment visit.
- T_{Amin}: date of the first admission record date.
- T_{Amax}: date of the last admission record date.
- T_{PCmin}: date of the first primary care record date.
- T_{PCmax}: date of the last primary care record date.

- T_{max}: $\max(T_{Umax}, T_{Amax}, T_{PCmax})$.
- T_{min}: $\min(T_{DM}, T_{Amin}, T_{PCmin})$.

7.3.2.2 Missing Data

The missing information of the observed confounders in the UKB data (n=23801) is given in the following.

Figure 7.20: Missing proportion of the observed confounders in the UKB data



The missing entries of SNPs are filled using mean value of that particular SNP, and the missing entries in the observed confounder matrix are filled using Multiple Imputation by Chained Equations (MICE) (Azur et al., 2011) with fully conditional specification (FCS).

7.3.2.3 Selection of SNPs

A total of 269 SNPs were chosen for IV analysis representing independent loci previously shown to be associated with mean SBP levels (Ko et al., 2022). These index variants were identified using PLINK 1.9; lead variants were chosen greedily starting with the SNPs with lowest p-value among those SNPs having p-value $< 5 \times 10^{-8}$. Sites that were < 250 kb away from an index variant and $r2 > 0:5$ with the index variant were assigned to that index variant's clump.

7.3.3 Slightly Informative Priors for Male and Female Cohorts of UKB Data

The slightly informative priors are estimated from 5% of the samples using the proposed DPMIV method.

Table 7.3: Priors of parameters in DPMIV for the analysis of UKB data

Parameters	Distribution
α_1	$N(0.015, 0.087)$
α_2	$N(0.017, 0.076)$
β_1	$N(-0.476, 1.401)$
β_2	$N(-0.025, 0.344)$
μ_{1c}	$N(4.804, 0.853)$
μ_{2c}	$N(4.583, 3.334)$
σ_{1c}	Inv-Gamma$(0.025, 1)$
σ_{2c}	Inv-Gamma$(0.025, 1)$
ρ	Uniform$(-1, 1)$
ν	$\underline{\nu} = 0.01, \overline{\nu} = 4.8, \omega = 2.4$

We use 5% of the samples as the training data to get posterior distributions of parameters and use them as priors for the the remaining 95% interval-censored data. For example, the prior of β_1 has a normal distribution with mean -0.476 and standard deviation 1.401. The prior of the first element of β_2 has a normal distribution with mean -0.025 and standard deviation 0.344.

7.3.4 The MCMC Algorithm for DPMIV Based on Neal's No-gaps

We follow notations defined in Sections 4.3.1 and 4.3.2. We develop an MCMC procedure to generate posterior samples of α_1, α_2, β_1, β_2, $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i\}$, $\vec{C} = \{c_1, \ldots, c_n\}$, $\theta_c = \{\mu_{1c}, \mu_{2c}, \sigma_{1c}^2, \sigma_{2c}^2, \rho_c\}$, and ν, where $i = 1, \ldots, n$, cluster indicators $\{c_1, \ldots, c_n\}$ are coded as values in $\{1, 2, \ldots, k\}$, k is the total number of clusters, $c = 1, \ldots, k$. In each iteration, we generate a new sample for each of the parameters listed above using the following algorithm.

- For α_1: We update vector α_1 by updating its elements one-by-one using the random walk Metropolis-Hasting (M-H) algorithm described in Section A2. For the j-th element α_{1j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ with large variance (e.g. $\mu_p = 0$, $\varsigma_p^2 = 100^2$), and a uniform proposal distribution Unif$(\alpha_{1j} - \omega_p, \alpha_{1j} + \omega_p)$ for the random

walk, where ω_p is a positive number chosen to give an appropriate acceptance rate (e.g. $20\% \sim 40\%$). A candidate sample α_{1j}^* is generated from the proposal distribution, and accepted as the current state of α_{1j} with probability $a(\alpha_{1j}, \alpha_{1j}^*)$. The log of acceptance probability is given by:

$$\log(a(\alpha_{1j}, \alpha_{1j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\alpha_{1j} - \mu_p)^2 - (\alpha_{1j}^* - \mu_p)^2\right)$$

where Θ^* is Θ with α_{1j} replaced by α_{1j}^*, $\ell(\cdot)$ is the log-likelihood function given by $\ell(\Theta) = log(\mathcal{L}(\Theta))$, and $\mathcal{L}(\Theta)$ is the likelihood function given by equation (7.3.6).

- For α_2: Similar procedure as for α_1 is used. Elements in vector α_2 is updated one-by-one using the M-H sampling algorithm. For the j-th element α_{2j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform proposal distribution Unif($\alpha_{2j} - \omega_p$, $\alpha_{2j} + \omega_p$) with appropriate width ω_p. A candidate sample α_{2j}^* is generated from the proposal distribution, and accepted as the current state of α_{2j} with probability $a(\alpha_{2j}, \alpha_{2j}^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\alpha_{2j}, \alpha_{2j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\alpha_{2j} - \mu_p)^2 - (\alpha_{2j}^* - \mu_p)^2\right)$$

where Θ^* is Θ with α_{2j} replaced by α_{2j}^*.

- For β_1: We update β_1 using the M-H sampling algorithm, similar to the procedure for α_{1j}. We propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform proposal distribution Unif($\beta_1 - \omega_p$, $\beta_1 + \omega_p$) with appropriate width ω_p. A candidate sample β_1^* is generated from the proposal distribution, and accepted as the current state of β_1 with probability $a(\beta_1, \beta_1^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\beta_1, \beta_1^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\beta_1 - \mu_p)^2 - (\beta_1^* - \mu_p)^2\right)$$

where Θ^* is Θ with β_1 replaced by β_1^*.

- For β_2: Similar procedure as for α_1 is used. Elements in vector β_2 is updated one-by-one using the M-H sampling algorithm. For the j-th element β_{2j}, we propose to use a vague normal prior distribution $N(\mu_p, \varsigma_p^2)$ and a uniform proposal distribution Unif($\beta_{2j} - \omega_p$, $\beta_{2j} + \omega_p$) with

appropriate width ω_p. A candidate sample β_{2j}^* is generated from the proposal distribution, and accepted as the current state of β_{2j} with probability $a(\beta_{2j}, \beta_{2j}^*)$. Similarly, the log of acceptance probability is given by:

$$\log(a(\beta_{2j}, \beta_{2j}^*)) = \ell(\Theta^*) - \ell(\Theta) + \frac{1}{2\varsigma_p^2}\left((\beta_{2j} - \mu_p)^2 - (\beta_{2j}^* - \mu_p)^2\right)$$

where Θ^* is Θ with β_{2j} replaced by β_{2j}^*.

- For $\vec{C}$: We update the cluster indicators $c_1, \ldots, c_n$, one-by-one. Let m be a prefixed number of auxiliary parameters. We use $m = 10$ in our simulation studies and real data examples in Chapter 4. For the base distribution H_0 of the Dirichlet process prior, we propose to use independent slightly informative priors $H_0 = \pi(\mu_{1i})\pi(\mu_{2i})\pi(\sigma_{1i}^2)\pi(\sigma_{2i}^2)\pi(\rho_i)$. Here 'slightly informative' means that the chosen priors spread out and properly cover the reasonable values for the parameters. We propose to use normal distributions for $\pi(\mu_{1i})$ and $\pi(\mu_{2i})$, inverse-gamma distributions for $\pi(\sigma_{1i}^2)$ and $\pi(\sigma_{2i}^2)$, and a uniform distribution Unif$(-1, 1)$ for $\pi(\rho_i)$. The following procedure is used to update cluster indicator c_i:

  1. For subject i: Let k^- be the number of distinct c_j for $j \neq i$. Let $h = k^- + m$, and $c^{-i} = \{c_j : j \neq i\}$.

  2. If $c_i = c_j$ for some $j \neq i$ (i.e. subject i is not a 'singleton'), draw m samples independently from H_0 as $\{\theta_{k^-+1}, ..., \theta_h\}$ (i.e. draw m independent samples from $\pi(\mu_{1i})$ as $\{\mu_{1,k^-+1}, ..., \mu_{1h}\}$, draw m independent samples from $\pi(\mu_{2i})$ as $\{\mu_{2,k^-+1}, ..., \mu_{2h}\}$, draw m independent samples from $\pi(\sigma_{1i}^2)$ as $\{\sigma_{1,k^-+1}^2, ..., \sigma_{1h}^2\}$, draw m independent samples from $\pi(\sigma_{2i}^2)$ as $\{\sigma_{2,k^-+1}^2, ..., \sigma_{2h}^2\}$, draw m independent samples from $\pi(\rho_i)$ as $\{\rho_k^- + 1, ..., \rho_h\}$).

  3. If $c_i \neq c_j$ for all $j \neq i$ (i.e. subject i is a 'singleton'), relabel these c_j with values in $\{1, ..., k^-\}$, and label c_i as $k^- + 1$. Draw $m - 1$ samples independently from H_0 as $\{\theta_{k^-+2}, ..., \theta_h\}$.

  4. Draw a new value for c_i from $\{1, ..., h\}$ with probabilities:
  $$P(c_i = c | c^{-i}, \theta_1, \ldots, \theta_h) = \begin{cases} b \cdot n_{-i,c} \cdot L_i(\theta_c) \ , & 1 \leq c \leq k^- \\ b \cdot \frac{\nu}{m} \cdot L_i(\theta_c) \ , & k^- \leq c \leq h \end{cases}$$
  where $n_{-i,c}$ is the number of subjects that are in $\{j : j \neq i, c_j = c\}$, and $L_i(\theta_c)$ is the

likelihood of subject i with parameter θ_c:

$$L_i(\theta_c) = \mathcal{L}(\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c \mid L_i, R_i, \delta_i, X_i, Z_i, G_i)$$

and b is a normalizing constant.

5. Update the total number of clusters k accordingly.

- For θ_c: We update cluster parameters θ_c, $c = 1, \ldots, k$, one-by-one. For each $c \in \{1, \ldots, k\}$, we update $\{\mu_{1c}, \mu_{2c}, \sigma^2_{1c}, \sigma^2_{2c}, \rho_c\}$ one-by-one, while keeping the other parameters at their current state, using the M-H sampling algorithm. We propose to use independent vague priors for the parameters: a normal distribution $N(\mu, \varsigma^2)$ with large variance (e.g. $\mu = 0$, $\varsigma^2 = 100^2$) for μ_{1c} and μ_{2c}; an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter and small scale parameter (e.g. $\gamma_1 = \gamma_2 = 0.001$) for σ^2_{1c} and σ^2_{2c}; and a uniform distribution Unif$(-1, 1)$ for ρ_c.

  - For μ_{1c}: We use a uniform proposal distribution Unif$(\mu_{1c}-\omega_p,\ \mu_{1c}+\omega_p)$ with appropriate width ω_p. A candidate sample μ^*_{1c} is generated from the proposal distribution, and accepted as the current state of μ_{1c} with probability $a(\mu_{1c}, \mu^*_{1c})$. The log of acceptance probability is given by:

$$\log(a(\mu_{1c}, \mu^*_{1c})) = \ell_c(\Theta^*) - \ell_c(\Theta) + \frac{1}{2\varsigma_p^2}\left((\mu_{1c} - \mu_p)^2 - (\mu^*_{1c} - \mu_p)^2\right)$$

    where Θ^* is Θ with μ_{1c} replaced by μ^*_{1c}, and $\ell_c(\cdot)$ is the log-likelihood function with subjects in cluster c only,

$$\ell_c(\Theta) = log(\mathcal{L}(\Theta \mid L_i, R_i, \delta_i, X_i, Z_i, G_i, i \in \{j : c_j = c\}))$$

  - For σ^2_{1c}: We use a uniform proposal distribution Unif$(\max(\sigma^2_{1c} - \omega_p, 0), \sigma^2_{1c} + \omega_p)$ with appropriate width ω_p. A candidate sample ${\sigma^2_{1c}}^*$ is generated from the proposal distribution, and accepted as the current state of σ^2_{1c} with probability $a(\sigma^2_{1c}, {\sigma^2_{1c}}^*)$. The log

of acceptance probability is given by:

$$\begin{aligned}\log(a(\sigma_{1c}^2, {\sigma_{1c}^2}^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) \\ &+ \log(\min(2\omega_p, \sigma_{1c}^2 + \omega_p)) - \log(\min(2\omega_p, {\sigma_{1c}^2}^* + \omega_p)) \\ &+ (\gamma_1 + 1)\left[\log(\sigma_{1c}^2) - \log({\sigma_{1c}^2}^*)\right] + \gamma_2\left(\frac{1}{\sigma_{1c}^2} - \frac{1}{{\sigma_{1c}^2}^*}\right)\end{aligned}$$

where Θ^* is Θ with σ_{1c}^2 replaced by ${\sigma_{1c}^2}^*$.

– For μ_{2c}: Similar to μ_{1c}, we use a uniform proposal distribution $\text{Unif}(\mu_{2c} - \omega_p,\ \mu_{2c} + \omega_p)$ with appropriate width ω_p. A candidate sample μ_{2c}^* is generated from the proposal distribution, and accepted as the current state of μ_{2c} with probability $a(\mu_{2c}, \mu_{2c}^*)$. The log of acceptance probability is given by:

$$\log(a(\mu_{2c}, \mu_{2c}^*)) = \ell_c(\Theta^*) - \ell_c(\Theta) + \frac{1}{2\varsigma_p^2}\left((\mu_{2c} - \mu_p)^2 - (\mu_{2c}^* - \mu_p)^2\right)$$

where Θ^* is Θ with μ_{2c} replaced by μ_{2c}^*.

– For σ_{2c}^2: Similar to σ_{1c}^2, we use a uniform proposal distribution $\text{Unif}(\max(\sigma_{2c}^2 - \omega_p, 0), \sigma_{2c}^2 + \omega_p)$ with appropriate width ω_p. A candidate sample ${\sigma_{2c}^2}^*$ is generated from the proposal distribution, and accepted as the current state of σ_{2c}^2 with probability $a(\sigma_{2c}^2, {\sigma_{2c}^2}^*)$. The log of acceptance probability is given by:

$$\begin{aligned}\log(a(\sigma_{2c}^2, {\sigma_{2c}^2}^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) \\ &+ \log(\min(2\omega_p, \sigma_{2c}^2 + \omega_p)) - \log(\min(2\omega_p, {\sigma_{2c}^2}^* + \omega_p)) \\ &+ (\gamma_1 + 1)\left[\log(\sigma_{2c}^2) - \log({\sigma_{2c}^2}^*)\right] + \gamma_2\left(\frac{1}{\sigma_{2c}^2} - \frac{1}{{\sigma_{2c}^2}^*}\right)\end{aligned}$$

where Θ^* is Θ with σ_{2c}^2 replaced by ${\sigma_{2c}^2}^*$.

– For ρ_c: We use a uniform proposal distribution $\text{Unif}(\max(\rho_c - \omega_p, -1), \min(\rho_c + \omega_p, 1))$ with appropriate width ω_p. A candidate sample ρ_c^* is generated from the proposal distribution, and accepted as the current state of ρ_c with probability $a(\rho_c, \rho_c^*)$. The log

of acceptance probability is given by:

$$\begin{aligned}\log(a(\rho_c, \rho_c^*)) =& \ell_c(\Theta^*) - \ell_c(\Theta) + \log(\min(\rho_c + \omega_p, 1)) - \log(\max(\rho_c - \omega_p, -1)) \\ & - \log(\min(\rho_c^* + \omega_p, 1)) + \log(\max(\rho_c^* - \omega_p, -1))\end{aligned}$$

where Θ^* is Θ with ρ_c replaced by ρ_c^*.

- For θ_i: After updating $\vec{C}$ and θ_c, $c = 1, \ldots, k$, the individual parameters $\theta_i = \{\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2, \rho_i\}$, $i = 1, \ldots, n$, can be derived.

- For ν: We update the strength parameter ν of the Dirichlet process prior using the M-H sampling algorithm. We propose to use prior distribution

$$P(\nu) \propto \left(\frac{\overline{\nu} - \nu}{\overline{\nu} - \underline{\nu}}\right)^{\omega} \cdot I(\underline{\nu} < \nu < \overline{\nu})$$

where $\underline{\nu}$ and $\overline{\nu}$ are chosen to give small k (e.g. mode of $k = 1$) and large k (e.g. mode of $k = 15$), respectively. ω is a constant chosen to control the shape of the prior (e.g. $\omega = 0.8$). We use a uniform proposal distribution $\text{Unif}(\max(\underline{\nu}, \nu - \omega_p), \min(\overline{\nu}, \nu + \omega_p))$. A candidate sample ν^* is generated from the proposal distribution, and accepted as the current state of ν with probability $a(\nu, \nu^*)$. The log of acceptance probability is given by:

$$\begin{aligned}\log(a(\nu, \nu^*)) =& \log(\min(\overline{\nu}, \nu + \omega_p) - \max(\underline{\nu}, \nu - \omega_p)) \\ & - \log(\min(\overline{\nu}, \nu^* + \omega_p) - \max(\underline{\nu}, \nu^* - \omega_p)) \\ & + \omega_p \left[\log(\overline{\nu} - \nu^*) - \log(\overline{\nu} - \nu)\right] \\ & + k(\log\nu^* - \log\nu) + \log(\Gamma(\nu^*)) - \log(\Gamma(\nu^* + n)) \\ & - \log(\Gamma(\nu)) + \log(\Gamma(\nu + n))\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function.

7.3.5 Extension of Li-Lu's PBIV to Arbitrary Censoring

In this subsection, we briefly describe how to extend the parametric Bayesian method in Li and Lu (2015) from right-censored data only to all four types of censoring.

Following the same notation, the likelihood of observing $(\vec{L}, \vec{R}, \vec{\delta})$ is:

$$\mathcal{L} = \prod_{i=1}^{n} (1 - S(L_i))^{I\{\delta_i=1\}} (S(L_i) - S(R_i))^{I\{\delta_i=2\}} S(R_i)^{I\{\delta_i=3\}} f(L_i)^{I\{\delta_i=4\}} \tag{7.3.11}$$

where $S(y) = Pr(Y > y)$ is the survival distribution function and $f(y) = -\frac{d}{dy}S(y)$ is the survival time density function.

The likelihood function of observing $(\vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ can be written as:

$$\mathcal{L}(\Theta \mid \vec{L}, \vec{R}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \;=\; P(\vec{X}, \vec{Z}, \vec{G} \mid \Theta) \cdot P(\vec{L}, \vec{R}, \vec{\delta} \mid \vec{X}, \vec{Z}, \vec{G}, \Theta) \tag{7.3.12}$$

where the first part is the marginal likelihood of the first-stage model, and the second part is the conditional likelihood of the second-stage model. For the first part: from the bivariate normality assumption of ξ_1 and ξ_2, the conditional distribution of ξ_{2i} given ξ_{1i} is:

$$\xi_{2i} \mid \xi_{1i} \sim N\left(\frac{\sigma_2}{\sigma_1}\rho\,\xi_{1i},\, (1-\rho^2)\sigma_2^2\right)$$

$i = 1, \ldots n$. Since $\xi_{1i} = X_i - \alpha_0 - \alpha_1{}'G_i - \alpha_2{}'Z_i$ from the first-stage model, the conditional distribution becomes:

$$\xi_{2i} \mid X_i, Z_i, G_i \sim N\left(\frac{\sigma_2}{\sigma_1}\rho(X_i - \alpha_0 - \alpha_1{}'G_i - \alpha_2{}'Z_i),\;\; (1-\rho^2)\sigma_2^2\right)$$

Therefore, given $\vec{X}$, $\vec{Z}$, $\vec{G}$, α_0, α_1 and α_2, the second-stage model has conditional survival function (T refers to L and R)

$$\begin{aligned}
S(T \mid X, Z, G) &= P(Y > T \mid X, Z, G) \\
&= P(\beta_0 + \beta_1 X + \beta_2{}'Z + \xi_2 > T) \\
&= P(\xi_2 > T - \beta_0 - \beta_1 X - \beta_2{}'Z) \\
&= 1 - \Phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}'Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}'G - \alpha_2{}'Z)}{\sqrt{(1-\rho^2)\sigma_2^2}}\right)
\end{aligned}$$

and conditional density function

$$\begin{aligned} f_1(T \mid X, Z, G) &= -\frac{\partial}{\partial t} S(T \mid X, Z, G) \\ &= \phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2{}' Z - \frac{\sigma_2}{\sigma_1}\rho(X - \alpha_0 - \alpha_1{}' G - \alpha_2{}' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}}\right) \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. Combine them together with (7.3.12), we have the full likelihood function for PBIV.

For the MH algorithm, independent diffuse priors are used for the parameters: a normal distribution $N(\mu, \varsigma^2)$ with large variance (e.g. $\mu = 0$, $\varsigma^2 = 100^2$) for each element in α_0, α_1, α_2, β_0, β_1 and β_2; an inverse-gamma distribution Inv-Gamma(γ_1, γ_2) with small shape parameter and small scale parameter (e.g. $\gamma_1 = \gamma_2 = 0.001$) for σ_1^2 and σ_2^2; and a uniform distribution Unif$(-1, 1)$ for ρ. Uniform proposal distributions are used for the random walk: Unif$(z-\omega,\, z+\omega)$ for each element in α_0, α_1, α_2, β_0, β_1 and β_2; Unif$(\max(z-\omega, 0), z+\omega)$ for σ_1^2 and σ_2^2; and Unif$(\max(z-\omega, -1), \min(z+\omega, 1))$ for ρ. Different positive ω is chosen for each parameter to obtain an appropriate acceptance rate (e.g. $20\% \sim 40\%$ depending on the sample size).

The detailed derivation of the log of acceptance probability for parameters in θ is as follows:

- α_0: Denote the current state and candidate sample as α_0 and α_0^*, respectively. With prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_0 - \omega,\, \alpha_0 + \omega)$, the log of acceptance probability:

$$\begin{aligned} \log(a(\alpha_0, \alpha_0^*)) = &\frac{1}{2\varsigma^2}\left((\alpha_0 - \mu)^2 - (\alpha_0^* - \mu)^2\right) + \sum_{i=1}^{n}\Bigg[I\{\delta_i = 4\}\left(\frac{1}{2}(q_i^2 - q_i^{*2})\right) \\ &+ I\{\delta_i = 3\}\Big(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Big) \\ &+ I\{\delta_i = 2\}\Big(\log(\Phi(q_i^*) - \Phi(p_i^*)) - \log(\Phi(q_i^*) - \Phi(p_i))\Big) \\ &+ I\{\delta_i = 1\}\Big(\log(1 - \Phi(p_i^*)) - \log(1 - \Phi(p_i))\Big) + \frac{1}{2}\Big(\nu_i^2 - \nu_i^{*2}\Big)\Bigg] \end{aligned} \tag{7.3.13}$$

where

$$p_i = \frac{1}{\sqrt{(1-\rho^2)\sigma_2^2}}\left[\log(L_i) - (\beta_0 + \beta_1 X_i + \beta_2{}' Z_i) - \frac{\sigma_2}{\sigma_1}\rho(X_i - \alpha_0 - \alpha_1{}' G_i - \alpha_2{}' Z_i)\right] \tag{7.3.14}$$

$$q_i = \frac{1}{\sqrt{(1-\rho^2)\sigma_2^2}}\left[\log(R_i) - (\beta_0 + \beta_1 X_i + \beta_2{}' Z_i) - \frac{\sigma_2}{\sigma_1}\rho(X_i - \alpha_0 - \alpha_1{}' G_i - \alpha_2{}' Z_i)\right] \tag{7.3.15}$$

$$\nu_i = \frac{1}{\sqrt{\sigma_1^2}}(X_i - \alpha_0 - \alpha_1{}' G_i - \alpha_2{}' Z_i) \tag{7.3.16}$$

p_i^*, q_i^* and ν_i^* are similar to p_i, q_i and ν_i, respectively, equations with all α_0 replaced by α_0^*.

- α_1: We update α_1 by updating its elements one-by-one. To update the j-th element α_{1j} with candidate sample α_{1j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_{1j} - \omega, \alpha_{1j} + \omega)$, the log of acceptance probability is similar to (7.3.13), with the first term replaced by $\frac{1}{2\varsigma^2}((\alpha_{1j} - \mu)^2 - (\alpha_{1j}^* - \mu)^2)$. p_i, q_i and ν_i stay the same as (7.3.14), (7.3.15) and (7.3.16). p_i^*, q_i^* and ν_i^* are similar to p_i, q_i and ν_i, respectively: All equations have all α_1 replaced by α_1^*, where α_1^* is α_1 with the j-th element replaced by α_{1j}^*.

- α_2: We update α_2 by updating its elements one-by-one. To update the j-th element α_{2j} with candidate sample α_{2j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\alpha_{2j} - \omega, \alpha_{2j} + \omega)$, the log of acceptance probability is similar to (7.3.13), with the first term replaced by $\frac{1}{2\varsigma^2}((\alpha_{2j} - \mu)^2 - (\alpha_{2j}^* - \mu)^2)$. p_i, q_i and ν_i stay the same as (7.3.14), (7.3.15) and (7.3.16). p_i^*, q_i^* and ν_i^* are similar to p_i, q_i and ν_i, respectively: All equations have all α_2 replaced by α_2^*, where α_2^* is α_2 with the j-th element replaced by α_{2j}^*.

- β_0: We update the current state β_0 with candidate sample β_0^*. With prior distribution

$N(\mu, \varsigma^2)$ and proposal distribution Unif$(\beta_0 - \omega,\ \beta_0 + \omega)$, the log of acceptance probability:

$$\begin{aligned}\log(a(\beta_0, \beta_0^*)) =& \frac{1}{2\varsigma^2}\left((\beta_0 - \mu)^2 - (\beta_0^* - \mu)^2\right) + \sum_{i=1}^{n}\Bigg[I\{\delta_i = 4\}\left(\frac{1}{2}(q_i^2 - q_i^{*2})\right) \\ &+ I\{\delta_i = 3\}\Big(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Big) \\ &+ I\{\delta_i = 2\}\Big(\log(\Phi(q_i^*) - \Phi(p_i^*)) - \log(\Phi(q_i^*) - \Phi(p_i))\Big) \\ &+ I\{\delta_i = 1\}\Big(\log(1 - \Phi(p_i^*)) - \log(1 - \Phi(p_i))\Big)\Bigg]\end{aligned} \tag{7.3.17}$$

where p_i, q_i stay the same as (7.3.14) and (7.3.15). p_i^*, q_i^* are similar to p_i, q_i, with β_0 replaced by β_0^*.

- β_1: We update the current state β_1 with candidate sample β_1^*. With prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\beta_1 - \omega,\ \beta_1 + \omega)$, the log of acceptance probability is similar to (7.3.17), with the first term replaced by $\frac{1}{2\varsigma^2}((\beta_1 - \mu)^2 - (\beta_1^* - \mu)^2)$. p_i and q_i stay the same as (7.3.14) and (7.3.15). p_i^* and q_i^* are similar to p_i and q_i, with β_1 replaced by β_1^*.

- β_2: We update β_2 by updating its elements one-by-one. To update the j-th element β_{2j} with candidate sample β_{2j}^*, and with prior distribution $N(\mu, \varsigma^2)$ and proposal distribution Unif$(\beta_{2j} - \omega,\ \beta_{2j} + \omega)$, the log of acceptance probability is similar to (7.3.17), with the first term replaced by $\frac{1}{2\varsigma^2}((\beta_{2j} - \mu)^2 - (\beta_{2j}^* - \mu)^2)$. p_i and q_i stay the same as (7.3.14) and (7.3.15). p_i^* and q_i^* are similar to p_i and q_i, with β_2 replaced by β_2^*, where β_2^* is β_2 with the j-th element replaced by β_{2j}^*.

- σ_1^2: We update the current state σ_1^2 with candidate sample σ_1^{2*}. With prior distribution Inv-Gamma(γ_1, γ_2) and proposal distribution Unif$(\max(\sigma_1^2 - \omega, 0), \sigma_1^2 + \omega)$, the log of acceptance

probability:

$$\begin{aligned}\log(a(\sigma_1^2, \sigma_1^{2^*})) = &\Big[\log(\sigma_1^2 + \omega - \max(0, \sigma_1^2 - \omega)) - \log(\sigma_1^{2^*} + \omega - \max(0, \sigma_1^{2^*} - \omega))\Big] \\ &+ \left[(\gamma_1 + 1)(\log\sigma_1^2 - \log\sigma_1^{2^*}) + \gamma_2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_1^{2^*}}\right)\right] \\ &+ \sum_{i=1}^{n}\Bigg[I\{\delta_i = 4\}\left(\frac{1}{2}(q_i^2 - q_i^{*2})\right) \\ &+ I\{\delta_i = 3\}\Big(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Big) \\ &+ I\{\delta_i = 2\}\Big(\log(\Phi(q_i^*) - \Phi(p_i^*)) - \log(\Phi(q_i^*) - \Phi(p_i))\Big) \\ &+ I\{\delta_i = 1\}\Big(\log(1 - \Phi(p_i^*)) - \log(1 - \Phi(p_i))\Big) \\ &+ \frac{1}{2}\Big((\log\sigma_1^2 - \log\sigma_1^{2^*}) + (\nu_i^2 - \nu_i^{*2})\Big)\Bigg]\end{aligned}$$

where p_i, q_i and ν_i stay the same as (7.3.14), (7.3.15) and (7.3.16). p_i^*, q_i^* and ν_i^* are similar to p_i, q_i and ν_i, respectively: All equations have all σ_1^2 replaced by $\sigma_1^{2^*}$.

- σ_2^2: We update the current state σ_2^2 with candidate sample $\sigma_2^{2^*}$. With prior distribution Inv-Gamma(γ_1, γ_2) and proposal distribution Unif$(\max(\sigma_2^2 - \omega, 0), \sigma_2^2 + \omega)$, the log of acceptance probability:

$$\begin{aligned}\log(a(\sigma_2^2, \sigma_2^{2^*})) = &\Big[\log(\sigma_2^2 + \omega - \max(0, \sigma_2^2 - \omega)) - \log(\sigma_2^{2^*} + \omega - \max(0, \sigma_2^{2^*} - \omega))\Big] \\ &+ \left[(\gamma_1 + 1)(\log\sigma_2^2 - \log\sigma_2^{2^*}) + \gamma_2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_2^{2^*}}\right)\right] \\ &+ \sum_{i=1}^{n}\Bigg[I\{\delta_i = 4\}\left(\frac{1}{2}(\log\sigma_2^2 - \log\sigma_2^{2^*} + q_i^2 - q_i^{*2})\right) \\ &+ I\{\delta_i = 3\}\Big(\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))\Big) \\ &+ I\{\delta_i = 2\}\Big(\log(\Phi(q_i^*) - \Phi(p_i^*)) - \log(\Phi(q_i^*) - \Phi(p_i))\Big) \\ &+ I\{\delta_i = 1\}\Big(\log(1 - \Phi(p_i^*)) - \log(1 - \Phi(p_i))\Big)\Bigg]\end{aligned}$$

where p_i and q_i stay the same as (7.3.14) and (7.3.15). p_i^*, q_i^* are similar to p_i, q_i, with all σ_2^2

replaced by σ_2^{2*}.

- ρ: We update the current state ρ with candidate sample ρ^*. With prior distribution Unif$(-1,1)$ and proposal distribution Unif$(\max(\rho-\omega,-1),\min(\rho+\omega,1))$, the log of acceptance probability:

$$\begin{aligned}\log(a(\rho^2,\rho^*)) =& \Big[\log(\min(\rho+\omega,1)-\max(\rho-\omega,-1))-\log(\min(\rho^*+\omega,1)-\max(\rho^*-\omega,-1))\Big]\\ &+\sum_{i=1}^{n}\Bigg[I\{\delta_i=4\}\frac{1}{2}\Big((\log(1-\rho^2)-\log(1-\rho^{*2})+q_i^2-q_i^{*2}\Big)\\ &+I\{\delta_i=3\}\Big(\log(1-\Phi(q_i^*))-\log(1-\Phi(q_i))\Big)\\ &+I\{\delta_i=2\}\Big(\log(\Phi(q_i^*)-\Phi(p_i^*))-\log(\Phi(q_i^*)-\Phi(p_i))\Big)\\ &+I\{\delta_i=1\}\Big(\log(1-\Phi(p_i^*))-\log(1-\Phi(p_i))\Big)\Bigg]\end{aligned}$$

where p_i and q_i stay the same as (7.3.14) and (7.3.15). p_i^*, q_i^* are similar to p_i, q_i, with all ρ replaced by ρ^*.

7.3.6 Ishwaran-James Block Gibbs Sampler

In additional to the MCMC algorithm developed in section 7.3.4, we also developed another MCMC algorithm based on Ishwaran-James truncated representation of a Dirichlet process (Ishwaran and James, 2001) (also known as block Gibbs sampler). The difference between this one and Neal's algorithm in section 7.3.4 is that we replace the step of updating $\vec{C}$ with a two-stage procedure. We summarize this step below.

- For $\vec{C}$: We update the cluster indicators $c_1, \ldots, c_n$ simultaneously. Let the truncated DP be

$$H \sim \sum_{k=1}^{N} \pi_k \delta_{\theta_k}$$

  where N is the truncation number and we augment the data by adding $(\pi_k, \theta_k), k = 1, \cdots, N$. For $i = 1, \cdots, n$, draw

$$P(c_i = c|\vec{\pi}, \vec{\theta}, \text{Data}) = \frac{\pi_c L_i(\theta_c)}{\sum_{k=1}^{N} \pi_k L_i(\theta_k)}$$

  where $L_i(\theta_c)$ is the likelihood of subject i with parameter θ_c:

$$L_i(\theta_c) = \mathcal{L}(\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_c \mid L_i, R_i, \delta_i, X_i, Z_i, G_i),$$

  note that the posterior of c_i does not involve c_{-i} as we truncate the DP at order N.

- For π_c: We update weights of each cluster using the following procedure.

  For $c = 1, 2, \cdots, N-1$, let $A_c = \sum_{i=1}^{n} I(c_i = c)$ and $B_h = \sum_{i=1}^{n} I(c_i > c)$. We generate

$$V_c \sim_{ind} Be(1 + A_c, \nu + B_c)$$

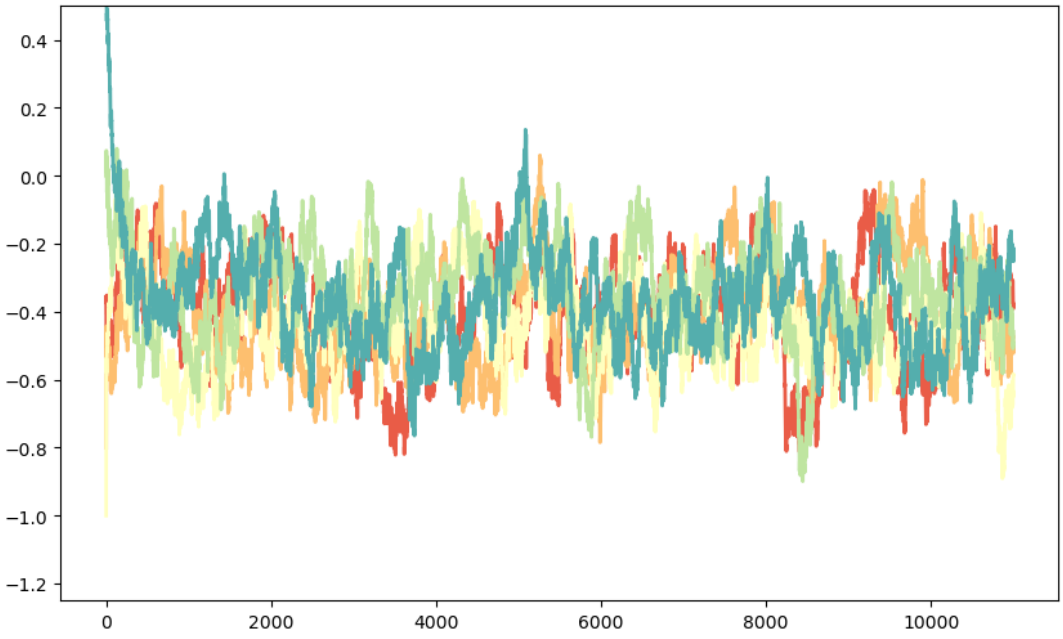
  and $V_N = 1$. Then we set

$$\pi_c = V_c \prod_{k=1}^{c-1} (1 - V_k), c = 1, \cdots, N.$$

In the UKB data analysis, we set $N = 5$ due to previous experiences and summarize the results in table 7.4. The trace plots are given in figure 7.21.

Figure 7.21: Trace plot of causal effect β_1 of the Dirichlet process mixture model for the UKB data.

(a) Female cohort



(b) Male cohort

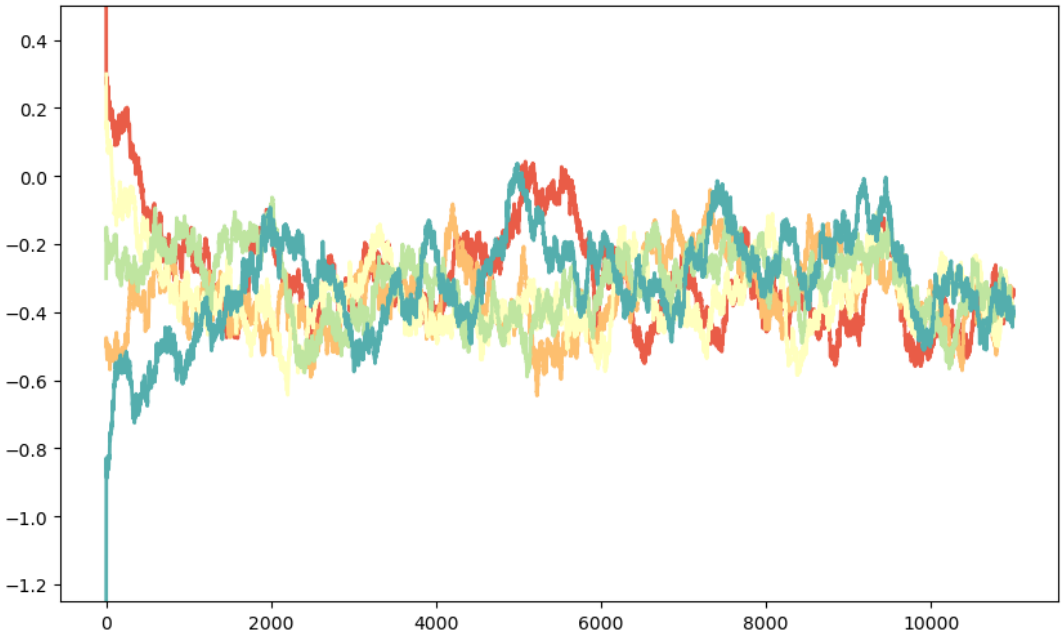


Table 7.4: Analysis of UKB data using Ishwaran-James block Gibbs sampler

	Estimate of β_1	SE	95% CI
Female Cohort	-0.399	0.141	(-0.677, -0.122)
Male Cohort	-0.328	0.111	(-0.523, -0.086)

7.3.7 Additional Simulations

In this section, we present additional simulation settings investigating the robustness of our method under different effect sizes, instrumental strengths and censoring rates (Table (7.5)-(7.10)). We also include a simulation setting (Table (7.11)) that mimics the UKB data in our paper. These simulation settings are similar to the one in the main paper (e.g., generation of covariates) except for the following:

- Zero effect size where $\beta_1 = 0$ in the DPMIV model (4.3.5)-(4.3.6) is presented in Table (7.5);
- Varying instrumental variable strength with partial R-squared equal to 2%, 15%, 35% and 50% (Table (7.6)-(7.9));
- The event rate is as low as 5% (Table (7.10)) so that the censoring rate is high;
- The effect size $\beta_1 = -0.363$ which resembles the UKB example in the paper and the standard deviation (SD) of X is scaled to 0.13, mimicking the SD of log systolic blood pressure (SBP) in the UKB example.

7.3.7.1 Zero effect size ($\beta_1 = 0$)

In this subsection, we set $\beta_1 = 0$ in the DPMIV model (4.3.5)-(4.3.6) and others remain the same as in the simulation section. The simulation result is consistent with nonzero effect size.

Table 7.5: Zero effect size. β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.290	0.060	0%	0.052	0.131	90%	0.206	0.174	88%	1.000
		500	0.268	0.058	0%	0.009	0.100	100%	0.134	0.130	90%	1.001
		1000	0.282	0.031	0%	0.055	0.069	92%	0.030	0.069	99%	1.000
2	Exponential	300	0.059	0.068	59%	0.005	0.050	100%	0.003	0.051	99%	3.129
		500	0.047	0.043	71%	0.005	0.038	90%	0.008	0.038	95%	3.885
		1000	0.031	0.032	98%	0.003	0.026	95%	0.007	0.028	99%	4.161
3	Normal Mixture I	300	0.162	0.063	12%	0.060	0.109	89%	0.079	0.117	89%	1.750
		500	0.163	0.052	0%	0.004	0.084	92%	0.027	0.087	90%	2.508
		1000	0.171	0.019	0%	0.007	0.063	99%	0.016	0.061	99%	2.890
4	Normal Mixture II	300	0.129	0.053	0%	0.006	0.085	100%	0.015	0.079	98%	2.090
		500	0.122	0.046	0%	0.003	0.062	100%	0.033	0.053	92%	2.863
		1000	0.141	0.036	3%	0.010	0.046	99%	0.000	0.031	100%	2.667
5	Normal Mixture III	300	0.123	0.175	65%	0.179	0.211	77%	0.048	0.205	90%	1.656
		500	0.036	0.171	89%	0.023	0.169	90%	0.040	0.133	99%	1.933
		1000	0.005	0.110	90%	0.048	0.127	95%	0.006	0.088	95%	3.050
6	Normal Mixture IV	300	0.077	0.186	68%	0.147	0.121	82%	0.249	0.114	36%	2.319
		500	0.072	0.105	73%	0.109	0.096	79%	0.173	0.094	56%	2.592
		1000	0.040	0.069	90%	0.037	0.072	100%	0.085	0.062	93%	2.982

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = 0$.
- k is the average number of clusters estimated by DPMIV method.

7.3.7.2 Varying Instrumental Variable Strength

In this subsection, we set the partial R-squared between G, the instruments and X, the exposure to be 2% (low strength), 15% (moderate strength), 35% (middle strength) and 50% (high strength) in the DPMIV model (4.3.5)-(4.3.6) and others remain the same as in the simulation section. The simulation results are consistent among different varying instrumental variable strengths.

Table 7.6: Low IV strength (partial R-squared is around 0.02). β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.185	0.139	27%	0.029	0.307	92%	0.038	0.348	91%	1.312
		500	0.177	0.099	33%	0.126	0.344	94%	0.055	0.307	100%	1.404
		1000	0.170	0.086	54%	0.029	0.307	91%	0.085	0.274	95%	1.659
2	Exponential	300	0.107	0.278	53%	0.187	0.211	32%	0.117	0.317	83%	2.589
		500	0.018	0.179	78%	0.141	0.273	61%	0.075	0.228	90%	3.708
		1000	0.028	0.116	82%	0.060	0.144	74%	0.045	0.124	99%	4.117
3	Normal Mixture I	300	0.557	0.098	0%	0.083	0.333	78%	0.117	0.324	82%	1.738
		500	0.602	0.064	0%	0.141	0.272	64%	0.095	0.317	88%	2.105
		1000	0.580	0.045	0%	0.035	0.216	90%	0.005	0.251	96%	3.288
4	Normal Mixture II	300	0.116	0.105	77%	0.064	0.255	80%	0.042	0.290	86%	1.901
		500	0.069	0.075	72%	0.326	0.258	69%	0.047	0.299	90%	2.807
		1000	0.165	0.093	46%	0.272	0.153	58%	0.141	0.210	82%	2.890
5	Normal Mixture III	300	1.114	0.658	26%	0.758	0.366	45%	0.502	0.234	65%	2.055
		500	0.805	0.360	29%	0.666	0.399	48%	0.459	0.238	67%	2.572
		1000	0.792	0.197	20%	0.656	0.308	39%	0.363	0.204	75%	3.313
6	Normal Mixture IV	300	1.161	0.404	10%	0.995	0.192	0%	0.135	0.162	87%	3.095
		500	1.181	0.441	9%	0.994	0.182	0%	0.055	0.125	90%	4.223
		1000	1.110	0.199	0%	0.981	0.178	0%	0.059	0.096	90%	5.706

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

Table 7.7: Moderate IV strength (partial R-squared is around 0.15). β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.675	0.077	0%	0.148	0.223	89%	0.039	0.221	87%	1.980
		500	0.654	0.066	0%	0.076	0.174	92%	0.023	0.173	91%	1.943
		1000	0.674	0.048	0%	0.024	0.135	96%	0.017	0.133	94%	1.998
2	Exponential	300	0.454	0.111	0%	0.077	0.192	86%	0.116	0.229	72%	3.521
		500	0.470	0.077	0%	0.054	0.145	90%	0.034	0.193	84%	2.980
		1000	0.479	0.049	0%	0.001	0.105	93%	0.001	0.112	94%	3.902
3	Normal Mixture I	300	0.089	0.078	38%	0.096	0.185	88%	0.099	0.178	91%	1.509
		500	0.092	0.045	33%	0.082	0.146	94%	0.038	0.131	96%	2.224
		1000	0.112	0.036	22%	0.078	0.101	81%	0.003	0.072	91%	3.043
4	Normal Mixture II	300	0.557	0.078	0%	0.076	0.166	93%	0.055	0.164	88%	2.367
		500	0.571	0.064	0%	0.029	0.128	90%	0.069	0.109	90%	2.823
		1000	0.563	0.046	0%	0.017	0.091	92%	0.049	0.068	89%	2.687
5	Normal Mixture III	300	0.819	0.560	62%	0.563	0.366	76%	0.224	0.333	92%	1.621
		500	0.907	0.353	22%	0.567	0.376	64%	0.077	0.282	90%	2.300
		1000	0.873	0.165	8%	0.464	0.336	71%	0.042	0.211	89%	3.140
6	Normal Mixture IV	300	0.999	0.237	9%	0.954	0.193	0%	0.001	0.158	97%	3.403
		500	0.967	0.250	10%	0.936	0.171	0%	0.006	0.125	96%	4.129
		1000	1.045	0.125	3%	0.894	0.162	0%	0.078	0.092	82%	5.205

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

Table 7.8: Medium IV strength (partial R-squared is around 0.35). β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.606	0.085	0%	0.038	0.135	93%	0.008	0.142	89%	1.861
		500	0.584	0.059	0%	0.014	0.107	98%	0.005	0.106	94%	1.828
		1000	0.600	0.027	0%	0.009	0.078	92%	0.001	0.077	95%	2.142
2	Exponential	300	0.425	0.133	2%	0.028	0.107	88%	0.032	0.122	92%	2.303
		500	0.415	0.092	0%	0.020	0.084	97%	0.021	0.092	94%	2.920
		1000	0.422	0.065	%	0.002	0.059	97%	0.006	0.053	91%	4.109
3	Normal Mixture I	300	0.164	0.064	33%	0.050	0.108	90%	0.044	0.112	96%	1.721
		500	0.168	0.051	10%	0.062	0.087	84%	0.046	0.081	87%	2.387
		1000	0.177	0.037	0%	0.060	0.059	85%	0.006	0.043	98%	2.647
4	Normal Mixture II	300	0.502	0.091	0%	0.019	0.099	90%	0.036	0.097	92%	3.222
		500	0.513	0.057	0%	0.012	0.079	92%	0.038	0.064	93%	2.823
		1000	0.508	0.039	0%	0.013	0.056	93%	0.029	0.040	91%	2.725
5	Normal Mixture III	300	0.657	0.436	62%	0.488	0.346	75%	0.201	0.332	94%	1.660
		500	0.777	0.352	41%	0.469	0.336	72%	0.077	0.282	91%	2.451
		1000	0.807	0.243	8%	0.361	0.315	78%	0.042	0.211	90%	3.225
6	Normal Mixture IV	300	0.759	0.233	26%	0.894	0.172	0%	0.078	0.158	79%	3.088
		500	0.908	0.253	6%	0.841	0.176	0%	0.006	0.125	96%	3.847
		1000	0.863	0.149	1%	0.767	0.162	0%	0.001	0.092	99%	4.980

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

Table 7.9: High IV strength (partial R-squared is around 0.50). β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.520	0.074	0%	0.001	0.103	88%	0.021	0.108	87%	1.640
		500	0.524	0.047	0%	0.016	0.079	94%	0.026	0.080	92%	1.652
		1000	0.528	0.027	0%	0.008	0.057	94%	0.013	0.057	93%	1.905
2	Exponential	300	0.363	0.093	0%	0.007	0.081	95%	0.030	0.088	94%	3.521
		500	0.375	0.073	0%	0.014	0.063	95%	0.003	0.061	100%	2.921
		1000	0.382	0.075	0%	0.008	0.043	90%	0.003	0.038	93%	4.024
3	Normal Mixture I	300	0.221	0.076	0%	0.042	0.085	96%	0.035	0.085	94%	1.545
		500	0.231	0.058	0%	0.047	0.065	86%	0.033	0.063	91%	2.500
		1000	0.229	0.029	0%	0.041	0.045	77%	0.003	0.033	96%	3.060
4	Normal Mixture II	300	0.463	0.068	6%	0.007	0.073	98%	0.016	0.070	94%	2.520
		500	0.450	0.042	0%	0.006	0.056	93%	0.020	0.044	97%	2.925
		1000	0.448	0.037	0%	0.007	0.039	92%	0.009	0.027	91%	2.884
5	Normal Mixture III	300	0.733	0.328	7%	0.445	0.332	77%	0.211	0.327	94%	1.724
		500	0.698	0.260	23%	0.331	0.310	83%	0.076	0.281	97%	2.346
		1000	0.669	0.165	8%	0.211	0.261	80%	0.019	0.184	98%	3.258
6	Normal Mixture IV	300	0.690	0.260	28%	0.856	0.175	0%	0.058	0.160	92%	3.341
		500	0.785	0.210	5%	0.807	0.170	0%	0.004	0.122	98%	3.789
		1000	0.794	0.145	0%	0.685	0.154	0%	0.014	0.092	98%	5.780

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

7.3.7.3 High Censoring Rates and Low Event Rate

In this subsection, we set the event rate (i.e., the number of observations with $L_i = R_i$) to be 5% in the DPMIV model (4.3.5)-(4.3.6) and others remain the same as in the simulation section. The simulation results are consistent among different censoring rates.

Table 7.10: High censoring rate (percentage of event is around 5%). β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
1	Normal	300	0.288	0.135	0%	0.008	0.174	100%	0.094	0.188	93%	1.000
		500	0.323	0.108	0%	0.059	0.137	98%	0.037	0.143	100%	1.050
		1000	0.279	0.067	0%	0.007	0.108	98%	0.002	0.099	96%	1.004
2	Exponential	300	0.476	0.114	0%	0.023	0.073	92%	0.086	0.125	91%	1.012
		500	0.431	0.140	0%	0.037	0.052	90%	0.038	0.061	97%	1.505
		1000	0.456	0.082	0%	0.005	0.041	92%	0.008	0.043	95%	1.699
3	Normal Mixture I	300	0.652	0.067	0%	0.054	0.136	90%	0.002	0.150	93%	1.575
		500	0.683	0.057	0%	0.029	0.108	90%	0.029	0.108	88%	2.566
		1000	0.684	0.039	0%	0.019	0.077	89%	0.014	0.078	96%	3.094
4	Normal Mixture II	300	0.466	0.124	0%	0.021	0.125	94%	0.039	0.163	93%	1.900
		500	0.486	0.090	0%	0.005	0.087	100%	0.008	0.099	95%	2.401
		1000	0.545	0.058	0%	0.034	0.067	96%	0.047	0.066	100%	2.998
5	Normal Mixture III	300	0.669	0.127	0%	0.110	0.197	80%	0.011	0.202	94%	2.003
		500	0.708	0.109	0%	0.018	0.153	99%	0.048	0.155	100%	1.985
		1000	0.726	0.062	0%	0.051	0.114	90%	0.066	0.111	99%	2.890
6	Normal Mixture IV	300	0.657	0.138	0%	0.479	0.138	0%	0.008	0.144	96%	3.602
		500	0.738	0.071	0%	0.461	0.118	0%	0.055	0.125	90%	3.732
		1000	0.706	0.048	0%	0.240	0.099	21%	0.059	0.082	92%	3.894

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -1$.
- k is the average number of clusters estimated by DPMIV method.

7.3.7.4 Mimicking the UKB Data

In this subsection, we set $\beta_1 = -0.363$ in the DPMIV model (4.3.5)-(4.3.6) and scales the standard deviation of X to be 0.13 and others remain the same as in the simulation section. The value 0.13 is calculated from the standard deviation of the log SBP in the UKB data.

Table 7.11: Specification of the bivariate distribution of $(\varepsilon_{1i}, \varepsilon_{2i})^T$ under new UK Biobank simulation scenario with $\beta_1 = -0.363$

Scenario 7		New UK Biobank				
Component	Proportion	μ_1	σ_1^2	μ_2	σ_2^2	ρ
1	75%	4.996	0.015	4.908	0.304	-0.015
2	10%	4.969	0.036	3.349	0.285	0.318
3	5%	5.012	0.084	5.525	0.481	0.789
4	5%	5.054	0.027	3.685	1.087	0.480
5	5%	4.972	0.099	5.028	0.839	-0.356

Table 7.12: New UK Biobank scenario with $\beta_1 = -0.363$. β_1 estimation with and without Instrumental Variable analysis mimicking UKB data with mixed censoring. Single-stage AFT estimate refers to the AFT model (Anderson-Bergman, 2017) without instrumental variables; PBIV refers to parametric Bayesian instrumental variable method; DPMIV refers to our proposed method.

Scenario	Error Distribution	n	Single-stage AFT estimate			PBIV estimate			DPMIV estimate			
			Bias	SD	CP	Bias	SD	CP	Bias	SD	CP	k
7	New UK Biobank	300	0.583	0.753	52%	0.923	0.184	0%	0.041	0.175	99%	2.449
		500	0.464	0.619	63%	0.923	0.183	0%	0.021	0.157	100%	3.158
		1000	0.455	0.393	53%	0.908	0.172	0%	0.018	0.129	100%	4.178

- Results of each scenario under each sample size are based on 100 simulation datasets.
- Mean and SD are the sample mean and sample standard deviation of the 100 posterior means, respectively.
- CP is the coverage probability: the proportion of 95% confidence intervals that cover $\beta_1 = -0.363$.
- k is the average number of clusters estimated by DPMIV method.

7.3.8 More Samples of the Imputed NPMLE from UKB Data

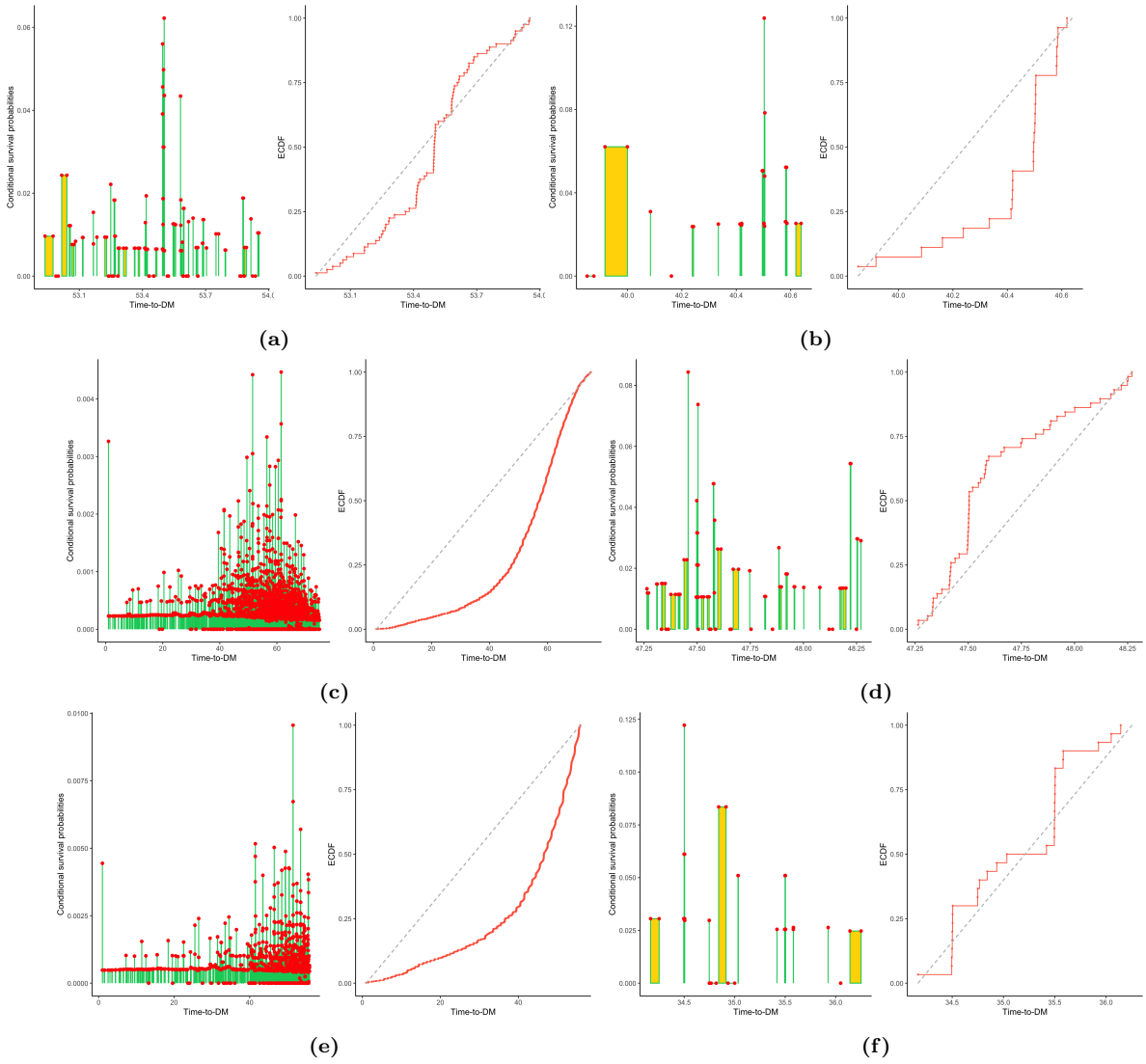


Figure 7.22: Six samples of the imputed NPMLE from the female cohort. For each sample figure, the left panel represents conditional survival probilities from the Turnbull's estimator. A single verticle line means it puts a probability mass at that particular age; a gold rectangle means it puts the probability on the interval. The right panel plots the empirical cumulative distribution function (ECDF) based on the left panels and the grey dashed lines are cdf of a uniform distribution.

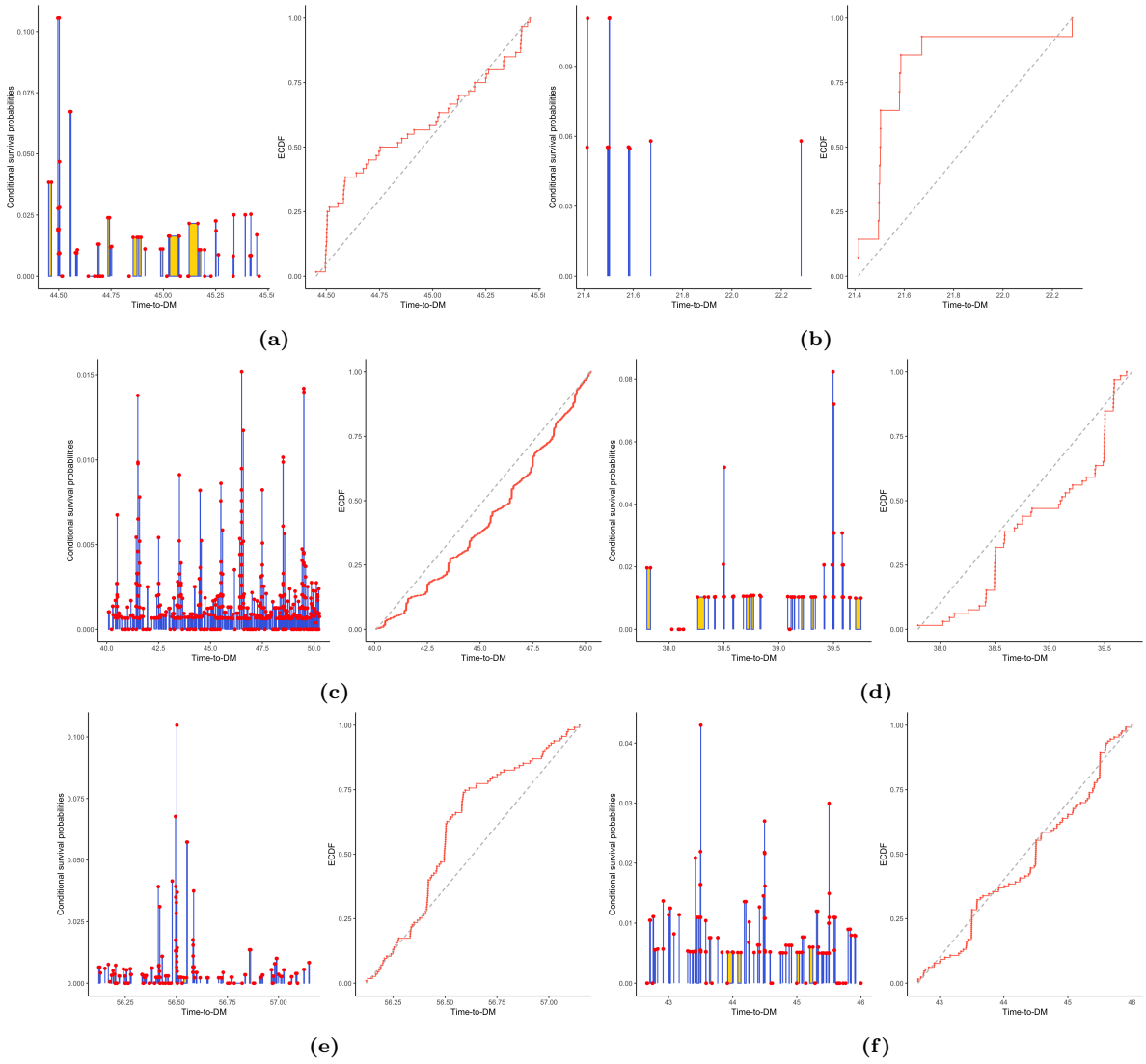


Figure 7.23: Six additional samples of the imputed NPMLE from the male cohort. For each sample figure, the left panel represents conditional survival probilities from the Turnbull's estimator. A single verticle line means it puts a probability mass at that particular age; a gold rectangle means it puts the probability on the interval. The right panel plots the empirical cumulative distribution function (ECDF) based on the left panels and the grey dashed lines are cdf of a uniform distribution.

7.4 Supplementary Information for Chapter 6

7.4.1 Some Theoretical Developments

Here we derive the equivalence theorem for other optimalities. Suppose

$$\Phi_A(M) = Tr\left(M^{-1}\right) = Tr\left((\alpha s_0 s_0^T + (1-\alpha)A_1)^{-1}\right).$$

Then its Fréchet derivative at M_1 in the direction of M_2 is

$$F_{\Phi_A}(M_1, M_2) = -(1-\alpha)Tr\left(M^{-1}(M_2 - M_1)M^{-1}\right).$$

The above Fréchet derivative coincides with those in conventional optimal design literature if we set $\alpha = 0$. Similar sensitivity function for a multiple objective optimal design can be derived similarly. For instance, set

$$\Phi(M) = \frac{1}{K}\sum_{i=1}^{K}\left(\lambda_i^1 Phi_D^i + \lambda_i^2 \Phi_A^i + (1-\lambda_i^1-\lambda_i^2)\Phi_c^i\right)$$

where K is the number of sets of nominal values and i is to emphasize the dependency of locally optimal design. The resulting design M is optimally optimal if the following holds for all x:

$$\sum_{i=1}^{K}\left(\frac{\lambda_i^1}{p}F_{\Phi_D}^i + \lambda_i^2 F_{\Phi_A}^i + (1-\lambda_i^1-\lambda_i^2)F_{\Phi_c}\right) \leq 0. \tag{7.4.1}$$

We next give some simulations on the above equivalence theorem and sensitivity functions.

7.4.2 Sensitivity Plots and Multiple Optimality

7.4.2.1 D-optimality

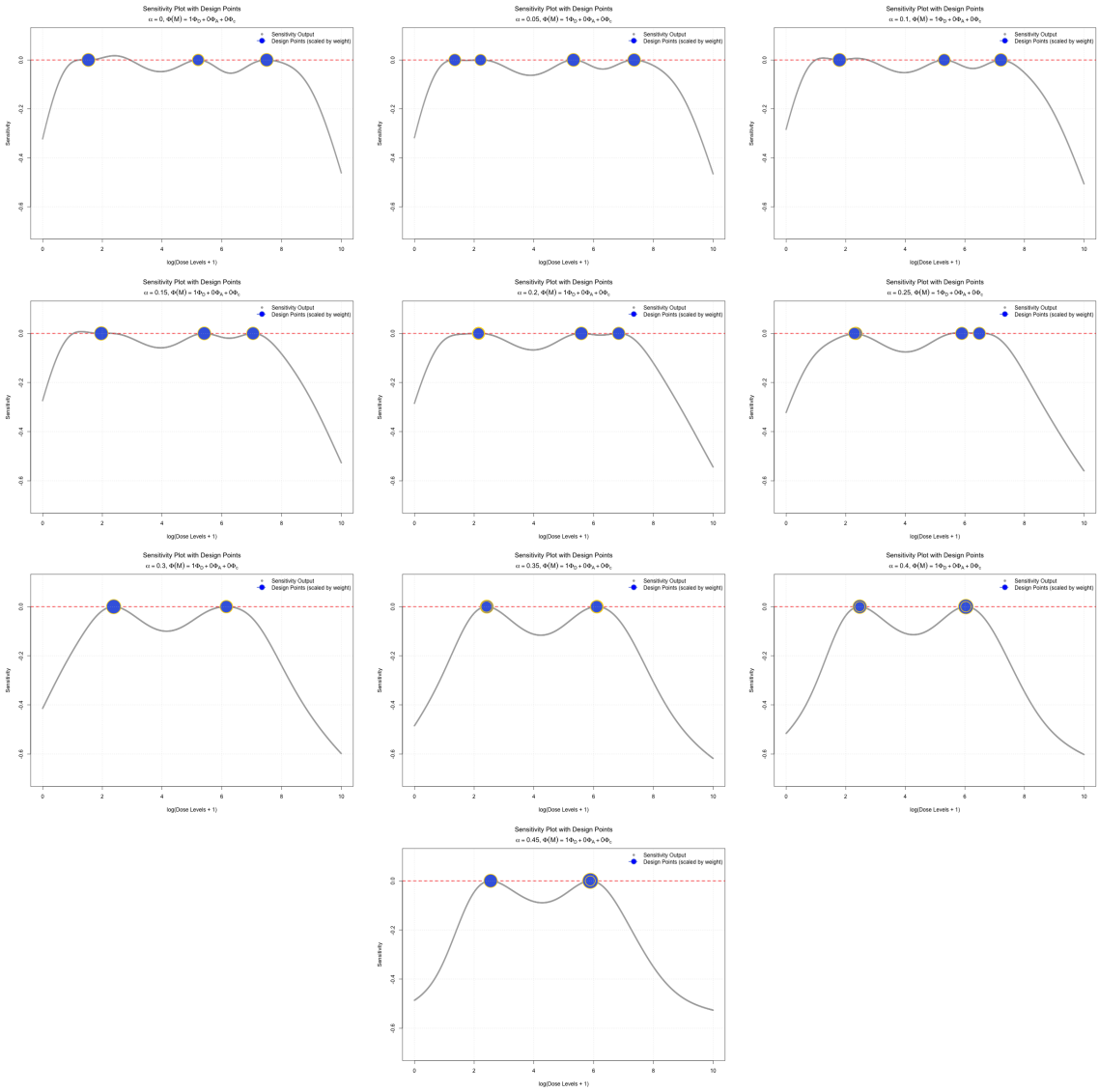


Figure 7.24: Sensitivity plots for D-optimality.

7.4.2.2 DA-optimality with equal weights

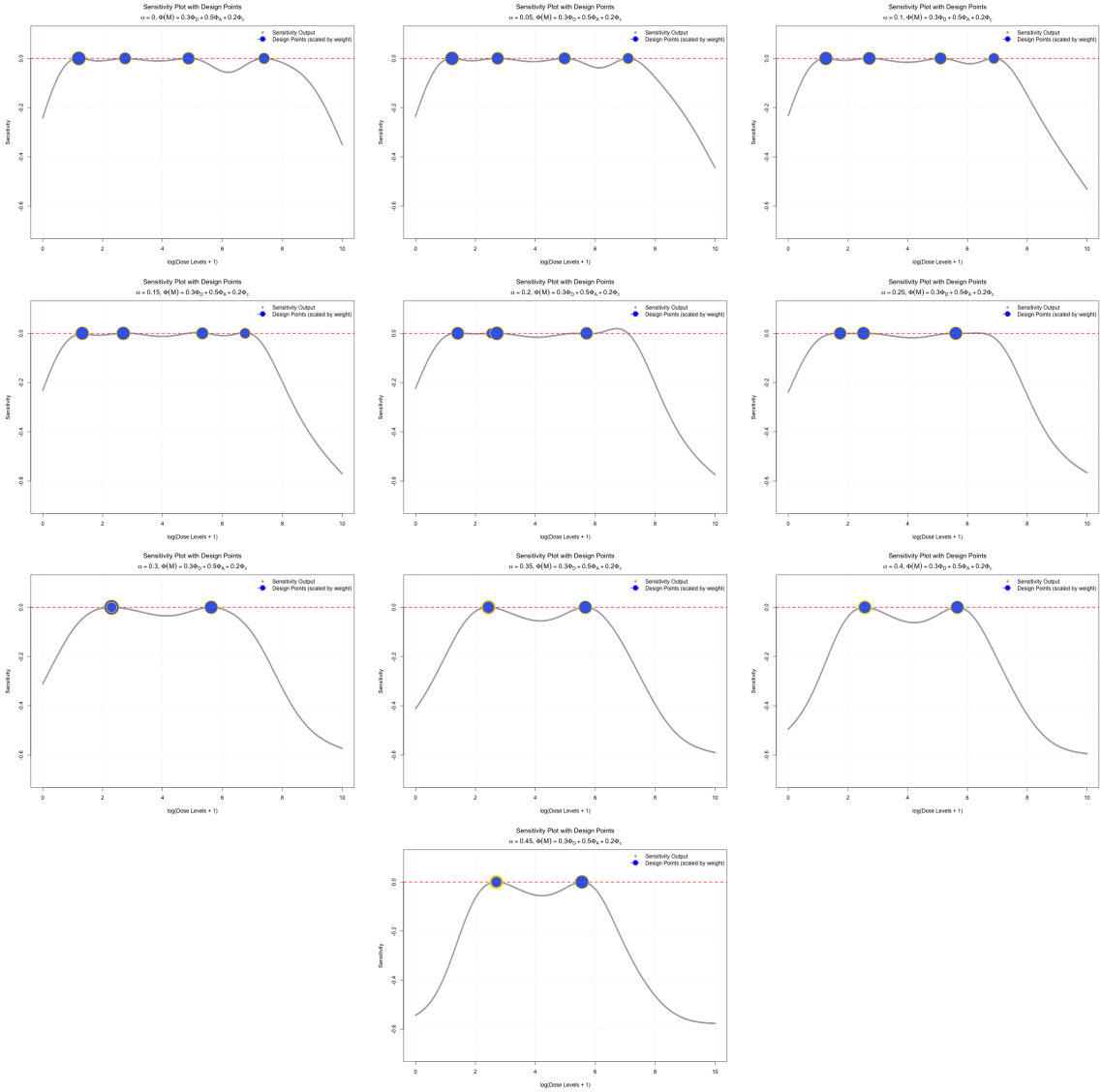


Figure 7.25: Sensitivity plots for DA-optimality.

7.4.2.3 Dc-optimality

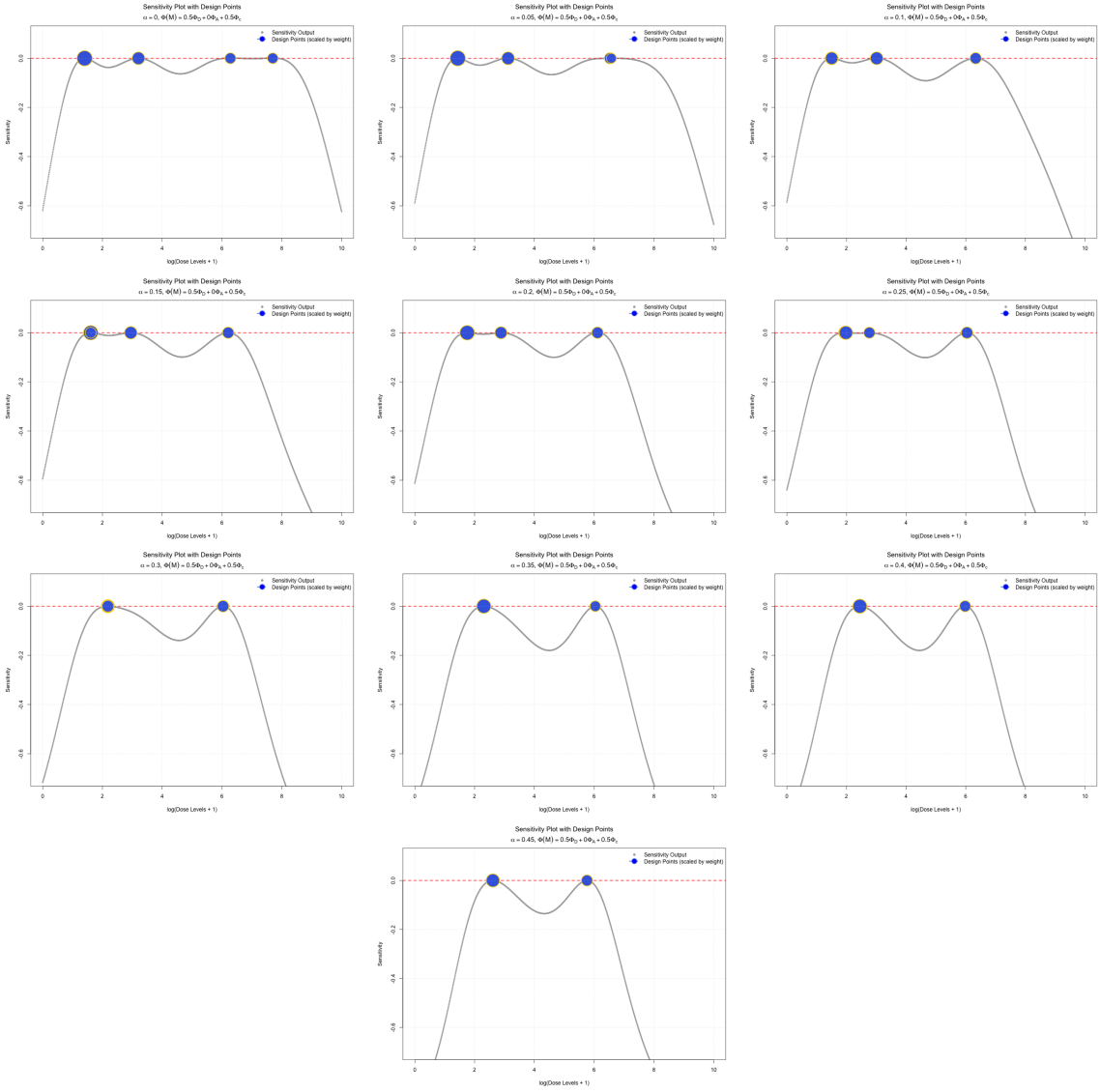


Figure 7.26: Sensitivity plots for Dc-optimality.

7.4.2.4 Ac-optimality

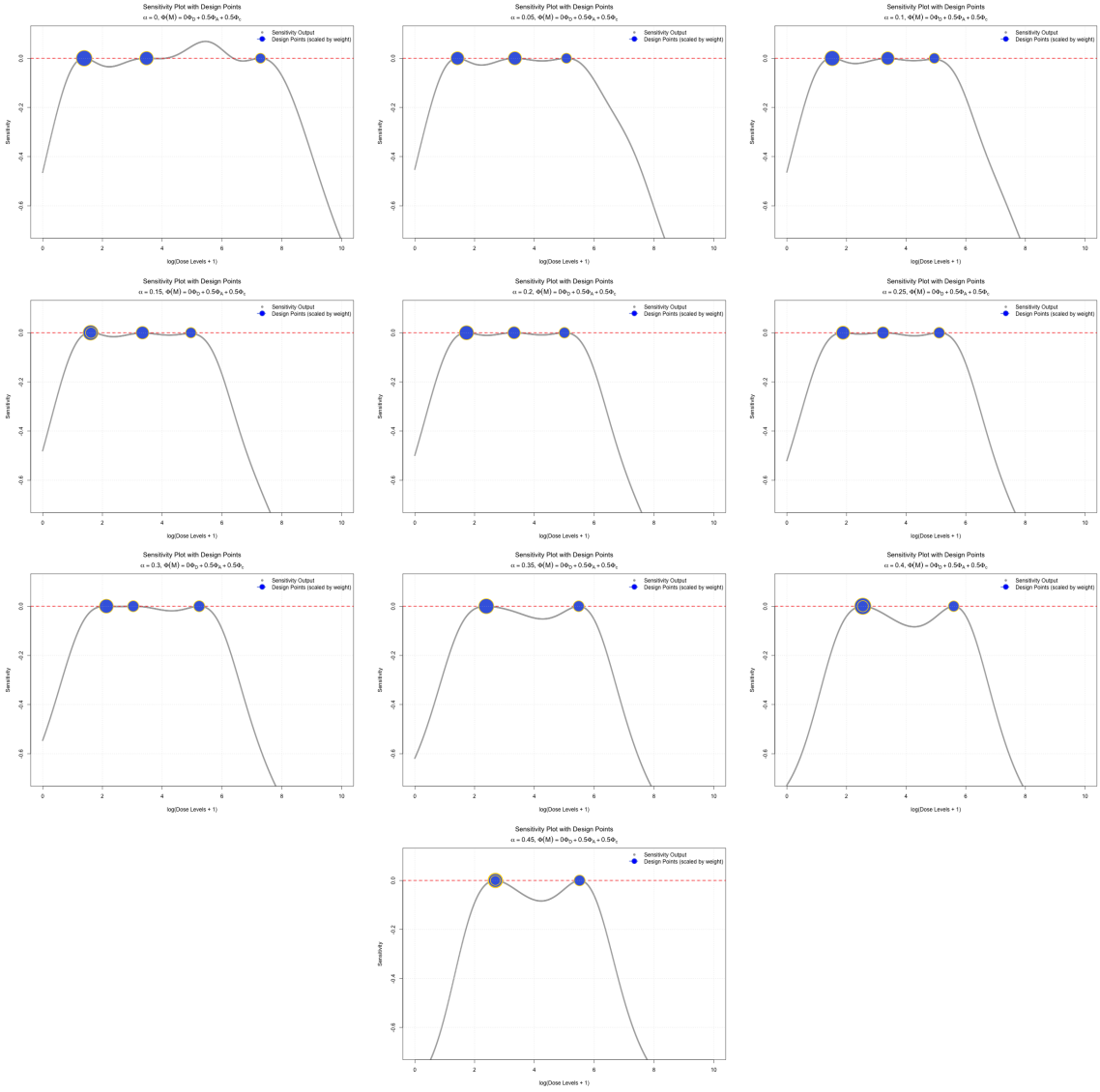


Figure 7.27: Sensitivity plots for Ac-optimality.

7.4.2.5 Multiple-optimality

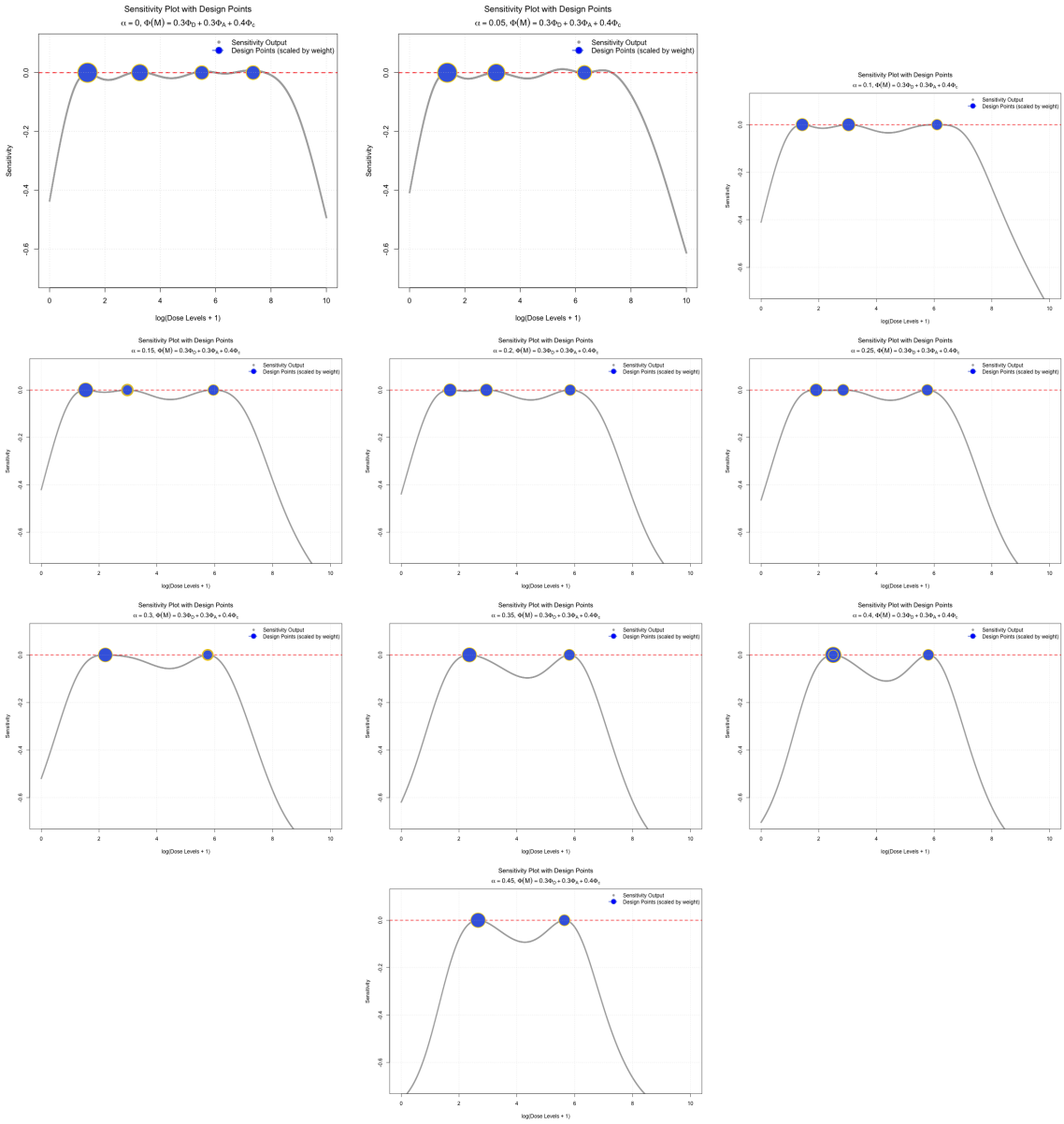


Figure 7.28: Sensitivity plots for Multiple-optimality.

7.4.2.6 A Nine-parameter Model

Finally, we plot the sensitivity function of a nine-parameter model of the following form:

$$\eta_1 = \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) = \beta_1 + \alpha_1 x + \gamma_1 x^2 + \tau_1 \sin(2x)$$
$$\eta_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right) = \beta_2 + \alpha_2 x + \gamma_2 x^2 + \tau_2 \sin(2x)$$
$$\eta_3 = \log(\pi_1 + \pi_2 + \pi_3) = 0$$

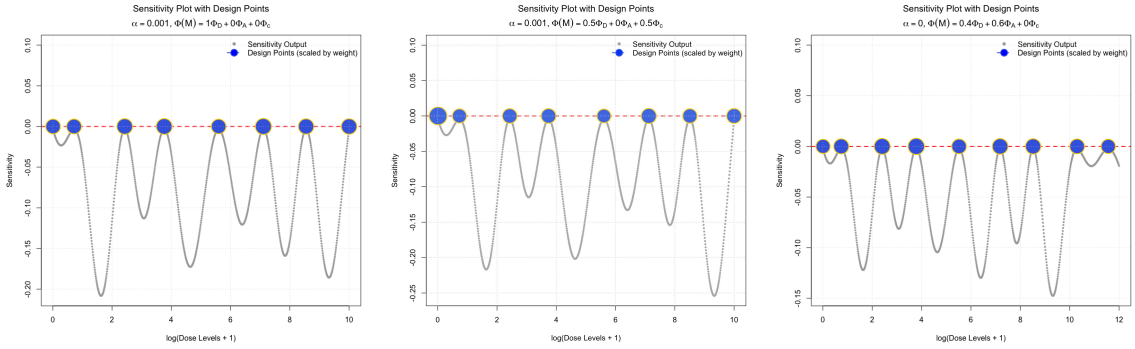


Figure 7.29: Sensitivity plots for the nine-parameter model.

7.5 Some Unpublished Proofs

7.5.1 A Proof on the Product Integral Representation of A Survival Function

Definition 7.5.1 (Product integral (Gill and Johansen, 1990)). Let $\Lambda(t), t \in \mathcal{T}$, be a cádlág function of locally bounded variation. We define

$$S = \prod(1 - \Lambda)$$

the product-integral of Λ over intervals of the form $[0, t], t \in \mathcal{T}$, as the following function:

$$S(t) = \prod_{s\in[0,t]} (1 - \Lambda(ds)) = \lim_{\max|t_i - t_{i-1}|\to 0} \prod(1 + \Lambda_{t_i} - \Lambda_{t_i-1})$$

where $0 = t_0 < t_1 < \cdots < t_n = t$ is a partition of $[0, t]$ and the matrix product is taken from left to right.

Lemma 4 (Cox's lemma (Cox, 1972; Cui, 2022)). *Let T be a nonegative random variable and $S(t) = \mathbb{P}(T \geq t)$ be its survival function. Let $\Lambda(t)$, $\Lambda(dt) = -S(dt)/S(t-)$, $\Lambda(0) = 0$ be the associated cumulative hazard. Then for any $t \leq \tau$, $S(\tau) > 0$, we have*

$$S(t) = \prod_{s\leq t} (1 - \Lambda(ds)) = \prod_{s\leq t} (1 - \Lambda(\Delta s)) \exp(-\Lambda_c(t))$$

where Λ_c is the continuous part of Λ.

Proof. $S(\tau) > 0$ implies Λ is of bounded variation on the interval $[0, \tau]$, which implies

$$\Lambda(t) = \Lambda_c(t) + \Lambda_d(t)$$

where $\Lambda_d(t) = \sum_{s\leq t} \Lambda(\Delta t)$ and $\Lambda_c(t) = \Lambda(t) - \Lambda_d(t) = \int_0^t \Lambda_c(ds)$, $\Lambda(\Delta t) = \Lambda(t) - \Lambda(t-)$. Hence, the product integral can be decomposed into

$$\prod_{s\leq t} (1 - \Lambda(ds)) = \prod_{s\leq t} (1 - \Lambda(\Delta s)) \prod_{s\leq t} (1 - \Lambda_c(ds)) .$$

But the first term is

$$\log \prod_{s \le t} (1 - \Lambda(\Delta s)) = \sum_{s \le t} \log \left(\frac{S(s)}{S(s-)} \right)$$

and the second term is ($\log(1-x) \approx -x$ when x small and use the definition of Riemann integral)

$$\begin{aligned} \prod_{s \le t} (1 - \Lambda_c(ds)) &= \exp \left(\sum_{s \le t} \log(1 - \Lambda_c(ds)) \right) \\ &= \exp \left(- \int_0^t \Lambda_c(ds) \right) \\ &= \exp \left(-\Lambda_c(t) \right). \end{aligned}$$

□

7.5.2 A Proof on the Predictable Variation of A Counting Process Martingale

Let $N(t), t \in [0, \tau]$ be a counting process and $\Lambda(t)$ be its compensator, i.e., $\Lambda(t)$ is predictable, càdlàg and finite variation such that $M(t) = N(t) - \Lambda(t)$ is a local-martingale (Andersen et al., 2012). We refer to $M(t)$ as the counting process martingale. By Doob-Meyer's theorem, there exists a unique finite variation càdlàg predictable process $\langle M \rangle(t)$ such that $M^2(t) - \langle M \rangle(t)$ is a local-martingale.

Lemma 5. *The predictable variation of $M(t)$ is*

$$\langle M \rangle(t) = \int_0^t (1 - \Lambda(\Delta s)\Lambda(ds)).$$

Proof. We have (note that $N(\Delta s)^k = N(\Delta s)$ for any integer k)

$$\begin{aligned} M(\Delta s)^2 &= (M(s) - M(s-))^2 \\ &= (N(s) - \Lambda(s) - N(s-) + \Lambda(s-))^2 \\ &= (N(\Delta s) - \Lambda(\Delta s))^2 \\ &= N(\Delta s) + \Lambda(\Delta s)^2 - 2N(\Delta s)\Lambda(\Delta s) \\ &= M(\Delta s)\left(1 - 2\Lambda(\Delta s)\right) + \Lambda(\Delta s)\left(1 - \Lambda(\Delta s)\right). \end{aligned}$$

Hence,

$$\sum_{s\leq t} M(\Delta s)^2 = \int_0^t (1-2\Lambda(\Delta s))M(ds) + \int_0^t (1-\Lambda(\Delta s))\Lambda(ds).$$

By integration-by-parts (Dabrowska, 2019),

$$\begin{aligned} M(t)^2 &= 2\int_0^t M(s-)M(ds) + \sum_{s\leq t} M(\Delta s)^2 \\ &= \left(2\int_0^t M(s-)M(ds) + \int_0^t (1-2\Lambda(\Delta s))M(ds)\right) + \int_0^t (1-\Lambda(\Delta s))\Lambda(ds). \end{aligned}$$

The first term is a local-martingale by the definition of Riemann-Stieltjes while the second term is predictable. Hence, by the uniqueness of Doob-Meyer's theorem (Dabrowska, 2019), we have $\langle M\rangle(t) = \int_0^t (1-\Lambda(\Delta s)\Lambda(ds))$.

□

7.5.3 A Proof on the Non-differentiability of Brownian Motion Paths

We start with a proposition that describes the peculiarities of the Brownian motion path, and then jump into pointwise and globalwise non-differentiability of B.

Lemma 6 (Liggett (2010); Karatzas and Shreve (1991))**.** *Almost surely the Brownian motion B is not monotone on any interval $[s,t]$.*

Proof. It is enough to show that the following set has probability 0:

$$\begin{aligned} A &= \bigcup_{s,t\in\mathbb{Q}^+} \{\omega\in\Omega : B(\omega) \text{ is monotone on [s,t]}\} \\ &= \bigcup_{s,t\in\mathbb{Q}^+} A_{st} \end{aligned}$$

By σ-additivity of a probability measure and stationarity of B, we only have to show that $A_{01} = \{B \text{ is monotonically decreasing on } [0,1]\}$ has probability 0. Set

$$B_n = \bigcap_{i=1}^{n} \left\{ B\left(\frac{i-1}{n}\right) - B\left(\frac{i}{n}\right) \geq 0 \right\}$$

so that $A_{01} = \cap_{n=1}^{\infty} B_n$. But

$$P^0(B_n) = 2^{-n} \to 0 \text{ as } n \to \infty.$$

□

Lemma 7 (Pointwise Nonsmoothness (Liggett, 2010))**.** *For a fixed $t \geq 0$,*

$$\mathbb{P}(B(\cdot, \omega) \text{ is not differentiable at } t) = 1. \tag{7.5.1}$$

Further, for $\mathbb{P} - a.s.\ \omega \in \Omega$,

$$m(A) = 0,\ A = \{t \geq 0 : B(\cdot, \omega) \text{ is differentiable at } t\}. \tag{7.5.2}$$

Proof. For $t = 0$, we have $\limsup_{t \downarrow 0} \frac{B(t)}{\sqrt{t}} = +\infty$ a.s. If B is differentiable at B, then $\exists K,\ \epsilon > 0$, and we have (by the mean value theorem)

$$|B(s) - B(0)| \leq K|s - 0|, s \in [0, \epsilon).$$

A contradiction. Next, for any $t > 0$, we note that $B(t) - B(t_0)$ has the same distribution as $B(t - t_0)$. For the second part, fix ω and define the random set

$$C = \left\{ t \geq 0 : \lim_{n \to \infty} \frac{B(t + \frac{1}{n}, \omega) - B(t, \omega)}{1/n} \text{ exists at } t. \right\}$$

By Fubini's theorem and joint measurability of $B(t, \omega)$

$$\mathbb{E}\left(m(A)\right) \leq \mathbb{E}m(C) = \mathbb{E}\left(\int_0^{\infty} \mathbb{I}_C(t)dt\right) = \int_0^{\infty} \int_{\Omega} \mathbb{I}_C(t)d\mathbb{P}dt = 0 \tag{7.5.3}$$

where the last equality follows from the first part. □

Theorem 7.5.1 (Nondifferentiability of B: Paley-Wiener-Zygmund (Liggett, 2010))**.** The Brownian motion $B(\cdot, \omega)$ is no where differentiable a.s.

Remarks: By the Markov property, $B(\cdot + s) - B(s)$ has the same distribution as $B(\cdot)$, so it is sufficient to show that $\mathbb{P}(B \text{ is differentiable on } (0, 1]) = 0$. Let D_s be the set such that the path $B(\cdot, \omega)$ is differentiable at s. By the previous lemma, we know $\mathbb{P}(D_s) = 0$ for any s. Here we want

to show that

$$D = \bigcup_{s\in(0,1]} D_s = \{\omega : B(s,\omega) \text{ is differentiable at some } s \in (0,1]\}$$

has $\mathbb{P}$-measure 0, but this is an uncountable union. The idea here is to bound this quantity by discretizing the sample path of B and each piece has a small oscillation. Then D can be written as a countable union of sets. Within each discretized interval, the sample path cannot be Lipschitz continuous and hence, cannot be differentiable.

Proof. Recall

$$D = \bigcup_{s\in(0,1]} D_s$$

where D_s is the set that the path $B(\cdot,\omega)$ is differentiable at s. Define

$$\begin{aligned}\Gamma &= \bigcup_{m=1}^{\infty} \liminf_{n\to\infty} \underbrace{\bigcup_{k=1}^{n-2} \bigcap_{j=k}^{k+2} \left\{ \left| B\left(\frac{j}{n}\right) - B\left(\frac{j-1}{n}\right) \right| \le \frac{3m}{n} \right\}}_{D_{mn}} \\ &= \bigcup_{m=1}^{\infty} \bigcup_{l=1}^{\infty} \bigcap_{n=l}^{\infty} D_{mn}\end{aligned} \tag{7.5.4}$$

If $D \subseteq \Gamma$, then we have

$$\mathbb{P}(D) \le \mathbb{P}(\Gamma) \le \sum_{m=1}^{\infty} \mathbb{P}\left(\liminf_{n\to\infty} D_{mn}\right) = \sum_{m=1}^{\infty} \mathbb{P}\left(\bigcup_{l=1}^{\infty} \bigcap_{n\ge l}^{\infty} D_{mn}\right)$$

and for any fixed m. The proof is completed if we can show each term on the right-hand side

is 0. By stationarity and independence of increments,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{l=1}^{\infty}\bigcap_{n=l}^{\infty} D_{mn}\right) &\leq \liminf_{n\to\infty} \mathbb{P}\left(D_{mn}\right) \\
&\leq \liminf_{n\to\infty} \sum_{k=1}^{n-2} \mathbb{P}\left(\left|B\left(\frac{1}{n}\right)\right| \leq \frac{3m}{n}\right)^3 \\
&\leq \liminf_{n\to\infty} n\left[\mathbb{P}\left(\left|B\left(1\right)\right| \leq \frac{3m}{\sqrt{n}}\right)\right]^3 \\
&\leq \liminf_{n\to\infty} n\left(\int_{-\frac{3m}{\sqrt{n}}}^{\frac{3m}{\sqrt{n}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx\right)^3 \\
&\leq \liminf_{n\to\infty} \left(\frac{6m}{\sqrt{2\pi}}\right)^3 \frac{1}{\sqrt{n}} \\
&= 0.
\end{aligned}$$

We only have $D \subseteq \Gamma$ left to show.

Lemma 8. *We have $D \subseteq \Gamma$ where $D = \bigcup_{s\in(0,1]} D_s$ and*

$$\Gamma = \bigcup_{m=1}^{\infty} \liminf_{n\to\infty} D_{mn}$$

where

$$D_{mn} = \bigcup_{k=1}^{n-2}\bigcap_{j=k}^{k+2} \left\{\left|B\left(\frac{j}{n}\right) - B\left(\frac{j-1}{n}\right)\right| \leq \frac{3m}{n}\right\}.$$

Proof. If B is differentiable at some $s \in (0,1]$, then it is Lipschitz continuous at s. Let

$$A_{mn} = \left\{\exists s \in (0,1],\ \left|B(t) - B(s)\right| \leq m|t-s| \ \forall t \text{ s.t. } |t-s| \leq \frac{2}{n}\right\}$$

so that clearly $D \subseteq \bigcup_m \bigcup_n A_{mn}$. We next show that for any $\omega \in D$, then $\omega \in \Gamma$.

If $\omega \in D$, then $\exists M, N$ so that $\omega \in A_{MN}$.

Pick k s.t. $\frac{k}{N} \leq s \leq \frac{k+1}{N}$, so $\omega \in \bigcap_{j=k}^{k+2} \left\{\left|B(\frac{j}{N}) - B(\frac{j-1}{N})\right| \leq \frac{3M}{N}\right\}$.

In other words, we can pick N_0 large and

$$\omega \in \bigcap_{N\geq N_0} \bigcup_{k=1}^{N-2} \bigcap_{j=k}^{k+2} \left\{ \left| B\left(\frac{j}{N}\right) - B\left(\frac{j-1}{N}\right) \right| \leq \frac{3M}{N} \right\} = \bigcap_{N\geq N_0} D_{MN}.$$

We have shown that if $\omega \in D$, then $\omega \in \Gamma = \bigcup_m \bigcup_l \bigcap_{n\geq l} D_{mn}$. □

□

7.5.4 A Proof on the Existence of Dirichlet Processes

Definition 7.5.2 (Dirichlet process (Ferguson, 1973))**.** Let α be a non-null finite measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We say P is a **Dirichlet process** on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with parameter α if for every $k = 1, 2, \cdots$, and measurable partition $(B_1, \cdots, B_k)$ of $\mathcal{X}$, we have

$$(P(B_1), \cdots, P(B_k)) \sim Dirichlet(\alpha(B_1), \cdots, \alpha(B_k)).$$

We write $P \sim \mathcal{D}(\alpha)$ to represent a Dirichlet process.

Theorem 7.5.2 (Existence of DP (Ferguson, 1973))**.** Let P be the DP defined above satisfying $H(\emptyset) = 0$. We assign probabilities to arbitrary measurable sets $A_1, \cdots, A_m \in \mathcal{B}(\mathcal{X})$ using the following rule:

$$P(A_i) = \sum_{(v_1,\cdots,v_m)\ni v_i=1} P(B_{v_1,\cdots,v_m})$$

where

$$B_{v_1,\cdots,v_m} = \cap_{j=1}^m A_j^{v_j}$$

for each $v_j = 0$ or 1 and $A_j^1 = A_j$ while $A_j^0 = A_j^c$. Then there **exists** a probability $\mathbb{P}$ on $([0,1]^{\mathcal{B}(\mathcal{X})}, \mathcal{B}\left([0,1]^{\mathcal{B}(\mathcal{X})}\right))$ where the σ-field is generated by cylinder sets.

Proof. Here we use a more explicit construction compared with the original proof in Ferguson (1973). We need the following theorem.

Theorem 7.5.3 (Kolmogorov's Theorem (Dabrowska, 2019))**.** Suppose that for each t, $(\Omega_t, \mathcal{F}_t)$ represents a complete separable metric space with its Borel σ-algebra. If $\mathbb{P}$ is a compatible family

of distributions then there exists a uniquely defined probability measure on the product space $(\prod_{t\in\mathbb{T}}\Omega_t, \otimes_{t\in\mathbb{T}}\mathcal{F}_t)$ such that its finite dimensional distributions are given by the family $\mathbb{P}$.

To check the Kolmogorov's consistency conditions, we must show that, for arbitrary m and measurable sets $A_1, \cdots, A_m$, the marginal distribution of $(P(A_1), \cdots, P(A_{m-1}))$ derived from marginalizing $(P(A_1), \cdots, P(A_m))$ is identical to the defined distribution of $(P(A_1), \cdots, P(A_{m-1}))$, i.e., the following are identical

$$\left(\sum_{(v_1,\cdots,v_m),v_1=1} P(B_{v_1,\cdots,v_m}), \cdots, \sum_{(v_1,\cdots,v_m),v_{m-1}=1} P(B_{v_1,\cdots,v_m}) \right)$$

$$\left(\sum_{(v_1,\cdots,v_{m-1}),v_1=1} P(B_{v_1,\cdots,v_{m-1}}), \cdots, \sum_{(v_1,\cdots,v_{m-1}),v_{m-1}=1} P(B_{v_1,\cdots,v_{m-1}}) \right).$$

Since $B_{v_1,\cdots,v_{m-1}} = B_{v_1,\cdots,v_{m-1},0} \uplus B_{v_1,\cdots,v_{m-1},1}$, we have

$$P(B_{v_1,\cdots,v_{m-1}}) =_d P(B_{v_1,\cdots,v_{m-1},0}) + P(B_{v_1,\cdots,v_{m-1},1})$$

by the properties of Dirichlet distribution (Ferguson, 1973). With this replacement, the two random vectors defined in the previous slide have the identical distribution, i.e., the definition of $P(\cdot)$ induces a compatible probability family on $([0,1]^{\mathcal{B}(\mathcal{X})}, \mathcal{B}\left([0,1]^{\mathcal{B}(\mathcal{X})}\right))$. Further, if there exists another partition (or disjointification of $A_1, \cdots, A_m$), then the same argument still applies. Hence, the Dirichlet process P actually defines a random process. □

7.5.5 A Proof on the Concentration of Kernel Density Estimate

This subsection derives a refined bound for Exercise 2.15 in Wainwright (2019). A convergence rate of order $O\left(\sqrt{\frac{\log n}{nh_n}}\right)$, where $h_n \to 0$ is the bandwidth, is provided in Example 10.14.3 of Dabrowska (2019).

Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of random variables drawn from a density f on the real line. A commonly used estimate of f is the *kernel density estimate*, defined as:

$$\hat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right),$$

where $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$ and $K : \mathbb{R} \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)\, dt = 1$, and $h > 0$ is a bandwidth parameter. Define the L^1-norm

$$\|\hat{f}_n - f\|_1 := \int_{-\infty}^{\infty} |\hat{f}_n(t) - f(t)|\, dt.$$

Lemma 9.

$$\mathbb{P}\left(\|\hat{f}_n - f\|_1 \geq \mathbb{E}[\|\hat{f}_n - f\|_1] + \delta\right) \leq e^{-n\delta^2/2}.$$

Proof. The random function $\|\widehat{f}_n - f\|_1$ satisfies the bounded difference property with a bound of 2. To demonstrate this, consider varying the first component X_1, and let X_1^* be an independent copy of X_1:

$$\begin{aligned}\int_{\mathbb{R}} \left(\left|\frac{1}{n}K_h(x - X_1) - f(x)\right| - \left|\frac{1}{n}K_h(x - X_1^*) - f(x)\right|\right) dx &\leq \frac{1}{n}\int_{\mathbb{R}} \left|K_h(x - X_1) - K_h(x - X_1^*)\right| dx \\ &\leq \frac{2}{n}.\end{aligned}$$

Hence, by McDiarmid's inequality (Wainwright, 2019), the claim follows.

□

Bibliography

Abdel-Basset, M., Abdel-Fatah, L., and Sangaiah, A. K. (2018). Metaheuristic algorithms: A comprehensive review. Computational intelligence for multimedia big data on the cloud with engineering applications, pages 185–231.

Abualigah, L., Shehab, M., Alshinwan, M., and Alabool, H. (2022). Salp swarm algorithm: a comprehensive survey. Neural Computing and Applications, 32(15):11195–11215.

Adams, D. C. and Otárola-Castillo, E. (2013). geomorph: an r package for the collection and analysis of geometric morphometric shape data. Methods in ecology and evolution, 4(4):393–399.

Afifi, A., May, S., and Clark, V. A. (2011). Practical multivariate analysis. Chapman and Hall/CRC.

Ahmadianfar, I., Heidari, A. A., Gandomi, A. H., Chu, X., and Chen, H. (2021). RUN beyond the metaphor: An efficient optimization algorithm based on Runge Kutta method. Expert Systems with Applications, 181:115079.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. technometrics, 16(1):125–127.

Allen, N. E., Sudlow, C., Peakman, T., Collins, R., and biobank, U. (2014). Uk biobank data: come and get it.

Amari, S.-I. (1982). Differential geometry of curved exponential families-curvatures and information loss. The Annals of Statistics, 10(2):357–385.

Amari, S.-I. (2006). Differential geometry of statistical inference. In Probability Theory and Mathematical Statistics: Proceedings of the Fourth USSR-Japan Symposium, held at Tbilisi, USSR, August 23–29, 1982, pages 26–40, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors-in-variables model. Journal of Econometrics, 28(3):273–289.

Amemiya, Y. (1990). Two-stage instrumental variables estimators for the nonlinear errors-in-variables model. Journal of Econometrics, 44(3):311–332.

Amini, M., Zayeri, F., and Salehi, M. (2021). Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. BMC public health, 21:1–12.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). Statistical models based on counting processes. Springer Science & Business Media.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The Annals of Statistics, pages 1100–1120.

Anderson-Bergman, C. (2017). icenreg: regression models for interval censored data in r. Journal of Statistical Software, 81:1–23.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444–455.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., et al. (2012). The'k'in k-fold cross validation. In ESANN, volume 102, pages 441–446.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174.

Aranha, C., Camacho Villalón, C. L., Campelo, F., Dorigo, M., Ruiz, R., Sevaux, M., Sörensen, K., and Stützle, T. (2021). Metaphor-based metaheuristics, a call for action: the elephant in the room. Swarm Intelligence, pages 1–6.

Archetti, F. and Schoen, F. (1984). A survey on the global optimization problem: general theory and computational approaches. Annals of Operations Research, 1(2):87–110.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. (2016). A locally adaptive normal distribution. Advances in Neural Information Processing Systems, 29.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. (2020). Geometrically enriched latent spaces. arXiv preprint arXiv:2008.00565.

Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. Biometrics, pages 535–546.

Askin, O. E., Inan, D., and Buyuklu, A. H. (2017). Parameter estimation of shared frailty models based on particle swarm optimization. Int J Stat Probab, 6(1):48–58.

Atiyat, M. (2011). Instrumental Variable Modeling in a Survival Analysis Framework. Ph.d. dissertation, The Pennsylvania State University.

Atkinson, A., Donev, A., and Tobias, R. (2007). Optimum experimental designs, with SAS, volume 34. Oxford University Press.

Atkinson, A. C., Fedorov, V. V., Herzberg, A. M., and Zhang, R. (2014). Elemental information matrices and optimal experimental design for generalized regression models. Journal of Statistical Planning and Inference, 144:81–91.

Atkinson, C. and Mitchell, A. F. (1981). Rao's distance measure. Sankhyā: The Indian Journal of Statistics, Series A, pages 345–365.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research, 20(1):40–49.

Bacher, R., Leng, N., Chu, L.-F., Ni, Z., Thomson, J. A., Kendziorski, C., and Stewart, R. (2018). Trendy: segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. BMC bioinformatics, 19(1):1–10.

Bahrami, M., Bozorg-Haddad, O., and Chu, X. (2018). Cat swarm optimization (cso) algorithm. In Advanced optimization by nature-inspired algorithms, pages 9–18. Springer.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. Statistics in medicine, 33(13):2297–2340.

Baker, F. B. and Kim, S.-H. (2004). Item response theory: Parameter estimation techniques. CRC press.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. Advances in Neural Information Processing Systems, 28.

Barndorff-Nielsen, O. E., Cox, D. R., and Reid, N. (1986). The role of differential geometry in statistical theory. International Statistical Review/Revue Internationale de Statistique, pages 83–96.

Bates, D. M. (2010). lme4: Mixed-effects modeling with r.

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., et al. (2020). Using social and behavioural science to support covid-19 pandemic response. Nature human behaviour, 4(5):460–471.

Beauchamp, J. J. and Cornell, R. G. (1966). Simultaneous nonlinear estimation. Technometrics, 8(2):319–326.

Bechtel, G. G. (1985). Generalizing the rasch model for consumer rating scales. Marketing Science. Institute for Operations Research and the Management Sciences (INFORMS), 4(1):62–73.

Bekele, B. N. and Thall, P. F. (2006). Dose-finding based on multiple ordinal toxicities in phase i oncology trials. Statistical Methods for Dose-Finding Experiments, pages 243–258.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6):1373–1396.

Belloni, A. and Chernozhukov, V. (2011). l1-penalized quantile regression in high-dimensional sparse models. Annals of Statistics, 39(1):82–130.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4):791–806.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. Cell, 157(3):714–725.

Bezruczko, N. (2005). Rasch measurement in health sciences. Maple Grove.

Bhattacharya, A. and Bhattacharya, R. (2012). Nonparametric inference on manifolds: with applications to shape spaces, volume 2. Cambridge University Press.

Bickel, P. J., Ritov, Y., Klaassen, J., and Wellner (1993). Efficient and adaptive estimation for semiparametric models, volume 4. Springer.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. Annals of Statistics, 37(4):1705–1732.

Bijwaard, G. E. (2008). Instrumental variable estimation for duration data. Tinbergen Institute Discussion Papers 08-032/4, Tinbergen Institute.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika, 46(4):443–459.

Boussaïd, I., Lepagnot, J., and Siarry, P. (2013). A survey on optimization metaheuristics. Information sciences, 237:82–117.

Bowden, J., Bornkamp, B., Glimm, E., and Bretz, F. (2021). Connecting instrumental variable methods for causal inference to the estimand framework. Statistics in Medicine, 40(25):5605–5627.

Boyd, S. P. and Vandenberghe, L. (2004). Convex optimization. Cambridge university press.

Bratton, D. and Kennedy, J. (2007). Defining a standard for particle swarm optimization. In 2007 IEEE swarm intelligence symposium, pages 120–127. IEEE.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4):434–455.

Bu, X., Majumdar, D., and Yang, J. (2020). D-optimal designs for multinomial logistic models. The Annals of Statistics, 48(2):983–1000.

Buckley, L. A., Bebenek, I., Cornwell, P. D., Hodowanec, A., Jensen, E. C., Murphy, C., and Ghantous, H. N. (2020). Drug development 101: A primer. International Journal of Toxicology, 39(5):379–396.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for mendelian randomization. Statistical methods in medical research, 26(5):2333–2355.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric bayesian model for randomised block designs. Biometrika, 83(2):275–285.

Calabrese, E. J. and Baldwin, L. A. (2003). Hormesis: the dose-response revolution. Annual review of pharmacology and toxicology, 43(1):175–197.

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. Methods in Ecology and Evolution, 12(4):665–680.

Campbell, K. R. and Yau, C. (2017). switchde: inference of switch-like differential expression along single-cell trajectories. Bioinformatics, 33(8):1241–1242.

Campbell, L. L. (1985). The relation between information theory and the differential geometry approach to statistics. Information sciences, 35(3):199–210.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745):496–502.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In Artificial intelligence and statistics, pages 73–80. PMLR.

Casarett, L. J. et al. (2008). Casarett and Doull's toxicology: the basic science of poisons, volume 71470514. McGraw-Hill New York.

Chan, I. I., Kwok, M. K., and Schooling, C. M. (2021). The total and direct effects of systolic and diastolic blood pressure on cardiovascular disease and longevity using mendelian randomisation. Scientific Reports, 11(1):21799.

Chang, K.-Y. and Ghosh, J. (1998). Principal curves for nonlinear feature extraction and classification. In Applications of artificial neural networks in image processing III, volume 3307, pages 120–129. International Society for Optics and Photonics.

Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. The Annals of Statistics, 18(1):391–404.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. Journal of the American Statistical Association, 106(494):608–625.

Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. Journal of Computational Biology, 16(2):279–290.

Chen, P.-Y., Chen, R.-B., and Wong, W. K. (2022). Particle swarm optimization for searching efficient experimental designs: A review. Wiley Interdisciplinary Reviews: Computational Statistics, page e1578.

Chen, Y., Lin, Z., and Müller, H.-G. (2021). Wasserstein regression. Journal of the American Statistical Association, pages 1–14.

Cheng, R. and Jin, Y. (2014). A competitive swarm optimizer for large scale optimization. IEEE transactions on cybernetics, 45(2):191–204.

Cheng, R. and Jin, Y. (2015). A competitive swarm optimizer for large scale optimization. IEEE Transactions on Cybernetics, 45(2):191–204.

Cheng, S., Chun Zhao, C., Wu, J., and Shi, Y. (2015). Particle swarm optimization in regression analysis: A case study. Conference Paper in Lecture Notes in Computer Science, pages DOI: 10.1007/978–3–642–38703–66.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. Biometrika, 85(2):347–361.

Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. Journal of Machine Learning Research, 17(229):1–20.

Chu, S. C. and Tsai, P. W. (2007). Computational intelligence based on the behavior of cats. International Journal of Innovative Computing, Information and Control, 3:163–173.

Chung, K. L. (1974). A course in probability theory. Academic press.

Clyde, M. and Chaloner, K. (1996). The equivalence of constrained and weighted designs in multiple objective design problems. Journal of the American Statistical Association, 91(435):1236–1244.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. Applied and computational harmonic analysis, 21(1):5–30.

Collins, M. D., Cui, E. H., Hyun, S. W., and Wong, W. K. (2022). A model-based approach to designing developmental toxicology experiments using sea urchin embryos. Archives of toxicology, pages 1–14.

Conley, T. G., Hansen, C. B., McCulloch, R. E., and Rossi, P. E. (2008). A semi-parametric bayesian approach to the instrumental variable problem. Journal of Econometrics, 144(1):276–305.

Cook, R. D. and Wong, W. K. (1994). On the equivalence of constrained and compound optimal designs. Journal of the American Statistical Association, 89(426):687–692.

Cook, R. J. and Lawless, J. F. (2007). The statistical analysis of recurrent events. Springer.

Costa, S., Barroso, M., Castañera, A., and Dias, M. (2010). Design of experiments, a powerful tool for method development in forensic toxicology: application to the optimization of urinary morphine 3-glucuronide acid hydrolysis. Analytical and bioanalytical chemistry, 396:2533–2542.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62(2):269–276.

Cui, E. H. (2022). A tutorial on statistical models based on counting processes. arXiv preprint arXiv:2210.07114.

Cui, E. H. and Shao, S. (2024). A metric-based principal curve approach for learning one-dimensional manifold. arXiv preprint arXiv:2405.12390.

Cui, E. H., Song, D., Wong, W. K., and Li, J. J. (2022). Single-cell generalized trend model (scgtm): a flexible and interpretable model of gene expression trend along cell pseudotime. Bioinformatics, 38(16):3927–3934.

Cui, E. H., Zhang, Z., Chen, C. J., and Wong, W. K. (2024a). Applications of nature-inspired metaheuristic algorithms for tackling optimization problems across disciplines. Scientific reports, 14(1):9403.

Cui, E. H., Zhang, Z., and Wong, W. K. (2024b). Optimal designs for nonlinear mixed-effects models using competitive swarm optimizer with mutated agents. Statistics and Computing, 34(5):156.

Dabrowska, D. M. (2012). Estimation in a semi-markov transformation model. The International Journal of Biostatistics, 8(1).

Dabrowska, D. M. (2019). Elements of real analysis and advanced probability, volume 1. Lecture Notes at UCLA.

Dabrowska, D. M., Sun, G.-W., and Horowitz, M. M. (1994). Cox regression in a markov renewal model: an application to the analysis of bone marrow transplant data. Journal of the American Statistical Association, 89(427):867–877.

Daley, D. J., Vere-Jones, D., et al. (2003). An introduction to the theory of point processes: volume I: elementary theory and methods. Springer.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to aids. Biometrics, pages 1–11.

de la Calle-Arroyo, C., Amo-Salas, M., López-Fidalgo, J., Rodríguez-Aragón, L. J., and Wong, W. K. (2023). A methodology to d-augment experimental designs. Chemometrics and Intelligent Laboratory Systems, 237:104822.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.

Dette, H. and Trampisch, M. (2012). Optimal designs for quantile regression models. Journal of the American Statistical Association, 107(499):1140–1151.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. Statistical methods in medical research, 16(4):309–330.

Do Carmo, M. P. and Flaherty Francis, J. (1992). Riemannian geometry, volume 6. Springer.

Dobson, A. J. and Barnett, A. G. (2018). An introduction to generalized linear models. Chapman and Hall/CRC.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences, 100(10):5591–5596.

Dorigo, M., Birattari, M., and Stutzle, T. (2006). Ant colony optimization. IEEE computational intelligence magazine, 1(4):28–39.

Dorigo, M., Maniezzo, V., and Colorni, A. (1991). Positive feedback as a search strategy. Technical Report.

Dragalin, V. and Fedorov, V. (2006). Adaptive designs for dose-finding based on efficacy–toxicity response. Journal of Statistical Planning and Inference, 136(6):1800–1823.

Dragalin, V., Fedorov, V., and Wu, Y. (2008a). Adaptive designs for selecting drug combinations based on efficacy–toxicity response. Journal of Statistical Planning and Inference, 138(2):352–373.

Dragalin, V., Fedorov, V. V., and Wu, Y. (2008b). Two-stage design for dose-finding that accounts for both efficacy and safety. Statistics in Medicine, 27(25):5156–5176.

Drigo, M. (1996). The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B, 26(1):1–13.

Dryden, I. L. and Mardia, K. V. (2016). Statistical shape analysis: with applications in R, volume 995. John Wiley & Sons.

Du, M., Zhou, Q., Zhao, S., and Sun, J. (2021). Regression analysis of case-cohort studies in the presence of dependent interval censoring. Journal of applied statistics, 48(5):846–865.

Dueck, G. and Scheuer, T. (1990). Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. Journal of computational physics, 90(1):161–175.

Eaton, M. L. (1981). On the projections of isotropic distributions. The Annals of Statistics, pages 391–400.

Eberhart, R. and Kennedy, J. (1995). A new optimizer using particle swarm theory. In MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pages 39–43. Ieee.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). The Annals of Statistics, pages 1189–1242.

Efron, B. (1978). The geometry of exponential families. The Annals of Statistics, pages 362–376.

Elashoff, R., Li, N., and Li, G. (2016). Joint modeling of longitudinal and time-to-event data. CRC press.

Embretson, S. E. and Reise, S. P. (2013). Item response theory. Psychology Press.

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. Jama, 318(19):1925–1926.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90(430):577–588.

Everitt, B. (2012). Introduction to optimization methods and their application in statistics. Springer science & business media.

Ezugwu, A. E., Agushaka, J. O., Abualigah, L., Mirjalili, S., and Gandomi, A. H. (2022). Prairie dog optimization algorithm. Neural Computing and Applications, 34(22):20017–20065.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. (2019). Reparameterizing distributions on lie groups. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 3244–3253. PMLR.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360.

Fanelli, D. and Piazza, F. (2020). Analysis and forecast of covid-19 spreading in china, italy and france. Chaos, Solitons & Fractals, 134:109761.

Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. Journal of statistical software, 36(2).

FDA (2003). Guidance for industry: exposure-response relationships-study design, data analysis, and regulatory applications. http://www. fda. gov/cber/gdlns/exposure. pdf.

Federov, V. (1972). Theory of Optimal Experiments, translated and edited by WJ. Elsevier.

Fedorov, V. (2010). Optimal experimental design. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5):581–589.

Fedorov, V. V. (1971). The design of experiments in the multiresponse case. Theory of Probability & Its Applications, 16(2):323–332.

Fedorov, V. V. and Hackl, P. (2012). Model-oriented design of experiments, volume 125. Springer Science & Business Media.

Fedorov, V. V. and Leonov, S. L. (2013). Optimal design for nonlinear response models. CRC Press.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2):209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In Recent advances in statistics, pages 287–302. Elsevier.

Fewell, Z., Davey Smith, G., and Sterne, J. A. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. American journal of epidemiology, 166(6):646–655.

Fischer, D. S., Theis, F. J., and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. Nucleic acids research, 46(20):e119–e119.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE transactions on medical imaging, 23(8):995–1005.

Flynn, C. J., Hurvich, C. M., and Simonoff, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. Journal of the American Statistical Association, 108(503):1031–1043.

Fogel, D. B. (1998). Artificial intelligence through simulated evolution. Wiley-IEEE Press.

Foster, E. M. (1997). Instrumental variables for logistic regression: an illustration. Social Science Research, 26(4):487–504.

Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of gaussian graphical model. Statistica Sinica, pages 1123–1146.

Garetto, M., Leonardi, E., and Torrisi, G. L. (2021). A time-modulated hawkes process to model the spread of covid-19 and the impact of countermeasures. Annual reviews in control, 51:551–563.

Gedon, D., Ribeiro, A. H., Wahlstörm, N., and Schön, T. B. (2023). Invertible kernel pca with random fourier features. IEEE Signal Processing Letters.

Geem, Z. W., Kim, J. H., and Loganathan, G. V. (2001). A new heuristic optimization algorithm: harmony search. simulation, 76(2):60–68.

Gentleman, R. and Vandal, A. (2010). Icens: Npmle for censored and truncated data. R package version, 1(0).

Gerber, S. and Whitaker, R. (2013). Regularization-free principal curve estimation. The Journal of Machine Learning Research, 14(1):1285–1302.

Gertsch, W. and Wong, W. K. (2024). An interactive tool for designing efficient toxicology experiments. Archives of Toxicology, 98(3):1015–1022.

Ghosal, S. and Van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, volume 44. Cambridge University Press.

Giles, J. (2006). Animal experiments under fire for poor design. Nature, 444(7122):981–982.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. The annals of statistics, pages 1501–1555.

Glonek, G. F. and McCullagh, P. (1995). Multivariate logistic models. Journal of the Royal Statistical Society: Series B (Methodological), 57(3):533–546.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. Computers & operations research, 13(5):533–549.

Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. Biometrics, 56(3):940–943.

Goggins, W. B., Finkelstein, D. M., and Zaslavsky, A. M. (1999). Applying the cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. Biometrics, 55(2):445–451.

Gogna, A. and Tayal, A. (2013). Metaheuristics: review and application. Journal of Experimental & Theoretical Artificial Intelligence, 25(4):503–526.

Goh, A. and Vidal, R. (2008). Clustering and dimensionality reduction on riemannian manifolds. In 2008 IEEE Conference on computer vision and pattern recognition, pages 1–7. IEEE.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. Econometrica, 40(6):979–1001.

Gollapudi, B., Johnson, G., Hernandez, L., Pottenger, L., Dearfield, K., Jeffrey, A., Julien, E., Kim, J., Lovell, D., Macgregor, J., et al. (2013). Quantitative approaches for assessing dose–response relationships in genetic toxicology studies. Environmental and molecular mutagenesis, 54(1):8–18.

Gómez, G., Espinal, A., and Lagakos, S. (2003). Inference for a linear regression model with an interval-censored covariate. Statistics in medicine, 22(3):409–425.

Gray, R. and Wheatley, K. (1991). How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplant, 7, Suppl 3:9–12.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. Epidemiology, pages 37–48.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. Scandinavian Journal of Statistics, 35(3):385–399.

Gu, S., Cheng, R., and Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 22(3):811–822.

Gumbel, E. J. (1960). Bivariate exponential distributions. Journal of the American Statistical Association, 55(292):698–707.

Gustafson, P. (2007). Measurement error modelling with an approximate instrumental variable. Journal of the Royal Statistical Society: Series B, 69(5):797–815.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. Econometrica, 11:1–12.

Haines, L. M., Kabera, G., and O'Brien, T. E. (2007). D-optimal designs for logistic regression in two variables. In mODa 8-Advances in Model-Oriented Design and Analysis, pages 91–98. Springer.

Haines, L. M., Perevozskaya, I., and Rosenberger, W. F. (2003). Bayesian optimal designs for phase i clinical trials. Biometrics, 59(3):591–600.

Hall, P., Lee, E. R., and Park, B. U. (2009). Bootstrap-based penalty choice for the lasso, achieving oracle performance. Statistica Sinica, pages 449–471.

Haouari, M. and Mhiri, M. (2021). A particle swarm optimization approach for predicting the number of covid-19 deaths. Scientific Reports, 11(1):1–13.

Hardin, J. W. (2002). The robust variance estimator for two-stage models. Stata Journal, 2(3):253–266(14).

Hardin, J. W. and Carroll, R. J. (2003). Variance estimation for the instrumental variables approach to measurement error in generalized linear models. Stata Journal, 3(4):342–350(9).

Hartigan, J. (1969). Linear bayesian methods. Journal of the Royal Statistical Society: Series B (Methodological), 31(3):446–454.

Hartung, T. and Rovida, C. (2009). Toxicology for the twenty-first century. Nature, 460:208–212.

Hashim, F. A. and Hussien, A. G. (2022). Snake optimizer: A novel meta-heuristic optimization algorithm. Knowledge-Based Systems, 242:108320.

Hastie, T. (1984). Principal curves and surfaces. Technical report, STANFORD UNIV CA LAB FOR COMPUTATIONALSTATISTICS.

Hastie, T., Qian, J., and Tay, K. (2021). An introduction to glmnet. CRAN R Repositary, 5:1–35.

Hastie, T. and Stuetzle, W. (1989). Principal curves. Journal of the American Statistical Association, 84(406):502–516.

Hastie, T. and Tibshirani, R. (1995). Discriminant adaptive nearest neighbor classification and regression. Advances in neural information processing systems, 8.

Hauberg, S., Freifeld, O., and Black, M. (2012). A geometric take on metric learning. Advances in Neural Information Processing Systems, 25.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1):83–90.

Heckman, J. J. (2008). Econometric causality. Working Paper 13934, National Bureau of Economic Research.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. Journal of Econometrics, 30(1-2):239–267.

Hernán, M. A. and Cole, S. R. (2009). Invited commentary: causal diagrams and measurement bias. American journal of epidemiology, 170(8):959–962.

Hoel, D. G. and Walburg Jr, H. (1972). Statistical analysis of survival experiments. Journal of the National Cancer Institute, 49(2):361–372.

Holland-Letz, T. and Kopp-Schneider, A. (2015). Optimal experimental designs for dose–response studies with continuous endpoints. Archives of toxicology, 89:2059–2068.

Homrighausen, D. and McDonald, D. J. (2018). A study on tuning parameter selection for the high-dimensional lasso. Journal of Statistical Computation and Simulation, 88(15):2865–2892.

Hoogerheide, L., Kleibergen, F., and van Dijk, H. K. (2007). Natural conjugate priors for the instrumental variables regression model applied to the Angrist Krueger data. Journal of Econometrics, 138:63–103.

Hsiao, C. (1983). Regression analysis with a categorized explanatory variable. In Studies in econometrics, time series, and multivariate statistics, pages 93–129. Elsevier.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics, pages 123–169. Springer.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the berkeley symposium on mathematical statistics and Probability.

Hughes-Oliver, J. M. and Rosenberger, W. F. (2000). Efficient estimation of the prevalence of multiple rare traits. Biometrika, 87(2):315–327.

Hui, F. K., Warton, D. I., and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using eric. Journal of the American Statistical Association, 110(509):262–269.

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M., and Visser, P. M. (2018). Cyanobacterial blooms. Nature Reviews Microbiology, 16(8):471–483.

Hussain, K., Salleh, M. N. M., Cheng, S., and Shi, Y. (2019). Metaheuristic research: a comprehensive survey. Artificial Intelligence Review, 52(4):2191–2233.

Hyun, S. W. and Wong, W. K. (2015). Multiple-objective optimal designs for studying the dose response function and interesting dose levels. The international journal of biostatistics, 11(2):253–271.

Hyun, S. W., Wong, W. K., and Yang, Y. (2018). Vnm: An r package for finding multiple-objective optimal designs for the 4-parameter logistic model. Journal of Statistical Software, 83:1–19.

Imbens, G. W. and Rubin, D. B. (2010). Rubin causal model. In Microeconometrics, pages 229–241. Springer.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for Stick-Breaking priors. Journal of the American Statistical Association, 96(453):161–173.

Izenman, A. J. (2008). Modern multivariate statistical techniques, volume 1. Springer.

Jacod, J. (1975). Multivariate point processes: predictable projection, radon-nikodym derivatives, representation of martingales. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 31(3):235–253.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). An introduction to statistical learning, volume 112. Springer.

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. Annals of applied statistics, 4(4):2126–2149.

Ji, S., Peng, L., Cheng, Y., and Lai, H. (2012). Quantile regression for doubly censored data. Biometrics, 68(1):101–112.

Ji, Z. and Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucleic acids research, 44(13):e117–e117.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. conference in modern analysis and probability (new haven, conn., 1982), 189–206. In Contemp. Math, volume 26.

Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and'race'. Nature genetics, 36(Suppl 11):S28–S33.

Jóźwiak, K. and Moerbeek, M. (2013). Podse: A computer program for optimal design of trials with discrete-time survival endpoints. Computer methods and programs in biomedicine, 111(1):115–127.

Kabera, M. G. and Haines, L. M. (2012). A note on the construction of locally d-and ds-optimal designs for the binary logistic model with several explanatory variables. Statistics & Probability Letters, 82(5):865–870.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. Journal of the American statistical Association, 111(513):132–144.

Karatzas, I. and Shreve, S. (1991). Brownian motion and stochastic calculus, volume 113. Springer Science & Business Media.

Kaso, A. W., Agero, G., Hurissa, Z., Kaso, T., Ewune, H. A., Hareru, H. E., and Hailu, A. (2022). Survival analysis of covid-19 patients in ethiopia: A hospital-based study. Plos one, 17(5):e0268280.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, volume 4, pages 1942–1948. IEEE.

Kiefer, J. (1959). Optimum experimental designs. Journal of the Royal Statistical Society: Series B (Methodological), 21(2):272–304.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). The annals of Statistics, pages 849–879.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. Canadian Journal of Mathematics, 12:363–366.

Kim, M. Y., De Gruttola, V. G., and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to aids. Biometrics, pages 13–22.

Kim, S. and Wong, W. K. (2018). Extended two-stage adaptive designs with three target responses for phase ii clinical trials. Statistical methods in medical research, 27(12):3628–3642.

Kjaersgaard, M. I. and Parner, E. T. (2016). Instrumental variable method for time-to-event data using a pseudo-observation approach. Biometrics, 72(2):463–472.

Kleibergen, F. and Van Dijk, H. K. (1998). Bayesian simultaneous equations analysis using reduced rank structures. Econometric Theory, 14:701–743.

Klein, J. P. and Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data, volume 1230. Springer.

Ko, S., German, C. A., Jensen, A., Shen, J., Wang, A., Mehrotra, D. V., Sun, Y. V., Sinsheimer, J. S., Zhou, H., and Zhou, J. J. (2022). Gwas of longitudinal trajectories at biobank scale. The American Journal of Human Genetics, 109(3):433–445.

Kochurov, M., Karimov, R., and Kozlukov, S. (2020). Geoopt: Riemannian optimization in pytorch. arXiv preprint arXiv:2005.02819.

Komárek, A. and Lesaffre, E. (2006). Bayesian semi-parametric accelerated failure time model for paired doubly interval-censored data. Statistical Modelling, 6(1):3–22.

Komárek, A. and Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. Journal of the American Statistical Association, 103(482):523–533.

Konikoff, J., Brookmeyer, R., Longosz, A. F., Cousins, M. M., Celum, C., Buchbinder, S. P., Seage III, G. R., Kirk, G. D., Moore, R. D., Mehta, S. H., et al. (2013). Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of hiv incidence in the united states. PloS one, 8(12):e82772.

Konopka, T. and Konopka, M. T. (2018). R-package: umap. Uniform Manifold Approximation and Projection.

Korani, W. and Mouhoub, M. (2021). Review on nature-inspired algorithms. In Operations Research Forum, volume 2, pages 1–26. Springer.

Koutra, V., Gilmour, S. G., and Parker, B. M. (2021). Optimal block designs for experiments on networks. Journal of the Royal Statistical Society Series C: Applied Statistics, 70(3):596–618.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., et al. (2020). Package 'caret'. The R Journal, 223(7).

Kume, A., Dryden, I. L., and Le, H. (2007). Shape-space smoothing splines for planar landmark data. Biometrika, 94(3):513–528.

Kwon, S., Lee, S., and Na, O. (2017). Tuning parameter selection for the adaptive lasso in the autoregressive model. Journal of the Korean Statistical Society, 46(2):285–297.

Lai, T. L. (2003). Stochastic approximation. The annals of Statistics, 31(2):391–406.

Lane, A. (2020). Adaptive designs for optimal observed fisher information. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(4):1029–1058.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. The Annals of Statistics, 24(2):726–752.

Lange, K. (2013). Optimization, volume 95. Springer Science & Business Media.

Larsen, R. B., Jouffroy, J., and Lassen, B. (2016). On the premature convergence of particle swarm optimization. In 2016 European control conference (ECC), pages 1922–1927. IEEE.

Laub, P. J., Lee, Y., and Taimre, T. (2021). The elements of Hawkes processes. Springer.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine, 27(8):1133–63.

Lazar, N. A. (2021). A review of empirical likelihood. Annual Review of Statistics and its Application, 8:329–344.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

Lederer, J. and Müller, C. (2015). Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the trex. In Proceedings of the AAAI conference on artificial intelligence, volume 29.

Lee, P. H. and Burstyn, I. (2016). Identification of confounder in epidemiologic data contaminated by measurement error in covariates. BMC medical research methodology, 16:1–18.

Lee, Y., Kennedy, E. H., and Mitra, N. (2023). Doubly robust nonparametric instrumental variable estimators for survival outcomes. Biostatistics, 24(2):518–537.

Levina, E. and Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. Advances in neural information processing systems, 17.

Levine, R. A. and Casella, G. (2001). Implementations of the monte carlo em algorithm. Journal of Computational and Graphical Statistics, 10(3):422–439.

Li, G. and Lu, X. (2015). A bayesian approach for instrumental variable analysis with censored time-to-event outcome. Statistics in medicine, 34(4):664–684.

Li, J., Fine, J., and Brookhart, A. (2015a). Instrumental variable additive hazards models. Biometrics, 71(1):122–130.

Li, S. and Peng, L. (2023). Instrumental variable estimation of complier causal treatment effect with interval-censored data. Biometrics, 79(1):253–263.

Li, W. V. and Li, J. J. (2018). Modeling and analysis of rna-seq data: a review from a statistical perspective. Quantitative Biology, 6(3):195–209.

Li, Y., Wei, Y., and Chu, Y. (2015b). Research on solving systems of nonlinear equations based on improved pso. Mathematical Problems in Engineering, 2015:1–10.

Liggett, T. M. (2010). Continuous time Markov processes: an introduction, volume 113. American Mathematical Soc.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics, pages 948–963.

Linacre, J. M. (2022). R statistics: survey and review of packages for the estimation of Rasch models. Int J Med Educ., 13:171–175.

Little, R. J. and Rubin, D. B. (2019). Statistical analysis with missing data, volume 793. John Wiley & Sons.

Liu, Y., Magnus, B., O'Connor, H., and Thissen, D. (2018). Multidimensional item response theory. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, pages 445–493.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. The annals of statistics, pages 351–357.

Luo, S. and Chen, Z. (2013). Extended bic for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. Journal of Statistical Planning and Inference, 143(3):494–504.

Maathuis, M. and Maathuis, M. M. (2022). Package 'mlecens'. menopause, 12:1.

Magwene, P. M., Lizardi, P., and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. Bioinformatics, 19(7):842–850.

Mahalanobis, P. C. (2018). On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-), 80:S1–S7.

Marcot, B. G. and Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete bayesian network analysis? Computational Statistics, 36(3):2009–2031.

Mardia, K. V., Jupp, P. E., and Mardia, K. (2000). Directional statistics, volume 2. Wiley Online Library.

Martins, L. F. and Gabriel, V. J. (2014). Linear instrumental variables model averaging estimation. Computational Statistics & Data Analysis, 71:709–724.

Martinussen, T., Nørbo Sørensen, D., and Vansteelandt, S. (2019). Instrumental variables estimation under a structural cox model. Biostatistics, 20(1):65–79.

Martinussen, T. and Vansteelandt, S. (2020). Instrumental variables estimation with competing risk data. Biostatistics, 21(1):158–171.

Martinussen, T., Vansteelandt, S., Tchetgen Tchetgen, E. J., and Zucker, D. M. (2017). Instrumental variables estimation of exposure effects on a time-to-event endpoint using structural cumulative survival models. Biometrics, 73(4):1140–1149.

Marx, K. (2000). Karl Marx: selected writings. Oxford University Press, USA.

Masci, J., Boscaini, D., Bronstein, M., and Vandergheynst, P. (2015). Geodesic convolutional neural networks on riemannian manifolds. In Proceedings of the IEEE international conference on computer vision workshops, pages 37–45.

Matabuena, M., Petersen, A., Vidal, J. C., and Gude, F. (2021). Glucodensities: a new representation of glucose profiles using distributional data analysis. Statistical methods in medical research, 30(6):1445–1464.

McCullagh, P. and Nelder, J. A. (2019). Generalized linear models. Routledge.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. Statistical Methods in Medical Research, 18(2):195–222.

Melis, G. G., Marhuenda-Muñoz, M., and Langohr, K. (2023). Regression analysis with interval-censored covariates. application to liquid chromatography. Emerging Topics in Modeling Interval-Censored Survival Data, page 271.

Mendes, J. M., Oliveira, P. M., Filipe Neves, F. N., and dos Santos, R. M. (2020). Nature inspired metaheuristics and their applications in agriculture: A short review. EPIA Conference on Artificial Intelligence EPIA 2019: Progress in Artificial Intelligence, pages 167–179.

Mingyue, Q., Tao, H., and Hengjian, C. (2020). Parametric estimation for the incubation period distribution of covid-19 under doubly interval censoring. Acta Mathematicae Applicatae Sinica, 43(2):200–210.

Miolane, N., Guigui, N., Le Brigant, A., Mathe, J., Hou, B., Thanwerdas, Y., Heyder, S., Peltre, O., Koep, N., Zaatiti, H., et al. (2020a). Geomstats: a python package for riemannian geometry in machine learning. The Journal of Machine Learning Research, 21(1):9203–9211.

Miolane, N., Guigui, N., Zaatiti, H., Shewmake, C., Hajri, H., Brooks, D., Le Brigant, A., Mathe, J., Hou, B., Thanwerdas, Y., et al. (2020b). Introduction to geometric learning in python with geomstats. In SciPy 2020-19th Python in Science Conference, pages 48–57.

Miranda, L. J. (2018). Pyswarms: a research toolkit for particle swarm optimization in python. Journal of Open Source Software, 3(21):433.

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69:46–61.

Mohanty, S. D. and Fahnestock, E. (2021). Adaptive spline fitting with particle swarm optimization. Computational Statistics, 36(1):155–191.

Mohapatra, P., Das, K. N., and Roy, S. (2017). A modified competitive swarm optimizer for large scale optimization problems. Applied Soft Computing, 59:340–362.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of hawkes processes. Advances in applied probability, 37(3):629–646.

Mondal, P. K., Saha, U. S., and Mukhopadhyay, I. (2021). Pseudoga: cell pseudotime reconstruction based on genetic algorithm. Nucleic Acids Research.

Morgan, M. S. (1991). The History of Econometric Ideas. Cambridge University Press.

Morrison, D., Laeyendecker, O., and Brookmeyer, R. (2022). Regression with interval-censored covariates: Application to cross-sectional incidence estimation. Biometrics, 78(3):908–921.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). Bayesian nonparametric data analysis, volume 1. Springer.

Murphy, K. M. and Topel, R. H. (1985). Estimation and inference in two-step econometric models. Journal of Business and Economic Statistics, 3(4):370–379.

Murray, M. K. and Rice, J. W. (1993). Differential geometry and statistics, volume 48. CRC Press.

Nagao, M. and Kadoya, M. (1971). Two-variate exponential distribution and its numerical table for engineering application. Bulletin of the Disaster Prevention Research Institute, 20(3):183–215.

Naimi, T. S., Brown, D. W., Brewer, R. D., Giles, W. H., Mensah, G., Serdula, M. K., Mokdad, A. H., Hungerford, D. W., Lando, J., Naimi, S., et al. (2005). Cardiovascular risk factors and confounders among nondrinking and moderate-drinking us adults. American journal of preventive medicine, 28(4):369–373.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265.

Normandin, M. E., Mohanty, S. D., and Weerathunga, T. S. (2018). Particle swarm optimization based search for gravitational waves from compact binary coalescences: Performance improvements. Physical Review D, 98(4):044029.

Ogata, Y. (1981). On lewis' simulation method for point processes. IEEE transactions on information theory, 27(1):23–31.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. Annals of the Institute of Statistical Mathematics, 50:379–402.

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. Annals of the Institute of Statistical Mathematics, 31:145–155.

Ozertem, U. and Erdogmus, D. (2011). Locally defined principal curves and surfaces. The Journal of Machine Learning Research, 12:1249–1286.

Pan, C., Cai, B., and Wang, L. (2020). A bayesian approach for analyzing partly interval-censored data under the proportional hazards model. Statistical methods in medical research, 29(11):3192–3204.

Pant, S., Kumar, A., and M., R. (2019). Solution of nonlinear systems of equations via metaheuristics. International Journal of Mathematical, Engineering and Management Sciences, 4:1108–1126.

Park, D. and Faraway, J. J. (1998). Sequential design for response curve estimation. Journal of Nonparametric Statistics, 9(2):155–164.

Park, J., Bakoyannis, G., and Yiannoutsos, C. T. (2019). Semiparametric competing risks regression under interval censoring using the r package intccr. Computer methods and programs in biomedicine, 173:167–176.

Pázman, A. (1986). Foundations of optimum experimental design, volume 14. Springer.

Pearl, J. (2000). Causality: models, reasoning, and inference, chapter 7. Cambridge University Press, New York, NY, USA.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). Causal inference in statistics: A primer. John Wiley & Sons.

Pelletier, B. (2006). Non-parametric regression estimation on closed riemannian manifolds. Journal of Nonparametric Statistics, 18(1):57–67.

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision, 25:127–154.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with euclidean predictors. The Annals of Statistics, 47(2):691–719.

Pincus, M. (1970). Letter to the editor—a monte carlo method for the approximate solution of certain types of constrained optimization problems. Operations research, 18(6):1225–1228.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. The Annals of Probability, pages 855–900.

Pshenichnyi, B. N. (2020). Necessary conditions for an extremum. CRC Press.

Qiu, J. and Wong, W. K. (2023). Nature-inspired metaheuristics for finding optimal designs for the continuation-ratio models. The New England Journal of Statistics in Data Science, 2(1):15–29.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nature methods, 14(10):979–982.

Rabadán, R. and Blumberg, A. J. (2019). Topological data analysis for genomics and evolution: topology in biology. Cambridge University Press.

Rabinowitz, D. and Jewell, N. P. (1996). Regression with doubly censored current status data. Journal of the Royal Statistical Society: Series B (Methodological), 58(3):541–550.

Rajwar, K., Deep, K., and Das, S. (2023). An exhaustive review of the metaheuristic algorithms for search and optimization: Taxonomy, applications, and open challenges. Artificial Intelligence Review, 56(11):13187–13257.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. Reson. J. Sci. Educ, 20:78–90.

Ravishanker, N., Melnick, E. L., and Tsai, C.-L. (1990). Differential geometry of arma models. Journal of Time Series Analysis, 11(3):259–274.

Ren, X. and Kuan, P.-F. (2020). Negative binomial additive model for rna-seq data analysis. BMC bioinformatics, 21(1):1–15.

Rényi, A. (1959). On measures of dependence. Acta mathematica hungarica, 10(3-4):441–451.

Ritz, C., Baty, F., Streibig, J. C., and Gerhard, D. (2015). Dose-response analysis using r. PLOS ONE, 10(12):e0146021.

Riza, L. S., Nugroho, E. P., et al. (2018). Metaheuristicopt: A R package for optimisation based on meta-heuristics algorithms. Pertanika Journal of Science & Technology, 26(3).

Riza, L. S., Nugroho, E. P., et al. (2019). Metaheuristicopt: Metaheuristic for optimization. R package version 1.0. 0, 2017.

Rizoiu, M.-A., Mishra, S., Kong, Q., Carman, M., and Xie, L. (2018). Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In Proceedings of the 2018 world wide web conference, pages 419–428.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, pages 400–407.

Robert, C. P., Casella, G., and Casella, G. (1999). Monte Carlo statistical methods, volume 2. Springer.

Robins, J. M. and Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in statistics-Theory and Methods, 20(8):2609–2631.

Robitzsch, A. (2021). A comprehensive simulation study of estimation methods for the Rasch model. Stats, 4(4):814–836.

Rockafellar, R. T. (1970). Convex analysis, volume 18. Princeton university press.

Roodman, D. (2011). Fitting fully observed recursive mixed-process models with cmp. The Stata Journal, 11(2):159–206.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500):2323–2326.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3):581–592.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. Annals of internal medicine, 127(8_Part_2):757–763.

Rudin, W. (1976). Principles of mathematical analysis, volume 3. McGraw-hill New York.

Rustagi, J. S. (2014). Optimization techniques in statistics. Elsevier.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell rna-seq: advances and future challenges. Nucleic acids research, 42(14):8845–8860.

Sampson, J. R. (1976). Adaptation in natural and artificial systems (john h. holland).

Sander, J., Schultze, J. L., and Yosef, N. (2017). Impulsede: detection of differentially expressed genes in time series data using impulse models. Bioinformatics, 33(5):757–759.

Schoen, E. (1996). Statistical designs in combination toxicology: a matter of choice. Food and chemical toxicology, 34(11-12):1059–1065.

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In International conference on artificial neural networks, pages 583–588. Springer.

Schwaab, M., Silva, F. M., Queipo, C. A., Barreto Jr, A. G., Nele, M., and Pinto, J. C. (2006). A new approach for sequential experimental design for model discrimination. Chemical engineering science, 61(17):5791–5806.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, pages 461–464.

Sengewald, M.-A., Steiner, P. M., and Pohl, S. (2019). When does measurement error in covariates impact causal effect estimates? analytic derivations of different scenarios and an empirical illustration. British Journal of Mathematical and Statistical Psychology, 72(2):244–270.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., et al. (2014). Single-cell rna-seq reveals dynamic paracrine control of cellular variation. Nature, 510(7505):363–369.

Sheehy, D. R. (2012). Linear-size approximations to the vietoris-rips filtration. In Proceedings of the twenty-eighth annual symposium on Computational geometry, pages 239–248.

Shi, Y., Zhang, Z., and Wong, W. K. (2019). Particle swarm based algorithms for finding locally and bayesian d-optimal designs. Journal of Statistical Distributions and Applications, 6(1):1–17.

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Enikolopov, G., Nauen, D. W., Christian, K. M., Ming, G.-l., et al. (2015). Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell stem cell, 17(3):360–372.

Shu, D. and Yi, G. Y. (2019). Causal inference with measurement error in outcomes: Bias analysis and estimation methods. Statistical methods in medical research, 28(7):2049–2068.

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. Computational and structural biotechnology journal, 18:2789.

Silvey, S. (2013). Optimal design: an introduction to the theory for parameter estimation, volume 1. Springer Science & Business Media.

Simo-Serra, E., Torras, C., and Moreno-Noguer, F. (2017). 3d human pose tracking priors using geodesic mixture models. International Journal of Computer Vision, 122:388–408.

Singh, G., Mémoli, F., Carlsson, G. E., et al. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. PBG@ Eurographics, 2:091–100.

Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., and Larmarange, J. (2021). Reproducible summary tables with the gtsummary package. The R journal, 13(1):570–580.

Skovgaard, L. T. (1984). A riemannian geometry of the multivariate normal model. Scandinavian journal of statistics, pages 211–223.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika, 12(1/2):1–85.

Song, D. and Li, J. J. (2021). Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. Genome biology, 22(1):1–25.

Sörensen, K. (2015). Metaheuristics—the metaphor exposed. International Transactions in Operational Research, 22(1):3–18.

Stacey, A. W. (2007). An adaptive Bayesian approach to Bernoulli-response clinical trials. Brigham Young University.

Stokes, Z., Mandal, A., and Wong, W. K. (2020). Using differential evolution to design optimal experiments. Chemometrics and Intelligent Laboratory Systems, 199:103955.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences, 102(36):12837–12842.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4):341–359.

Straub, J., Chang, J., Freifeld, O., and Fisher III, J. (2015). A dirichlet process mixture model for spherical data. In Artificial Intelligence and Statistics, pages 930–938. PMLR.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics, 19(1):1–16.

Sun, C., Ding, J., Zeng, J., and Jin, Y. (2016). A fitness approximation assisted competitive swarm optimizer for large scale expensive optimization problems. Memetic Computing, pages 1–12.

Sun, J. (2001). Nonparametric test for doubly interval-censored failure time data. Lifetime Data Analysis, 7(4):363.

Sun, J. (2006). The statistical analysis of interval-censored failure time data, volume 3. Springer.

Sun, J., Liao, Q., and Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to aids studies. Biometrics, 55(3):909–914.

Sun, L., Kim, Y.-j., and Sun, J. (2004). Regression analysis of doubly censored failure time data using the additive hazards model. Biometrics, 60(3):637–643.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. Biometrika, 99(4):879–898.

Sverdlov, O., Ryeznik, Y., and Wong, W. K. (2020). On optimal designs for clinical trials: an updated review. Journal of Statistical Theory and Practice, 14(1):1–29.

Swanson, S. A. and Hernán, M. A. (2013). Commentary: how to report instrumental variable analyses (suggestions welcome). Epidemiology, 24(3):370–374.

Talbi, E.-G. (2009). Metaheuristics: from design to implementation. John Wiley & Sons.

Tapp, K. (2016). Differential geometry of curves and surfaces. Springer.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. Epidemiology (Cambridge, Mass.), 26(3):402.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. science, 290(5500):2319–2323.

Theil, H. (1958). Economic forecasts and policy. North-Holland, Amsterdam.

Tian, T. and Sun, J. (2022). Variable selection for nonparametric additive cox model with interval-censored data. Biometrical Journal.

Tian, Z. and Fong, S. (2016). Survey of meta-heuristic algorithms for deep learning training. Optimization algorithms—methods and applications.

Tibshirani, R. (1992). Principal curves revisited. Statistics and computing, 2:183–190.

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. Statistical applications in genetics and molecular biology, 8(1).

Tong, X. T., Choi, K. P., Lai, T. L., and Wong, W. K. (2021). Stability bounds and almost sure convergence of improved particle swarm optimization methods. Research in the Mathematical Sciences, 8(2):30.

Topp, R. and Gómez, G. (2004). Residual analysis in linear regression models with an interval-censored covariate. Statistics in medicine, 23(21):3377–3391.

Townsend, J., Koep, N., and Weichwald, S. (2016). Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. arXiv preprint arXiv:1603.03236.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology, 32(4):381–386.

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology, 102(6):e03336.

Tse-Tung, M. (2014). Selected Works of Mao Tse-Tung: Volume 5, volume 5. Elsevier.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological), 38(3):290–295.

U.S. Department of Health and Human Services (2015). Product Development Under the Animal Rule: Guidance for Industry. Food and Drug Administration (FDA), Silver Spring, MD. Available online: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

Valian, E., Mohanna, S., and Tavakoli, S. (2011). Improved cuckoo search algorithm for feedforward neural network training. International Journal of Artificial Intelligence & Applications, 2(3):36–43.

Valian, E., Tavakoli, S., Mohanna, S., and Haghi, A. (2013). Improved cuckoo search for reliability optimization problems. Computers & Industrial Engineering, 64(1):459–468.

Van den Berge, K., De Bezieux, H. R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. Nature communications, 11(1):1–13.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. Journal of machine learning research, 9(11).

Van Loan, C. F. and Golub, G. (1996). Matrix computations (johns hopkins studies in mathematical sciences). Matrix Computations, 53.

VanderWeele, T. J., Rothman, K. J., and Lash, T. L. (2021). Confounding and confounders. Modern epidemiology, pages 263–286.

VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., and Kraft, P. (2014). Methodological challenges in mendelian randomization. Epidemiology (Cambridge, Mass.), 25(3):427.

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press.

Wan, E. Y. F., Fung, W. T., Schooling, C. M., Au Yeung, S. L., Kwok, M. K., Yu, E. Y. T., Wang, Y., Chan, E. W. Y., Wong, I. C. K., and Lam, C. L. K. (2021). Blood pressure and risk of cardiovascular disease in uk biobank: a mendelian randomization study. Hypertension, 77(2):367–375.

Wang, F.-K. and Huang, P.-R. (2014). Implementing particle swarm optimization algorithm to estimate the mixture of two weibull parameters with censored data. Journal of Statistical Computation and Simulation, 84(9):283–300.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. Journal of the Royal Statistical Society Series B: Statistical Methodology, 71(3):671–683.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. Annual Review of Statistics and its application, 3:257–295.

Wang, K., Yang, F., Porter, D. W., and Wu, N. (2013). Two-stage experimental design for dose–response modeling in toxicology studies. ACS sustainable chemistry & engineering, 1(9):1119–1128.

Wang, L., Peng, B., Bradic, J., Li, R., and Wu, Y. (2020a). A tuning-free robust and efficient approach to high-dimensional regression. Journal of the American Statistical Association, 115(532):1700–1714.

Wang, L., Tchetgen Tchetgen, E., Martinussen, T., and Vansteelandt, S. (2023a). Instrumental variable estimation of the causal hazard ratio. Biometrics, 79(2):539–550.

Wang, Q., Wang, L., and Wang, L. (2023b). Bayesian instrumental variable estimation in linear measurement error models. Canadian Journal of Statistics.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. Journal of Multivariate Analysis, 102(7):1141–1151.

Wang, W., Vilella, F., Alama, P., Moreno, I., Mignardi, M., Isakova, A., Pan, W., Simon, C., and Quake, S. R. (2020b). Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. Nature Medicine, 26(10):1644–1653.

Wang, Y. and Mohanty, S. D. (2010). Particle swarm optimization and gravitational wave data analysis: Performance on a binary inspiral testbed. Physical Review D, 81(6):063002.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics: The official journal of the International Environmetrics Society, 16(3):275–289.

Wasserman, L. (2018). Topological data analysis. Annual Review of Statistics and Its Application, 5:501–532.

Wehby, G. L., Ohsfeldt, R. L., and Murray, J. C. (2008). 'mendelian randomization' equals instrumental variable analysis with genetic instruments. Statistics in Medicine, 27:2745–2749.

Whitacre, J. M. (2011a). Recent trends indicate rapid growth of nature-inspired optimization in academia and industry. Computing, 93:121–133.

Whitacre, J. M. (2011b). Survival of the flexible: Explaining the recent dominance of nature-inspired optimization within a rapidly evolving world. Computing, 93:135–146.

White, H. (1980). A Heteroskedasticity-Consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4):817–838.

White, L. V. (1973). An extension of the general equivalence theorem to nonlinear models. Biometrika, 60(2):345–348.

Wickham, H. (2021). Mastering shiny. " O'Reilly Media, Inc.".

Wiesenfarth, M., Hisgen, C. M., Kneib, T., and Cadarso-Suarez, C. (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. Journal of Business & Economic Statistics, 32(3):468–482.

Wild, C. and Seber, G. (1989). Nonlinear regression. John Wiley & Sons: New York, NY, USA, 46:86–88.

Wong, K. Y., Zhou, Q., and Hu, T. (2023). Semiparametric regression analysis of doubly-censored data with applications to incubation period estimation. Lifetime Data Analysis, 29(1):87–114.

Wong, W. K. (2021). Lecture notes for biostat 279. Unpublished Manuscript at UCLA.

Wong, W. K. and Lachenbruch, P. A. (1996). Designing studies for dose response. Statistics in Medicine, 15(4):343–359.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1):3–36.

Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

Wright, P. G. et al. (1928). Tariff on animal and vegetable oils. The Macmillan Co.

Wright, S. E. and Bailer, A. J. (2006). Optimal experimental design for a nonlinear response in environmental toxicology. Biometrics, 62(3):886–892.

Wu, Y., Fedorov, V. V., and Propert, K. J. (2005). Optimal design for dose response using beta distributed responses. Journal of Biopharmaceutical Statistics, 15(5):753–771.

Wu, Y. and Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. Annual review of statistics and its application, 7:209–226.

Xiong, G. and Shi, D. (2018). Orthogonal learning competitive swarm optimizer for economic dispatch problems. Applied Soft Computing.

Xu, W., Wong, W. K., Tan, K. C., and Xu, J.-X. (2019). Finding high-dimensional d-optimal designs for logistic models via differential evolution. IEEE access, 7:7133–7146.

Yang, F.-C. and Wang, Y.-P. (2007). Water flow-like algorithm for object grouping problems. Journal of the Chinese Institute of Industrial Engineers, 24(6):475–488.

Yang, X.-S. (2009). Firefly algorithms for multimodal optimization. In International symposium on stochastic algorithms, pages 169–178. Springer.

Yang, X.-S. (2010). Firefly algorithm, levy flights and global optimization. In Research and development in intelligent systems XXVI, pages 209–218. Springer.

Yang, X.-S. (2012). Flower pollination algorithm for global optimization. In International conference on unconventional computing and natural computation, pages 240–249. Springer.

Yang, X.-S. (2017). Nature-inspired algorithms and applied optimization, volume 744. Springer.

Yang, X.-S. and Deb, S. (2009). Cuckoo search via lévy flights. In 2009 World congress on nature & biologically inspired computing (NaBIC), pages 210–214. Ieee.

Yang, X.-S. and Gandomi, A. H. (2012). Bat algorithm: a novel approach for global engineering optimization. Engineering computations.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015). Functional and structural methods with mixed measurement error and misclassification in covariates. Journal of the American Statistical Association, 110(510):681–696.

Yi, G. Y. and Yan, Y. (2021). Estimation and hypothesis testing with error-contaminated survival data under possibly misspecified measurement error models. Canadian Journal of Statistics, 49(3):853–874.

Yin, M.-Z., Zhu, Q.-W., and Lü, X. (2021). Parameter estimation of the incubation period of covid-19 based on the doubly interval-censored data model. Nonlinear Dynamics, 106(2):1347–1358.

Yu, B. (2010). A bayesian mcmc approach to survival analysis with doubly-censored data. Computational statistics & data analysis, 54(8):1921–1929.

Yu, Y. and Feng, Y. (2014). Modified cross-validation for penalized high-dimensional linear regression models. Journal of Computational and Graphical Statistics, 23(4):1009–1027.

Zang, C., Friswell, M., and Mottershead, J. (2005). A review of robust optimal design and its application in dynamics. Computers & structures, 83(4-5):315–326.

Zeng, D., Mao, L., and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika, 103(2):253–271.

Zhang, M. and Fletcher, T. (2013). Probabilistic principal geodesic analysis. Advances in neural information processing systems, 26.

Zhang, Q., Cheng, H., Ye, Z., and Wang, Z. (2017). A competitive swarm optimizer integrated with cauchy and gaussian mutation for large scale optimization. In Control Conference (CCC), 2017 36th Chinese, pages 9829–9834. IEEE.

Zhang, W. X., Chen, W. N., and Zhang, J. (2016). A dynamic competitive swarm optimizer based-on entropy for large scale optimization. In Advanced Computational Intelligence (ICACI), 2016 Eighth International Conference on, pages 365–371. IEEE.

Zhang, Y. (2020a). K-means principal geodesic analysis on riemannian manifolds. In Proceedings of the Future Technologies Conference (FTC) 2019: Volume 1, pages 578–589. Springer.

Zhang, Z. (2020b). Using Competitive Swarm Optimizer with Mutated Agents to Find Optimal Experimental Designs. University of California, Los Angeles.

Zhang, Z., Wong, W. K., and Tan, K. C. (2020). Competitive swarm optimizer with mutated agents for finding optimal designs for nonlinear regression models with multiple interacting factors. Memetic Computing, 12(3):219–233.

Zhou, X.-D., Wang, Y.-J., and Yue, R.-X. (2021). Optimal designs for discrete-time survival models with random effects. Lifetime Data Analysis, 27(2):300–332.

Zocchi, S. S. and Atkinson, A. C. (1999). Optimum experimental designs for multinomial logistic models. Biometrics, 55(2):437–444.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. Annals of Statistics, 35(5):2173–2192.