Title
Bayesian Modeling of Interactions in Structured Heterogeneous Data

Author
Yajima, Masanao

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Bayesian Modeling of Interactions in Structured Heterogeneous Data

(Towards Applications in Integrative Biology)

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

by

Masanao Yajima

2013

Abstract of the Dissertation

Bayesian Modeling of Interactions in Structured Heterogeneous Data

(Towards Applications in Integrative Biology)

by

Masanao Yajima

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2013

Professor Jan de Leeuw, Co-chair

Professor Donatello Telesca, Co-chair

We propose Bayesian models tailored to infer complex patterns of dependence among heterogeneous sets of data. We consider highly structured information and illustrate modeling of flexible multivariate distributions using the formalism of graphical models. Motivating applications guiding our methodological developments come from the field of integrative biology. In particular, we tackle two fundamental problems: the detection of causal SNPs in pharmacogenetics studies and the assessment of differential patterns of interactions characterizing the activity of biomolecular pathways. We discuss inference based on Markov Chain Monte Carlo simulation and apply our methods to several synthetic data sets, as well as case study data from cancer genomics. In these settings, we show how the flexibility of the Bayesian framework is especially attractive, since it allows for the integration of scientific information by means of prior distributions, while

also soundly characterizing the problem of multiple comparisons as a decision problem.

The dissertation of Masanao Yajima is approved.

Mark Handcock

Ying Nian Wu

Donatello Telesca, Committee Co-chair

Jan de Leeuw, Committee Co-chair

University of California, Los Angeles

2013

To my wife Marin...

who has always kept me in balance

with cheerful smiles and insightful comments.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the help and the support of many people who have guided me in various ways through my years of PhD.

First and foremost, I would like to thank Professor Jan de Leeuw, with his big heart, has funded me throughout my years of studies. He has exposed me to various projects and taught me many skills that expanded my capabilities as a statistician, which I am truly grateful for.

Deepest gratitude is due to Professor Donatello Telesca who has guided me through the projects that lead to this dissertation. Finding Donatello was the most serendipitous encounter during my time at UCLA. With his sharp thought provoking comments, has intellectually entertained me even beyond the realm of statistics. I am truly grateful for his patience and support throughout the thesis.

In addition, I would also like to give a special appreciation to Professor Peter Müller, who was a part of the projects that led to this thesis. His precise comments based on the deep understanding of the materials have helped me refine this thesis to a higher standard.

Great deal of thought provoking comments were provided by my committee members Professor Mark Handcock and Professor Ying Nian Wu. They have given me guidance at some critical points of this thesis, which I am truly grateful for. Moreover, I was fortunate to get valuable inputs from many professor and friends at variety of levels regarding the details of this thesis. I would like to show my appreciation by listing there names here: Professor Gary Rosner, Professor Yuan Ji, Professor Rafael Becerril, and Krishna Bhogaonker.

I would also like to acknowledge my appreciation toward my fellow graduate students, professors, and staffs of both the Department of Statistics and the

Department of Biostatistics for all the help and support they have given me, the University of California Los Angeles for the Dissertation Year Fellowship that carried me through the last year of the study, and the UCLA graduate writing center for providing accessible resources for improving the writing of the thesis.

Furthermore, I would like to thank AJ and the staffs of Conservatory for Coffee, Tea, and Cocoa for countless cups of coffee they have provided me while I wrote this thesis. Last but not least, I would like to thank Professor Jason Tsou, Arthur Shoenfeld, and Professor Cheng-Chieh Yu for maintaining my health through Tai-chi practice during the last stretch of my study. Above all, I would like to thank my family especially my wife Marin who has at all times inspired me with her exuberant sprit and supported me with her benevolent patience.

Vita

2000	B.A. in Policy Studies at the Kwansei Gakuin University, Japan
2007	M.A. in Statistics at the Columbia University, Department of Statistics
2012	UCLA Dissertation Year Fellowship
2013	WNAR 2013 student paper competition 1st prize (1/29)

Publications

(Submitted) Masanao Yajima, Donatello Telesca, Garry Rosner and Peter Muller. "Bayesian Pharmacogenetics Modeling"

(Submitted)Masanao Yajima, Donatello Telesca, Yuan Ji, and Peter Muller. "Detecting Differential Patterns of Interaction in Molecular Pathways"

(2011) A. Gelman, J. Hill, M.Yajima, "Why we (usually) don't worry about multiple comparisons" *Journal of Research on Educational Effectiveness*

(2009) Y. Su, A. Gelman, J. Hill, and M. Yajima "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box". *Journal of Statistical Software.*

(2006) L. Paninski, M. Yajima, "Undersmoothed kernel entropy estimators" *IEEE Transactions on Information Theory*

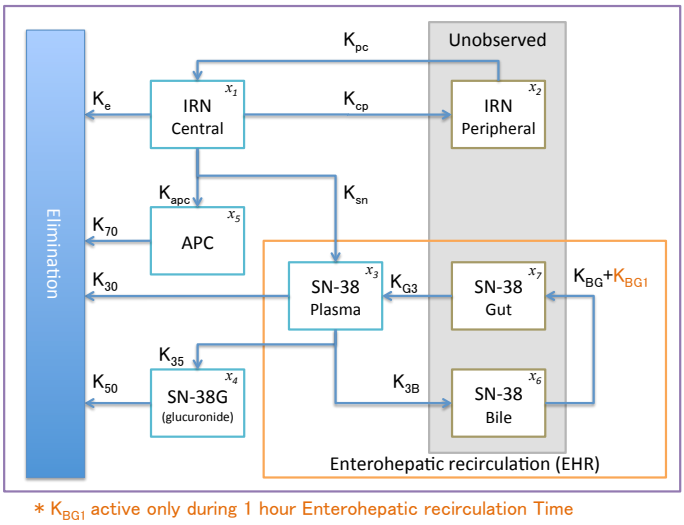
CHAPTER 1

INTRODUCTION

The ever-growing ability to measure and quantify key quantities related to the manifestation of complex biological phenomena allows for the rigorous investigation of finely detailed scientific hypotheses. However, the dynamic development of data generation technologies has not been paralleled by a corresponding development of sound statistical methodology. In this dissertation, we propose statistical models to infer interactions among structured heterogeneous data sources. Motivating applications come from the field of integrative biology, where the immediate scientific challenge is in combining information across many data sources. This includes several specific sub-disciplines like: genomics, proteomics, transcriptomics, metabonomics, etc.

To illustrate the intricacy characterizing some of these studies, we consider a chemotherapy dose determination study. Figure 1.1 shows the systematic mechanism of absorption, distribution, metabolization, and elimination (ADME) of a chemotherapeutic agent irinotecan; each box represents a different state of the metabolized substance, some of which are measured as in figure 1.2. The data is multivariate, temporally correlated, and the joint sampling distribution is reasonably expected to obey a structured dependence pattern. The fundamental substantive question is to assess how the ADME process of a drug is affected by genetics (Rosner et al., 2008). The time course metabolite study is therefore completed with the measurement of single nucleotide polymorphisms (SNPs), whose

Figure 1.1: Pharmacokinetic model with enterohepatic recirculation (EHRT)



joint distribution of genetic mutation is to be related to the drug pharmacokinetics.

A second example relates to the inference of association structures that are expectedly heterogeneous across different subclasses of subjects. This type of problem arises in studies aimed to assess differences in the molecular association between subgroups in targeted disease populations, such as diabetes (Valcárcel et al., 2011) or lung cancer (Danaher et al., 2011). We analyze a specific data set

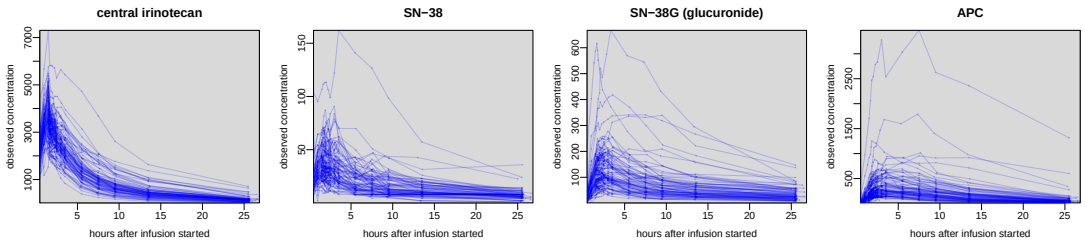


Figure 1.2: Observed concentration values for irinotecan, SN-38, SN-38G (glucuronide), and APC. Each solid (blue) line represents observed concentration for a patient plotted over time.

involving protein expression in Acute Myeloid Leukemia (AML) patients. Figure 1.3 provides a sample correlation of protein expression for refractory vs. relapsed patients. The two correlation matrices are mostly similar, however, a closer examination will reveal local differences in parts of the empirical correlation. The scientific concern in this case is that of identifying the differential association structure that distinguishes the end result of therapy for subgroups of patients.

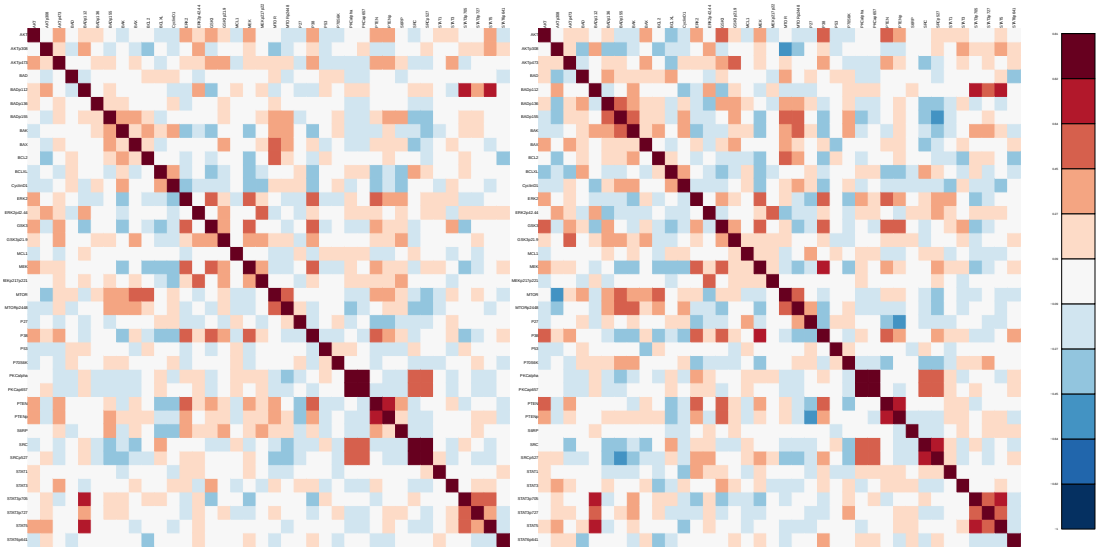


Figure 1.3: Sample correlations for two sub-groups of patients with Acute Myeloid Leukemia.

Although we focus on issues involving molecular biology, the demand for models that can identify interaction structures in complex heterogeneous data is flourishing across scientific disciplines. This includes fields as diverse as marketing (Steenburgh et al., 2003; Albuquerque and Bronnenberg, 2012; Srinivasan et al., 2010), atmospheric science (Saatchi et al., 2011), crime modeling (Mohler and Short, 2012; Hegemann et al., 2011), computer science (Thiesson et al., 1997; Guo et al., 2011), etc. Biomedical research, nevertheless, poses some of the most interesting, yet challenging questions. Often, while the number of samples is lim-

ited, the number of different biological and phenotypical dimensions that can be quantified each patient is increasing rapidly. At the same time, the accumulation of scientific studies has given rise to numerous collaborative databases that are publicly available to the research community. This balance of data scarcity and large amount of information available a priori makes the employment of Bayesian methods attractive since they allow us to fold more information into our inferences and decisions (Gelman, 2014).

We propose Bayesian hierarchical modeling frameworks to address some of the issues concerning structural inference that combine multiple sources of information. We make use of the graphical models to assess the conditional independence structure embedded in the data under the assumption of sparsity At the same time, we address the multiple comparisons issues inherent in these analyses through decision theoretic arguments.

The thesis will be structured as follows. In the following sections we cover some of the basic building blocks that will be used in the remainder of the thesis. In chapter 2 we address the problem of structural association in Bayesian pharmacogenetics. In chapter 3 we further consider the issue of structural comparisons by modeling the differential association structure of protein expression networks. Finally we end with conclusions and future directions in chapter 4.

1.1 Marginal Independence and Conditional Independence

1.1.1 Marginal independence

Independence between two distinct stochastic entities in statistics refers to a relationship where knowledge of one event does not inform us of the probability of another event. If we use $P(\cdot)$to denote a probability measure, then independence

of two random quantities Z and Y implies

$$P(Y \in \mathcal{Y} \mid Z \in \mathcal{Z}) = P(Y \in \mathcal{Y}).$$

This also translates to

$$P(Z \in \mathcal{Z} \cup Y \in \mathcal{Y}) = P(Z \in \mathcal{Z})P(Y \in \mathcal{Y}).$$

For any continuous random variable Y and Z, with PDF $p_Y(y)$ and $p_Z(z)$ respectively, then Y and Z are independent if and only if

$$p_{YZ}(y, z) = p_Y(y)p_Z(z), \tag{1.1}$$

and we denote this relationship as $Y \perp\!\!\!\perp Z$.

1.1.2 Conditional independence

The conditional independence relationship of random variables Y, Z, and X is;

$$p_{YZ}(y, z \mid x) = p_Y(y \mid x)p_Z(z \mid x). \tag{1.2}$$

We say Y and Z are independent given X and we denote this relationship as $Y \perp\!\!\!\perp Z \mid X$.

Conditional independence has following properties (Lauritzen, 1996) for random variables W, X, Y, Z

(C1) If $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) If $X \perp\!\!\!\perp Y \mid Z$ and $U = h(Y)$, then $X \perp\!\!\!\perp U \mid Z$;

(C3) If $X \perp\!\!\!\perp Y \mid Z$ and $U = h(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

(C5) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z)$ given that joint density $p(w, x, y, z)$ w.r.t product measure is positive and continuous.

1.2 Representing Dependence Through Graphical Models

In this section we briefly review essential graphical modeling notation and concepts. For a comprehensive review we refer to Lauritzen (1996).

A graphical model is a mathematical model used to express conditional independence of a set of random variables. A graph is characterized by an algebraic structure $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, composed of a set of *vertices* $\mathcal{V}$, and a set of *edges* $\mathcal{E} \subseteq \{(v_i, v_j), v_i \in \mathcal{V}\}$. Vertices v_i and v_j are *adjacent* or are *neighbors* if $(v_i, v_j) \in \mathcal{E}$ and this relationship is denoted as $v_i \sim v_j$. An adjacency matrix $\mathcal{A}_d$ is a matrix where element in ith row jth column is 1 if $v_i \sim v_j$ and 0 otherwise. Graphs are either *undirected* $(v_i, v_j) = (v_j, v_i)$, *directed* $(v_i, v_j) \neq (v_j, v_i)$, or a combination of the two. For clarity, we also denote a directed edge from v_i to v_j as $v_i \to v_j$.

A *path* is defined as a set of vertices $\{v_1, \cdots, v_k\}$ such that $v_i \sim v_{i+1}$ for each $i = 1, \cdots, k-1$. A *directed path* is a same set of vertices but it requires $v_i \to v_{i+1}$ for each $i = 1, \cdots, k-1$. If $v_1 = v_k$ for a directed path, then it is called a *directed cycle*.

A graph is *complete* if all vertices are joined by an arrow or a line so that $(v_i, v_j) \in \mathcal{E}, \forall v_i, v_j \in \mathcal{V}$ and otherwise it is called *incomplete* . A *complete subset* induces subgraph that is complete. A complete subset that is maximal is called

a *clique* . A set S is called a *separator* of (A, B) if every path from A to B goes through S.

For a directed graph, if $v_i \to v_j$ then v_i is called a *parent* and v_j is called a *child* . The set of parents of v_i is denoted as $pa(v_i)$ and the set of children as $ch(v_i)$. If there exists a directed path from v_i to v_k then v_i is an *ancestor* of v_k and v_k is a *descendant* of v_i. The set of *ancestors* of v_k is denoted as $an(v_k)$ and the set of *descendants* of v_i will be denoted as $de(v_i)$. All the vertices that are not descendants of v_i are called *non-descendant* and will be denoted as $nd(v_i)$. An *ancestral matrix* $\mathcal{A}$ is a matrix where element in ith row jth column is 1 if $v_i \to v_j$ and 0 otherwise. A *boundary* of a vertex denoted as $bd(v_i)$ is a set of vertices in $\mathcal{V} \setminus v_i$ that are either parent or neighbor of v_i. Using $ne(A)$ to denote the neighboring set of v_i, $bd(v_i) = pa(v_i) \cup ne(v_i)$. We will also define the *closure* of v_i to be denoted as $cl(v_i) = v_i \cup bd(v_i)$. We should also note that for subset $A \subseteq V$ we expand the notation of $pa(A)$ to be short hand for $pa(A) = \cup_{a \in A} pa(a) \setminus A$ and similarity for relational sets such as $ch(A)$ and $bd(A)$ etc.

A decomposition of graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$ is defined by partitioning $\mathcal{G}$ into non-overlapping sets A and B with separator S such that $\mathcal{V} = A \cup B$ and $S = A \cap B$ is complete. The decomposition is *proper* if neither A nor B is empty. A sequence of subgraphs that cannot be decomposed further are called the *prime components* of a graph. When every prime component is complete the graph is called a *decomposable graph*.

For a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, an undirected graph $\mathcal{G}^M$ that satisfy the following conditions will be called a *moral graph* $\mathcal{G}$.

1. $\mathcal{G}$ and $\mathcal{G}^M$ share all the vertices.

2. All the vertices that have an edge in $\mathcal{G}$ will have edges in $\mathcal{G}^M$.

3. There is an edge between any pair of vertices that share a common child. If $ch(v_i) \cap ch(v_k) \neq \emptyset \Rightarrow (v_i, v_k) \in E$

1.2.1 Markov properties

The conditional independence statements encoded in a graphical model are often defined as Markov properties. Depending on the scope of the implied conditional independence, Markov properties are subdivided into global, local, and pairwise Markov properties in the order of implication. For more details on the Markov properties, we refer the readers to Lauritzen (1996).

We will consider a situation where we have a collection of random variables $\{Y_v\}_{v\in\mathcal{V}}$ taking values in probability space $\{\mathcal{Y}_v\}_{v\in\mathcal{V}}$. We use the set subscript $V \subseteq \mathcal{V}$ to denote a subspace of $\mathcal{Y}$ so that $\mathcal{Y}_V = \times_{v\in V}\mathcal{Y}_v$ and similarly for Y_V. Also following the convention we use $A \perp\!\!\!\perp B \mid C$ in place of $Y_A \perp\!\!\!\perp Y_B \mid Y_C$ for notational clarity.

1.2.2 Undirected Graphs

For an undirected graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ and a set of random variables $\{Y_v\}_{v\in\mathcal{V}}$, a probability measure P on $\mathcal{Y}$ is said to obey

(P) the pairwise Markov property if

$$v_i \perp\!\!\!\perp v_j \mid \mathcal{V} \setminus \{v_i, v_j\} \text{ for } v_i, v_j \in \mathcal{V} \text{ and } v_i \nsim v_j$$

(L) the local Markov property if

$$v \perp\!\!\!\perp \mathcal{V} \setminus cl(v) \mid bd(v), \text{ for } v \in V$$

(G) the global Markov property if, for disjoint subsets of $\mathcal{V}$, (A, B, S), S separates A and B then

$$A \perp\!\!\!\perp B \mid S$$

For any undirected graph, it is shown that (G)⇒(L)⇒(P).

1.2.3 Directed Acyclic Graphs

A *Directed Acyclic Graph* (DAGs), is a directed graph without any directed cycle. Given a DAG, the implied Markov properties characterizing a set of random variables $\{Y_v\}_{v\in\mathcal{V}}$ follows similar rules as the undirected case

(DP) the directed pairwise Markov property if

$$v \perp\!\!\!\perp u \mid nd(v) \setminus \{u\} \text{ for } v, u \in \mathcal{V} \text{ and } v \nsim u, u \in nd(v)$$

(DL) the directed local Markov property if

$$v \perp\!\!\!\perp nd(v) \mid pa(v) \text{ for } v \in \mathcal{V}$$

(DG) the directed global Markov property if, for disjoint subsets of $\mathcal{V}$, (A, B, S), S separates A and B in $(\mathcal{G}_{an(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$ then.

$$A \perp\!\!\!\perp B \mid S$$

The order of implication follows the similar ruled as the undirected graph;

$$\text{(DG)} \Rightarrow \text{(DL)} \Rightarrow \text{(DP)}.$$

The (DG) has an equivalent definition under the name of *d-separation* criterion Pearl (1985, 2000).

Definition Let A, B, and S be disjoint subsets of a directed, acyclic graph $\mathcal{G}$. Then S *d-separates* A from B if and only if S separates A from B in $\mathcal{G}^M_{an(A\cup B\cup S)}$. Lauritzen (1996)

An equivalent definition using the idea of collider, where a collider C is defined as a set of vertices where two directed edges meet, is defined as

Definition If $\mathcal{G}$ is a directed graph in which A, B and S are disjoint sets of vertices, then A and B are *d-connected* by S in $\mathcal{G}$ if and only if there exists an undirected path U between some vertex in A and some vertex in B such that for every collider C on U, either C or a descendent of C is in S, and no non-collider on U is in S. A and B are *d-separated* by S in $\mathcal{G}$ if and only if they are not d-connected by S in $\mathcal{G}$.

DAGs are appealing from the modeling perspective since the joint distribution of the vertices is simply expressed as the product of conditional densities of each of the vertices conditioned on their parents. This makes for great flexibility of modeling (Edwards, 2000). The acyclicity restriction could represent a drawback in some applications. However, when dealing with a network where association is usually sparse, this restriction is often not critical.

1.2.4 Chain Graphs

Chain graphs also have a Markov properties similar to that of undirected and directed graphs defined as

(PB) the pairwise block-recursive Markov property if for a concurrent set

$C(t) = V(1) \cup \cdots \cup V(t)$ where $V(t)$ is a partition of $\mathcal{V}$ such that each of the sets $V(t)$ has lines between vertices, and arrows point from vertices in sets with a lower number to those with a higher number, then

$$v \perp\!\!\!\perp u \mid C(t^*) \setminus \{v, u\}$$

where $t*$ is the smallest t such that $v, u \in C(t)$.

(PC) the pairwise chain Markov property if

$$v \perp\!\!\!\perp u \mid nd(v) \setminus \{vu\} \text{ for } v, u \in \mathcal{V} \text{ and } v \nsim u, u \in nd(v)$$

(LC) the local chain Markov property if

$$v \perp\!\!\!\perp nd(v) \mid pa(v) \text{ for } v \in \mathcal{V}$$

(GC) the global chain Markov property if, for disjoint subsets of $\mathcal{V}$, (A, B, S), S separates A and B in $(\mathcal{G}_{an(A \cup B \cup S)})^m$ then

$$A \perp\!\!\!\perp B \mid S$$

The direction of implication is the same as the other graphs except that (PB) is also implied by (PC). Therefore the implication relationship is ordered as:

$$\text{(GC)} \Rightarrow \text{(LC)} \Rightarrow \text{(PC)} \Rightarrow \text{(PB)}.$$

1.3 Gaussian Graphical Models

Gaussian Graphical Models (GGMs), also known as Covariance Selection Models (Dempster, 1972) or Gaussian concentration graph models (Cox and Wermuth, 1996), are a class of multivariate Gaussian models that obey the pairwise Markov property defined by a graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$. For an n–sample of p–variate observations Y and a given set A, we use the set subscript A to denote $Y_A = \{Y_j\}_{j \in A}$. When Y has a full covariance matrix $\Sigma = [\sigma_{i,j}]$, partition of covariance matrix corresponding to a subset Y_A will be denoted by $\Sigma_{AA} = [\sigma_{i,j}]_{i \in A, j \in A}$.

Let $M(\mathcal{G})$ denote the symmetric matrices $\mathcal{A}$ with $\mathcal{A}_{AB} = 0$ if and only if $(A, B) \notin \mathcal{E}$. We use $M^+(G)$ to denote the positive definite subset of $M(\mathcal{G})$. Then the GGM is described as

$$\boldsymbol{Y} = \{Y_v, v \in \mathcal{V}\} \mid \mathcal{G}(\mathcal{E}, \mathcal{V}) \sim N_{\mathcal{G}}(\mu, \Sigma), \quad \Sigma^{-1} = \Omega \in M^+(\mathcal{G})$$

where Σ is a positive semidefinite covariance matrix and Ω is the inverse covariance matrix or a precision matrix Lauritzen (1996).

In other words, when $\boldsymbol{Y}$ follows a GGM with respect to $\mathcal{G}(\mathcal{E}, \mathcal{V})$, if we define the concentration matrix of the conditional distribution of (Y_A, Y_B) given $Y_{\mathcal{V} \setminus \{A,B\}}$ to be

$$\Omega_{\{A,B\}} = \begin{pmatrix} \Omega_{AA} & \Omega_{AB} \\ \Omega_{BA} & \Omega_{BB} \end{pmatrix},$$

then $\Omega_{AB} = 0 \Leftrightarrow (A, B) \notin \mathcal{E}$.

GGMs are popular from both an estimation and inferential perspective. They allow for efficient estimation of the covariance matrix by exploiting the zeros in the off diagonal of the inverse covariance matrix. At the same time the esti-

mated graphical model is a useful tool in making inference about the dependence structure amongst the variables.

Decomposable GGMs Conditionally on a decomposable graph $\mathcal{G}$, the likelihood of the GGM model factorizes as

$$p(Y|\Sigma,\mathcal{G}) \;=\; \frac{\prod_{P\in\mathscr{P}} p(Y_P|\Sigma_P)}{\prod_{S\in\mathscr{S}} p(Y_S|\Sigma_S)} \tag{1.3}$$

where $\mathscr{P}$ and $\mathscr{S}$ denote the set of prime components and separators. The density $p(Y_P|\Sigma_P)$ corresponds to clique-marginal models s.t.

$$p(Y_P|\Sigma_P) \;=\; (2\pi)^{-\frac{n|P|}{2}} \det(\Sigma_P)^{-\frac{n}{2}} \mathrm{etr}\left[-\frac{1}{2}\{R_P(\Sigma_P)^{-1}\}\right]$$

where $R_P = Y_P^{\mathrm{T}} Y_P$ and similarly for separators $p(y_S|\Sigma_S)$.

For a decomposable graph, each of $P \in \mathscr{P}$ is complete thus it is a clique. Since every separator $S \in \mathscr{S}$ is a subset of some clique $P \in \mathscr{P}$, based on the conditional independence relation defined by $\mathcal{G}$, collection of the clique-marginal covariances $\{\cup_{P\in\mathscr{P}}\Sigma_P\}$ fully determines Σ.

Non-decomposable GGMs Non-decomposable GGMs factors out in similar fashion as (1.3), yet the prime components are not necessarily complete and additional constraints need to be considered.

1.3.1 Bayesian conjugate analysis

Bayesian Analyses of the GGM are for the large part based on the exploitation of the conjugacy principle. Bjerg and Nielsen (1993) showed that the D-Y conjugate prior distribution for decomposable GGM is the hyper inverse wishart (HIW)

prior (Dawid and Lauritzen, 1993). When Σ follows a HIW distribution with parameters b and Φ,

$$p(\Sigma \mid b, \Phi, \mathcal{G}) =_d \frac{\prod_{P \in \mathscr{P}} Inv\text{-}Wishart_P\left(\Sigma_P | b, \Phi_P\right)}{\prod_{S \in \mathscr{S}} Inv\text{-}Wishart_S\left(\Sigma_S | b, \Phi_S\right)}$$

where $Inv\text{-}Wishart(\Sigma|b, \Phi)$ is an inverse Wishart Density. Ergo, the HIW distribution is just a way to define the standard Inverse Wishart distribution prior on each of the decomposed components in a clever manner. By the definition of conjugacy, the resulting posterior distribution is again $HIW(b+n, \Phi + H(Y^{\mathrm{T}}Y))$, where $H(Y^{\mathrm{T}}Y)$ is the hyper-Markov sum-of-squares matrix corresponding to the cliques and separators of $\mathcal{G}$.

For more general classes of graphs including those that are not decomposable Roverato (2002) showed the existence of D-Y conjugate G-Wishart prior (Atay-Kayis and Massam, 2005) denoted as

$$p(\Omega \mid \mathcal{G}) = C_{\mathcal{G}}(b, \Phi)^{-1} |\Omega|^{(b-2)/2} \mathrm{etr}\left\{\frac{1}{2}\Phi\Omega\right\} 1_{\{\Omega \in M^+(\mathcal{G})\}}$$

where $b > 2$ is the degree of freedom parameter, D is a symmetric positive definite matrix. The term $C_{\mathcal{G}}(b, D)$ is a normalizing constant

$$C_{\mathcal{G}}(b, \Phi) = \int_{M^+(\mathcal{G})} |\Omega|^{(b-2)/2} \mathrm{etr}\left\{\frac{1}{2}\Phi\Omega\right\} 1_{\{\Omega \in M^+(\mathcal{G})\}} d\Omega$$

$M^+(\mathcal{G})$ is the cone of symmetric positive definite matrices defined by the graph $\mathcal{G}$ such that $\Omega_{ij} = 0, \{i, j; (i, j) \notin \mathcal{E}\}$. This formulation incorporates the decomposable case where it reduces to the HIW distribution and the complete case where it is simply an inverse Wishart distribution (Atay-Kayis and Massam, 2005).

1.3.2 Graphical Lasso

Alternatively, rather than trying to estimate the entire posterior distribution of zeros in the inverse covariance matrix, one can reduce the problem to point estimation using a regularized maximum likelihood estimator (MLE). The sparsity constraint can be introduced by use of the penalized likelihood (Meinshausen and Büllmann, 2006; Peng et al., 2009) or the penalized log likelihood (Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008) that takes a form

$$\max_{\Theta}\{\log\det\Theta - \text{trace}(S\Theta) - \lambda\|\Theta\|\}, \tag{1.4}$$

where S is the empirical covariance matrix and Θ is the estimator for the inverse covariance matrix. The function (1.4) is maximized wrt Θ. When $\|\Theta\|$ is L_1 norm, the penalty is called the lasso (Tibshirani, 1996) and λ is the tuning parameter that controls the level of shrinkage that is usually chosen based on cross validation. Because the problem is reduced to a single optimization, it scales to large size problems, making them popular in the data heavy applications.

1.3.3 Computation

Computation of GGMs is an active area of research that has many challenges. A comprehensive survey of the techniques can be found in (Wang and Li, 2012).

1.3.4 Other priors for transformations of a covariance matrix

Beyond the conjugate Inverse Wishart prior and Jeffery's prior $p(\Sigma) \propto 1/\left|\Sigma\right|^{(J+1)/2}$ there are herds of priors proposed for decompositions of covariance matrices in the literature. The spectral decomposition method used to be popular for the dimension reduction and computational ease it provided. Examples of spectral

decomposition priors are the following.

- The reference prior (Yang and Berger, 1994),

$$p(\Sigma) \propto 1/\left\{|\Sigma| \prod_{i<j}(d_i - d_j)\right\}$$

  where d_i is eigenvalue of covariance matrix Σ

- The log matrix prior (Leonard and Hsu, 1992; Chiu et al., 1996) logarithmic transformation of the eigenvalue/eigenvector decomposition of Σ and allows for hierarchical shrinkage to be done with the eigenvalues.

- (Daniels and Kass, 1999) proposes the following couple of separate schemes for hierarchical prior.

  - hierarchical extension of the inverse-Wishart prior
  - normal prior for Fisher's z transform of the correlation coefficients
  - eigenvalue/eigenvector parameterization,orthogonal eigenvector matrix parameterised in terms of the Givens angles

despite their convenience, parametrization based on spectral decomposition were usually criticized for the lack of interpretability and difficulty in the incorporation of the prior information.

Cholesky decomposition is another popular type of decomposition that has a regression interpretation. Examples are as follows.

- Rue-Held-2005-GaussianCholesky decomposition of precision (Smith and Kohn, 2002)

$$\Sigma^{-1} = HDH^{\mathrm{T}}$$

discrete mixture prior on elements of lower triangular matrix H for sparsity.

- Cholesky decomposition of precision (Cai and Dunson, 2006)

$$\Sigma^{-1} = \text{diag}(\Delta)\Gamma\Gamma^{\text{T}}\text{diag}(\Delta)^{\text{T}}$$

  discrete mixture prior on elements of diagonal matrix Δ as well as lower triangular matrix Γ

Parameterizations based on a correlation matrix or partial correlation matrix have become popular in recent years due to the interpretability of the parameters and ease of incorporating the prior information. The examples of such parametrization are following.

- Defining separate priors on correlation and standard deviation (Barnard et al., 2000),

$$\Sigma = \text{diag}(S)\ R\ \text{diag}(S)$$

  Independent priors on S and for the correlation matrix R;

  - marginally uniform prior: independent beta distribution on each off diagonal $R_{ij}, i \neq j$
  - jointly uniform prior: R uniformly distributed over all possible correlation matrices

- Defining priors for clustered correlation (Liechty et al., 2004). The proposed parametrization is similar to (Barnard et al., 2000) such that,

$$\Sigma = \text{diag}(S)\ R\ \text{diag}(S)$$

where R is correlation matrix and S is the diagonal standard deviation. They define mixture prior on R so that the correlation can be grouped into common correlation values or by variables.

- Defining priors on the partial correlation matrix (Wong et al., 2003)

$$\Omega = \Sigma^{-1} = \text{diag}(T)\ C\ \text{diag}(T)$$

where T is a diagonal precision matrix and C is the negative of partial correlation matrix. They define a gamma distribution for the diagonal entry of Ω by defining $T_i| \propto T_i^{2\alpha-1} exp(-\beta T_i^2)$. The off diagonals of Ω is defined as a mixture distribution on C. The main difficulty in this method is computational, since there is no direct way to simulate from the desired distribution. The problem is evaded through normal approximations.

1.4 Gaussian DAG Models

1.4.1 Likelihood

The Gaussian DAG (GDAG) models proposed by Fronk and Giudici (2004) are defined as the product of conditional regression models for each variable given their parents. Let $pa(i)$ be parent index for node i for DAG $\mathcal{G}$, the likelihood for the Y_i given all it's parents $Y_{pa(i)}$ is defined as

$$Y_i|Y_{pa(i)}, \beta_{i|pa(i)}, \sigma^2_{i|pa(i)}, \mathcal{G} \sim N\left(\beta_{i0} + \sum_{y_l \in pa(i)} \beta_{il} y_l, \sigma^2_{i|pa(i)}\right)$$

for $i = 1, \ldots, n$, $\beta_{i|pa(i)} = \beta_{i,j}; j \in pa(i)$ and $\sigma^2_{i|pa(i)} = \sigma^2_{ii} - \Sigma_{i,pa(i)}\Sigma^{-1}_{pa(i)}\Sigma_{pa(i),i}$. Then the joint likelihood is defined as the product over all the p variables

$$p(Y|\beta, \sigma^2, \mathcal{G}) = \prod_{i=1}^{p} p(y_i, |y_{pa(i)}, \beta_{i|pa(i)}, \sigma^2_{i|pa(i)}, \mathcal{G}).$$

1.4.2 Conjugate inference

As with regular regression models, conjugate analysis is attractive for its tractability. The conjugate prior distribution for $(\beta_{i|pa(i)}, \sigma^2_{i|pa(i)})$ is the Normal Inverse Gamma prior defined as

$$\begin{aligned} \beta_{i|pa(i)}|\sigma^2_{i|pa(i)}, \mathcal{G} &\sim N_{|pa(i)|+1}\left(b_{i|pa(i)}, \frac{1}{\alpha}\sigma^2_{i|pa(i)}I\right) \\ \sigma^2_{i|pa(i)}|\mathcal{G} &\sim Inv\text{-}Ga(\delta_{i|pa(i)}, \lambda_{i|pa(i)}) \end{aligned}$$

The joint distribution is then

$$\begin{aligned} p(Y, \beta, \sigma^2, d) &= p(Y|\beta, \sigma^2, d)p(\beta|\sigma^2, d)p(\sigma^2|d)p(d) \\ &= \prod_{i=1}^{p} p(y_i, |y_{pa(i)}, \beta_{i|pa(i)}, \sigma^2_{i|pa(i)}, d) \\ &\quad \prod_{i=1}^{p} p(\beta_{i|pa(i)}|\sigma^2_{i|pa(i)}, d) \prod_{i=1}^{p} p(\sigma^2_{i|pa(i)}|d)p(d) \end{aligned}$$

1.5 Chain Graph Models

Large classes of models fit under the general framework of chain graphs, including: factor analysis models, latent class model, path regression model, linear structural equation model, regression model, etc. A overview and classification of these models can be found in Wermuth and Lauritzen (1990). The most popular form of

the model is the multivariate regression model, where the partitioning between the dependent variable and covariates, and consequently directionality between them, is a product of model specification and not a conclusion to be drawn (Whittaker, 1990). Edwards (2000) proposes CG-regression model where assumption is made on the covariates to have fully connected graph. More detailed treatment of chain graph regression models can be found in Whittaker (1990).

1.6 Prior Distribution on Graphs

1.6.1 Non-informative prior on graphs

Prior distributions on the space of graphs are necessary in order to conduct posterior inference on the graph structures. Uniform distribution over set of all possible graphs is an intuitive and popular choice (Giudici and Green, 1999; Fronk and Giudici, 2004).

$$p(\mathcal{G}) \propto \frac{1}{|D|},$$

where $|D|$ represents number of possible graphs. Although this choice seems desirable for it's objectivity, it is shown that these priors tend to favor "medium" sized graphs and suggested to be inappropriate for large graphs (Jones et al., 2005) .

Another popular choice in the literature is a prior on graph that models edge inclusion as exchangeable Bernoulli trials (Dobra et al., 2004; Jones et al., 2005). Let $|\mathcal{E}_k|$ be the number of edges in graph $\mathcal{G}_k$, then

$$p(\mathcal{G}_k \mid \psi_k) = \psi_k^{|\mathcal{E}_k|}(1-\psi_k)^{M-|\mathcal{E}_k|}. \tag{1.5}$$

When the inclusion probabilities ψ_k is modeled hierarchically using a beta

distribution $Beta(v_1, v_2)$, this class of stochastic schemes is know to provide automatic multiplicity correction in the posterior $p(\mathcal{G}_k \mid \boldsymbol{Y})$ (Scott and Berger, 2010; Carvalho and Scott, 2009).

The marginal prior distribution for $\mathcal{G}_k$ is available in closed form as

$$\begin{aligned} p(\mathcal{G}_k) &\propto B(v_1 + |\mathcal{E}_k|, v_2 + M - |\mathcal{E}_k|) \\ &= \frac{\Gamma((v_1 + |\mathcal{E}_k|)\Gamma(v_2 + M - |\mathcal{E}_k|)}{\Gamma(v_1 + v_2 + M)}, \end{aligned}$$

which simplifies to $p(\mathcal{G}_k) = \frac{1}{(M+1)}\binom{M}{|\mathcal{E}_k|}$, if $\psi_k \sim U(0, 1)$.

These priors are often categorized as non-informative or objective and are popular mostly for their convenience, since not much consideration needs to be given in setting up the model. When prior information on interaction structures is available, informative priors have been suggested by Mukherjee and Speed (2008); Telesca et al. (2012b).

1.6.2 Informative prior on graphs

A general form of informative log linear prior distribution over graph is proposed by Mukherjee and Speed (2008) as,

$$p(\mathcal{G}|\mathcal{G}_p, \psi) \propto \exp\left\{\lambda \sum_t w_t f_t(\mathcal{G})\right\}. \tag{1.6}$$

The concordance function $f_t(\mathcal{G})$ measures the degree of concordance with the prior knowledge for feature t. Hyper parameters w_t define the weighting amongst the features ($w_1 = 1$) and λ controls the strength of belief. They provide several types of concordance functions based on the

- concordance of individual edge

$$f_t(\mathcal{G}) = |\mathcal{E}(\mathcal{G}) \cap \mathcal{E}_+| - |\mathcal{E}(\mathcal{G}) \cap \mathcal{E}_-|$$

  Where $\mathcal{E}_+$ denotes a set of edges expected to be present ("positive edge-set") and $\mathcal{E}_-$ denotes a set of edges expected to be absent ("negative edge-set").

- edges between classes of vertices.

  Similar functional form as the individual edge however the negative edge set is defined as

$$\mathcal{E}_- = \{e = (v_l v_m) : C(v_l) = C_i, C(v_m) = C_j\}$$

  $C(v)$ denotes the class to which vertex $v \in \mathcal{V}$ belongs.

They also give examples of a formulation to impose network sparsity, capture the concordance of higher-level network features, and degree distributions.

Moon et al. (2013) suggests a similar model

$$p(\mathcal{G}|\mathcal{G}_p, \psi) = \exp\left\{-\sum_t w_t f_t(\mathcal{G})\right\} / Z(w). \tag{1.7}$$

They defin f_t to be a distance measure between the prior information matrix $\mathcal{B}_t$ and the adjacency matrix $\mathcal{A}$ for each of multiple prior information source t and puts a independent exponential prior on w_t. They provide a practical recipe to elicit $\mathcal{B}$ from scientific databases for transcription factor and DNA binding, protein-protein interaction and gene ontology annotations.

Telesca et al. (2012b) defines the prior on the graph structure to be exponentially decaying function centered around the prior information

$$p(\mathcal{G}) \propto \psi^{d(\mathcal{G},\mathcal{G}^\star)}, (\psi \in [0,1])$$

where distance function is defined as $d(\mathcal{G},\mathcal{G}^\star) = |\mathcal{E}^c \cap \mathcal{E}^\star| + \delta|\mathcal{E} \cap \mathcal{E}^{\star c}|, \delta \geq 1$. In the framework of Mukherjee and Speed (2008), this is a two feature model for a missing edge and an extra edge where the extra edge is penalized more by weight δ and $\lambda = log(\psi)$. A hierarchical prior on ψ is defined by Scott and Berger (2006) to control for the multiplicity. By defining the prior distribution of ψ to be a beta distribution with shape parameters s_1 and s_2 that has mean at $s_1/(s_1+s_2)$ and variance of $(s_1 s_2)/[(s_1+s_2)^2(s_1+s_2+1)]$, ψ can be integrated out to get a marginal prior distribution for $\mathcal{G}$.

$$p(\mathcal{G}) \propto \frac{beta(s_1 + d(\mathcal{G}_p,\mathcal{G}), s_2)}{beta(s_1, s_2)} = \frac{\Gamma(s_1 + d(\mathcal{G}_p,\mathcal{G}))\Gamma(s_1+s_2)}{\Gamma(s_1)\Gamma(s_1+s_2+d(\mathcal{G}_p,\mathcal{G}))}$$

where $beta(.,.)$ stands for the $Beta$ function. For a special case when $s_1 = 1$ and $s_2 = 1$ this becomes

$$p(\mathcal{G}) \propto \frac{1}{(d(\mathcal{G}_p,\mathcal{G})+1)}$$

1.7 Multiple Comparisons

In statistics, the problem of multiplicity has roots in both modeling and decision/inference settings. In frequentist literature, the burden is often addressed inferentially considering controlling family wise errors (FWE) or more recently the false discovery rates (FDR). For a Bayesian, part of this problem can be dealt with by the use of prior distributions. When dealing with a continuous model

space, regularization imposed by the prior distribution adjusts for multiplicity in the model by shrinking the estimates closer to their mean (Gelman et al., 2012), hence adjustment for multiplicity in the decision is simplified. This is not the case for discrete model spaces, where part of decision is encoded into the specification of the prior distribution, and a formulation that does not account for the number of comparisons has shown to have detrimental results (Scott and Berger, 2010). A prior distributions as such, however, do not completely relieve the practitioner of decision that still needs to account for the multiplicity.

When there agreeable domain specific standards for the level of "significance" such as 0.05 in social science or 10^{-4} in microbiology, Bayesians for the most part do not have to worry about multiplicity in practice (Gelman et al., 2012). However, when this is not the case one need to decide on how much loss one is willing to incur, which gives a rise to a loss function or utility function leading to the idea of optimal decision making. The use of decision theory to correct for multiplicity is traced back to Duncan (1965). A great review of the matter is summarized in Berry and Hochberg (1999). Connections between these loss functions to the results obtained from the classical perspective have been recently explored. These include p-value (Rice, 2010), confidence intervals (Thulin, 2012), and FDR (Efron and Tibshirani, 2006; Genovese and Wasserman, 2003). In terms of multiple comparison, Müller et al. (2006) lay out a foundation for methods to explicitly controls for the Bayesian posterior expected FDR and proposes options for several loss functions.

1.7.1 False discovery rates

The idea of controlling for false discovery rates (FDR) was introduced by Benjamini and Hochberg (1995) to address the problem of massively multiple compar-

isons. For m exchangeable hypotheses, let the underlying truth indicator $r_i = 1$, $(i = 1, \ldots, m)$ when effect is present and $r_i = 0$ otherwise. Corresponding to each hypothesis, we also define a decision indicator d_i, so that $d_i = 1$ when effect i is decided to be present and 0 otherwise. Then FDR is defined as

$$FDR = \frac{\sum d_i(1 - r_i)}{\sum d_i}.$$

Because r is unknown, in classical setting, one proceeds by taking the expectation over repeated samples. Although controlling for FDR is most commonly done, we can also control for other error rates such as the false negative rates (FNR) defined in similar fashion as

$$FNR = \frac{\sum(1 - d_i)r_i}{m - \sum d_i}.$$

In the Bayesian setting, posterior expectation equivalent are defined as

$$\overline{FDR} = \int FDR(d, r)dp(r|y) = \frac{\sum d_i(1 - v_i)}{\sum d_i}$$

$$\overline{FNR} = \int FNR(d, r)dp(r|y) = \frac{\sum(1 - d_i)v_i}{m - \sum d_i}$$

where $v_i = P(r_i = 1 \mid Y)$.

CHAPTER 2

BAYESIAN MODELING OF POPULATION PHARMACOGENETICS

2.1 Introduction

It is estimated that about a quarter of all FDA approved drugs get later relabeled for dose reduction due to safety concerns (Wacheck, 2010). Drug metabolism is thought to vary by person, according to characteristics that potentially escape standard covariate information (age, weight, gender, etc.) and are often related to more subtle variations involving the individual's genetic makeup (Ring and Kroetz, 2002). In fact, a considerable body of evidence suggests that single nucleotide polymorphisms (SNPs) in genes encoding drug transporters, drug-metabolizing enzymes, enzymes involved in DNA biosynthesis and repair might determine drug efficacy and toxicity (Shastry, 2005).

In this context, we propose a Bayesian modeling framework to assess interactions between inherited genetic traits, measured by SNPs, and drug metabolization dynamics. This concept of using hereditary genetic information to improve our understanding of pharmacokinetics is referred to as pharmacogenetics.

A typical pharmacogenetics study provides data in the form of: absorption dynamic data $\boldsymbol{Y}$, including measurement of one or multiple metabolites over time; genotype information $\tilde{\boldsymbol{Z}}$, usually in the form of SNPs; and standard baseline pa-

tients characteristics $\boldsymbol{X}$. Our motivating case study (Innocenti et al., 2004b; Iyer et al., 2002), for example, includes concentration trajectories for four metabolites of the drug irinotecan (Fig. 2.3), about 40 target SNPs and several baseline patient characteristics (Section 2.5).

The biomedical understanding of drug absorption dynamics is often represented as a compartment model. This representation of a biological process is in itself of great scientific interest. However, the pharmacokinetics literature is too vast to review here and we maintain our focus on its statistical aspects in population pharmacokinetics (PopPK). Given an expert elicited compartment model, PopPK deals with the task of producing statistical inference for the variation of PK parameters across a population. In the PopPK literature, hierarchical nonlinear models (Davidian and Giltinan, 2003) have been a popular choice (Sheiner and Steimer, 2000), as they allow for natural modeling of individual- and population-level variability. Bayesian approaches to PopPK inference are developed in Wakefield (1996) and Gelman et al. (1996). Other influential works includes modeling of pharmacokinetic and pharmacodynamic (PK/PD) data (Wakefield et al., 1999) and extensions to nonparametric inference Rosner and Müller (1997).

The idea of incorporating genetic information in the analysis of PopPK models was already introduced by Wakefield et al. (1999). The field has since then considered approaches focusing on a candidate gene (Ring and Kroetz, 2002), with a promise of direct clinical application, and larger exploratory genome-wide association studies (GWAS) (Klein et al., 2005). Whereas the former approach can be criticized for excessive reductionism, the latter point of view meets enormous challenges, both from a sample size perspective (Wu and Lin, 2010; Uher et al., 2010) and from a computational/inferential prospective (He and Lin, 2011).

More recently, a compromise between these two strategies has focused on

subsets of genes, selected *a priori* as candidate pathways that potentially explain differential drug metabolism (Johnson et al., 2013). Although the selection itself is a difficult scientific challenge, there is evidence supporting the validity of such procedures (Yang et al., 2005). For successful examples of pharmacogenetics research see Yiannakopoulou (2013), Rosner et al. (2008) and Bertrand and Balding (2013).

In this paper we aim to provide a solid methodological foundation to the statistical analysis of pharmacogenetics data. In particular, we extend the PopPK model of Wakefield (1996) to include dependence on hereditary genetic information in the form of SNPs. We utilize the idea of sparsity as in Bertrand and Balding (2013) in order to select meaningful gene-PK parameter interactions. However, rather than including SNPs as fixed covariates, we model gene set ordinal probit scores in a multivariate fashion, explicitly accounting for the dependence structure between SNPs. Without doing so, the sparsity induced by the model may understate the association of correlated SNPs to relevant PK parameters. Also, this feature allows for natural adaption of missing data and measurement error under the Bayesian framework.

We show that the joint distribution of PK trajectories and SNPs variation can be modeled according to the Markov laws of a chain graph (Lauritzen, 1996). Under this general and intuitive framework, we discuss posterior inference and the basis of a Bayesian decision-theoretic approach to control for the false discovery rate (FDR).

The remainder of this chapter is organized as follows. In the following section we briefly summarize the background material necessary to understand the contents in this section. In section 2.3 we introduce a joint Bayesian pharmacogenetics model, followed by a description of estimation and inference in section 2.4.

In section 2.5 we illustrate the use of the proposed method through a case study involving the pharmacogenetics of the anti-cancer drug irinotecan. We conclude with a brief discussion 2.6.

2.2 Background

2.2.1 Brief summary of biological terms

The heredity information from our parents is stored in the nucleus of our cells as *chromosomes*. Chromosomes are comprised of deoxyribonucleic acid (DNA) and proteins. Each DNA molecule codes the heredity information as a sequence of *nucleotide* or *basepair* (bp). Specific positions on DNA are called *loci* (singular *locus*) and a variant on a specific locus is called *allele*. A *genotype* at a particular locus is defined by two alleles at the locus on two strands of DNA. When the two alleles are the same on both of the DNA strands, genotype is referred to as a *homozygous*, whereas if they are different, it is referred to as *heterozygous*. *Mutation* is a permanent alternation of loci/locus on DNA sequence(s) that happens by substitution, insertion, deletion, duplication, or some combinations of these alterations of the basepairs on DNA. Mutations that occur in more than 1% of the population are referred to as *polymorphisms*.

2.2.2 Single nucleotide polymorphisms

A specific polymorphism that is characterized by alteration on a locus of DNA is referred to as a *single nucleotide polymorphism* (SNP). In the strict definition, minor allele frequency (MAF), which is the rarer allele, must be prevalent in at least 1% of the population. SNPs occur frequently throughout the genome and tend to be relatively stable genetically, making them particularly suitable as indicators

for a person's genetical variability. One thing to keep in mind is that, although SNPs are indicator of a person's genetic disposition that signifies person's hereditary factors, they are not necessarily the cause of phenotypic differences. There are ongoing projects to identify and catalog SNPs and this information is becoming available through public databases (Phillips, 2007), including dbSNP from NCBI (Sherry et al., 2001).

2.2.3 Problems with dosage determination regimen

Under the current clinical trials protocol, dosage is determined using the "maximum tolerated dose" (MTD) concept. "Maximum tolerated dose" (MTD) is defined as the maximum dose at which subjects do not exhibit "dose limiting toxicities" (DLT) as outlined in the study protocol in the "first in men" phase I clinical trial. Then this MTD is used as the "recommended phase two dose" (RPTD) for subsequent phase II clinical trials where the dose-response study is conducted.

This paradigm originates from the field of oncology, where it is believed that toxicities acts as proxies for the activity of a drug. MTD is optimal protocol only when there is a strong linear correlation between the toxicity and the effect. When the relationship is non-linear and complex, it is easy to be in a situation where the MTD is suboptimal choice, as shown in the hypothetical figure 2.1. Especially in the modern pharmacology, where a molecular therapeutics is designed for specific biological target in mind with side effects not necessary correlated with the targeted, not only plausibility but also the ethical validity of such crude method is questionable. (Wacheck, 2010).

Another problem surrounding dosage determination in clinical trials is population discrepancy between the phases. In a typical trial, the dosage is determined

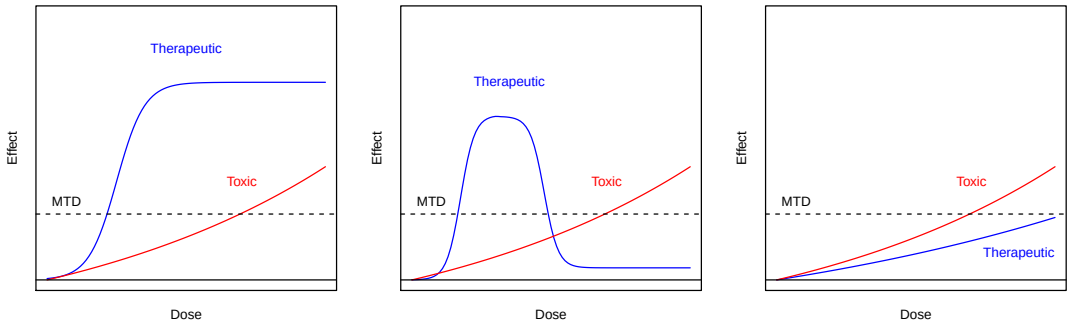


Figure 2.1: Hypothetical dose-effect relationship scenarios that maximum tolerance dose (MTD) regime is suboptimal choice, adopted from (Wacheck, 2010). The two figures on the right shows cases where the maximum therapeutic effect is achieved much faster than the toxicity limit. The central figure is the worst-case scenario for MTD where the therapeutic effect vanishes close to the toxicity limit. The right figure is a case where the trial should be canceled because the toxicity is consistently larger then the therapeutic effect.

in the phase I stage of the trial using healthy young subjects who for the most part do not resemble the actual patients that the drug is intend for. Some measurable information such as age or the body characteristics are used to adjust for the differences but validity of such procedure is unknown.

2.2.4 Pharmacokinetics and Pharmacodynamics

The pharmacological process of drug metabolization is often modeled using the two partitioned processes of pharmacokinetics and pharmacodynamics. The pharmacokinetics models (PK) are used to model processes involving drug absorption, distribution and elimination of a drug after its introduction to the body. On the other hand, pharmacodynamics models (PD) are used to model the relationship between drug concentrations and surrogate biological responses and possibly this response's effect on the clinical outcomes.

The division of PK and PD allows separation in factors that influence the inter-individual variability in pharmacokinetics from those that influence pharmacodynamics (Wakefield et al., 1999), under the implicit assumption that these processes are unlikely to be causally interrelated. Furthermore, the fact that in clinical trials, dose-concentration study is separate from and usually precedes large-scale efficacy studies, gives practical motives in the choice.

Pharmacokinetics is usually simplified into a set of compartment model where compartments are defined by grouping together parts of the body that share similar kinetics for a particular drug. The flow of drug between the compartments is often assumed to be proportional to the drugs in the donor compartment, leading to a system of first-order differential equations (Wakefield, 1996). Linear equations are popular in the literature since it is straightforward to obtain expressions for the amounts of drug in the different compartments as a function of time (Wakefield et al., 1999).

For an overview of PK/PD modeling following literature are well known: Wakefield et al. (1999) summarizes hierarchical PK/PD model with relation to the clinical trial process; Davidian and Giltinan (2003) reviews various PK models embedded in the nonlinear models.

2.2.5 Pharmacokinetics (PK) Model

Let $g_{ik}(t, \boldsymbol{\theta_i}., D_i)$, be the expected concentration associated with a compound of interest in compartment k, $(k = 1, ..., K)$, at time $t \in [t_0, t_n]$, for the i^{th} individual, $(i = 1, ..., N)$. In the forgoing formulation $g_i(\cdot) = (g_{i1}(\cdot), \cdots, g_{iK}(\cdot))'$ represents a functional, usually arising as a particular solution to a K-dimensional system of differential equations, given a history of doses D_i, a set of parameters $\boldsymbol{\theta_i}. = (\theta_{i1}, ..., \theta_{iv}, ..., \theta_{iV})'$, and initial values $g_i(t_0, \boldsymbol{\theta_i}., D_i) = g_i(t_0, D_i)$.

The observed concentration $y_{ik}(t)$ for subject i, compartment k at time t is modeled as regression with constant coefficient of variation as:

$$y_{ik}(t) = g_{ik}(t, \boldsymbol{\theta_i}., D_i) + \varepsilon_{ik}(t), \tag{2.1}$$

where $\varepsilon_{ik}(t) \sim \mathcal{N}(0, g_{ik}(t, \boldsymbol{\theta_i}., D_i)^{\mathfrak{r}}\sigma^2_{\varepsilon_{ik}})$, for $\mathfrak{r} \geq 0$.

Equivalently, it is popular to model the relation ships as a log-log linear formulation as:

$$\log\{y_{ik}(t)\} = \log\{g_{ik}(t, \boldsymbol{\theta_i}., D_i)\} + \varepsilon_{ik}(t), \tag{2.2}$$

where $\varepsilon_{ik}(t) \sim \mathcal{N}(0, \sigma^2_{\varepsilon_{ik}})$.

2.2.6 Pharmacodynamics (PD) Model

The specification of PD model is similar to PK model. Let $z_{i\mathfrak{h}}(t)$ be measured response for subject i, response $\mathfrak{h}$, at time t. For a response $\mathfrak{h}$, $(\mathfrak{h} = 1, ..., \mathfrak{H})$, at time $t \in [t_0, t_n]$, for the i^{th} individual, $(i = 1, ..., N)$, expected effect is denoted as $f_{i\mathfrak{h}}(t, \boldsymbol{\phi_i}, \boldsymbol{\theta_i})$, given $\mathfrak{U}$ dimensional PD parameters $\boldsymbol{\phi_i}. = (\phi_{i1}, ..., \phi_{iv}, ..., \phi_{iV})'$ and V dimensional PK parameters $\boldsymbol{\theta_i}. = (\theta_{i1}, ..., \theta_{iv}, ..., \theta_{i\mathfrak{U}})'$. If we use $f_i(\cdot) = (f_{i1}(\cdot), \cdots, f_{i\mathfrak{H}}(\cdot))'$ to denote the $\mathfrak{H}$ dimensional function, the PD model is defined as.

$$z_i(t) = f_i(t, \boldsymbol{\phi_i}, \boldsymbol{\theta_i}) + \varphi_{i\mathfrak{h}}(t), \tag{2.3}$$

where $\varphi_{i\mathfrak{h}}(t) \sim \mathcal{N}(0, \sigma^2_{\varphi_{i\mathfrak{h}}})$.

2.2.7 Population PK model

Suppose that for each individual we observe a set of $\mathcal{P}_\theta$ covariate measurements $(X_{i1}, ..., X_{i\mathcal{P}_\theta})$. A popular covariate model (Wakefield 1996, Gelman et al. 1996)

assumes a log-linear relationship between individual Pk parameter and covariates

$$\log(\theta_{i\nu}) = \boldsymbol{X}_i\boldsymbol{\beta}_\nu + \delta_{i\nu}; \tag{2.4}$$

where $\boldsymbol{X}_i = (1, X_{i1}, ..., X_{i\mathcal{P}_\theta})'$, $\boldsymbol{\beta_\nu} = (\beta_{\nu\,0}, \beta_{\nu\,1}, ..., \beta_{\nu\,\mathcal{P}_\theta})'$ and $\delta_{i\nu}$ is a subject specific random error.

2.2.8 Pharmacogenetics or Pharmacogenomics?

Pharmacogenetics (PKGx) and Pharmacogenomics (PKGm) both refer to the field of study on how variant in genetic polymorphism makes a difference in a response to drugs. Two terms tend to be used interchangeably, although some discipline use them to differentiate between germ line and somatic mutation types (Roses, 2000) or size and complexity of the genes involved (Roden et al., 2006) there is no overarching consensus definition prior to this thesis.

Since we have no interest in further complicating the issue by proposing a new definition, nor do we believe we can sort out the discrepancies, we take the same approach as Senn (2008):

> In this chapter I shall take the Humpty Dumpty line on linguistics: words mean what I say they mean. That is, whatever the difference between pharmacogenetics and pharmacogenomics may or may not be, this chapter is about differences (presumed or real) in the effect of treatment due to genetic variation between patients, how to detect such differences and what to do with them. From now on I shall use the term pharmacogenetics to describe this field.

Henceforth we will use the word Pharmacogenetics to indicate the study of how genetic differences influence the variability in patients' responses to drugs (Roses,

2000).We hope the reader will accept this crude definition at least for the extent of this thesis.

2.3 Model Formulation

We consider pharmacokinetics in the form of a data array denoted as $\boldsymbol{Y}$, hereditary genetic information in the form of SNPs coded as ordinal trinary vectors $\tilde{\boldsymbol{Z}}$, and baseline clinical information denoted by $\boldsymbol{X}$ and $\boldsymbol{U}$, possibly overlapping. Let $g(\boldsymbol{\theta})$, be a functional of a compartment models's parameters and $\boldsymbol{Z}$ be latent ordinal probit scores for the trinary mutation vector $\tilde{\boldsymbol{Z}}$, we define the following probability model:

$$p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{U}) = \underbrace{p(\boldsymbol{Y} \mid g(\boldsymbol{\theta}))}_{\text{PK}} \underbrace{p(\boldsymbol{Z} \mid \boldsymbol{U})}_{\text{Gx}} \underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{X})}_{\text{PopPKGx}}. \tag{2.5}$$

Equation (2.5) implies a model $p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{U})$ for the observed SNPs after integrating w.r.t. the latent probit scores over the relevant subintervals (see later). In (2.5) we assume that drug absorption dynamics $\boldsymbol{Y}$ are conditionally independent of genetic information $\tilde{\boldsymbol{Z}}$, given PK parameters $\boldsymbol{\theta}$.

2.3.1 A population pharmacokinetics model

Let $g_{ik}(t, \boldsymbol{\theta_i}, D_i)$, be the expected concentration associated with a metabolite k, $(k = 1, ..., K)$, at time $t \in [t_0, t_n]$, for the i^{th} individual, $(i = 1, ..., N)$. We also let $g_i(\cdot) = (g_{i1}(\cdot), \cdots, g_{iK}(\cdot))'$ represent a functional, usually arising as a particular solution to a K-dimensional system of differential equations, given a history of doses D_i, a set of parameters $\boldsymbol{\theta_i} = (\theta_{i1}, ..., \theta_{iv}, ..., \theta_{iV})'$, and initial values $g_i(t_0, \boldsymbol{\theta_i}, D_i) = g_i(t_0, D_i)$.

The observed concentration $Y_{ik}(t)$ for subject i, metabolite k at time t is modeled as:

$$\log\{Y_{ik}(t)\} = \log\{g_{ik}(t, \boldsymbol{\theta_i}, D_i)\} + \varepsilon_{ik}(t), \tag{2.6}$$

where $\varepsilon_{ik}(t) \sim \mathcal{N}(0, \sigma^2_{\varepsilon_{ik}})$. An example of compartmental model defining $g_{ik}(t, \boldsymbol{\theta}_i, D_i)$ for the pharmacokinetics of the drug Irinotecan is reported in Appendix 5.2.1.

Suppose that for each individual we observe a set of p_X covariates $(X_{i1}, ..., X_{ip_X})$. A popular model (Wakefield, 1996; Gelman et al., 1996) assumes a log-linear relationship between individual PK parameter and covariates

$$\log(\theta_{i\nu}) = \boldsymbol{X}_i'\boldsymbol{\beta}_\nu + \delta_{i\nu}; \tag{2.7}$$

where $\boldsymbol{X}_i = (1, X_{i1}, ..., X_{ip_X})'$, $\boldsymbol{\beta_\nu} = (\beta_{\nu\,0}, \beta_{\nu\,1}, ..., \beta_{\nu\,p_X})'$ and $\delta_{i\nu}$ is a subject-specific effect and $v = 1, \ldots, V$ indexes the PK parameters. In section 2.3.3 we will introduce a model for $\delta_{i\nu}$ that explains subject-specific effects as a regression on SNPs. This is where the genetics enters the pharmacogenetics.

2.3.2 Modeling single nucleotide polymorphism (SNP) array

Let $\tilde{Z}_{iq}$ be a trinary polymorphism indicator, for gene q $(q = 1, ..., Q)$ and subject i $(i = 1, ..., N)$; taking values $\tilde{Z}_{iq} = 1$ if gene q has a polymorphism on one of the chromosomes $\tilde{Z}_{iq} = 2$ if gene q has polymorphism on both chromosomes and $\tilde{Z}_{iq} = 0$ otherwise. We model the observed polymorphism $\tilde{Z}_{iq}$ as an ordinal trinomial random variable with probabilities $p^{(1)}_{iq} = P(\tilde{Z}_{iq} = 1)$, $p^{(2)}_{iq} = P(\tilde{Z}_{iq} = 2)$ and $p^{(0)}_{iq} = P(\tilde{Z}_{iq} = 0) = 1 - (p^{(1)}_{iq} + p^{(2)}_{iq})$.

These ordinal probit scores can be related to a set of covariates $\boldsymbol{U}_i = (1, U_{i1}, ..., U_{ip_U})'$, which may or may not coincide with the covariate set in equation (2.7), through a probit link as in Albert and Chib (1993). Using the notation of Chen and Dey

(2000), we assume there exists a continuous random quantity $Z_i \sim N_Q(\boldsymbol{U}_i'\boldsymbol{\gamma}, \Sigma_Z)$, where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_Q)$ is a $(p_U + 1) \times Q$-dimensional matrix of unknown parameters and the matrix Σ_Z maps the dependence structure between genes via polychoric correlations (Ronning and Kukuk 1996). Then without loss of generality, we observe

$$\tilde{Z}_{iq} = \begin{cases} 0 & if \quad Z_{iq} < 0 \\ 1 & if \quad 0 \leq Z_{iq} < 1 \\ 2 & if \quad 1 \leq Z_{iq} \end{cases} ; \tag{2.8}$$

for all $q = 1, \ldots, Q$ and $i = 1, \ldots, N$. The implied ordinal scale on Z is justified by the assumption that risk conferred by a heterozygous SNP (having one risk allele) lies somewhere between baseline risk and a homozygous SNP (having two risk alleles) that is, between the risk of a recessive and dominant traits (Wittkowski et al., 2013). In summary, Z are latent ordinal probit scores for the trinary SNPs $\tilde{Z}$. The scores Z are essentially a continuous version of $\tilde{Z}$. In the following we will refer to Z simply as SNPs, keeping in mind that the mapping (2.8) links Z to the actually observed trinary SNP.

Let $\boldsymbol{Z} = (\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_N')'$ be the $(N \times Q)$ matrix of ordinal probit scores and let $\boldsymbol{m}$ be an $(N \times Q)$ mean matrix, with entries $m_{iq} = \boldsymbol{U}_i'\boldsymbol{\gamma}_q$. The multivariate characterization of the scheme in (2.8) is completed with a prior $p(\boldsymbol{Z}|\boldsymbol{U}, \boldsymbol{\gamma})$ having matrix normal distribution as

$$(\boldsymbol{Z} - \boldsymbol{m}) \sim \mathcal{MN}(I_N, \Sigma_Z), \tag{2.9}$$

where I_N is an $(N \times N)$ identity matrix (Chib and Greenberg 1998; Chen and Dey 2000).

For each subject $(i = 1, \ldots, N)$, let $\tilde{\boldsymbol{Z}}_i = (\tilde{Z}_{i1}, ..., \tilde{Z}_{iq})$ denote a collection of polymorphism trinary indicators on all Q genes. For subject i, the candidate

gene set mutation score $P(\tilde{\boldsymbol{Z}}_i)$ can be expressed in terms of the latent scores $\boldsymbol{Z}_i$ as

$$P(\tilde{\boldsymbol{Z}}_i) = \int_{\mathcal{A}_{iQ}} \cdots \int_{\mathcal{A}_{i1}} (2\pi)^{Q/2} |\Sigma_Z|^{1/2} \exp\{-1/2(\boldsymbol{Z}_i - \boldsymbol{m}_i)' \Sigma_Z^{-1} (\boldsymbol{Z}_i - \boldsymbol{m}_i)\} \, d\boldsymbol{Z}_i, \tag{2.10}$$

where the event $\mathcal{A}_{iq}$ is the interval $(-\infty, 0)$ if $\tilde{Z}_{iq} = 0$, the interval $[0, 1)$ if $\tilde{Z}_{iq} = 1$ and the interval $[1, \infty)$ if $\tilde{Z}_{iq} = 2$.

2.3.3 A Bayesian pharmacogenetics (PKGx) model

In (2.6) and (2.8-2.9) we defined a sampling model for SNPs and drug concentrations. We now complete the model by relating the PK parameters $\boldsymbol{\theta}$ with the SNP ordinal probit scores $\boldsymbol{Z}$ by way of modeling $\delta_{i\nu}$ in (2.7) as a regression on $\boldsymbol{Z}$.

Let $\boldsymbol{\theta}$ denote an $(N \times V)$ matrix of subject-level PK parameters, where the matrix entry θ_{iv}, $(i = 1, ..., N;\ v = 1, ..., V)$ denotes the v^{th} PK parameter characterizing a certain aspect of the concentration curves observed for subject i. We elaborate the prior model (2.7)

$$\log(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\beta} + Z\boldsymbol{\rho} + \boldsymbol{\eta}; \tag{2.11}$$

where $\boldsymbol{X}$ is an $(N \times (p_X + 1))$ matrix of baseline covariates and $\boldsymbol{\beta}$ is a $((p_X + 1) \times V)$ matrix of unknown regression coefficients. The dependence between PK parameters and SNP genotypes are modeled as a linear regression on $\boldsymbol{Z}$ with a $(Q \times V)$ matrix of unknown regression parameters $\boldsymbol{\rho}$.

The conditional distribution of $log(\boldsymbol{\theta})$ is specified through $\boldsymbol{\eta}$ as a matrix nor-

mal distributions.

$$p(\boldsymbol{\eta}) =_d \mathcal{MN}_{NV}(I_N, \Sigma_\theta). \tag{2.12}$$

When extra variability is present in the data, one can account for it by expanding $\boldsymbol{\eta}$ to be a scale mixture of multivariate Normal distributions (West 1984). More precisely, let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)'$ be a vector of random variance inflation parameters, the joint distribution

$$p(\boldsymbol{\eta}, \boldsymbol{\tau}) =_d \mathcal{MN}_{NV}(D_\tau, \Sigma_\theta) p(\boldsymbol{\tau} \mid \upsilon), \quad D_\tau = diag(\boldsymbol{\tau}), \tag{2.13}$$

admits the marginal $p(\boldsymbol{\eta} \mid \Sigma) =_d \mathcal{MT}_{NV}(\upsilon, I_N, \Sigma_\theta)$, whenever $\tau_i \sim_{iid} \mathcal{G}(\upsilon/2, \upsilon/2)$; a matrix $\mathcal{T}$ with degrees of freedom υ and column covariance Σ_θ.

The formulation in (2.11), together with (2.6) and (2.8),defines a joint model for the SNPs $\tilde{\boldsymbol{Z}}$ and the concentrationtrajectories $\boldsymbol{Y}$ as:

$$p(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\theta}) \propto p(\boldsymbol{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{Z})\, p(\boldsymbol{Z}), \tag{2.14}$$

suppressing dependence on the fixed covariates $\boldsymbol{X}$ and $\boldsymbol{U}$ in the notation. Dependence between $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Z}}$ is introduced hierarchically by defining the joint distribution $p(\boldsymbol{\theta}, \boldsymbol{Z}) \propto p(\boldsymbol{\theta} \mid \boldsymbol{Z})\, p(\boldsymbol{Z})$ and the deterministic mapping $\tilde{\boldsymbol{Z}} \sim f(Z)$ in (2.8). The matrix $\boldsymbol{\rho}$ introduces stochastic dependence between the genetic ordinal probit scores $p(\boldsymbol{Z})$ and the PK parameters $\boldsymbol{\theta}$. The between-column covariance matrix of $\boldsymbol{\theta}$. Σ_θ, on the other hand, defines the dependence structure within PK parameters, given $\boldsymbol{\rho}$.

In the following section, we show how structural restrictions on Σ_θ, Σ_Z and $\boldsymbol{\rho}$ correspond to specific assumptions about the Markov structure of the pharmacogenetics distribution $p(\boldsymbol{Y}, \tilde{\boldsymbol{Z}})$. However, we will use the notion of a graphical model only to highlight and summarize the proposed model structure. In other

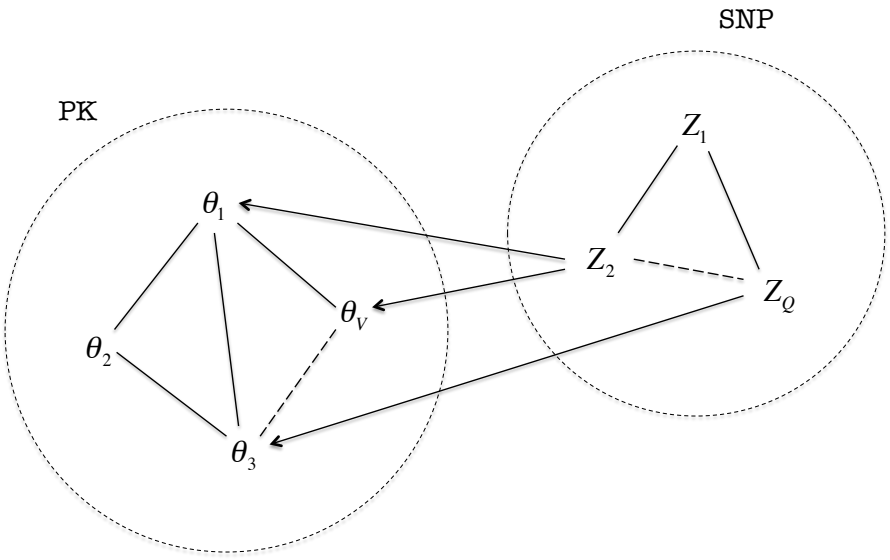


Figure 2.2: Chain graph representation for the PKGx Markov structure.

words, the graph $\mathcal{G}$ will only be a summary $\mathcal{G}(\Sigma_\theta, \Sigma_Z, \boldsymbol{\rho})$ of the already defined other parameters. The prior model $p(\Sigma_\theta, \Sigma_Z, \boldsymbol{\rho})$ will implicitly define a prior $p(\mathcal{G})$ on the graph, and there is no separate definition of $p(\mathcal{G})$.

2.3.4 PKGx Markov structure and chain graphs

The Markov structure of the model proposed in (2.11) can be represented as a graphical model (Lauritzen 1996). A graph is a pair $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, composed of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V}\}$. When the set of nodes $\mathcal{V}$ represents a collection of random quantities, the edges $\mathcal{E}$ can be used to identify

the full set of conditional independence relations between the components of $\mathcal{V}$ via, what is often called, the *global Markov property* associated with $\mathcal{G}$ (Lauritzen 1996).

The conditional dependence fabric relating the random quantities modeled in (2.11) is best characterized in terms of a chain graph (Lauritzen and Wermuth 1989). A chain graph is a graph with both directed and undirected edges, but no partially directed cycles. See Lauritzen and Richardson (2002) for a technical discussion of chain graph models.

In our model, undirected edges describe conditional dependence within the PK parameters $\boldsymbol{\theta}$ and within the latent mutation scores $\boldsymbol{Z}$, while directed edges describe a potential causal pathway between PK parameters and mutations. Figure 2.2 illustrates the structure of our model for a hypothetical configuration of $\mathcal{G}$.

The graphical topology, arising from a representation of (2.11) as a chain graph, has an appealing causal interpretation of genetic mutation on the PK dynamic of a compound of interest. Meanwhile, the dependence structure of $\boldsymbol{Z}$ defines implicit clustering of the SNPs that should be taken into account when considering this regression.

Under the normal distributions in (2.9) and (2.11) different configurations of $\mathcal{G}$ are defined by the appropriate placement of 0's in the concentration matrices $\Omega_\theta = \Sigma_\theta^{-1}$ and $\Omega_Z = \Sigma_Z^{-1}$ and in the coefficient matrix $\boldsymbol{\rho}$. In particular, setting $\Omega_{(i,j)} = 0$ in one of the foregoing concentration matrices corresponds to the absence of the undirected edge between "θ_i—θ_j", or similarly to the vanishing of the edge ("Z_i— Z_j"). On the other hand, setting the entry $\boldsymbol{\rho}_{(q,k)} = 0$ corresponds to the vanishing of the directed edge ("$Z_q \longrightarrow \theta_k$").

2.3.5 Model determination and prior distributions

Let $\boldsymbol{\Theta}$ be the full set of parameters defining the probabilistic scheme introduced in § 2.3.4. We seek inference for $\boldsymbol{\Theta}, \mathcal{G})$ given $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Z}}$ and consider lasso-type regularization (Tibshirani, 1996) for the estimation of the graphical models (Friedman et al., 2008) by placing a sparcity inducing prior (Park and Casella, 2008; Hans, 2009; Wang, 2012) on $\boldsymbol{\Theta}$ to infer the association structure $\mathcal{G}$ in the posterior distribution of $\boldsymbol{\Theta}$.

We note that this type of prior is absolutely continuous with respect to the Lebesque measure. Therefore, the posterior probability of any edge of $\mathcal{G}$ not existing is exactly zero. This feature can be criticized for not facilitating automatic model selection. We discuss the issue in more detail in §2.4.2, where we propose a decision-theoretic procedure aimed at controlling expected posterior error rates.

2.3.5.1 Prior distributions on concentration matrices

We follow the approach of Wang (2012) and place independent graphical lasso priors on Ω_θ and Ω_Z. Specifically, a concentration matrix Ω is distributed according to a graphical lasso distribution if:

$$p(\Omega \mid \{\lambda_{ij}\}_{i\leq j}) \propto C_{\{\lambda_{ij}\}_{i\leq j}} \prod_{i<j} \left\{ \tfrac{\lambda_{ij}}{2} \exp(-\lambda_{ij}|\omega_{ij}|) \right\} \prod_{i=1}^{p} \left\{ \tfrac{\lambda_{ii}}{2} \exp\left(-\tfrac{\lambda_{ii}}{2}\omega_{ii}\right) \right\} 1_{\Omega \in M^+},$$

and

$$p(\{\lambda_{ij}\}_{i<j} \mid \{\lambda_{ii}\}_{i=1}^{p}) \propto C_{\{\lambda_{ij}\}_{i\leq j}} \prod_{i<j} \lambda_{ij}^{r-1} \exp(-s\lambda_{ij}).$$

Here, M^+ is the cone of positive definite matrices in $\mathbb{R}^{p\times p}$. The prior assumes a double exponential distribution for the off-diagonal elements and an exponential distribution distribution for the diagonal entries. The hyperparameters for λ_{ij} are given independent gamma distributions. Finally, the term $C_{\{\lambda_{ij}\}_{i\leq j}}$ is an

intractable normalizing constant. Fortunately, $C_{\{\lambda_{ij}\}_{i\leq j}}$ cancels out in the acceptance probabilities for the Metropolis-Hastings acceptance probabilities.

2.3.5.2 Prior distribution on $\boldsymbol{\rho}$

The regression coefficients $\boldsymbol{\rho}$ represent the strength of association between SNP and PK parameters in the graph $\mathcal{G}_{Z\to\theta}$. Brown et al. (1998) first discussed multivariate Bayesian variable selection assuming separability in the row and column covariance structure. This assumption is unwarranted in our case as it would impose too many restrictions on how SNP variability affects drug absorption.

Instead we extend the adaptive lasso penalization introduced in the concentration matrix prior to a prior for $\boldsymbol{\rho}$ (Sun et al., 2010). Specifically, we model each component ρ_{qv} independently as

$$p(\rho_{qv} \mid \kappa_{qv}) = -\frac{1}{\kappa_{qv}} \exp\left(-\frac{|\rho_{qv}|}{\kappa_{qv}}\right), \quad p(\kappa_{qv} \mid \delta, \phi) = \frac{\phi^{\delta}}{\Gamma(\delta)} \kappa_{qv}^{-1-\delta} \exp\left(-\frac{\phi}{\kappa_{qv}}\right); \tag{2.15}$$

where $\delta > 0$ and $\phi > 0$ are hyperparameters. We follow the suggestion of Sun et al. (2010) and model $(\delta, \phi) \propto \frac{1}{\phi}$.

The prior $p(\Omega_\theta, \Omega_Z, \boldsymbol{\rho})$ implies an approximate prior on $\mathcal{G}$ by defining the prior probability for the coefficients ω_{ij} and ρ_{ij} being small and practically equivalent to structural zeros. We will still discuss a formal criterion for judging what is "small".

2.4 Estimation and Iinference

2.4.1 Markov Chain Monte Carlo estimation

Whenever possible, we exploit conjugacy so that posterior simulation for most parameters can be based on Gibbs sampling. We briefly describe the algorithm in this section, (See table 2.4.1 for a schematic outline). Detailed calculations are reported in the appendix.

Parameter	Algorithm	Section
$\boldsymbol{\theta}$	Metropolis Hasting	2.4.1.1
Ω_Z, Ω_θ	Block Gibbs Sampling	2.4.1.2
$\boldsymbol{\rho}$	Gibbs Sampling	2.4.1.3
$\boldsymbol{Z}$	Gibbs Sampling	2.4.1.4
$\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2, τ	Gibbs Sampling	Appendix

2.4.1.1 Updating of PK parameters $\boldsymbol{\theta}$

The conditional posterior distribution of $\boldsymbol{\theta}$ is usually unavailable in closed form, as it depends on the compartmental model structure and its solution $g(\cdot)$. It is, however, straightforward, in principle, to simulate from such distribution via Metropolis-Hastings. In practice, the evaluation of $g(\cdot)$ may be costly. We therefore use a joint proposal of $\boldsymbol{\theta}_i$, $(i = 1, \ldots, N)$ with a V-variate proposal.Our implementation is based on adaptive Metropolis, (Haario et al., 2001; Roberts and Rosenthal, 2009). Specifically, we generate a proposal from a V-variate normal distribution centered at the current $\boldsymbol{\theta}_i$, using the empirical covariance matrix for the scaling parameter. Detailed calculations appear in the appendix.

2.4.1.2 Updating Ω_θ, Ω_Z

To sample from conditional distribution of the inverse covariance matrix given the adaptive graphical lasso prior, we follow the algorithm of Wang (2012) and use the data-augmented block Gibbs sampler. The algorithm sequentially updates each column of Ω while maintaining the positive definiteness. For Ω_Z and Ω_θ respectively, conditional posterior distributions depend on the data through summary statistics

$$\begin{aligned} S_Z &= [\boldsymbol{Z} - \boldsymbol{U}\boldsymbol{\gamma}]^T [\boldsymbol{Z} - \boldsymbol{U}\boldsymbol{\gamma}], \text{ and} \\ S_\theta &= [\log(\boldsymbol{\theta}) - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\rho})]^T \boldsymbol{D}_\tau [\log(\boldsymbol{\theta}) - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\rho})] + [\boldsymbol{\beta} - \mathbf{b}_0]^T \boldsymbol{B}_0 [\boldsymbol{\beta} - \mathbf{b}_0]. \end{aligned}$$

Block Gibbs sampler (Wang, 2012) The idea is based on the fact that each column (equivalently row) of the inverse covariance matrix conditioned on all other parameters have a standard distribution. In this section we sketch the algorithm, for more detailed derivation see Wang (2012).

The joint posterior distribution $p(\Omega, \varkappa, \lambda \mid \boldsymbol{Y})$ for a graphical lasso model specified as in section 2.3.5.1 is

$$\begin{aligned} p(\Omega, \varkappa, \lambda \mid \boldsymbol{Y}) \propto\ & |\Omega|^{n/2} \exp\left\{-tr\left(\frac{1}{2}\boldsymbol{S}\Omega\right)\right\} \\ & \prod_{i<j} \left\{\varkappa_{ij}^{-1/2} \exp\left(-\frac{\omega_{ij}^2}{2\varkappa_{ij}}\right) \exp\left(\frac{\lambda_{ij}^2}{2}\varkappa_{ij}\right)\right\} \\ & \prod_{i=1}^{p} \left\{\exp\left(-\frac{\lambda_{ij}}{2}\omega_{ii}\right)\right\} 1_{\Omega \in M^+} \\ & \prod_{i<j} \lambda_{ij}^{r_1 - 1} \exp\{-r_2 \lambda_{ij}\} \end{aligned}$$

It is straight forward to show that

$$\lambda_{ij} \mid \Omega \sim Ga(1 + r_1, |\omega_{ij}| + r_2)$$

and

$$\frac{1}{\varkappa_{ij}} \mid \Omega, \lambda \sim Inv\text{-}N\left(\sqrt{(\lambda_{ij}/\omega_{ij})}, \lambda^2\right)$$

where $Inv\text{-}N$ stands for inverse Gaussian density.

To derive the full conditional distribution of $p \times p$ inverse covariance matrix Ω, we first define Υ to be symmetric matrix with diagonal element set to 0 and off diagonal element to be $\varkappa_{ij}$. Given sum of squares matrix $\boldsymbol{S}$ and Υ the full conditional density is defined iteratively for each column j,$j = 1, \ldots, p$. For each j let index set I to be defined as $I = \{1, \ldots, p\} \setminus j$. Given a partition for Ω, $\boldsymbol{S}$ and Υ s.t.

$$\Omega = \begin{pmatrix} \Omega_{II} & \omega_{Ij} \\ \omega_{jI} & \omega_{jj} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{II} & s_{Ij} \\ s_{jI} & s_{jj} \end{pmatrix}, \text{ and } \Upsilon = \begin{pmatrix} \Upsilon_{II} & \varkappa_{Ij} \\ \varkappa_{jI} & 0 \end{pmatrix}$$

The full conditional distribution for a column j of Ω is defined as

$$\begin{aligned} p(\omega_{Ij}, \omega_{jj} \mid \Omega_{II}, \Upsilon, \boldsymbol{Y}, \lambda) \quad &\propto \quad (\omega_{jj} - \omega_{Ij}^T \Omega_{II}^{-1} \omega_I j)^{n/2} \\ &\exp\left[-\frac{1}{2}\left\{\omega_{Ij}^T D_{\tau_{Ij}}^{-1} \omega_{Ij} + 2 s_{Ij}^T \omega_{Ij} + (s_{jj} + \lambda_{jj})\omega_{jj}\right\}\right] \end{aligned}$$

where $D_{\tau_{Ij}} = diag(\tau_{Ij})$.

By change of variable as

$$(\omega_{Ij}, \omega_{jj}) \rightarrow \left(\beta = \omega_{Ij}, \gamma = \omega_{jj} - \omega_{Ij}^T \Omega_{II}^{-1} \omega_{Ij}\right)$$

equation (2.16) can be rewritten as

$$p(\beta,\gamma \mid \Omega_{II},\Upsilon,\boldsymbol{Y},\lambda_{jj}) \propto \gamma^{n/2}\exp\left(-\frac{s_{jj}+\lambda_{jj}}{2}\gamma\right) \exp\left(-\frac{1}{2}\left[\beta^T\left\{D^{-1}_{\tau_{Ij}}+(s_{jj}+\lambda_{jj})\Omega^{-1}_{II}\right\}\beta+2s^T_{Ij}\beta\right]\right).$$

Which shows we can sample from independent Gamma and Gaussian distributions

$$(\beta,\gamma) \mid (\Omega_{II},\Upsilon,\boldsymbol{Y},\lambda) \sim Ga(n/2+1,(s_{jj}+\lambda_{jj})/2)N(-Cs_{jI},C)$$

where $C=\{D^{-1}_{\tau_{Ij}}+(s_{jj}+\lambda_{jj})\Omega^{-1}_{II}\}^{-1}$.

2.4.1.3 Updating Pharmacogenetics parameter $\boldsymbol{\rho}$

PKGx parameters $\boldsymbol{\rho}$, along with the hyperparameters in (2.15), can be sampled using the Gibbs sampler described in Sun et al. (2010). Part of the algorithm is based on adaptive rejection sampling (Gilks, 1992). Details are discussed by Sun et al. (2010).

2.4.1.4 Updating Probit Scores $\boldsymbol{Z}$

The full conditional distribution of $\boldsymbol{Z}_i$ is $N(\tilde{m}_{z_i},\tilde{S}^{-1}_{z_i})\prod_{q=1}^{Q} I\{z_{iq}\in\mathcal{A}_{iq}\}$ (Chen and Dey, 2000), where

$$\tilde{m}_{z_i}=\tilde{S}^{-1}_{z_i}\left[\Omega_Z\boldsymbol{\gamma}\boldsymbol{U}_i+\tau_i\boldsymbol{\rho}^T\Omega_\theta\left(\log(\boldsymbol{\theta}_i)-\boldsymbol{\beta}\boldsymbol{X}_i\right)\right], \quad \tilde{S}_{z_i}=\left(\Omega_Z+\tau_i\boldsymbol{\rho}^T\Omega_\theta\boldsymbol{\rho}\right),$$

and

$$\mathcal{A}_{iq} = \begin{cases} (-\infty, 0) & \text{if } \tilde{z}_{iq} = 0 \\ [0, 1) & \text{if } \tilde{z}_{iq} = 1 \\ [1, \infty) & \text{if } \tilde{z}_{iq} = 2 \end{cases} .$$

Sampling from a multivariate truncated Gaussian distribution difficult. Instead we generate each $\boldsymbol{Z}_q$ conditional on $\boldsymbol{Z}_{-q}$. See Appendix 5.2.3 for details.

2.4.2 Posterior inference from Monte Carlo samples

The posterior distribution $p\left(\Omega_\theta, \Omega_Z, \boldsymbol{\rho} \mid \boldsymbol{Y}, \tilde{\boldsymbol{Z}}\right)$ jointly encompasses all information concerning interactions between the SNP and the pharmacokinetic parameters. Nevertheless, in order to make sense of all the interactions, noteworthy effects need to be screened.

We are testing multiple statistical hypotheses simultaneously. A popular screening strategy is based on controlling error rates, like the false discovery rate (FDR) (Benjamini and Hochberg, 1995). From a Bayesian perspective, Müller et al. (2006) illustrate a decision-theoretic procedure aimed at minimizing the posterior expected false negative rate ($\overline{FNR}$), while controlling for the posterior expected false discovery rate ($\overline{FDR}$) at a level α. Some care is needed in our model, because applying this procedure separately to each of the parameters Ω_θ, Ω_Z, and $\boldsymbol{\rho}$, may lead to inflation in the overall FDR (Cai and Sun, 2009).

Specifically, let K be the number of non-exchangeable parameter sub vectors. We consider the sub vectors corresponding to Ω_θ, Ω_Z, and $\boldsymbol{\rho}$, i.e. $K = 3$. Also let m_k denote the number of comparison within each parameter set and $m = \sum_k m_k$ be the overall number of comparisons. Let $I\{\}$ be an indicator function taking a value 1 when true and otherwise 0. For each parameter set we define the follow-

ing local indicators: local truth indicators $r_i^k = I\{\text{Genuine effect or interaction}\}$, local posterior evidence $v_i^k = P(r_i^k = 1 \mid Y)$ and finally a local decision indicator $d_i^k = I\{\text{Effect or interaction screened as genuine}\}$, $(i = 1, \ldots, m_k)$. The FDR and FNR specific to each of the parameter sub vectors, are defined as

$$fdr_k = \frac{\sum d_i^k (1 - r_i^k)}{\sum^{m_k} d_i^k + \mathfrak{E}}, \quad fnr_k = \frac{\sum^{m_k} (1 - d_i^k) r_i^k}{m_k - \sum d_i^k + \mathfrak{E}};$$

with $\mathfrak{E} \approx 0$ to avoid zero denominator. The local expected FDR and FNR are defined as

$$\overline{fdr}_k = \int fdr_k(d^k, r^k) dp(r^k|y) = \frac{\sum^{m_k} d_i^k (1 - v_i^k)}{\sum^{m_k} d_i^k + \mathfrak{E}},$$

$$\overline{fnr}_k = \int fnr_k(d^k, r^k) dp(r^k|y) = \frac{\sum^{m_k} (1 - d_i^k) v_i^k}{m_k - \sum^{m_k} d_i^k + \mathfrak{E}}.$$

Let $fdr(s)$ denote the fdr under the rule $\delta = I(v > s)$. Integrating results of Cai and Sun (2009) into the formulation above, we find that optimal decisions must follow the rule:

$$d_i^k = I(v_i^k > t_{2R}), \text{ where } t_{2R} = min\{s : \overline{FRD}(s, y) < \alpha\}. \tag{2.16}$$

Several options are available in the evaluation of $v_i = P(\text{effect } i \text{ is notable} \mid Y)$. When information is available about a meaningful effect size, hard thresholds around 0 may be appropriate (Berger, 1985). Example of hard thresholding in conjunction with Bayesian FDR procedure can be found in Telesca et al. (2009). Alternatively, one can use $ev = 1 - \overline{ev}$, where $\overline{ev} = P(0 \in S|Y)$ for S defined to be highest posterior probability (HPD) credible interval that is adjacent to 0 (de Bragança Pereira and Stern, 1999; Thulin, 2012). The procedure closely resembles that of one-sided test yet it is fully Bayesian as it is described in Pereira et al. (2008). We use this last procedure in our the case study (§2.5).

2.5 Case Study

2.5.1 Pharmacogenetics of irinotecan

We apply our model to explore relationships between pharmacokinetic pathways and polymorphisms in genes associated with metabolism and transport of irinotecan. Irinotecan is a chemotherapeutic agent that has regulatory approval in several countries for the treatment of colorectal cancer and is also an active agent for other various solid tumors.

Irinotecan requires activation to a potent topoisomerase I inhibitor SN-38 (7-ethyl-10-hydroxycamptothecin) through hydrolysis by the high affinity carboxylesterase-2 (CES-2) in order for it to take effect. Hence SN-38 formation within a tumor is an important bio-marker for anti-tumor activity. Details of the pharmacokinetics of irinotecan can be found in Rosner et al. (2008).

Although irinotecan is the standard treatment for colorectal cancer, it is known to have severe adverse effect such as severe diarrhea and neutropenia (20% to 35%) and fatal events (up to 5.3%) (Innocenti et al., 2004a). Numerous studies link genetic variants to the unwanted side effects, with UGT1A1 being a prime suspect since it impedes the glucuronidation of SN-38 (Innocenti et al., 2004a; O'Dwyer and Catalano, October 1, 2006). However, a more holistic analysis of the the complex interactions has not been fully explored.

2.5.2 The data

We analyze data from a study that enrolled 86 patients with advanced solid tumors treated at the University of Chicago Innocenti et al. (2004b); Iyer et al. (2002). The patients received single-agent irinotecan at doses of 300 mg/m^2 (20 patients) or 350 mg/m^2 (66 patients) infused over 90 minutes every three

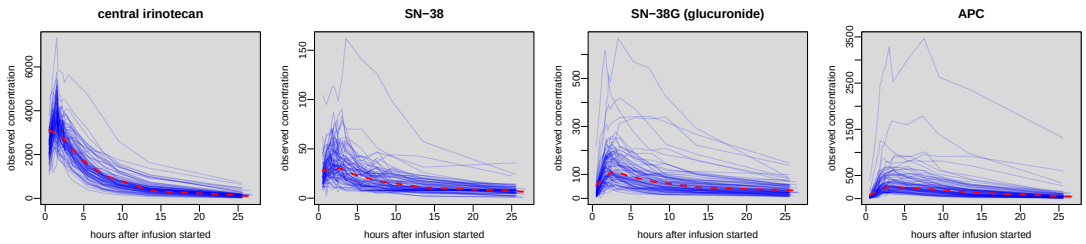


Figure 2.3: Observed concentration values for irinotecan, SN-38, SN-38G (glucuronide), and APC. Each solid line represents observed concentration for a patient plotted over time. A overlaid dashed line is a Lowess curve to show the overall trend amongst the curves.

weeks. Detailed description of the patient population can be found in Innocenti et al. (2004b); Iyer et al. (2002). Sampling of venous blood (7 ml) into sodium heparinized evacuated tubes for pharmacokinetic analysis occurred after the first irinotecan administration (at cycle 1). Sampling times were day 1 of cycle 1 (prior to irinotecan infusion) and at 0.5, 1.0, 1.5, 1.67, 1.83, 2.0, 2.25, 2.5, 3.0, 3.5, 5.5, 7.5, 13.5, 25.5 hours after the start of the infusion. For our analysis we included only patients without missing observations, leaving $n = 83$ patients out of 86 total patients.

Observed concentration-time curve for irinotecan, SN-38, SN-38G (glucuronide), and APC are plotted in figure 2.3. Each solid line corresponds to a patient and a dashed line showing the overall trend. In all 4 panels we see an overall decaying trend within the first 10 hours. However, there are signs of individuals with relatively slower elimination time, which is problematic for SN-38 since it may result in diarrhea and/or neutropenia.

2.5.3 Compartment model

We follow the work of Rosner et al. (2008) and fit a seven-compartment pharmacokinetic model with enterohepatic recirculation to concentrations of irinotecan and its metabolites SN-38, SN-38G (glucuronide), and APC. The model is expressed as a 15 parameters differential equation model (5.11) in the appendix. See Appendix 5.2.1 for the detail specification of the compartment model.

The enterohepatic recirculation time (EHRT) was treated separately and given an informative prior based on previous studies. (Rosner et al., 2008). Other hyperparameters were specified to define vague priors.

2.5.4 Data analysis

Inference for the proposed model is implemented in the R package "bppkgx". Using the R-package "bppkgx", we fit the PopPKGx model to the irinotecan data. We ran the simulation for 2 million iterations saving every 200 samples. Figure 2.4 displays a posterior concentration-time curve in irinotecan, SN-38, SN-38G (glucuronide), and APC for a sample patient with observed concentration value superimposed on top as a solid thick line (blue). The dotted line (red) corresponds to the posterior median and the dashed line (orange) is the 95% credible interval for the particular patient. Figure 2.5 is an estimate of the chain graph. The estimated is determined using the rule (2.16), with FDR control at $\alpha = 0.001$. The figure only shows PK parameters and SNPs that are included in at least one edge of the estimated graph.

Much of the result is consistent with prior research (Rosner et al., 2008). As expected from previous studies, UGT1A1 3156 mutation is linked to K_{35} along with the mutation in HNF1α not often discussed in the literatures.

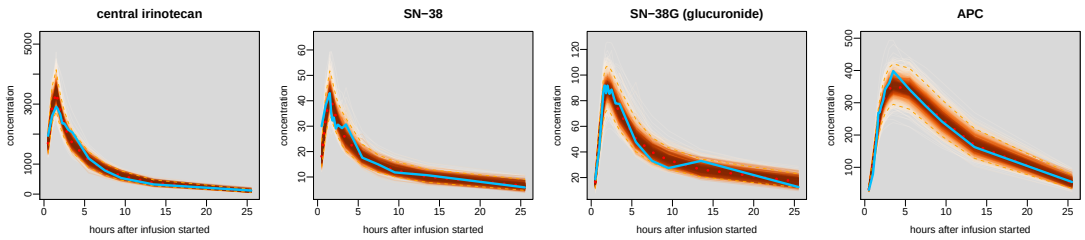


Figure 2.4: Posterior samples of concentration-time curve for irinotecan, SN-38, SN-38G (glucuronide), and APC for a particular patient plotted over time. A solid line represents observed concentration. A dotted line is a mean trajectory, the dashed lines are ±95% credible interval, and the remaining trajectories are colored in descending intensity based on the distance from the mean trajectory.

Figure 2.6 shows the posterior log concentration-time curve (mean ±95% credible interval) for patients with increasing level of UGT1A1 3156 mutation, that is $\tilde{\boldsymbol{Z}} = 0$, 1 or 2, shown as dotted, dashed, and solid lines respectively. The following interpretations are speculative interpretations of the reported inference, short of formally validated hypotheses. Higher level of mutation in UGT1A1 3156 seem to be associated with slower decay in the level of SN-38. This is further explored in figure 2.7, K_{35}. This parameter governs the rate of the glucuronidation of SN-38, is incrementally lower for patients with higher level of mutation in UGT1A1 3156. At the same time K_{30}, a parameter that govern the rate of elimination of SN-38, is not affected by the mutation, which explains the slower decay of SN-38 concentration. Also note that patients with high mutation in UGT1A1 3156 are estimated to have low level of K_{3B}, which is important in delivering the SN-38 to the tumor. In summary, patients with high level of UGT1A1 3156 mutation may not be getting as much expected chemotherapeutic effect of irinotecan, while perhaps being more likely to experience high-dose side effects.

A more comprehensive picture is obtained if we consider interactions amongst the genes. Figure 2.8 displays the same posterior log concentration trajectory

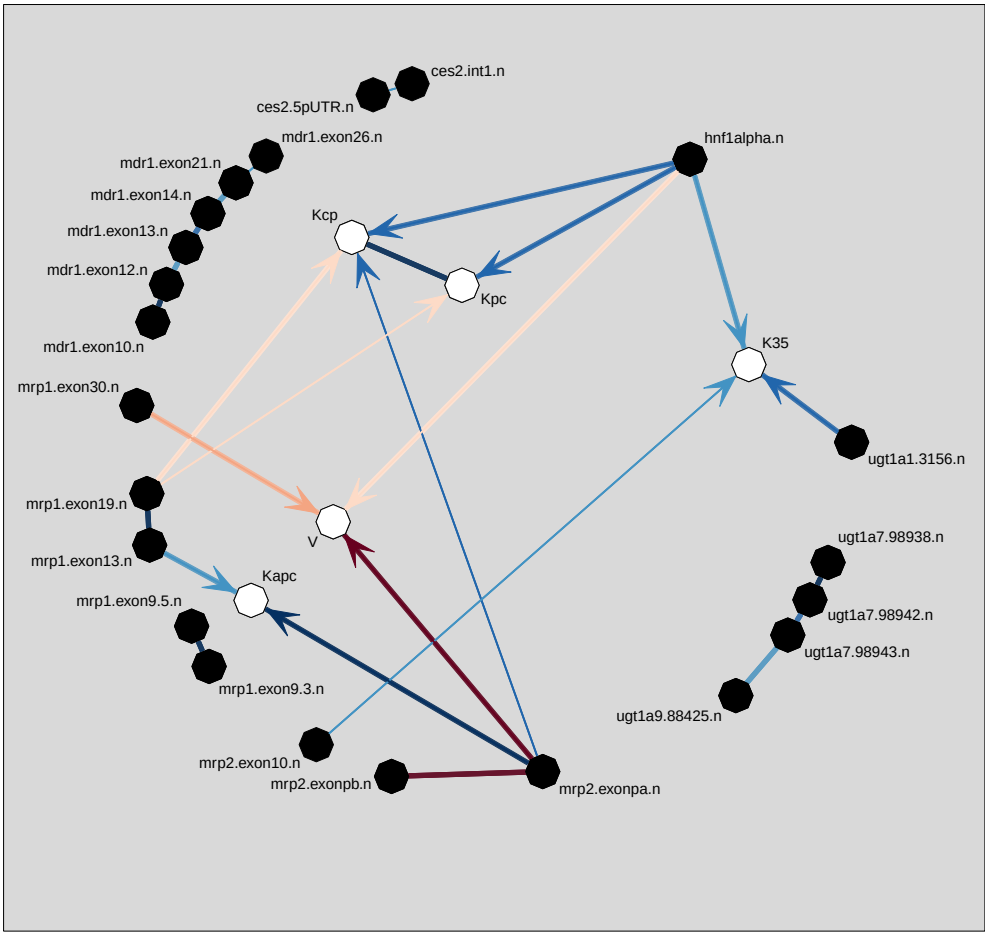


Figure 2.5: Representative chain graph for irinotecan pharmacogenetics chosen as described in section 2.5.4. The thicker edge is chosen at the expected FDR of 0.001 and thinner line is chosen at 0.0025. Only vertices with any association are shown. By simultaneously clustering the genes, it is possible to see latent associations that otherwise will be harder to see.

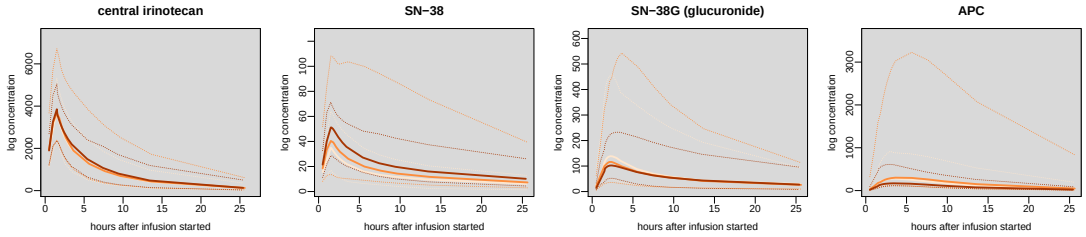


Figure 2.6: Posterior log concentration-time curve (median ±95% credible interval) for irinotecan, SN-38, SN-38G (glucuronide), and APC with level of polymorphism in UGT1A1 3156 represented in incremental order as dotted, dashed, and solid lines respectively. SN-38 elimination is slower for patients with higher level of mutation on UGT1A1 3156.

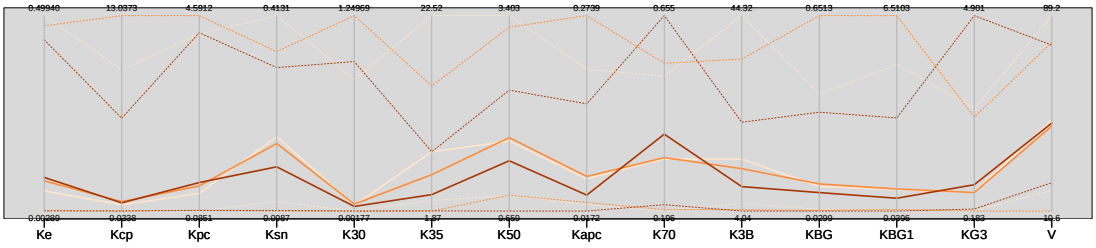


Figure 2.7: Posterior median ±95% credible interval for log PK parameters plotted as a parallel coordinate plot. Each line corresponds to a line in figure 2.6 representing the level of polymorphism in UGT1A1 3156. Increased level of UGT1A1 3156 mutation seems to reduce the level of K35, which governs the glucuronidation of SN-38 to SN-38G.

stratified over the mutation in HNF1α. Each row corresponds to each line in figure 2.7 depicting incremental level of mutation in UGT1A1 3156. Within each figure each line correspond to level of mutation in HNF1α with color intensity displaying the increase in mutation level. What is notable from this figure is that higher level of HNF1α seem to lower the level of SN-38 for the high UGT1A1 3156 group (row 3). Figure 2.9 shows the effect of UGT1A1 3156 on K_{35} is consistent as before but mutation in HNF1α for high UGT1A1 3156 mutation group seems to increase the level of K_{3B}, leading to lower concentration of SN-38 in the plasma and perhaps getting more of the expected chemotherapeutic effect.

These results are speculative and need confirmation through further research. The example shown above displays the possible importance of considering interactions among genes when considering effects of multiple genes on the pharmacokinetics of a substance. When the pharmacokinetics dynamics become highly interwind, small changes in a parameter could make a substantial difference in the expected outcome. This can be achieved, for example, using the approach proposed proposed in this paper.

2.6 Discussion

We proposed a Bayesian pharmacogenetics model to jointly model the interaction between population pharmacokinetic and SNP by use of chain graphs to formalize the dependence structure. One distinct feature of the model is the fact that we modeled the random association in the SNP as a graphical Gaussian model instead of a fixed covariate. We believe this is important when coupled with variable selection procedure for SNPs because sparsity induced by the prior distribution may understate the clustering amongst the SNPs that should be considered simultaneously. It also allows for natural Bayesian treatment of measurement-error or

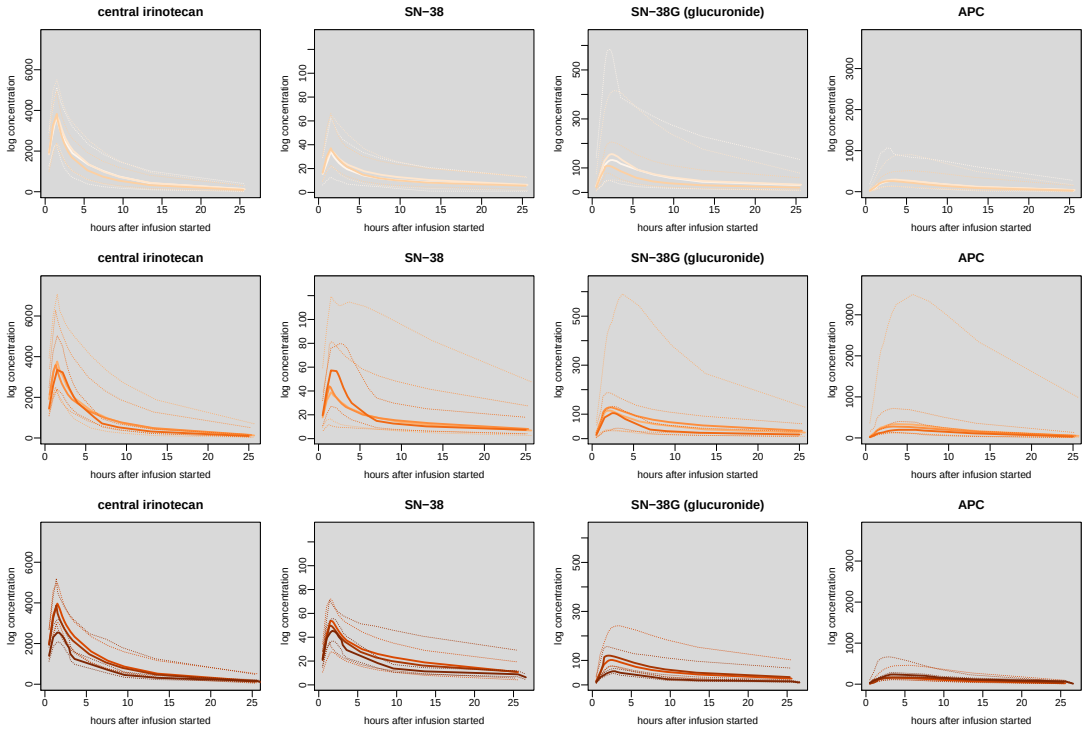


Figure 2.8: Posterior log concentration-time curve (median ±95% credible interval) for irinotecan, SN-38, SN-38G (glucuronide), and APC. Each row of the plot represents level of polymorphism in UGT1A1 3156 in incremental order from top to bottom, and each line with in a plot represents level of polymorphism in HNF1α represented in incremental order as dotted, dashed, and solid lines respectively. Effect of UGT1A1 3156 on SN-38 elimination is mitigated for patients higher level of mutation on HNF1α showing a sign of interaction effect between the two genes.

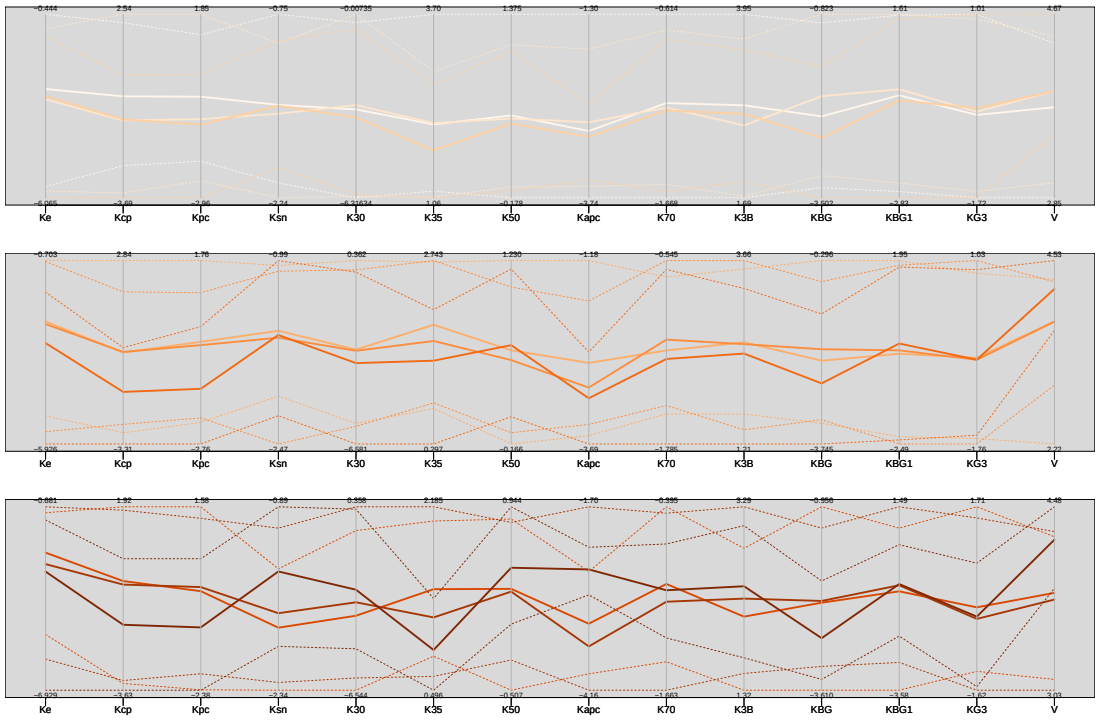


Figure 2.9: Posterior median ±95% credible interval for log PK parameters plotted as a parallel coordinate plot. Each row of the plot corresponds to the level of polymorphism in HNF1α, and the color scheme for the lines are same as the figure 2.8. The level of UGT1A1 3156 mutation seems to reduce the level of K_{35} across mutations in HNF1α, yet for high level of mutation on UGT1A1 3156 mutation on HNF1α seems to have a reversal effect on K_{3B} compared with figure 2.7.

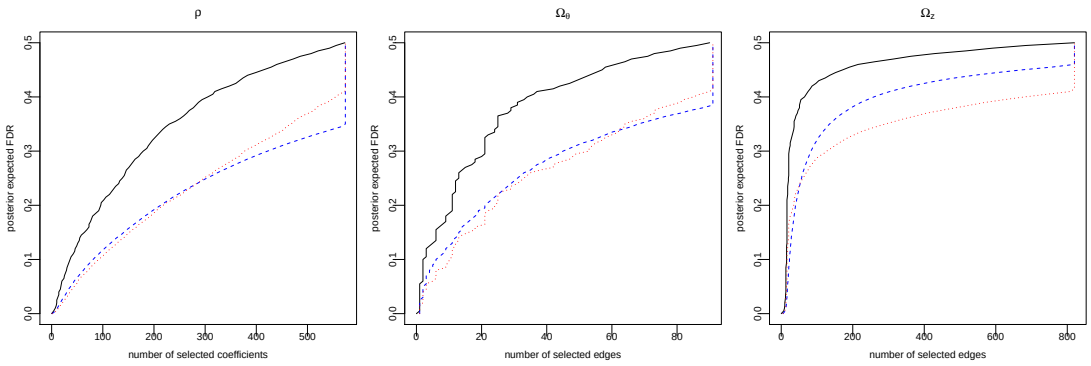


Figure 2.10: Expected posterior FDR by number of differential selected effects. The solid line corresponds to choice made using Lindley's procedure using one sided $1 - \alpha\%$ CI, the dashed line corresponds to choice made using Bayesian FDR control but without pooling across the parameters, and the dotted line corresponds to choice made using simultaneous Bayesian FDR across the parameters. FDR procedure allows for a more aggressive decision and by pooling across the parameters, more claims are made on the ρ and Ω_θ by sacrificing Ω_z for lower level of posterior expected FDR.

missing data prevalent in pharmacogenetics studies, without adding extra layers of ad-hoc adjustments.

The choice of Laplacian shrinkage priors, does not directly allow for the definition of a posterior distribution over the model space $p(\mathcal{G} \mid \boldsymbol{Y}, \tilde{\boldsymbol{Z}})$. A common alternative is based on placing mixture priors over model selection parameters, such as the G-Wishart prior (Dawid and Lauritzen, 1993; Roverato, 2002). Even though this last approach is theoretically appealing, we acknowledge a growing number of criticisms to point mass mixture priors, both from a theoretical (Rice, 2010; Thulin, 2012) and a practical perspective (Jones et al., 2005). Our choice of shrinkage priors, is indeed motivated by pragmatism. When the analysis objective is exploratory, as long as proper levels of shrinkage are achieved, one may in fact argue if the gain associated with the use of explicit model selection priors outweighs the immense computational cost.

There are possible extensions to this model that we have not explored such as allowing the association structure to vary amongst subset of the patients. Parametric approaches such as covariance regression (Hoff and Niu, 2012) or nonparametric approaches such as Rodriguez et al. (2011) may be used to determine the difference in the structure of associations for some known subset of patients or to determine potential heterogeneity in the patients that differ in the pharmacogenetics mechanism.

Extending beyond Gaussian assumptions to allow for higher-order interaction is a possibly important next step. However, to allow for elaborate extensions, improvement in the way we address differential equations within a MCMC must be dealt with. We believe active research in MCMC computation will allow for such extensions to bring us closer to the science of the problem in the near future.

The analysis in this manuscript was carried out using an R-package "bppkgx", which is a general purpose R package we created to implement the model proposed in this paper. The package is available on CRAN and is free for public use. It is designed so that a user can specify the functional form of the pharmacokinetics model in the same way a user would specify them if they were using the standard differential equation solver package deSolve in R (Soetaert K, 2010).

CHAPTER 3

DETECTING DIFFERENTIAL PATTERNS OF INTERACTION IN MOLECULAR PATHWAYS

3.1 Introduction

We propose a methodological framework to assess heterogeneous patterns of association amongst components of a random vector. Figure 3.1 (left) is a toy example illustrating what happens when one tries to determine the association between two variables, without accounting for heterogeneity subsumed in the data. The issue becomes obvious as the information on known sample subsets is revealed as in figure 3.1 (center); two conflicting effects as shown in figure 3.1 (right) cancel out when integrating over the subsets. Despite the simplicity of the scenario, it highlights the danger of failing to account for subset labels, which is often available in most comparative studies. One such example is the case of estimating molecular interactions from large scale genomic or proteomic studies, where there is substantive interest in understanding whether disease progression in patient subgroups exhibits differential regulatory patterns. This chapter is indeed partially motivated by a study on Acute Myeloid Leukemia (AML) patients (section 3.6), where interest centers on comparing refractory vs. relapsed patients. Figure 3.2 shows the targeted protein expression level for AML patients,

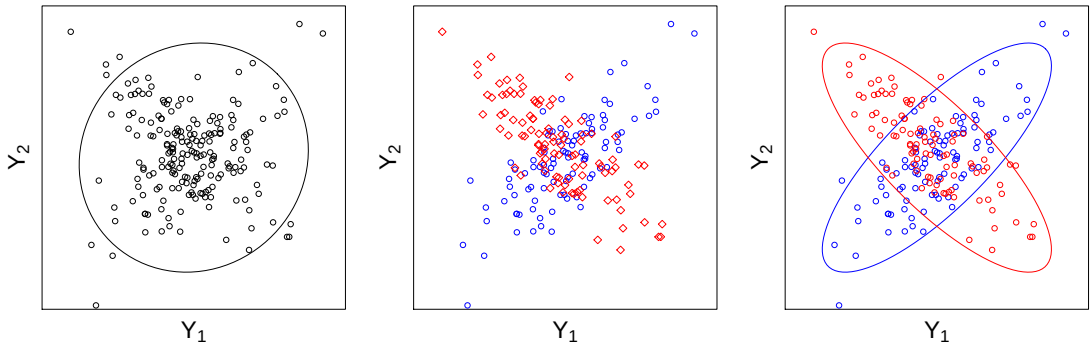


Figure 3.1: Illustrative example of the differential effects. Without taking into account of the subgrouping figure on the left shows no strong association between Y_1 and Y_2. But if we knew the data comes from two different source as shown in central figure, we could use that information and see that there is actually 2 strong effects in the opposite direction as shown in the right figure.

measured using a high-throughput proteomic technology called reverse phase protein array (RPPA) (Tibes et al., 2006), along with the sample partial correlation matrix of the expression levels for both refectory and relapsed patients in the upper and lower triangle respectively. The two sample partial correlation mostly agree with each other, yet there are clear discrepancies that may signify differing interactions mechanism. The proposed methodology is designed to account for subset-specific heterogeneity, while uncovering the hidden differential association structure in a multivariate setting.

Inference and estimation algorithms for structured inverse covariance matrices in the multivariate Gaussian framework have been described by Dempster (1972). More recently, focus has shifted to using graphical models to represent the conditional dependence structure of a multivariate vector. Several authors have contributed to the development of graphical model classes as instruments of statistical inference: decomposable graphs (Giudici and Green, 1999; Jones et al., 2005; Wang and West, 2009), non-decomposable graphs (Roverato, 2002; Atay-

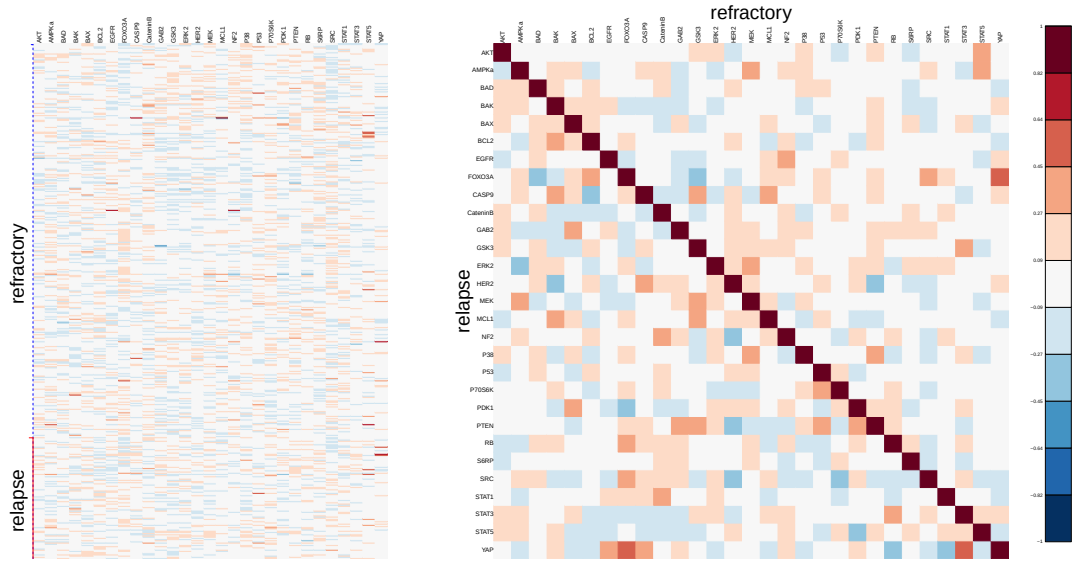


Figure 3.2: The observed expression levels of targeted proteins for AML patients quantified using RPPA (left) and a image plot of the sample partial correlation coefficients for refractory patients in the upper triangle and relapsed patients in the lower triangle (right).

Kayis and Massam, 2005), Directed Acyclic Graphs (Madigan et al., 1995; Dobra et al., 2004; Fronk, 2002; Fronk and Giudici, 2004), and the computation associated with such models (Scott and Carvalho, 2007; Barker et al., 2010). To our knowledge, however, limited attention has been given to cases where the Markov structure describing the multivariate distribution of interest depends on known subgroup indicators. In the computer science literature early work by Thiesson et al. (1997) had a similar concept under the name of mixtures of DAGs models, though the implementation was limited to very small graphs and inference was based on heuristic arguments.

In statistical literature, Guo (2002) proposed a method that makes use of penalized likelihood to estimate jointly several graphical models. The proposed procedure was shown to be scalable to large graphs, with estimators that enjoy asymptotic consistency. A recent applied paper by Valcárcel et al. (2011) considered a closely related problem, regarding inference on differential networks. The Authors discuss inference about differences in the molecular association between normal and the prediabetic patients, using permutation arguments.

Both methods are of great practical relevance, since they scale to large networks and may prove to be an important tool in data exploration. At the same time, both procedures are based on several ad-hoc corrections and heuristic choices, which raise methodological and theoretical questions regarding multiplicity correction and final inference validity.

In the Bayesian nonparametric literatures models that aims to use multiple graphical structure for clustering purpose has been proposed by Ickstadt et al. (2010) for differential edges and Rodriguez et al. (2011) for differential association structures. These nonparametric models are exploratory in nature that does not provide means to do cluster specific statistical inference.

We contrast these approaches proposing a probability model that provides a coherent framework for estimation as well as inference for differential patterns of association, described as multiple graphical models. We show how, from a Bayesian perspective, principled inference can be carried out using sound decision theoretic principles, without the need to resort to ad-hoc arguments.

To facilitate exposition and notation we consider the case of two known subsets in the sample. We will call one group the baseline group and the other one the differential group. In a symmetric fashion, we will define a baseline network/graph and a differential network/graph. Extensions to k subsets are straightforward. We propose a full Bayesian model which follows the original development of Fronk and Giudici (2004), with the additional consideration of structural constraints defined by the differential network. We will jointly estimate the baseline graph and the differential graph as well as the strength of association by the use of stochastic simulation technique called Reversible Jump Markov Chain Monte Carlo (Green, 1995). Then turn to decision theoretic framework proposed by Müller et al. (2006) to decide on the meaningful association.

The modeling approach proposed in this manuscript highlights several novel contributions. We describe a coherent probability model of differential association. We provide a computational framework for the simultaneous estimation of several graphical structures and associated parametric forms of structured multivariate Gaussian vectors. Finally, we propose a decision theoretic framework aimed at the definition of posterior estimates, which account for considerations of multiplicity.

This chapter is structured as follows. In section 3.3 we propose a Gaussian Differential DAG model followed by computational detail in section 3.4. We illustrate the method further with a simulated example and an application to the

Reverse Phase Protein Array (Tibes et al., 2006) data on Acute Myeloid Leukemia patients. We conclude the manuscript with a critical discussion in section 3.7.

3.2 Representing Dependence Through Graphical Models

For the rest of the chapter, we focus on *Directed Acyclic Graphs* (DAGs). A DAG is directed graph with no directed cycles. The acyclicity restriction could represent a drawback in some applications. However, when dealing with a network where association is usually sparse, this restriction is often not critical. Furthermore, we find that structural computational advantages of DAG-based models far outweigh small gains in flexibility, obtained dropping the acyclicity restriction.

Finally, we should be clear that our use of DAGs is not intended to code any causal relationship (Pearl, 2000), but is strictly based on theoretical and computational convenience.

3.3 A Model for Differential Interactions

We consider data in the form of an $n \times p$ matrix $Y = [y_{ij}]$, such that $E[y_{ij}] = 0$, for all $i = 1, ..., n,\ j = 1, ..., p$. Without loss of generality, we consider the case of two known subgroups and assume that the rows of Y are labelled by a subgroup indicator $s_i = I\{\text{differential group}\}$. The sampling model for Y depends on a graph $\mathcal{G}_s$, describing the dependence structure between columns of Y. The strength of this dependence is indexed by two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The key feature of the proposed model is that the the graph $\mathcal{G}_s$ is indexed by subgroups indicators $\mathbf{s} = (s_1, ..., s_n)'$. Let $\mathcal{G} = \{\mathcal{G}_s,\ s = 0, 1\}$ denote the set of graphs. In

summary the joint probability model is defined as:

$$p(Y, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathcal{G} \mid \mathbf{x}) = \underbrace{p(Y \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathcal{G}; \mathbf{s})}_{3.3.1} \underbrace{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathcal{G}; \mathbf{s})}_{3.3.2} \underbrace{p(\mathcal{G} \mid \mathbf{s})}_{3.3.3}.$$

The model includes two separate graphs, $\mathcal{G}_0 = \{\mathcal{V}, \mathcal{E}_0\}$ for the baseline samples ($s_i = 0$) and $\mathcal{G}_1 = \{\mathcal{V}, \mathcal{E}_1\}$ for the differential sample ($s_i = 1$). Our inference will focus on identifying a set of differential interactions partially indexed by the set $\{(\mathcal{E}^{(0)})^c \cap \mathcal{E}^{(1)}\} \cup \{\mathcal{E}^{(0)} \cap (\mathcal{E}^{(1)})^c\}$. For clarity of notation, the foregoing formulation in (3.3) integrates over nuisance parameters completing the coherent definition of sampling and prior models. In the following sections we discuss each component of the model in more detail. Under-braced section numbers in (3.3) indicate where each submodel is discussed. Nuisance parameters are described in section 3.3.4.

3.3.1 Sampling model:

We have data in the form of a $n \times p$ matrix Y. We assume that Y can be subdivided into two groups as $Y^{(0)}$ and $Y^{(1)}$ each of size n_0 and n_1, where $n_0 + n_1 = n$. We will refer to the former as the baseline group and latter as the differential group. Throughout this chapter we will assume the baseline is stacked on top of the differential group for notational convenience, i.e., $Y = (Y_0, Y_1)$.

The Gaussian Differential DAG model for Y is defined as the product of conditional Gaussian DAG models for $Y^{(0)}$ and $Y^{(1)}$, given the graphical structures $\mathcal{G}_0$ and $\mathcal{G}_1$. Let $pa_k(j)$ denote the parent nodes of vertex j, induced by graph $\mathcal{G}_k$. Let $Y_j = (y_{1j}, \ldots, y_{nj})^T$, $j = 1, \ldots, p$, the joint likelihood is defined as

$$p(Y \mid \cdot) \;=\; \prod_{k=0}^{1} \prod_{j}^{p} p(Y_j^{(k)} \mid Y_{pa_k(j)}^{(k)}, \mathcal{G}_k, \cdot),$$

where $p(Y_j^{(k)} \mid Y_{pa_k(j)}^{(k)}, \mathcal{G}_k, \cdot) = \prod_{i=1}^{n_k} p(y_{ij}^{(k)} \mid Y_{pa_k(j)}^{(k)}, \mathcal{G}_k, \cdot)$. In the multivariate Gaussian framework, we can express each of $p(y_{ij}^{(k)} \mid Y_{pa_k(j)}^{(k)}, \mathcal{G}_k, \cdot)$ as a conditional regression of the form

$$y_{ij}^{(k)} \mid Y_{pa_k(j)}^{(k)}, \alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2, \mathcal{G}_k \quad \sim \quad N\left(\alpha_j + \sum_{l \in pa_k(j)} y_{il}^{(k)} \left(\beta_{lj} + \gamma_{lj}\, I\{s_i = 1\}\right), \sigma_j^2\right), \quad (3.1)$$

for $i = 1, \ldots, n$, $j = 1, \ldots, p$, and $k = 0, 1$. Here α_j is a nuisance parameter for the mean value and σ_j^2 is a variance parameter. In (3.1) we let $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{(j-1)j}, 0, \beta_{(j+1)j}, \ldots, \beta_{pj})^T$ and define $\boldsymbol{\gamma}_j$ in a similar fashion (we include the 0 for the $j-$th element to simplify later expressions). We also use $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to denote $p \times p$ matrices $[\beta_{lj}]$ and $[\gamma_{lj}]$, and define $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$.

In vector form, we define Y_{-j} as the $n \times p$ matrix comprising all data, repalcing the j-th column with all 0's. The conditional distribution of the random vector Y_j, given its parents can be written as

$$Y_j \mid Y_{-j}, \boldsymbol{\gamma}_j, \alpha_j, \boldsymbol{\beta}_j, \sigma_j^2, \mathcal{G}_0, \mathcal{G}_1 \sim N\left(\tilde{X}_j B_j,\, \sigma_j^2 \mathbf{I}_n\right) \text{ for } j = 1, \ldots, p \qquad (3.2)$$

where

$$B_j = (\alpha_j, \boldsymbol{\beta}_j^T, \boldsymbol{\gamma}_j^T)^T, \text{ and } \tilde{X}_j = \left(\begin{array}{c|c|c} \mathbf{1}_{n_0} & Y_{-j}^{(0)} & \mathbf{0}_{n_0 \times (p-1)} \\ \hline \mathbf{1}_{n_1} & Y_{-j}^{(1)} & Y_{-j}^{(1)} \end{array}\right).$$

In the previous formula $\mathbf{1}_{n_k}$ is a column vector of 1s with length n_k and $\mathbf{0}_{n_k \times p}$ is a $n_k \times p$ matrix of 0s. Furthermore, restrictions to structural zeros in $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ assure that $\boldsymbol{y}_j$ is regressed only on the set of parent nodes $pa(j)$, as indexed by $\mathcal{G}_0$ and $\mathcal{G}_1$.

For any random vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$ in the baseline or differential group, constructions (3.1) or (3.2) define the joint sampling distribution in closed form

as

$$\boldsymbol{y}_i^{(0)} \sim N\left((\Lambda_0^{-1})^T\boldsymbol{\alpha}, (\Lambda_0^{-1})^T\Omega\Lambda_0^{-1}\right) \text{ and } \boldsymbol{y}_i^{(1)} \sim N\left((\Lambda_1^{-1})^T\boldsymbol{\alpha}, (\Lambda_1^{-1})^T\Omega\Lambda_1^{-1}\right),$$

where $\Omega = diag(\sigma_1^2, \cdots, \sigma_p^2)$ and

$$[\Lambda_0]_{lj} = \begin{cases} 1 & (l = j) \\ -\beta_{lj} & (l \to j \in \mathcal{E}_0) \\ 0 & (o.w.) \end{cases}, \quad [\Lambda_1]_{lj} = \begin{cases} 1 & (l = j) \\ -(\beta_{lj} + \gamma_{lj}) & (l \to j \in \mathcal{E}_1) \\ 0 & (o.w.) \end{cases}.$$

In the foregoing formulation, β_{lj} indexes the strength of association between $y_{il}^{(0)}$ and $y_{ij}^{(0)}$, with the convention $\beta_{lj} = 0$ when $l \to j \notin \mathcal{E}_0$. The strength of association between $y_{il}^{(1)}$ and $y_{ij}^{(1)}$ is defined by $\beta_{lj} + \gamma_{lj}$, with $(\beta_{lj} + \gamma_{lj}) = 0$ whenever $l \to j \notin \mathcal{E}_1$. In this setting, the parameter γ_{lj} becomes the main quantity of interest as it directly informs the differences in association between subgroup random quantities. Details about how γ is used to index the differences between $\mathcal{E}_0$ and $\mathcal{E}_1$ and final inference about differential interactions are discussed in section 3.3.2.

3.3.2 Priors on interaction parameters

The strength of association between random quantities in the baseline group is parametrized through β_{lj} coefficients. Conditioned on the baseline graph $\mathcal{G}_0$, we define a conjugate Gaussian distribution for β_{lj} similar to Fronk and Giudici (2004), so that

$$\beta_{lj} \mid \sigma_j^2, \mathcal{G}_0 \sim \begin{cases} \delta_0 & \texttt{if } l \notin pa_0(j) \\ N\left(b_{lj}, \frac{1}{\omega_j}\sigma_j^2\right) & \texttt{if } l \in pa_0(j) \end{cases}. \tag{3.3}$$

Here δ_0 denotes a Dirac mass at 0. Hyperparameters b_{lj} are usually set to 0 unless we have information otherwise. Integrating over the model space $\mathcal{G}_0$, this prior is marginally equivalent to defining mixture of a conjugate Gaussian distribution and a point mass at zero, in a fashion that is similar to standard Bayesian variable selection strategies (Kuo and Mallick, 1998; Brown et al., 1999; George and McCulloch, 1993).

Differential parameters γ_{lj} distinguish the strength of association between the baseline and differential groups. Intuitively, when γ_{lj} is close to 0, partial correlations in the baseline and differential groups are about the same size. We are interested in answering two main questions. First, are there differences in patterns of conditional dependence between baseline and differential groups? This question relates to the identification of the set $\{(\mathcal{E}^{(0)})^c \cap \mathcal{E}^{(1)}\} \cup \{\mathcal{E}^{(0)} \cap (\mathcal{E}^{(1)})^c\}$.Second, when considering edges that are shared between both baseline and differential groups, are there significant differences in the way these edges are defining conditional dependence patters? Here we consider the set $(\mathcal{E}_0 \cap \mathcal{E}_1)$, but we are specifically interested in the size of γ_{lj}.

These inferential goals are coded directly into the prior distribution for γ_{lj}, which is defined conditionally on the baseline association strength β_{lj} as well as conditionally on the graphs $\mathcal{G}_0$ and $\mathcal{G}_1$. We define

$$\gamma_{lj} \mid \mathcal{G}_0, \mathcal{G}_1, \beta_{lj}, \sigma_j^2 \sim \begin{cases} N\left(\nu_{lj}, \frac{1}{\omega_j}\sigma_j^2\right) & \text{if } (l \notin pa_0(j),\ l \in pa_1(j)) \\ \pi_{lj}\delta_0 + (1-\pi_{lj})N\left(\nu_{lj}, \frac{1}{\omega_j}\sigma_j^2\right) & \text{if } (l \in pa_0(j),\ l \in pa_1(j)) \\ \delta_{-\beta_{lj}} & \text{if } (l \in pa_0(j),\ l \notin pa_1(j)) \\ \delta_0 & \text{if } (l \notin pa_0(j),\ l \notin pa_1(j)) \end{cases} \tag{3.4}$$

where δ_d is a Dirac mass at d, ν_{lj} and ω_j are known hyper parameters, and π_{lj} are unknown mixing proportions. The last two lines of (3.4) formalize the

convention $\gamma_{lj} = 0$ for an excluded edge. In this formulation, the full set of differential interactions is identified by γ_{ij} being sampled from $\delta_{-\beta_{ij}}$ or $N\left(\nu_{lj}, \frac{1}{w_j}\sigma_j^2\right)$. Equivalently, identical interactions between baseline and differential groups are indexed by a Dirac mass at 0 for γ_{lj}.

In the later discussion it will be convenient to introduce latent indicators $\boldsymbol{z} = [z_{lj}], z_{lj} \in \{0, 1, 2\}$ that allow us to replace (3.4) by a hierarchical model $p(\boldsymbol{z} \mid \ldots) \cdot p(\boldsymbol{\gamma} \mid \boldsymbol{z}, \ldots)$. Specifically

$$z_{lj} \mid \mathcal{G}_0, \mathcal{G}_1, \beta_{ij} = \begin{cases} 0 & \text{if } (l \notin \text{pa}_0(j),\ l \notin \text{pa}_1(j)) \\ \pi_{lj}\delta_0 + (1 - \pi_{lj})\delta_2 & \text{if } (l \in \text{pa}_0(j),\ l \in \text{pa}_1(j)) \\ 1 & \text{if } (l \in \text{pa}_0(j),\ l \notin \text{pa}_1(j)) \\ 2 & \text{if } (l \notin \text{pa}_0(j),\ l \in \text{pa}_1(j)) \end{cases}$$

and

$$\gamma_{lj} \mid z_{lj}, \beta_{lj}, \sigma^2 \quad \sim \quad \begin{cases} \delta_0 & \text{if } z_{lj} = 0 \\ \delta_{-\beta_{lj}} & \text{if } z_{lj} = 1 \\ N\left(\nu_{lj}, \frac{1}{\omega_j}\sigma_j^2\right) & \text{if } z_{lj} = 2 \end{cases} .$$

Given this parametrization, posterior inference over differential patterns of interaction focuses directly on $p(\gamma_{lj} \mid Y)$, informing about the size of differences in partial correlation, and $p(z_{lj} \neq 0 \mid Y)$, informing about the significance of such differences.

3.3.3 Model space priors

Our inference depends on obtaining posterior draws from the model space spanned by DAGs $\mathcal{G}_0$ and $\mathcal{G}_1$. For simplicity, we will model $\mathcal{G}_0$ and $\mathcal{G}_1$ independently, so that $p(\mathcal{G}_0, \mathcal{G}_1) = p(\mathcal{G}_0)p(\mathcal{G}_1)$.

As for the priors on each graph $\mathcal{G}_k$, $(k = 0, 1)$, we model edge inclusion probabilities as exchangeable Bernoulli trials (Giudici and Green, 1999; Fronk and Giudici, 2004). Let $|\mathcal{E}_k|$ be the number of edges in graph $\mathcal{G}_k$, then $p(\mathcal{G}_k \mid \psi_k) = \psi_k^{|\mathcal{E}_k|}(1-\psi_k)^{M-|\mathcal{E}_k|}$.

For a class of $Beta$ prior distribution on inclusion probabilities $\psi_k \sim Beta(v_1, v_2)$, this class of stochastic schemes is know to provide automatic multiplicity correction in the posterior $p(\mathcal{G}_k \mid \boldsymbol{Y})$ (Scott and Berger, 2006; Carvalho and Scott, 2009). The marginal prior distribution for $\mathcal{G}_k$ is available in closed form as

$$\begin{aligned} p(\mathcal{G}_k) &\propto B(v_1 + |\mathcal{E}_k|, v_2 + M - |\mathcal{E}_k|) \\ &= \frac{\Gamma((v_1 + |\mathcal{E}_k|)\Gamma(v_2 + M - |\mathcal{E}_k|)}{\Gamma(v_1 + v_2 + M)}, \end{aligned}$$

which simplifies to $p(\mathcal{G}_k) = \frac{1}{(M+1)}\binom{M}{|\mathcal{E}_k|}$, if $\psi_k \sim U(0,1)$.

When prior information on interaction structures is available, informative priors may be defined following the approaches of Mukherjee and Speed (2008); Telesca et al. (2012b). Finally, the model space prior is completed specifying mixture probabilities π_{lj} for the case $(l \in pa_0(j),\ l \in pa_1(j),\ \gamma_{lj} \neq \beta_{lj})$. We exploit conditional conjugacy and assume $\pi_{lj} = \pi \sim Beta(v_1, v_2)$.

3.3.4 Priors on nuisance parameters α_j and σ_j^2

For dispersion parameters σ_j^2 we model each of the σ_j^2, $j = \{0, \ldots, p\}$ as a conjugate Inverse Gamma prior with hyper parameters $\frac{\delta_j}{2}$ and $\frac{\tau_j}{2}$, so that $\sigma_j^2 \mid \mathcal{G} \sim Inv\text{-}Ga\left(\frac{\delta_j}{2}, \frac{\tau_j}{2}\right)$. In a similar fashion, we exploit conditional conjugacy and place a Gaussian prior on the intercept terms α_j, so that $\alpha_j \mid \sigma_j^2 \sim N\left(a, \frac{1}{\omega}\sigma_j^2\right)$.

3.4 Posterior Inference

To obtain draws from the posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathcal{G}_1, \mathcal{G}_0 \mid Y)$, we use reversible jumps Markov Chain Monte Carlo (RJMCMC) (Green, 1995). More precisely, we extend the approach of Fronk and Giudici (2004) to differential Gaussian DAGs. Fronk and Giudici's algorithm moves through the model space spanned by a DAG $\mathcal{G}$ by proposing the addition, deletion, or switch in direction for one individual edge at the time. Acyclicity is assessed online and, for a given graph $\mathcal{G}$, remaining variables in the model are updated component wise via Gibbs sampling.

The addition of a differential graphical structure and differential parameters is, in principle, easily treated with a small modification to the simulation scheme proposed by Fronk and Giudici (2004). The only change is in the consideration of an additional structure $\mathcal{G}_1$, together with the baseline $\mathcal{G}_0$.

We note that, in our formulation, $\mathcal{G}_1$ is fully determined by $\mathcal{G}_0$ and latent components z_{lj}. It follows that, systematic or random scans through the following transition sequence define an ergodic Markov chain, we can use to sample from

posterior quantities of interest. We consider the following transition sequence

$$\begin{aligned}
\mathcal{G}_0 \quad &| \quad \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}, \sigma^2 \quad (3.4.1) \\
\boldsymbol{z} \quad &| \quad \mathcal{G}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 \quad (3.4.2) \\
\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \quad &| \quad \mathcal{G}_0, \boldsymbol{z}, \sigma^2 \quad (3.4.3) \\
\sigma^2 \quad &| \quad \mathcal{G}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z} \quad (3.4.4)
\end{aligned}$$

Details about each transition are explained in the corresponding sections.

3.4.1 Updating the baseline DAG $\mathcal{G}_0$

To update $\mathcal{G}_0$, we select an edge $(l \to j)$ at random, i.e., using a uniform distribution over all possible edges $l \to j$. If $(l \to j) \notin \mathcal{E}_0$ we propose its addition to $\mathcal{E}_0$ (birth); if $(l \to j) \in \mathcal{E}_0$ we propose removal (death); if $(l \leftarrow j) \in \mathcal{E}_0$ but $(l \to j) \notin \mathcal{E}_0$ then propose to remove $(l \leftarrow j)$ and add $(l \to j)$ (switch). This is the algorithm proposed by Fronk and Giudici (2004), with the added caveat that changes in $\mathcal{G}_0$ may also affects $\mathcal{G}_1$.

3.4.1.1 Birth move

Adding the edge $(l \to j)$ in $\mathcal{E}_0$ results in augmenting the parameter space with one extra coefficient β'_{lj}, which will also define changes in $\mathcal{E}_1$. To maintain local moves and protect $\mathcal{E}_1$ from being affected, we also propose a state transition for γ_{lj} and z_{lj}. A birth move then consists of the following proposal $(\mathcal{G}_0,\ \beta_{lj} = 0,\ z_{lj}, \gamma_{lj}) \Rightarrow (\mathcal{G}'_0,\ \beta'_{lj},\ z'_{lj},\ \gamma'_{lj})$, where $\beta'_{lj} \sim q_b(\beta_{lj})$ and $(z'_{lj}, \gamma'_{lj}) \sim q_g(z'_{lj}, \gamma'_{lj};\ z_{lj})$. Let $\boldsymbol{\theta} = (\mathcal{G}_0,\ \beta_{lj} = 0,\ z_{lj}, \gamma_{lj})$ and let $\boldsymbol{\theta'} = (\mathcal{G}'_0,\ \beta'_{lj} \neq 0,\ z'_{lj}, \gamma'_{lj})$ denote the current state vector and the joint proposal. In particular:

- If $z_{lj} = 0$, propose $z'_{lj} = 1$ and $\gamma'_{lj} \sim \delta_{-\beta'_{lj}}$. The reversible jump ratio is

$$R_{B_0}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\beta'_{lj} \mid \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right)}{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\mathcal{G}_0\right) q\left(\beta'_{lj}\right)},$$

- If $z_{lj} = 2$, propose one of the following moves with equal probability $\frac{1}{2}$:
  - propose $z'_{lj} = 0$ and $\gamma'_{lj} \sim \delta_0$, with reversible jump ratio

$$R_{B_1}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\beta'_{lj} \mid \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right) q_g\left(\gamma_{lj}\right) \pi_j}{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\boldsymbol{\gamma}_{lj} \mid z_{lj}, \beta_{lj}, \sigma_j^2, \mathcal{G}_0\right) p\left(\mathcal{G}_0\right) q_b\left(\beta'_{lj}\right) \left(\frac{1}{2}\right)},$$

  - or propose $z'_{lj} = 2$ and $\gamma'_{lj} = \gamma_{lj}$, with reversible jump ratio

$$R_{B_2}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\beta'_{lj} \mid \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right) \left(1 - \pi_j\right)}{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\mathcal{G}_0\right) q_b\left(\beta'_{lj}\right) \left(\frac{1}{2}\right)}.$$

In the calculations above, $\boldsymbol{\beta}'_j$ refers to $\boldsymbol{\beta}_j$ with the $l-$th element set to β'_{lj} and $\boldsymbol{\gamma}'_j$ refers to $\boldsymbol{\gamma}_j$ with the $l-$th element set to γ'_{lj}. The acceptance probability for each move is calculated as $A_{B_i} = \min\{1, R_{B_i}\}$. Note that $p(\mathcal{G}'_0) = 0$ if the proposed graph $\mathcal{G}'_0$ were to include directed cycles, i.e., $\mathcal{G}'_0$ is not a DAG. A test of acyclicity was proposed by Fronk and Giudici (2004).

Test of acyclicity (Fronk and Giudici, 2004): Given an ancestral matrix $\mathcal{A}$ corresponding to a DAG $\mathcal{G}$, a DAG $\mathcal{G}$ is not acyclic if

$$diag(\mathcal{A}^i) = 0, \forall i = \{1, \cdots, min(G, |\mathcal{G}|)\} \tag{3.5}$$

Where $\mathcal{A}^i$ is matrix exponent, $diag()$ is the diagonal elements of the matrix, $|\mathcal{G}|$ is the number of edges in a graph $\mathcal{G}$ and G is the number of vertices.

This proposal transition scheme is designed to define symmetry with respect to the reverse (death) move. Details are discussed in supplemental appendix B. In our implementation we consider $q_b(\beta'_{lj}) =_d N(0, \zeta^2)$ and $q_g(\gamma'_{lj}) =_d N(0, \zeta^2)$. When adding $l \to j$ to $\mathcal{E}_0$ defines a cycle in $\mathcal{G}'_0$, we evaluate $p(\mathcal{G}'_0) = 0$ and thus $R_{B_i} = 0$ and the proposal is discarded.

3.4.1.2 Death move

Deletion of an edge $l \to j$ is equivalent to forcing $\beta'_{lj} = 0$. In order to maintain detailed balance we design these transitions as the inverse of those proposed in the birth step. In more detail:

- If $z_{lj} = 1$, propose $z'_{lj} = 0$ and $\gamma'_{lj} \sim \delta_0$, with reversible jump ratio $R_{D_0}(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1/R_{B_0}(\boldsymbol{\theta}', \boldsymbol{\theta})$.
- If $z_{lj} = 0$, propose $z'_{lj} = 2$ and $\gamma'_{lj} \sim q_g(\gamma'_{lj})$, with $R_{D_1}(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1/R_{B_1}(\boldsymbol{\theta}', \boldsymbol{\theta})$.
- If $z_{lj} = 2$, propose $z'_{lj} = 2$ and $\gamma'_{lj} \sim q_g(\gamma'_{lj})$, with $R_{D_2}(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1/R_{B_2}(\boldsymbol{\theta}', \boldsymbol{\theta})$.

The acceptance probability for each move is then $A_{D_i} = \min\{1, R_{D_i}\}$. Detailed calculations are reported in supplemental appendix B.

3.4.1.3 Switch move

Proposing the switch of an edge implies a death move on $j \to l$, as well as a birth move on $l \to j$. Hence the acceptance is determined by the combination of reversible jump ratios noted earlier for birth and death, R_{B_0}, R_{B_1}, or R_{B_2} and R_{D_0}, R_{D_1}, or R_{D_2} according to the current values of z_{jl} and z_{lj} respectively. The

$\mathcal{G}_0$	current z_{lj}	proposed z_{lj}	move type	probability	move #
$(i,j) \notin \mathcal{E}_0$	$z_{lj} = 0$	$z'_{lj} = 2$	RJ birth	1	1
	$z_{lj} = 2$	$z'_{lj} = 0$	RJ death	1	2
$(i,j) \in \mathcal{E}_0$	$z_{lj} = 0$	$z'_{lj} = 1$	MH	1/2	3
		$z'_{lj} = 2$	RJ birth	1/2	4
	$z_{lj} = 1$	$z'_{lj} = 0$	MH	1/2	5
		$z'_{lj} = 2$	RJ birth	1/2	6
	$z_{lj} = 2$	$z'_{lj} = 0$	RJ death	1/2	7
		$z'_{lj} = 1$	RJ death	1/2	8

Table 3.1: Proposal transition scheme for exploration of the differential model space to update z_{lj}. The transition probabilities 1 through 8 include four pairs of moves that are each other's inverse: (1,2), (3,5), (4,7) and (6,8).

acceptance probability of a switch is calculated as $A_{S_{ij}} = \min\left\{1, \left(R_{D_i} R_{B_j}\right)\right\}$. As in the birth move, if adding $l \to j$ to $\mathcal{E}_0$ defines a cycle in $\mathcal{G}'_0$, we set $A_{S_{ij}} = 0$.

3.4.2 Updating the differential model space through latent indicators z_{lj}

Given the baseline graph $\mathcal{G}_0$ we propose to move over the differential model space updating the latent variables z_{lj}. Updates in the state of $\boldsymbol{z} = [z_{lj}]$ will also define changes in $\mathcal{G}_1$.

We select an edge $l \to j$ at random. Depending on the current state of $\mathcal{G}_0$ and z_{lj}, we consider the proposal transitions summarized in Table 3.1.

Acceptance probabilities for the proposed transitions are detailed in the following sections. As before, let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ denote the current state and the proposed new state. Note that the probabilities of selecting one of the transition probabilities, numbered 1 through 8 in Table 3.1, are exactly matched. Therefore these probabilities do not appear in the Metropolis-Hastings acceptance probabilities stated below.

3.4.2.1 Birth move

When z'_{lj} is proposed to be 2 it results in increase of dimension in γ. We follow the principles of RJMCMC and augment the γ by proposing γ'_{lj} from $N(0, \zeta^2)$. Then the Jacobian matrix is 1 and proposal is symmetric so they also cancel out and we are left with the acceptance probability $A_{B_z} = min\,(1, S_B)$ with

$$S_B(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\gamma'_{lj} \mid \nu_{lj}, \sigma_j^2, z'\right) p\left(\mathcal{G}'_1\right) p(z' \mid \mathcal{G}_0, \mathcal{G}'_1)}{q\left(\gamma'_{lj}\right) p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\gamma_{lj} \mid \nu_{lj}, \sigma_j^2, z\right) p\left(\mathcal{G}_1\right) p(z \mid \mathcal{G}_0, \mathcal{G}_1)}$$

3.4.2.2 Death move

When the current $z_{lj} = 2$ then move to 0 or 1 will result in a reduction in dimension. Using the same argument as Giudici and Green (1999), this is nothing more than inverse of the birth move. Hence the acceptance probability becomes $A_{D_z} = min\,\{1, 1/S_B(\boldsymbol{\theta}', \boldsymbol{\theta})\}$.

3.4.2.3 Moving z_{lj} between 0 and 1

The transition $z_{lj} \in \{0, 1\} \longrightarrow z'_{lj} \in \{0, 1\}$ does not involve changes in the dimension of γ. The acceptance probability, in this case, is obtained via ordinary Metropolis Hastings calculations as

$$A_{D_z} = min\left\{1, \frac{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\mathcal{G}'_1\right) p(z' \mid \mathcal{G}_0, \mathcal{G}'_1)}{p\left(Y_j \mid \tilde{X}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\mathcal{G}_1\right) p(z \mid \mathcal{G}_0, \mathcal{G}_1)}\right\}.$$

Priors density for $\boldsymbol{\gamma}$ cancel out since they are both 1.

3.4.3 Updating other parameters

Component-wise updates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are amenable to Gibbs sampling. This strategy may however lead to poor mixing and slow convergence (Geyer, 2010b). We will use the fact that a closed form solution for constrained MLE is available for maximum likelihood estimation (MLE) if we define a linear equality constraint based on $\mathcal{G}_0$ and z (Golub, 1965; Stirling, 1981; Neytchev, 1995) . Using this peak in the likelihood we can jointly propose $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\gamma}_j$ for each $j = 1, \ldots, p$, by the method of over relaxation (Neal, 1995).

Overrelaxation Algorithm: We propose a new set of values for $\boldsymbol{B}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j$ by the method of over-relaxation (Adler, 1981). We partition $\boldsymbol{B}_j$ into three,

- the 0 constrained group ($l \notin pa_0(j)$ or $z_{lj} = 0$),
- the equality constrained group ($z_{lj} = 1$), and
- the remaining group.

Proposal for the first 2 groups are trivial, the proposed value of the first group is 0 and second group is $-\beta_{lj}$. The proposal of the third group is done in two steps, first we get the constrained MLE then using that MLE we move the center from the current location to the other side of the MLE than propose a new set of values from a joint distribution. For the ease of notation, for the remainder of this section we will use $\boldsymbol{B}_j$ to denote only the set of parameters that belong in the third group.

Constrained MLE: For a given set of constraints $\mathcal{G}_0$ and z, we can construct a linear constraint matrix L_j explicitly as having a row for each of the constraint

imposed by the combination of $\mathcal{G}_0$ and z so that $L_j B_j = 0$. This translates to defining the entries of L_j as

$$L_j = \begin{cases} \text{for each } l \notin pa_0(j): & l_{\beta_{lj}} \to 1 \\ \text{for each } z_{lj} = 0: & l_{\gamma_{lj}} \to 1 \\ \text{for each } z_{lj} = 1: & l_{\beta_{lj}} \to 1 \text{ and } l_{\gamma_{lj}} \to 1 \end{cases}$$

where $l_{\beta_{lj}}$ and $l_{\gamma_{lj}}$ are entry in L_j with position corresponding to β_{lj} and γ_{lj}.

Then, given

$$y_j \;=\; \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j + \epsilon, \qquad \epsilon \sim N(0, \Sigma) \text{ and } L_j \boldsymbol{B}_j = 0$$

the maximum likelihood estimation (MLE) for $\hat{\boldsymbol{B}}_j = (\hat{\boldsymbol{\alpha}}_j, \hat{\boldsymbol{\beta}}_j^T, \hat{\boldsymbol{\gamma}}_j^T)^T$ has a closed form solution (Golub, 1965; Stirling, 1981; Sallas, 1988; Neytchev, 1995) .

$$\hat{\boldsymbol{B}}_j = \breve{\boldsymbol{B}}_j - (\tilde{\boldsymbol{X}}_j^T \tilde{\boldsymbol{X}}_j)^{-1} L_j^T (L_j (\tilde{\boldsymbol{X}}_j^T \tilde{\boldsymbol{X}}_j)^{-1} L_j^T)^{-1} L_j \breve{\boldsymbol{B}}_j \text{ where } \breve{\boldsymbol{B}}_j = (\tilde{\boldsymbol{X}}_j^T \tilde{\boldsymbol{X}}_j)^{-1} \tilde{\boldsymbol{X}}_j^T y_j$$

The unconstrained posterior distribution: For all the parameters defined in the complementary space of L_j, the proposal can be made from joint Gaussian distribution after over relaxation move, which is a benefit of working with a jointly Gaussian model.

Since the joint prior distribution of $\boldsymbol{B}_j$ is

$$\boldsymbol{B}_j \sim N(\mu_b = (a, b_j^T, \nu_j^T)^T, \Gamma_j) \text{ and } \Gamma_j = \frac{1}{\omega_j} \sigma_j^2 I \tag{3.6}$$

without the structural constraint, the unconstrained posterior is distributed as $N(\tilde{\boldsymbol{B}}_j, \tilde{\Sigma}_j)$ where $\tilde{\Sigma}_j = \left(\tilde{\boldsymbol{X}}_j^T \Sigma^{-1} \tilde{\boldsymbol{X}}_j + \Gamma_j^{-1} \right)^{-1}$ and $\tilde{\boldsymbol{B}}_j = \tilde{\Sigma}_j \left(\tilde{\boldsymbol{X}}_j^T \Sigma^{-1} \tilde{\boldsymbol{X}}_j \breve{\boldsymbol{B}}_j + \Gamma_j^{-1} \mu_b \right)$.

Overrelaxation Algorithm: Algorithm for updating $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ proceeds as following

1. Start with the current value of the estimate $\boldsymbol{B}_j$
2. We partition $\boldsymbol{B}_j$ into 3; $(B_j^{(0)}, B_j^{(1)}, B_j^{(2)})$;
   - For parameters $B_j^{(0)}$ corresponding to $l \notin pa_0(j)$ and $z_{lj} = 0$,
     - propose $B_j^{(0)'} = 0$
   - For parameters $B_j^{(2)}$ corresponding to $l \in pa_0(j)$ or $z_{lj} = 2$,
     - Propose new values $B_j^{(2)'}$ from $N\left(B_j^{(2)} + 2(\breve{\boldsymbol{B}}_j^M - B_j^{(2)}), \frac{1}{\varphi}\tilde{\Sigma}_j\right)$
   - For parameters $\gamma_{lj}^{(1)'} \in B_j^{(1)}$ corresponding to $z_{lj} = 1$
     - propose $\gamma_{lj}^{(1)'} = -\beta_{lj}^{(2)'}$
3. Acceptance probability is calculated as

$$A_o = min\left(1, \frac{N(B_j^{(2)'}; \tilde{\boldsymbol{B}}_j^{(2)}, \tilde{\Sigma}_j^{(2)})}{N(B_j^{(2)}; \tilde{\boldsymbol{B}}_j^{(2)}, \tilde{\Sigma}_j^{(2)})}\right) \tag{3.7}$$

4. Set $\boldsymbol{B}_j = \boldsymbol{B}'_j = (B_j^{(0)'}, B_j^{(1)'}, B_j^{(2)'})$ if $u \leq A$ where $u \in U[0,1]$, otherwise set it to $\boldsymbol{B}_j$

3.4.4 Updating σ^2

We use Gibbs sampling to update σ^2. The conditional posterior distribution for σ_j^2 $(j = 1, \ldots, p)$ is available in closed form as an Inverse Gamma distribution

$Inv\text{-}Ga(\tilde{\delta}_j, \tilde{\tau}_j)$ where

$$\begin{aligned}
\tilde{\delta}_j &= \frac{1}{2}\left(\delta_j + n + 1 + \sum_l I\{\mathcal{G}_{lj} = 1\} + \sum_l I\{z_{lj} = 2\}\right) \\
\tilde{\tau}_j &= \frac{1}{2}\tau_j + \frac{1}{2}\left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right)^T \left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right) \\
&\quad + \frac{1}{2}\omega_j \left((\alpha_j - a_j)^2 + \sum_l (\beta_{lj} - b_{lj})^2 I\{l \in pa_0(j)\} + \sum_l (\gamma_{lj} - \nu_{lj})^2 I\{z_{lj} = 2\}\right)
\end{aligned}$$

Detailed calculations are reported in supplemental appendix 5.3.2.

3.4.5 Other computational concerns

Although the above algorithm is straightforward to implement, the computation of MCMC on the space of graphs requires extra considerations. Several Authors pointed out how the model space may be characterized by many local modes (Scott and Carvalho, 2007; Barker et al., 2010). Furthermore regions of high posterior probability could get extremely peaky as the sample size increases, making it difficult for a naïve Monte Carlo simulation scheme to effectively transition between highly likely alternative models.

To deal with this problem Scott and Carvalho (2007) suggested using a stochastic search method, which combines a local as well as a global move. Their method is devised for decomposable undirected graphs and it is not directly applicable to our model. Alternatively Barker et al. (2010) recently proposed the MC^4 algorithm on DAGs by expanding the MC^3 algorithm (Madigan et al., 1995) with a parallel tempering (Geyer, 1991) step and showed improved performance.

In this regards, to increase the efficiency of our sampler, we expanded our sampler to perform parallel tempering (Geyer, 1991) on RJMCMC as suggested by Jasra et al. (2007) and Barker et al. (2010).

3.4.5.1 Parallel tempering move

Parallel tempering (Geyer, 1991) is a population Monte Carlo technique where the target distribution is augmented with an indicator that specify a level of smoothing applied to each of the target distribution. The new joint distribution is the product of each of the distribution over the indicators since each of the density is independent of each other given the indicator. Markov chains at different temperatures are run in parallel and the neighboring states are exchanged between the chains with a predefined rate.

For the case of RJMCMC Jasra et al. (2007) proposes on adding an additional delayed rejection (Green and Mira, 2001) step that increases the efficiency of the algorithm by allowing swaps between the non-neighboring temperatures.

3.4.6 Tempering move with delayed rejection for RJMCMC

Choose a set of temperatures $\{1, \ldots, T\}$ (Geyer and Thompson, 1995) and for each temperature, replicate the parameters $\boldsymbol{\theta}$ for T times;

$$\theta_t = \left\{ \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, z^{(t)}, \sigma^{2(t)}, \mathcal{G}_0^{(t)} \right\}, \; t \in \{1, \ldots, T\}.$$

1. For a preset probability P_t, perform a switch temperature move.

   (a) Choose two temperatures $i_1, i_2 \in \{1, \ldots, T\}$

   (b) Exchange θ_{i_1} with θ_{i_2} with probability

$$P_1(\boldsymbol{\theta}, \boldsymbol{\theta}') \;\; = \;\; min \left\{ 1, \frac{\pi_{i_1}(\theta_{i_2})\pi_{i_2}(\theta_{i_1})}{\pi_{i_1}(\theta_{i_1})\pi_{i_2}(\theta_{i_2})} \right\} \tag{3.8}$$

   where $\pi_t(\theta_t)$ is a posterior density at temperature t evaluated at θ_t.

   (c) If rejected perform delayed rejection step by choosing neighboring tem-

peratures $i_3, i_4 \in \{1, \ldots, T\}$ and exchange θ_{i_3} with θ_{i_4} with probability

$$P_2(\boldsymbol{\theta}, \boldsymbol{\theta}'') \;=\; min\left\{1, \frac{\pi_{i3}(\theta_{i4})\pi_{i4}(\theta_{i4})(1 - P_1(\boldsymbol{\theta}'', \boldsymbol{\theta}^*)}{\pi_{i3}(\theta_{i3})\pi_{i4}(\theta_{i4})(1 - P_1(\boldsymbol{\theta}, \boldsymbol{\theta}'))}\right\} \quad (3.9)$$

were θ^* is the hypothetical θ if $P_1(\boldsymbol{\theta}, \boldsymbol{\theta}')$ was accepted.

2. Perform regular RJMCMC with probability $(1 - P_t)$ for each temperature T.

For details of the method and other suggestions we refer you to Jasra et al. (2007) and Barker et al. (2010).

3.4.7 Posterior summaries

Posterior probabilities $p(\mathcal{G}_0 \mid Y)$, $p(\mathbf{z} \mid Y)$ and corresponding MCMC samples characterize our knowledge about baseline and differential interactions in light of the data. Based on these quantities, the main inferential goal is to select representative baseline and differential graphs, say $\mathcal{G}_0^*$ and $\mathcal{G}_1^*$. While posterior probabilities do summarize evidence about interaction structures, selection a point estimate in the models space requires further decision theoretic considerations.

Given a joint model on edge and parameter inclusion probabilities, in the Bayesian framework, selection of point estimators for interaction structures $\mathcal{G}_0$ and $\mathcal{G}_1$ usually translates into the appropriate definition of a cutoff value for posterior inclusion probabilities (Scott and Berger, 2006; Müller et al., 2006). A cutoff threshold is often determined in order to ensure optimization of a chosen loss function. For example, a loss function that equally weigh false positives and the false negatives would threshold inclusion probabilities at 0.5. This choice coincides with the median probability model proposed by Barbieri and Berger (2004). They justify the median probability model by the optimal predictive

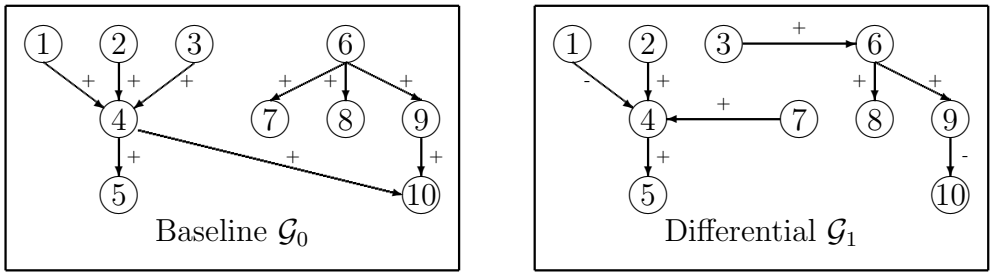


Figure 3.3: The true graphs used to generate the data for the simulation.

performance (under some additional assumptions).

An alternative common strategy is to select a point estimator, on the basis of classical multiple comparison arguments. An often used error rate is the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Rules as discussed in Benjamini and Hochberg (1995) control the frequentist expectation of the error rate across repeat experimention. Several authors chose instead to control the posterior expectation of the same error rate. See, for example, Newton (2004).

The rest of this chapter is based on results obtained under median model selection (Barbieri and Berger, 2004) and controlling explicitly the posterior expected FDR. Alternative decision theoretic arguments and possible loss functions are discussed in Müller et al. (2006).

3.5 Simulation Study

We tested the proposed method on synthetic data, by generating observations from graphs configured as in figure 3.3. There are 10 vertices and 9 directed edges in the baseline graph, all with positive weights β_{lj} on the edges. The differential graph has 8 directed edges that is the result of 3 cancelation, 2 additional edges,

and 2 edges with negative effect sizes $\gamma_{lj} + \beta_{lj}$. We simulated 50 baseline samples and 30 differential sample.

Figure 3.5 shows the estimated edge inclusion probabilities for the baseline graph and for the differential graph. The barplot in row ℓ and column j corresponds to the edge $v_\ell \to v_j$. Edges that are present in the simulation truth are marked with an asterix. The estimated inclusion probability is high for edges that were included in the simulation truth, as desired. There is, however, some uncertainty, especially in the upper portion of the graph. Figure 3.6 shows the bar plot of the posterior estimates of mixing proportions for the differential edge: $z_{ij} = 0$ as left white bar, $z_{ij} = 1$ as central blue bar, and $z_{ij} = 2$ as right red bar. Again, edges that are present in the simulation truth are marked by a red surrounding box and the true value is indicated by an asterisk below each plot. The proposed method identifies differential interactions quite accurately, defining strong control over false negatives (row 6 column 7) and false positives (row 1 column 5).

Figure 3.7 shows the posterior mean and standard deviation of the effect size for each of the edges in the baseline graph (left) and the differential graph (right). The true value marked with asterisk below the density is covered by the posterior samples, indicating that the model provides accurate recovery of true effects size.

We compare results over two decision criteria: varying the threshold of the posterior inclusion probability and varying the threshold value for the q-value in FDR procedure on the posterior inclusion probability for the baseline graph and the differential graph. We evaluate the operative characteristics of different decision criteria in our simulated experiment on the basis of two quantities: the False Discovery Rate (FDR) and the Missed Detection Rate (MDR). Letting TP indicate true positives, FP false positives, and FN is false negatives, we defing

the FDR and MDR are defined as follows

$$FDR = \frac{FP}{FP + TP}, \qquad MDR = \frac{FN}{FN + TP}. \tag{3.10}$$

Figure 3.4 shows a comparison of two decision criteria in relation to these quantities.

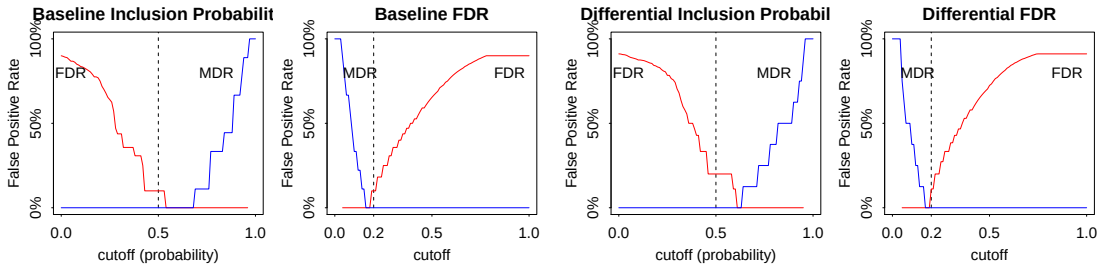


Figure 3.4: Comparison of the decision criteria for False Positive Rate (FPR) and Missed Detection Rate (MDR) for the baseline and the differential group. A dotted line on the inclusion probability plots corresponds to the choice made by median probability model. The FDR plots have a dotted line at 0.2 corresponding to suggestion by Efron (2007).

The dotted line on the inclusion probability corresponds to the choice made by median probability model (Barbieri and Berger, 2004). The FDR has a dotted line at the threshold value of 0.2 corresponding to suggestion by Efron (2007). For this particular simulation, both median graph criteria and the criteria of Efron (2007) are performing equally well.

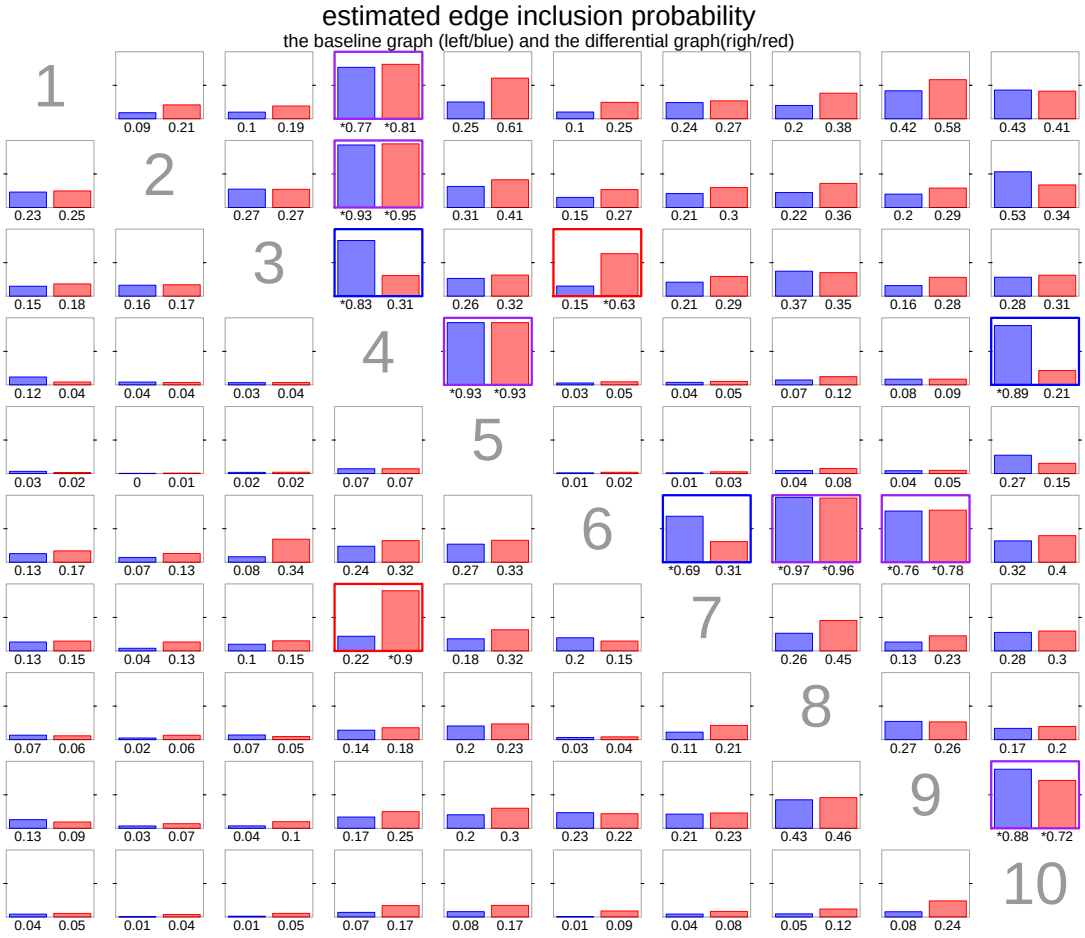


Figure 3.5: Barplot of the estimated edge inclusion probabilities for the baseline graph (left) and the differential graph (right) for each edge. The barplot in row ℓ and column j corresponds to the edge $v_\ell \to v_j$. Edges that are present in the simulation truth are marked with an asterix below the corresponding bar, and have a thick colored box.

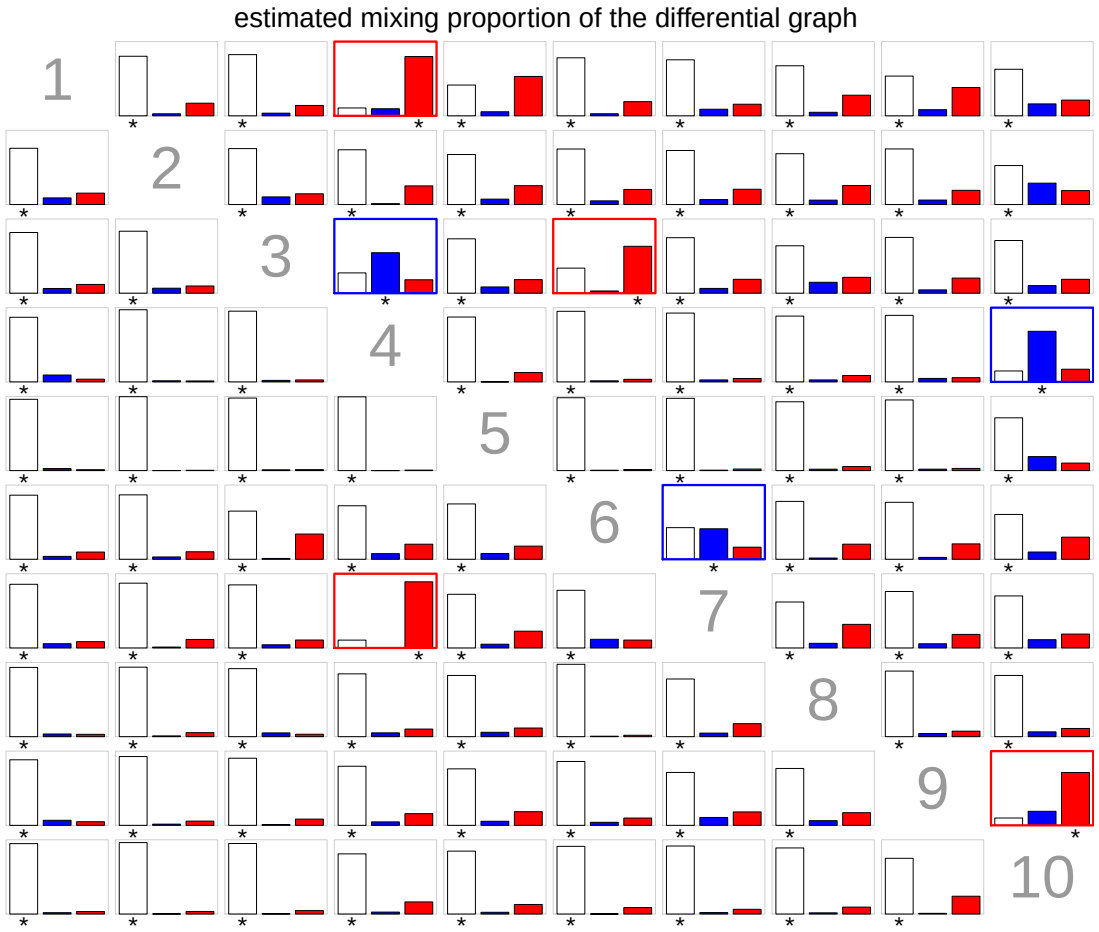


Figure 3.6: Barplot of the posterior estimates of the mixing proportions π_{lj}^0, π_{lj}^1, and $1 - \pi_{lj}^0 - \pi_{lj}^1$ for each edge. The true value is marked with an asterisk below the density and the true signal has thick colored surrounding line.

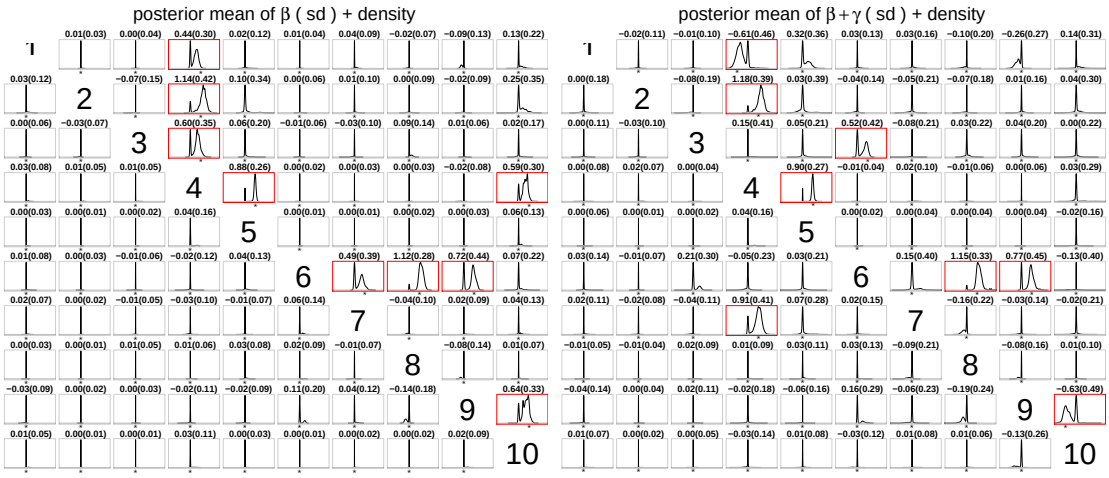


Figure 3.7: Marginal posterior distributions $p(\beta_{\ell j} \mid Y)$ for the baseline coefficients and $p(\gamma_{\ell j} + \beta_{\ell j} \mid Y)$ for the differential coefficients. All densities are plotted over the same range, for easy comparison. The number above the density are the posterior mean and standard deviation. The true value of the estimate is marked with an asterisk below the density and the true edges have thick red surrounding box. The posterior density covering the true estimate indicates that the model is tracking the effect size accurately.

3.6 Case Study

We apply our model to the data from a study of Acute Myeloid Leukemia (AML) obtained using the reverse phase protein arrays (RPPA) (Tibes et al., 2006) . RPPA is a high-throughput proteomic technology that provides a quantification of the expression for specifically targeted proteins selected from molecular pathways (figure 3.8).

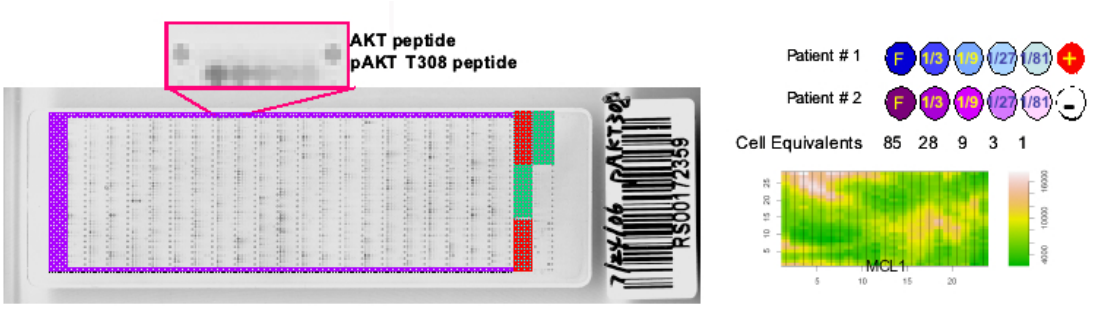


Figure 3.8: Image of the actual reverse phase protein arrays (RPPA).

We use data from a large AML study based on RPPA. We consider 435 AML patients; 332 primary refractory patients and 103 relapsed patients. We will call the refractory patients the baseline group and the relapsed patients the differential group. The objective of this study is to investigate the difference in interactions of important protein markers related to AML for the refractory patients and the relapsed patients. We selected 38 proteins in signal transduction, apoptosis, and cell cycle regulatory pathways and studied their expression profiles in all 435 samples. An attractive feature of the AML data under study is that the number of samples (n = 435) is much greater than the number of proteins (p = 38), which provides an opportunity for principled inference about differential interaction structures on the basis of a highly structured stochastic system.

The prior distributions on the parameters were selected as vague as possible to show that this method does not require strong prior information, which makes it suitable for initial studies since the likelihood will dominate the posterior when the sample size is large. Mean parameters for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ were set to 0. The two parameters of the dispersion parameter σ_l^2 where set to 0.5 and 0.5. The prior on ψ_k is set to $Beta(1, 1)$. For the temperatures of parallel tempering, we selected them uniformly spaced between 1 to 100 on the log scale. We ran our algorithm for 20000 iteration saving every 20th sample.

For the decision rule, since we have no reason to weigh either false discovery nor false negative more than one another, we chose a equal weight loss function $L_N = \overline{FD} + \overline{FN}$. The corresponding decision rule for this loss function thresholds the inclusion probability at 0.5 (Müller et al., 2004) which is the median graph proposed by Barbieri and Berger (2004).

Figure 3.9 is a network representation of the estimated graph for the refractory and relapsed patients. The network of the relapsed patients is sparse compared to the refractory patients; the baseline network had 99 edges whereas the differential network only had 83 edges. Table 3.2 lists the differential edges that differed between the two networks.

While we maintain that our findings are purely exploratory, we have found that selected differential interaction patterns have been confirmed in the literature as potential indicators of more aggressive forms of AML. For example Kornblau et al. (2011) report that signaling changes affecting the AKT-S6 pathway are associated with relapse after chemotherapy in AML patients (see our corresponding result in Table 3.2, Cancelled Edges). On the differential activation side (Extra edges, Table 3.2), our results agree with (Ozawa et al., 2008) who reported how SRC family kinases regulate STAT transcription factors in AML cells, which are

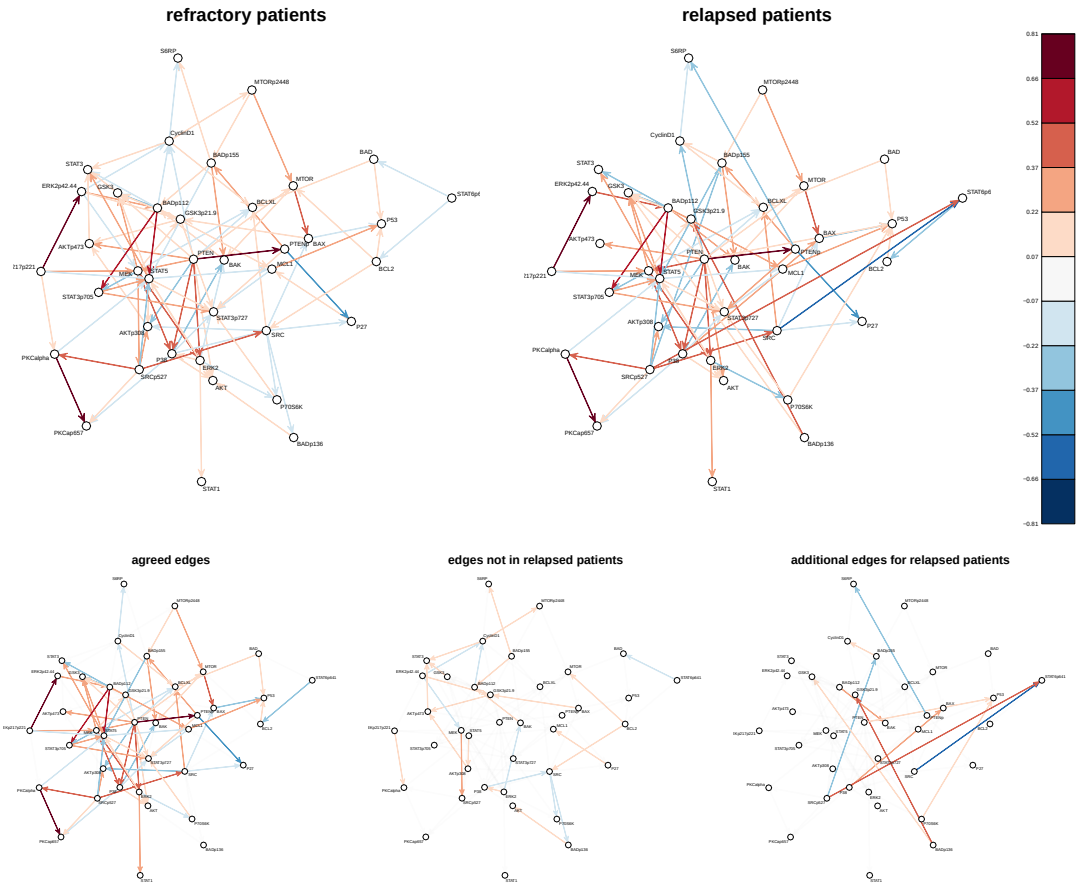


Figure 3.9: Network representation of the estimated protein network for refectory patients and relapsed patients. The strength of association is shown as the intensity of the color; the red is for positive association and blue is for negative association, as shown in the thermometer bar on the right. The bottom three plots classify the edges into three categories: the edges that two groups agree on, the edges that does not exist in the differential graph, and edges that only exist in the differential graph. The differential graph is more sparse compared with the baseline network.

Extra Edges	Canceled Edges
SRCp527→BADp155	BADp136→ AKT
BADp112→BAK	STAT5→ AKTp308
P38→BAX	AKTp308→ AKTp473
BADp155→CyclinD1	GSK3p21.9→ AKTp473
BADp136→GSK3	STAT6p641→ BAD
BADp136→GSK3p21.9	BADp155→ BADp112
P70S6K→P53	BAK→ BADp112
PTEN→P53	SRC→ BADp136
PTENp→S6RP	ERK2→ BAK
SRC→STAT6p641	BADp112→ CyclinD1
SRCp527→STAT6p641	ERK2p42.44→ CyclinD1
	BAX→ GSK3p21.9
	ERK2p42.44→ GSK3p21.9
	P27→ MCL1
	ERK2p42.44→ MEK
	BCL2→ MTOR
	CyclinD1→ MTORp2448
	ERK2→ P38
	SRC→ P70S6K
	MEKp217p221→ PKCalpha
	BADp155→ S6RP
	BCL2→ SRC
	P38→ SRC
	MEK→ SRCp527
	CyclinD1→ STAT3
	GSK3→ STAT3
	STAT5→ STAT3

Table 3.2: The list of differential edges.

known to play a fundamental role in growth and proliferation processes.

Figure 3.10 is the estimated posterior inclusion probability. The figures for the estimated mixing proportion and the posterior density plot of the coefficients can also be found in the supplementary materials. A comprehensive bio-medical interpretation of our findings is perhaps out of the scope or this chapter, but it is our hope that our illustration shows the potential and practical relevance of the proposed method.

3.7 Discussion

We proposed a novel probability model for inference on differential interaction in Gaussian DAGs. The proposed framework is likely to be particularly useful when primary interest focuses on potential contrasts characterizing the association structure between known subgroups of a given sample. Although we only worked on a case where there are only two subgroups, the method is directly

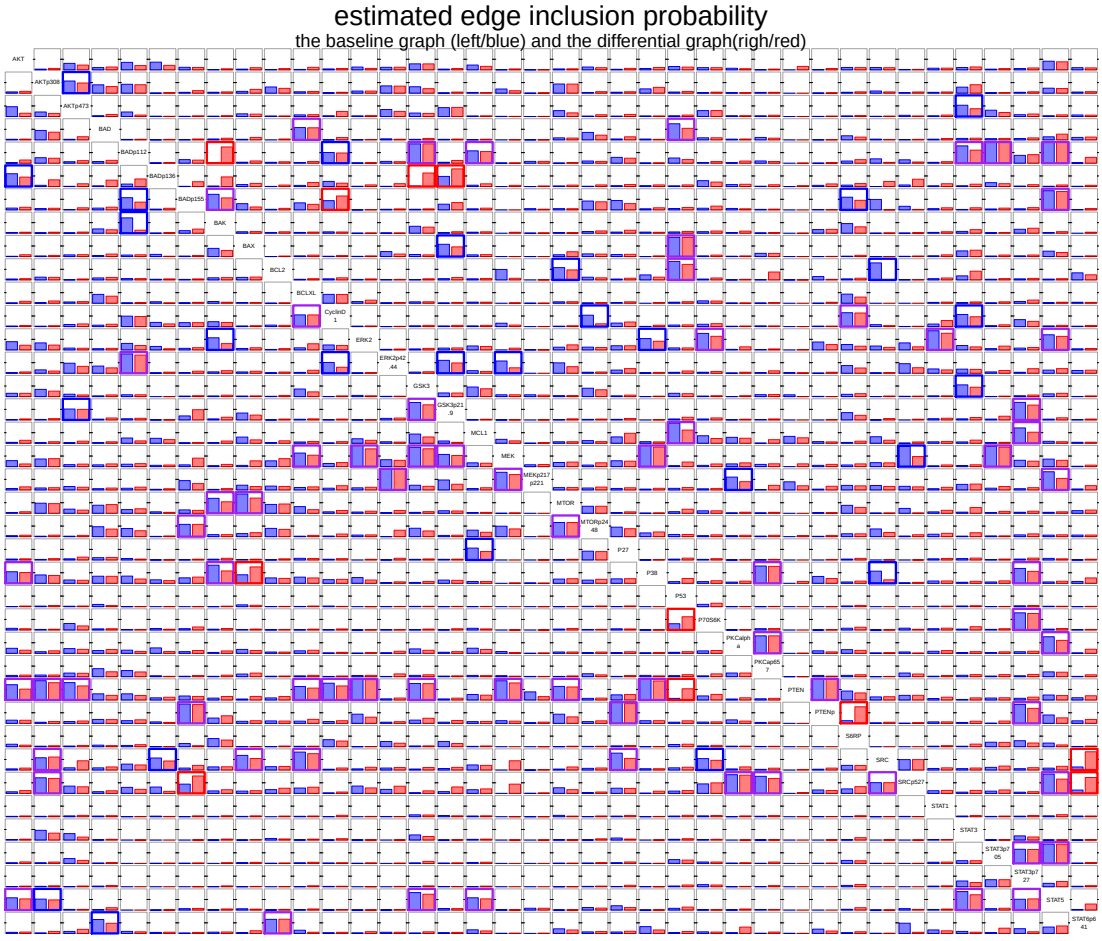


Figure 3.10: Barplot of the estimated edge inclusion probability for the refractory patients (left) and the relapsed patients (right) for each edge.

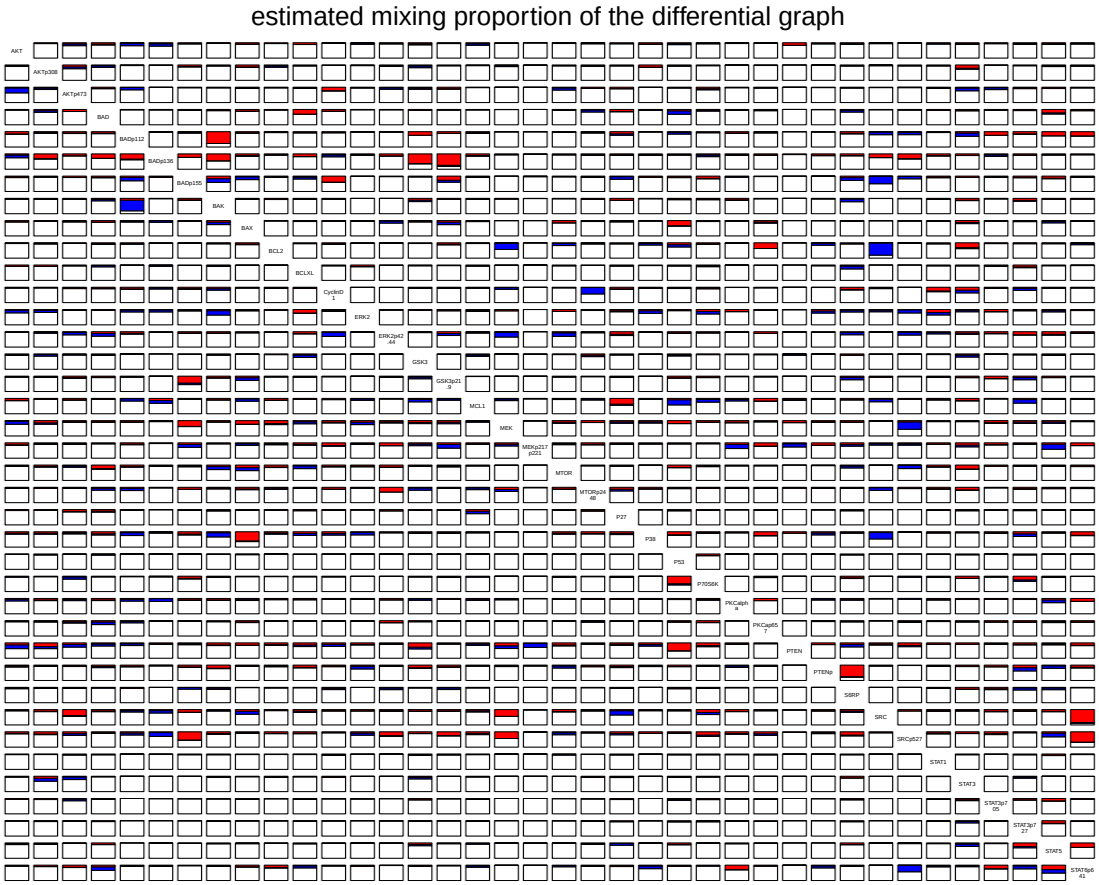


Figure 3.11: Stacked barplot of the posterior estimates of the mixing proportions for z_{lj} defined for each differential edge: $z_{lj} = 0$ is white, $z_{lj} = 1$ is blue, and $z_{lj} = 2$ is red.

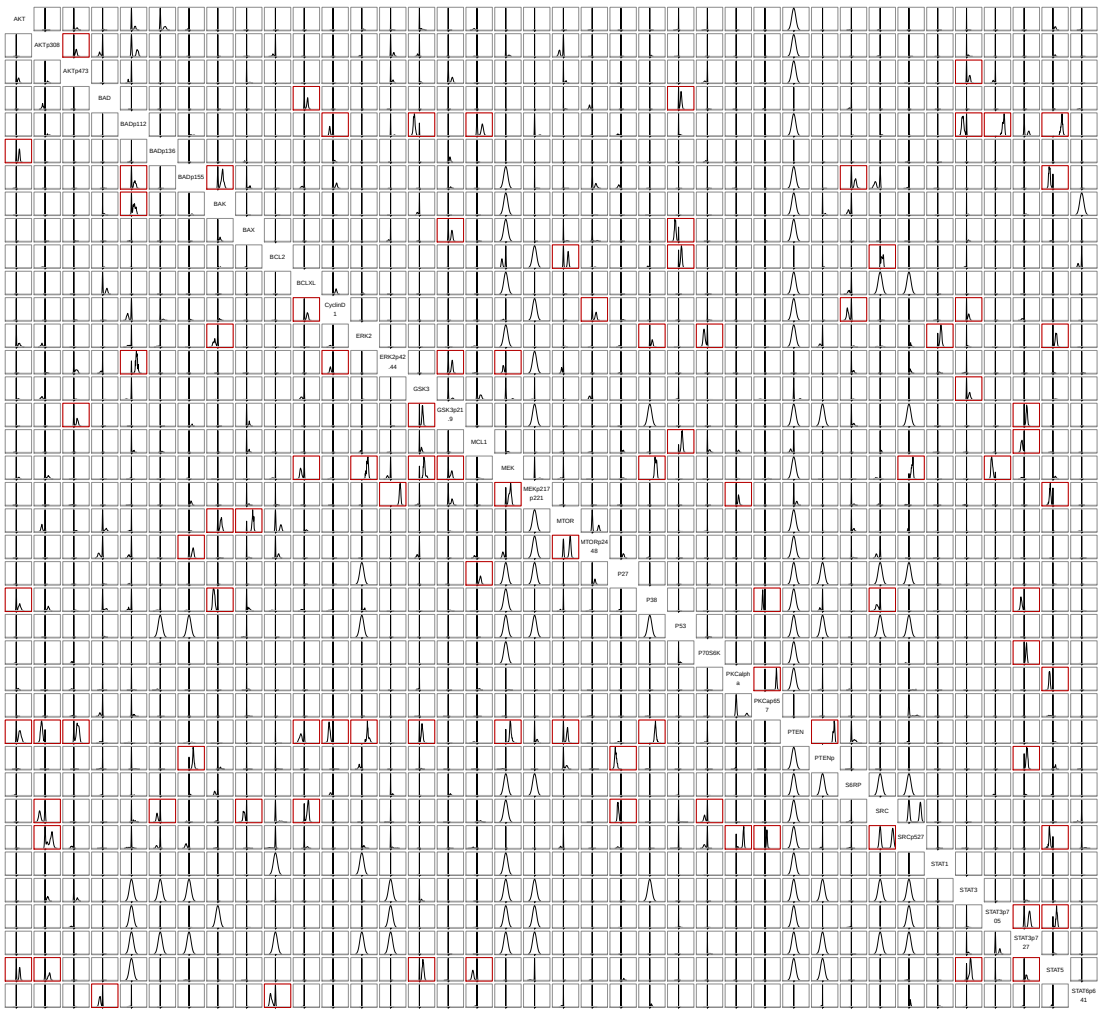


Figure 3.12: The density plot of the estimated posterior distribution for the baseline coefficients plotted on same horizontal range. The edges in median graph has thick red surrounding box.

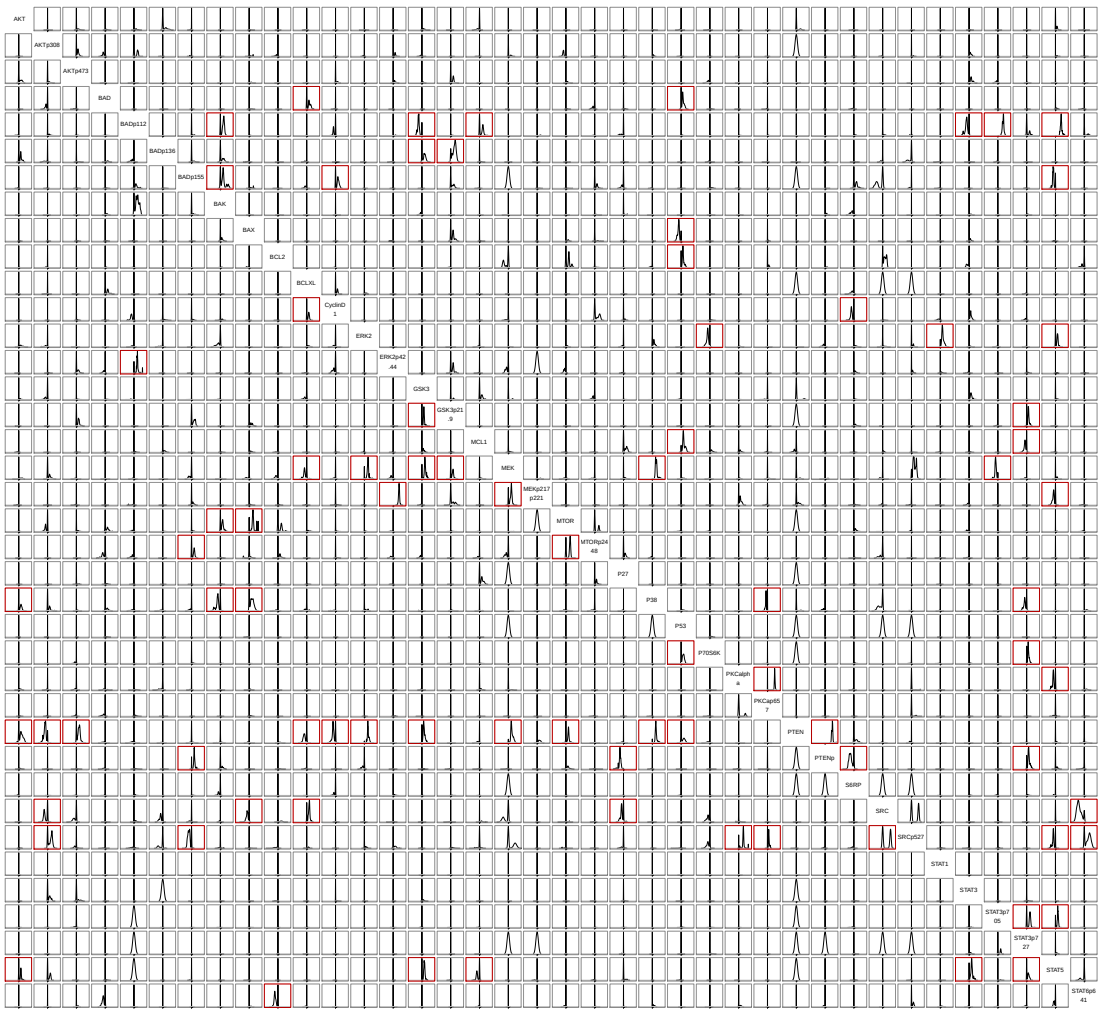


Figure 3.13: The density plot of the estimated posterior distribution for the differential coefficients plotted on same horizontal range. The edges in median graph has thick red surrounding box.

generalizable to the case of k subgroups. We evaluated our method analyzing data generated from a synthetic experiment and showed that our inferences have desirable operative characteristics. The application of the proposed model to the analysis of RPPA data in AML identified interesting differential regulation patterns, distinguishing refractory from relapsed patients. While we are well aware that our model belongs to the class of hypothesis generation tools, we remark that the proposed methodology avoids the use of step-wise analyses and ad-hoc penalization choices, providing a principled tool for inference on differential networks.

The conjugate Gaussian setting, provides several algebraic and computational advantages. However, there are costly steps, associated with the proposal of coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, requiring several matrix inversions. While this is not an issue, as long as the sampled graph is sparse; the proposed computation could be computationally demanding, when dealing with large and dense graphs. In these cases one may need to consider alternative proposal strategies.

The propose framework of differential network inference could be extended beyond the multivariate Gaussian distribution. Our prior on models space and interaction parameters could, for example be applied to the approach of Telesca et al. (2012b), who show how to incorporate heavy tails in the observations by the use of a mixture model. As for the case of discrete and mixed data, the copula Gaussian graphical model framework proposed by Dobra and Lenkoski (2011) could be easily expanded using a modeling strategy similar to the one proposed in this chapter.

Extension beyond DAGs may be desirable in many applied settings. Fore example, in the setting of Reciprocal Graphs Koster (1996), used in Telesca et al. (2012a) one may allow baseline and differential models, to be defined in terms

of undirected edges as well as the directed ones, with the possibility of including cycles and reciprocal relations. We should also point out that the same idea could of course be applied to undirected graphical models. While these extension are conceptually trivial, coherent multivariate representation and computational constraints may require extensive additional work.

CHAPTER 4

CONCLUSION AND FUTURE DIRECTIONS

4.1 Conclusion

We have presented a Bayesian hierarchical modeling framework to address complex interactions across heterogeneous structured data. Our contributions are especially directed toward integrative biology applications. With regards to statistical modeling, our proposal can be itemized into:

- the introduction of genuine a chain graph modeling framework in the Bayesian hierarchical nonlinear regression literature;
- an inferential model for structural comparison of Gaussian Graphical Models along with novel computational scheme;
- a simultaneous posterior expected FDR control across sets of parameters that are not exchangeable;
- a posterior credible interval thresholding method for FDR control of continuous parameters;
- a set of visualization techniques to make sense of the complex interactions;
- a computational package for R.

The two fundamental models we developed in this thesis by no means span the entire possibility of complex interactions that occur in integrative biology applications, however we believe they address large classes of problems. For the technical details we refer you to the specific chapters.

4.2 Future Directions

In the remainder of this thesis we discuss classes of models that we did not address in this thesis along with other possible future extensions to the proposed models. We also include preliminary work on the approximation of discrete models using continuous models discussed in 4.2.8 in section 4.3.

4.2.1 Bayesian computation

The Bayesian hierarchical modeling framework that we have proposed provides flexibility that allows for heterogeneous structured data to be integrated into the model. However, there are immediate challenges in the computation associated with our proposed modeling framework and that Bayesian statistics in general must address in order for it to keep up with the demands of the scientific community. The MCMC method is computationally costlier than optimization methods. Hamiltonian Monte Carlo approaches (Neal, 2012) are showing promising directions with software becoming available (Stan Development Team, 2013; Hoffman and Gelman, 2011). These methods have not been extended for nonlinear differential equation models, which are something that would increase the practicality of our proposed pharmacogenetics model. The idea of integrated nested Laplace approximations (Rue et al., 2009) is growing in its presence as an alternative to MCMC methods for Gaussian processes. Although there are models available for

fixed network structure, it is not fit for models with unknown graph structure at the moment. It may be interesting to see how these ideas may be combined to give reasonable approximations as a possible direction for future research.

4.2.2 Computation of graphical models

The computation for graphical models has seen much improvements while this thesis was being written. The use of double metropolis hastings (Liang, 2010) provides an appealing solution to the computation of intractable normalizing constants in the simulation of undirected graphs (Wang and Li, 2012; Cheng and Lenkoski, 2012). For the computation of DAGs, sampling from set of a partial directed acyclic graphs (PDAGs) or essential graphs He et al. (2012), instead of sampling the whole space of DAGs, seems to be a promising direction that would make the idea of differential Gaussian DAGs more scalable. Nevertheless, in terms of computational convenience, these methods are not even close to penalized regression methods (Witten et al., 2011). In all fairness, however, frequentist regularized estimation is not concerned at all with the estimation of uncertainty. Considering how to assess the uncertainty in approximate but scalable computation frameworks (Jones et al., 2005; Scott and Carvalho, 2007; Murray and Ghahramani, 2004) may be a fruitful direction in the short term to compete with these methods.

4.2.3 Structural inference

With regards to making statistical inference on the structures of graphs, there are still unresolved methodological questions. The nonparametric Bayesian inference of Rodriguez et al. (2011) that uses Dirichlet Process mixture of GGMs tries to address the issue of identifying the global subgroups by avoiding making

explicit inference on the graph structures between the groups. Therefore as the byproduct of the discretization of the Dirichlet Process, one obtains clustering in the posterior. However, it is not possible to make cluster-specific inference. When subpopulations are identified in a sample, the natural question to ask is how are the associations different between subgroups. The differential graphical model we propose, in theory can be extended to multiple graphs that would allow for subpopulation level inference. Nevertheless it is unclear how to achieve global inference on the graph structures between the subpopulation after posterior distribution on the edges have been collected. To achieve the middle ground between the two approaches the work of Rodriguez et al. (2011) can be extended by using product partition models with partitions induced by regression on covariates as Müller and Quintana (2010), which would allow for some level of inference at the covariate level.

4.2.4 Chain graph dependence structure

The use of chain graphs as dependence structures in the Bayesian modeling is an idea that can be explored much further. Temporally evolving networks and directed spatial-temporal models would naturally fall under this category. Together with the idea of higher order interactions, such models may be able to capture the time evolving characteristics of the higher order interaction.

4.2.5 Incorporating informative graphical prior based on biological databases

Modeling the biological association structure in a multivariate Gaussian fashion is at it's most primitive form and is something that should be improved upon. A rigorous approach would incorporate the information available in the scientific

databases such as Gene Oncology (GO) database (http://www.geneontology.org/) in a reasonable fashion. However, in graphical models in general, the likelihood surface is likely to be highly spiky, even for a moderately sized dataset and any vague information will be ignored in the posterior. Therefore, quantifying the amount of information associated with current cumulative scientific knowledge, while letting the model learn from a current experiment data is not an easy task.

4.2.6 Higher order interactions

Incorporating higher order interactions in the association structure is important especially for biological applications since it is usually not a single gene that creates a functioning protein but multiple genes or multiple proteins that combine to create some phenotypic reaction. The general formulation is laid out by Besag (1974) however, implementation of such modeling poses computational challenges. Incorporation of scientific information, such as relative distance between the SNPs, to confine the number of higher order interactions to be considered, will be crucial in realistic applications.

4.2.7 Differential Gaussian DAG models

For the differential GDAG model, depending on the application, Gaussian assumptions may not be viable. In those cases, the simplest amendment is to introduce an additional layer of latent variable so that the problem is reduced to a Gaussian. When the data is continuous but non-Gaussian, a mixture of Gaussian model could be an alternative. For example to extend the Gaussian DAG

model the likelihood will look like

$$Y_i|Y_{pa_{s_i}(i)}, \beta, \sigma^2, d_{s_i}, \rho = \sum \pi_k N \left(\beta_{i0}^{(k)} + \sum_{y_l \in pa_{s_i}(i)} \beta_{il}^{(k)} y_l, \sigma^{2(k)}_{i|pa_{s_i}(i)} \right).$$

For binary case we may use the probit model (Chib and Greenberg, 1996) or autologistic model (Besag, 1974) as in Telesca et al. (2012a)

$$p(y_{ij} = 1|y_{i,pa_{s_j(j)}}, d_j, \rho) \quad = \quad \frac{exp(\alpha_j + \sum \beta_{j|pa_{s_j}(j)} y_{i,pa_{s_j(j)}})}{1 + exp(\alpha_j + \sum \beta_{j|pa_{s_j}(j)} y_{i,pa_{s_j(j)}})}.$$

When variables are not measured on a same scale, use of copula Gaussian model (Dobra and Lenkoski, 2011)

$$\begin{aligned} y_{ij} &= F_j^{-1}[\Phi(z_{ij})] \\ Z_i \mid \cdot &\sim N_p(0, \Sigma_k) \end{aligned}$$

for $i = 1, \ldots, n$. Σ_k could be specified by mixture model of the covariance structure, is a possibility.

4.2.8 Approximation of discrete models using continuous models

One issue that came up repeatedly during the course of this thesis was the issue involving the computational cost of discrete modeling. Without getting into the argument on the theoretical validity of discrete modeling and given the fact that novel algorithms are being proposed to improve the computation, the immense size of discrete model space still poses computational challenge. Discrete modeling is intended to assist the decision, however, the criteria of median graph is confounded by the quality of the estimates of edge inclusion probabilities. For the time being, it may make sense to turn the argument around and use an ap-

proximate continuous model and use a better criteria to make decision in the posterior since after all, a model is a model and it is always wrong. But we can adjust for its short coming by making reasonable decisions. With this idea in mind we have incomplete work we will attach in appendix 4.3. Although work is preliminary, for many of the variable selection problems, using continuous models with educated decision out performs the other rigorous variable selection models or penalized regression approaches.

4.3 Model Reporting in Continuous Models

Reporting "significant" coefficients after fitting a regression model is a screening procedure that is consistent with the Fisher's notion of statistical testing (Rice, 2010). When a coefficient is reported as "significant" we are not claiming anything about the alternative but simply stating that a regression coefficient has small chance of being "insignificant". The term "insignificant" is a subject dependent notion. In domains such as genetics a gene having zero effect on some protein formulation is well accepted, whereas in social science the idea of zero effect may be refutable(Gelman and Rubin, 1995).

In general, the screening process should be conservative since reported "significant" coefficients usually must go under the scrutiny of science, which may be expensive depending on the domain of study. Yet often times researchers want to maximize the probability of selecting an association that no one has found before at the cost of flagging a few "insignificant" effects as begin significant. These antagonistic goals have given rise to the idea of false discovery rate (FDR) control, which has become popular in many areas of studies. Again, there is much entanglement with the idea of FDR and the hypothesis-testing framework. However, it can be justified from the perspective of maximizing utility under the

Bayesian decision theoretic framework (Müller et al., 2006) without getting into the hypothesis testing argument.

Leaving aside the theoretical issues, the practical side of the matter with regards to variable screening is not completely resolved. Most of the argument centers around two schools of thoughts. On one hand are people who prefer to use point-mass prior to calculate the posterior inclusion probabilities and on the other end are people who use continuous priors and look at the posterior credible interval (CI) to see if the $1 - \alpha\%$ interval covers 0 or not. We will not get into the argument of whether the use of point-mass prior is appropriate; there are references that deal with this issue (Berger and Delampady, 1987). We focus on the later case, especially limited to the case of linear regression, where a continuous prior is used and we need a reasonable way to screen noteworthy coefficients amongst them.

The idea is derived from Gelman et al. (2012), in which given the posterior distribution of the coefficients, we calculate the probability that that posterior distribution is bigger or smaller than a point mass distribution at 0. This may sound disturbing at first, yet, it is in fact almost identical to the procedure based on $1 - \alpha\%$ CI. The procedure we propose has an additional merit of allowing for explicit control of FDR under the framework of Müller et al. (2006). As we will show through simulation that the FDR control is accurate under this procedure.

The rest of the section is structured as follows; in section 4.3.1 we show a procedure that is equivalent to variable selection based on $1 - \alpha\%$ CI. We also look at the frequentist property of FDR control under the use of this (p-value) and show some simulation result in section 4.3.2. We end the paper with discussion.

4.3.1 Bayesian "p-value" like statistics

The model we consider is a regression model of predictor $X = [x_{ij}]_{n\times p}$ on the outcome variable $y = [y_i]_n$. Without loss of generality, we further assume y and X are standardized so that they both have mean 0 and standard deviation of 1 so that

$$y \sim N(X\beta, I_p) \tag{4.1}$$

where β is the standardized regression coefficient. In modern Bayesian analysis, β will be given a prior and M posterior samples b are drawn using MCMC machinery.

After fitting this model, what is commonly done is to look at the posterior CI of β and if 0 is not included in the $1-\alpha\%$ posterior CI, we claim that the probability of β being 0 is less than α. If we formally define a decision indicator $d_j = 1, (j = 1, \ldots, p)$ to indicate decision to report the coefficient β_j as significant, then decision rule take the form

$$\begin{aligned} d_j &= 1\left\{0 \notin \left[b\left(M\frac{\alpha}{2}\right), b\left(M\left(1-\frac{\alpha}{2}\right)\right)\right]\right\} \\ &= 1\left\{0 < b\left(M\frac{\alpha}{2}\right) \text{ or } b\left(M\left(1-\frac{\alpha}{2}\right)\right) < 0\right\} \end{aligned}$$

where $b(\cdot)$ is the ordered b so that $b(1) \leq b(2) \leq \cdots \leq b(p)$. For the simplicity of the argument we assume the multiplicity in the coefficients is taken care of as in Gelman et al. (2012). The use of CI to choose coefficients can also be justified from decision theoretic framework in Thulin (2012).

An equivalent procedure can be derived using a similar argument used in Gelman et al. (2012) by converting the question from "is the posterior probability that the coefficient is 0 less than α?", to "is the posterior distribution bigger or smaller than a point mass at 0?". If the $1-\alpha\%$ of the posterior distribution

of regression coefficients are bigger/smaller than 0, we claim with $1-\alpha$ level confidence that the the regression coefficient is significant. Specifically, using a statistic of the form

$$p_j^B = \frac{1}{M}\sum_i^M 1_{b_i>0}$$

if $p_j^B < \alpha$ or $p_j^B > 1-\alpha$ we claim the coefficient to be significantly smaller or larger than 0 at $1-\alpha$ level confidence. The decision rule take the form

$$\begin{aligned} d_j &= 1\{p^B < \alpha \text{ or } 1-\alpha < p^B\} \\ &= 1\left\{\sum_i^M 1_{v_i>0} < M\alpha \text{ or } M(1-\alpha) < \sum_i^M 1_{v_i>0}\right\} \\ &= 1\left\{0 < b(M\alpha) \text{ or } b(M(1-\alpha)) < 0\right\} \end{aligned}$$

You may identify it as the one sided equivalence of $1-\alpha\%$ level CI procedure in the frequentist literatures. As with the one sided p-value, the (pvalue) has more power at α level then the $1-\alpha\%$ level CI since we can use the fact that we are only interested in the case where the posterior is larger or smaller than 0.

Choosing regression coefficient based on $1-\alpha$ level CI has good frequentist properties for regression, especially under the situation where the predictors are uncorrelated (illustrated on table 4.1 taken from Celeux et al. (2012)). It is also surprisingly robust agains various anomalies in the data.

4.3.2 Explicitly controlling for FDR using continuous priors

Controlling for False Discovery Rate in the reporting process is popular in some fields of study. The Bayesian equivalent of such procedure is discussed in Müller

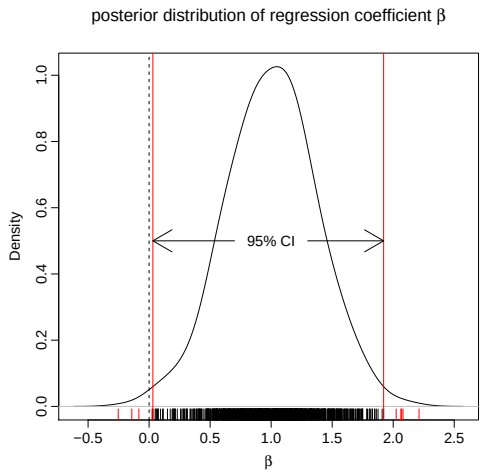


Figure 4.1: Illustration of 95% posterior credible interval.

et al. (2006) along with a couple of suggestions for possible loss function. From a Bayesian perspective, the FDR procedure is no more than a decision problem based on one's utility. For $\mathfrak{m}$ decisions to be made, let truth indicator $r_i = 1$ when a coefficient should be flagged and 0 otherwise ($i = 1, \ldots, \mathfrak{m}$). We also define an accompanying decision indicator d_i to be 1 when a coefficient i is flagged and 0 otherwise . The loss functions proposed in (Müller et al., 2006) also require the definition of probability that $r_i = 1$ such that $v_i = P(r_i = 1 \mid Y)$. When using discrete prior distribution, v_i is simply the posterior inclusion probability. Under an absolutely continuous prior, we can substitute the (pvalue) for v_i to obtain an optimal decision criteria. Figure 4.2 shows the actual FDR for loss function

$$L_{2R} = (\overline{\text{FDR}}, \overline{\text{FNR}})$$

(Müller et al., 2006) used on 4 examples from Celeux et al. (2012). The actual FDR is controlled at the level of expected FDR.

Table 4.1: Comparison with the example 1 from Celeux et al. (2012), simple case of 10 uncorrected predictors with 4 true effects with sample size of 200. The Bayesian regression method performs as well as the oracle under the prior distribution for the regression coefficient chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	HITS	FP
ORACLE	4:00(0:00)	0:00(0:00)
99% CI	4:00(0:00)	0:00(0:00)
95% CI	4:00(0:00)	0:00(0:00)
90% CI	4:00(0:00)	0:00(0:00)
AIC	3:94(0:02)	2:78(0:17)
BIC	3:90(0:03)	2:29(0:17)
BRIC	3:75(0:05)	0:65(0:09)
EB-L	3:80(0:04)	0:66(0:09)
EB-G	3:78(0:04)	0:65(0:09)
ZS-N	3:78(0:04)	0:65(0:09)
ZS-F	3:90(0:03)	1:73(0:14)
OVS	3:63(0:06)	0:54(0:09)
HG-3	3:75(0:05)	0:55(0:09)
HG-4	3:65(0:05)	0:54(0:08)
HG-2	3:75(0:05)	0:59(0:09)
NIMS	3:75(0:05)	0:57(0:08)
LASSO	3:89(0:03)	2:68(0:20)
DZ	3:72(0:07)	2:41(0:15)
ENET	3:89(0:04)	2:79(0:29)

Table 4.2: Comparison with the example 2 (sparse correlated design) from Celeux et al. (2012),10 correlated ($\rho = 0.9$) predictors with 4 true effects with sample size of 200. The Bayesian regression method performs well relative to the other methods in terms of low false discovery. The prior distribution for the regression coefficient was chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	HITS	FP
ORACLE	4.00(0.00)	0.00(0.00)
99% CI	2.48(0.72)	0.00(0.00)
95% CI	3.57(0.56)	0.00(0.00)
90% CI	3.91(0.29)	0.01(0.00)
AIC	3.12(0.08)	2.75(0.16)
BIC	2.97(0.09)	2.39(0.16)
BRIC	2.44(0.10)	0.99(0.10)
EB-L	2.43(0.10)	1.03(0.10)
EB-G	2.42(0.10)	0.95(0.10)
ZS-N	2.43(0.10)	1.03(0.10)
ZS-F	2.97(0.08)	2.18(0.10)
OVS	2.16(0.11)	1.09(0.09)
HG-3	2.32(0.11)	0.96(0.10)
HG-4	2.35(0.10)	0.86(0.09)
HG-2	2.35(0.10)	0.81(0.09)
NIMS	2.42(0.10)	0.96(0.09)
LASSO	3.35(0.09)	2.95(0.15)
DZ	2.83(0.09)	2.23(0.10)
ENET	3.70(0.07)	4.36(0.17)

Table 4.3: Comparison with the example 3 (sparse noisy correlated design) from Celeux et al. (2012). The prior distribution for the regression coefficient was chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	HITS	FP
ORACLE	3.00(0.00)	0.00(0.00)
99% CI	2.79(0.41)	0.00(0.00)
95% CI	2.97(0.17)	0.00(0.00)
90% CI	2.99(0.10)	0.04(0.20)
AIC	2.11(0.07)	2.06(0.14)
BIC	1.97(0.07)	1.68(0.14)
BRIC	1.66(0.07)	0.53(0.08)
EB-L	1.84(0.07)	0.79(0.09)
EB-G	1.88(0.07)	0.83(0.09)
ZS-N	1.81(0.07)	0.76(0.09)
ZS-F	2.10(0.07)	1.26(0.11)
OVS	1.78(0.07)	0.64(0.09)
HG-3	1.81(0.07)	0.77(0.09)
HG-4	1.84(0.07)	0.78(0.09)
HG-2	1.80(0.08)	0.73(0.10)
NIMS	1.83(0.07)	0.77(0.09)
LASSO	2.33(0.07)	1.61(0.16)
DZ	2.20(0.11)	2.06(0.16)
ENET	2.38(0.06)	2.04(0.16)

Table 4.4: Comparison with the example 4 (saturated correlated design) from Celeux et al. (2012). The prior distribution for the regression coefficient was chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	HITS	FP
ORACLE	8.00(0.00)	0.00(0.00)
99% CI	1.40(0.89)	0.00(0.00)
95% CI	3.97(1.15)	0.00(0.00)
90% CI	5.59(1.06)	0.00(0.00)
AIC	6.32(0.11)	0.00(0.00)
BIC	5.99(0.12)	0.00(0.00)
BRIC	4.35(0.11)	0.00(0.00)
EB-L	4.39(0.10)	0.00(0.00)
EB-G	4.34(0.10)	0.00(0.00)
ZS-N	4.38(0.10)	0.00(0.00)
ZS-F	5.37(0.10)	0.00(0.00)
OVS	3.82(0.10)	0.00(0.00)
HG-3	4.32(0.10)	0.00(0.00)
HG-4	4.19(0.09)	0.00(0.00)
HG-2	4.18(0.11)	0.00(0.00)
NIMS	4.39(0.10)	0.00(0.00)
LASSO	7.13(0.12)	0.00(0.00)
DZ	6.82(0.11)	0.00(0.00)
ENET	7.53(0.08)	0.00(0.00)

Table 4.5: Comparison with the example 5 from Celeux et al. (2012). The prior distribution for the regression coefficient was chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	HITS	FP
ORACLE	2.00(0.00)	0.00(0.00)
99% CI	2.00(0.00)	0.00(0.00)
95% CI	2.00(0.00)	0.00(0.00)
90% CI	2.00(0.00)	0.03(0.17)
AIC	1.93(0.02)	2.88(0.19)
BIC	1.94(0.02)	2.04(0.18)
BRIC	1.93(0.02)	0.50(0.09)
EB-L	1.93(0.02)	0.58(0.10)
EB-G	1.93(0.02)	0.60(0.10)
ZS-N	1.93(0.02)	0.57(0.10)
ZS-F	1.94(0.02)	1.84(0.14)
OVS	1.89(0.03)	0.76(0.08)
HG-3	1.93(0.02)	0.53(0.09)
HG-4	1.93(0.02)	0.54(0.09)
HG-2	1.93(0.02)	0.36(0.09)
NIMS	1.93(0.02)	0.57(0.10)
LASSO	1.99(0.01)	2.93(0.21)
DZ	1.91(0.03)	2.70(0.18)
ENET	1.96(0.02)	3.25(0.20)

Table 4.6: Comparison with the example 6 (null model) from Celeux et al. (2012). The prior distribution for the regression coefficient was chosen to be normal with mean 0 and Zellner's G prior as the covariance matrix.

	FP
ORACLE	0.00(0.00)
99ci	0.09(0.40)
95ci	0.30(0.61)
90ci	0.70(0.89)
AIC	3.16(0.21)
BIC	2.24(0.19)
BRIC	0.59(0.11)
EB-L	2.87(0.15)
EB-G	1.54(0.19)
ZS-N	1.02(0.17)
ZS-F	2.51(0.17)
OVS	2.10(0.17)
HG-3	2.18(0.18)
HG-4	2.54(0.17)
HG-2	2.17(0.15)
NIMS	0.99(0.13)
LASSO	1.79(0.22)
DZ	2.49(0.20)
ENET	2.23(0.23)

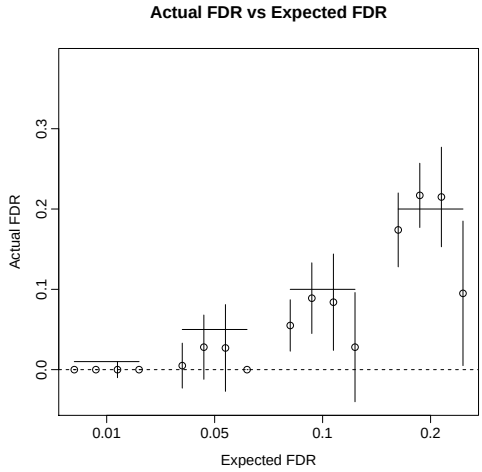


Figure 4.2: The expected and actual FDR for 4 simulations from Celeux et al. (2012) with 4 different value of expected FDR values.

CHAPTER 5

APPENDIX

5.1 Appendix 1: Technical Supplements

5.1.1 Markov Chain Monte Carlo

We review some elaborate techniques for Markov Chain Monte Carlo simulation used in this thesis. For an overview of the Markov Chain Monte Carlo techniques Brooks et al. (2011) provides a great coverage of the material.

5.1.2 Reversible Jump MCMC

Reversible jumps Markov Chain Monte Carlo (RJMCMC) (Green, 1995) or Metropolis-Hastings-Green with Jacobians (MHGJ) is a special case of Metropolis-Hastings-Green elementary update algorithm (Geyer, 2010b) that allows moves between parts of the state space that are Euclidean spaces of different dimension.

We assume the underlying states pace is disjoint union of Euclidian spaces S_m for $m = 1, \dots M$, each with dimension d_m. Let U_m and U_n be Euclidian spaces such that $S_m \times U_m$ is the same dimension as $S_n \times U_n$. The basic idea of the elementary update algorithm moves between the spaces $S_m \times U_m$ to $S_n \times U_n$ that have the same dimension. For a simple example, when proposing a move from y to y' in higher dimensional space, one would draw a vector of continuous random variable u, independent of y so that dimension of (u, y) matches that of y'. Then

a proposal is made by setting $y' = y'(y, u)$ with $y'(.,.)$ an invertible deterministic function. The move from y to y' is accepted with a probability

$$\alpha_m(y, y') \quad = \quad min\left\{1, \frac{\pi(y')}{\pi(y)} \times \frac{r_m(y')}{r_m(y)q(u)} \times \left|\frac{\partial y'}{\partial(y, u)}\right|\right\}$$

- where $\pi(y)$ is the target density evaluated at y,
- $r_m(y)$ is the probability of a move of type m, evaluated at y, and
- $q(u)$ is the density function of u.

For the specifics of the algorithm see Green (1995); Geyer (2010b).

5.1.3 Over-relaxation algorithm

Method of over-relaxation is a technique to improve the proposal of MCMC originally proposed by Adler (1981) for Gaussian case and generalized by Neal (1995) as ordered over-relaxation method, applicable in place of Gibbs sampling. It is particularly useful when there is a strong correlation between the parameters that makes the Gibbs sampling inefficient. When y_i follows $N(\mu, \sigma^2)$ and if current value of y_i in the Markov Chain is $y_i^{(t)}$ Adler's method proposes $y_i^{(t+1)}$ to be

$$y_i^{(t+1)} = \mu + \alpha(y_i^{(t)} - \mu) + (1 - \alpha^2)^{1/2}\sigma\nu$$

where $\nu \sim N(0, 1)$ and $\alpha \in [-1, 1]$ is a tuning parameter. When $\alpha < 0$ method is called over-relaxation and $\alpha > 0$ it is called under-relaxation. Intuitive illustration of why over-relaxation proposal is efficient is displayed in figure 5.1. When strong correlation is present, both MH and Gibbs Sampling suffer from the locality of proposal yet this is not a problem for over-relaxation proposal.

In ordered over-relaxation of Neal (1995), K samples from $p(y_i|y_{-i})$ is gen-

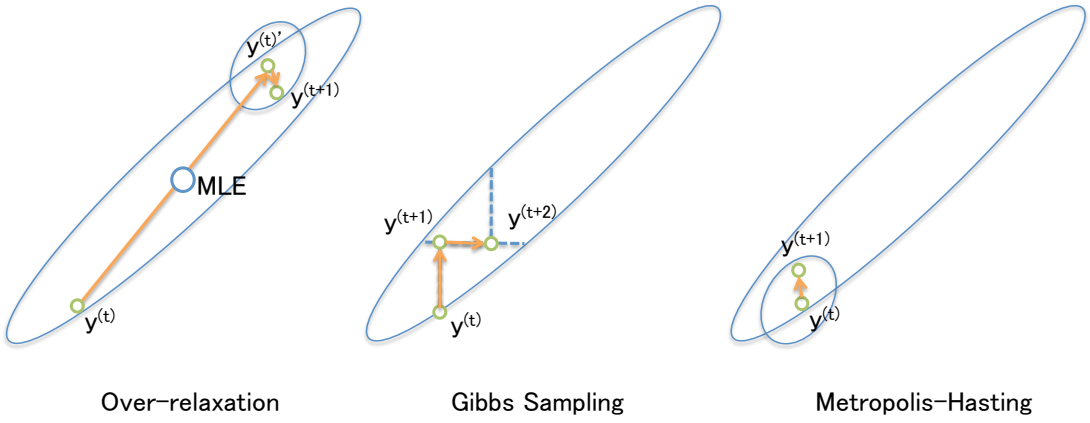


Figure 5.1: Illustrative example of the different MCMC strategies.

erated, which are ranked along with the current value y_i. If we let the rank of current value to be k, the proposal is chosen amongst the $K+1$ samples, who's ranking is $K-k$. Note that when $K=1$ this is simply the Gibbs sampler.

5.1.3.1 Parallel tempering

Parallel tempering (Geyer, 1991) is a population Monte Carlo technique where the target distribution is augmented with an indicator that specify a level of smoothing applied to each of the target distribution. The new joint distribution is the product of each of the distribution over the indicators since each of the density is independent of each other given the indicator. Markov Chains at different temperatures are run in parallel and the neighboring states are exchanged between the chains with a predefined rate. Given initial sets of parameters $\theta^{(0)}$, with T replications of $\theta^{(0)}$ so that new parameter space is $\boldsymbol{\theta} = \{\theta_1^{(0)}, \ldots, \theta_T^{(0)}\}$. Each set of parameter $\theta_i^{(s)}$ is updated according to an update regime define as following For $i = 1, \ldots T$,

- With probability p, update $\theta_i^{(s+1)}$ based on a Monte Carlo update scheme

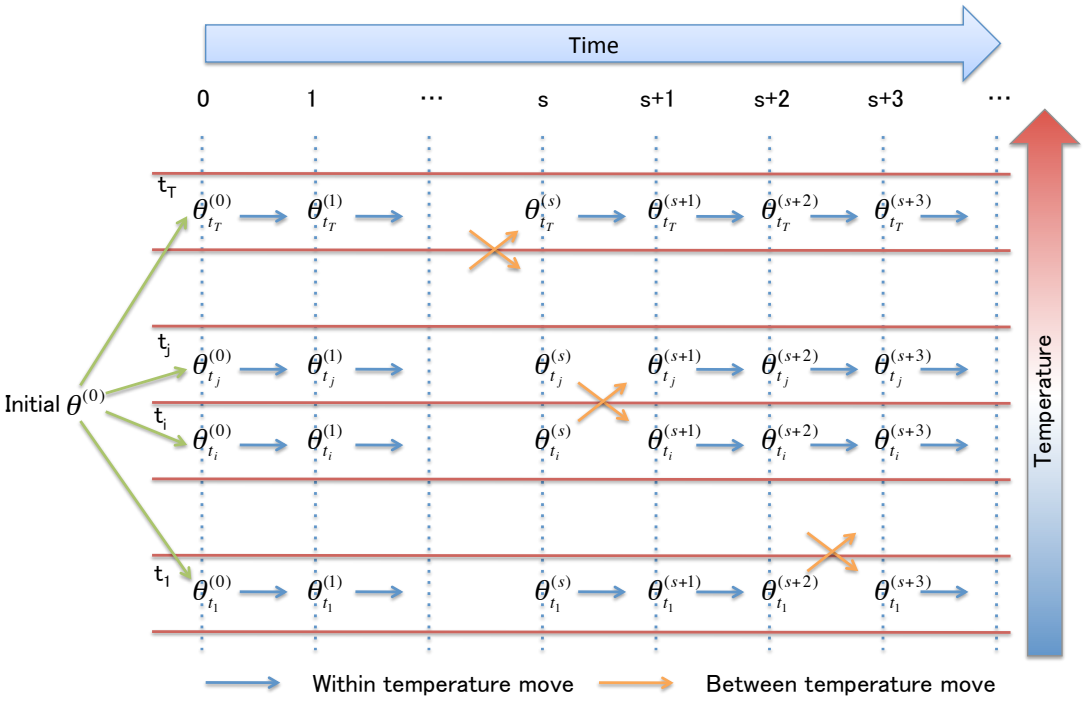


Figure 5.2: Illustration of parallel tempering algorithm.

$f(\theta_i^{(s)})$.

- With probability $1-p$, propose swap move between temperatures i_1 and i_2 by setting $\theta_{i_1}^{(s+1)} = \theta_{i_2}^{(s)}$ and $\theta_{i_2}^{(s+1)} = \theta_{i_1}^{(s)}$ with probability

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') \quad = \quad min\left\{1, \frac{\pi_{i_1}(\theta_{i_2})\pi_{i_2}(\theta_{i_1})}{\pi_{i_1}(\theta_{i_1})\pi_{i_2}(\theta_{i_2})}\right\} \tag{5.1}$$

where $\pi_t(\cdot)$ defines a smooth version of the target density indexed by t. For details about the smoothing scheme as well as other details about the parallel tempering see Geyer (2010a).

5.1.4 Statistical distributions

We list a few of the standard distribution in this section.

5.1.4.1 Multivariate Gaussian Distribution

When p variate random variable $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ follows Gaussian distribution, with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and covariance $\Sigma = [\sigma_{ij}]_{p\times p}$, we denote it as $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$. The pdf of $\boldsymbol{Y}$ is defined as

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\mu}, \Sigma\right) \quad =_d \quad (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\boldsymbol{Y}-\boldsymbol{\mu}\right]^T \Sigma^{-1}\left[\boldsymbol{Y}-\boldsymbol{\mu}\right]\right\} \tag{5.2}$$

We can express the same relationship using the inverse covariance matrix or sometimes called the precision matrix $\Omega = \Sigma^{-1} = [\omega_{ij}]_{p\times p}$

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\mu}, \Omega^{-1}\right) \quad =_d \quad (2\pi)^{-\frac{p}{2}} |\Omega^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\boldsymbol{Y}-\boldsymbol{\mu}\right]^T \Omega\left[\boldsymbol{Y}-\boldsymbol{\mu}\right]\right\} \tag{5.3}$$

5.1.4.2 Matrix variate Gaussian distribution

$n \times p$ random matrix $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T = [y_{ij}]_{n\times p}$ is said to follow matrix variate Gaussian distribution, with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_n^T)^T$, row covariance $\Phi = [\phi_{ij}]_{n\times n}$, and column covariance $\Sigma = [\sigma_{ij}]_{p\times p}$, we denote it as $\boldsymbol{Y} - \boldsymbol{\mu} \sim \mathcal{MN}_{n\times p}(\Phi, \Sigma)$. The pdf of $\boldsymbol{Y}$ is defined as

$$p\left(\boldsymbol{Y}-\boldsymbol{\mu} \mid \Phi, \Sigma\right) \quad =_d \quad (2\pi)^{-\frac{np}{2}} |\Phi|^{-\frac{p}{2}} |\Sigma|^{-\frac{n}{2}} \operatorname{etr}\left\{-\frac{1}{2}\Phi^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)\Sigma^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)^{\mathrm{T}}\right\} \tag{5.4}$$

5.1.4.3 Inverse Gaussian distribution

A continuous random variable $y > 0$ following inverse Gaussian distribution with mean $\mu > 0$ and shape parameter $\lambda > 0$ is denoted as $y \sim Inv\text{-}N(\mu, \lambda)$. The pdf of y is defined as

$$p(y \mid \mu, \lambda) \quad =_d \quad \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right\}. \tag{5.5}$$

$E(y) = \mu$ and $var(y) = \mu^3/\lambda$.

5.1.4.4 Matrix Student T distribution

A random $n \times p$ matrix Y distributed according to a central Matrix Student T distribution with parameters υ,Φ, and Σ is denoted as $Y \sim \mathcal{MT}(\upsilon, \Phi, \Sigma)$. The pdf of Y is defined as

$$p(y \mid \Sigma, \upsilon) \propto \left|\Phi + \frac{1}{\upsilon} Y \Sigma^{-1} Y^T\right|^{-\frac{\upsilon+p}{2}}. \tag{5.6}$$

5.1.4.5 Wishart Distribution

When $k \times k$ matrix follow a Wishart distribution with degrees of freedom $\nu > 0$ and symmetric positive definite $k \times k$ matrix S, we denote it as $W \sim Wishart_\nu(S)$.

$$p\left(W \mid \nu, S\right) \quad =_d \quad \left(2^{\frac{\nu k}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^{k} \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right)^{-1} \tag{5.7}$$

$$\times \quad |S|^{-\frac{\nu}{2}} |W|^{\frac{(\nu-k-1)}{2}} \mathrm{etr}\left(-\frac{1}{2} S^{-1} W\right) \tag{5.8}$$

$E(W) = \nu S$

5.1.4.6 Inverse Wishart Distribution

When $k \times k$ matrix follow a inverse Wishart distribution with degrees of freedom $\nu > 0$ and symmetric positive definite $k \times k$ matrix S, we denote it as $W \sim Inv\text{-}Wishart(\nu, S)$.

$$p\left(W \mid \nu, S\right) =_d \left(2^{\frac{\nu k}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^{k} \Gamma\left(\frac{\nu + 1 - i}{2}\right)\right)^{-1} \tag{5.9}$$

$$\times \quad |S|^{\frac{\nu}{2}} |W|^{\frac{-(\nu+k+1)}{2}} \text{etr}\left(-\frac{1}{2} S W^{-1}\right) \tag{5.10}$$

$E(W) = (\nu - k - 1)^{-1} S$

5.1.4.7 Hyper Inverse Wishart Distribution Dawid and Lauritzen (1993)

When psd matrix Σ follows a Hyper Inverse Wishart Distribution with parameters α and Φ, it is denoted as $\Sigma \sim HIW_g(\alpha, \Phi)$.

$$p(\Sigma \mid \mathcal{G}, \alpha, \Phi) =_d \frac{\prod_{C \in \mathscr{C}} Inv\text{-}Wishart_C\left(\Sigma_C | \alpha, \Phi\right)}{\prod_{S \in \mathscr{S}} Inv\text{-}Wishart_S\left(\Sigma_S | \alpha, \Phi\right)}$$

Where $\mathscr{C}$ and $\mathscr{S}$ denote the set of cliques and separators and $Inv\text{-}Wishart(\Sigma|\alpha, \Phi)$ is a density function for inverse Wishart distribution.

5.1.4.8 Exponential Distribution

A positive real variable $\theta > 0$ with rate parameter $\lambda > 0$ following Exponential distribution is denoted as $\theta \sim Exp(\lambda)$

$$p(\theta \mid \lambda) =_d \lambda \exp(-\lambda\theta)$$

$E(\theta) = \frac{1}{\lambda}$ and $var(\theta) = \frac{1}{\lambda^2}$

5.1.4.9 Gamma Distribution

A positive real variable $\theta > 0$ with shape $\alpha > 0$ and inverse scale $\beta > 0$ following Gamma distribution is denoted as $\theta \sim Ga(\alpha, \beta)$

$$p(\theta \mid \alpha, \beta) =_d \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

$E(\theta) = \frac{\alpha}{\beta}$ and $var(\theta) = \frac{\alpha}{\beta^2}$

5.1.4.10 Inverse Gamma Distribution

A positive real variable $\theta > 0$ with shape $\alpha > 0$ and scale $\beta > 0$ following Inverse Gamma distribution is denoted as $\theta \sim Inv\text{-}Ga(\alpha, \beta)$

$$p(\theta \mid \alpha, \beta) =_d \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-\alpha-1}e^{-\beta/\theta}$$

$E(\theta) = \frac{\alpha}{\beta}$ and $var(\theta) = \frac{\alpha}{\beta^2}$

5.1.4.11 Beta Distribution

A positive real variable $\theta \in [0,1]$ with parameters $\alpha > 0$ and $\beta > 0$ following beta distribution is denoted as $\theta \sim Beta(\alpha, \beta)$

$$p(\theta \mid \alpha, \beta) =_d \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$E(\theta) = \frac{\alpha}{\alpha+\beta}$ and $var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

5.1.4.12 Binomial Distribution

A positive real variable $\theta \in [0,n]$ with sample size $n \in \mathbb{N} \setminus \{0\}$ and success probability $p \in [0,1]$ following binomial distribution is denoted as $\theta \sim Bin(n,p)$,

$$p(\theta \mid n, p) =_d \binom{n}{\theta} p^{\theta}(1-p)^{n-\theta}$$

$E(\theta) = np$ and $var(\theta) = np(1-p)$

5.1.4.13 Sweep operation for solution to linear regression with linear equality constraint

Sweep operation can be used to give solution to linear regression with linear equality constraint problem. Starting with augmented sums of square cross product matrix, we fist sweep on the upper left corner $\tilde{X}_g^T\tilde{X}_g$ which gives

$$\begin{pmatrix} \tilde{X}_g^T\tilde{X}_g & \tilde{X}_g^T y_g & L_g^T \\ y_g^T\tilde{X}_g & y_g^T y_g & 0^T \\ L_g & 0 & 0 \end{pmatrix} \overset{sweep}{\rightarrow} \begin{pmatrix} (\tilde{X}_g^T\tilde{X}_g)^{-1} & (\tilde{X}_g^T\tilde{X}_g)^{-1}\tilde{X}_g^T y_g & (\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T \\ -y_g^T\tilde{X}_g(\tilde{X}_g^T\tilde{X}_g)^{-1} & y_g^T y_g - \tilde{X}_g^T y_g(\tilde{X}_g^T\tilde{X}_g)^{-1}y_g^T\tilde{X}_g & -y_g^T\tilde{X}_g(\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T \\ -L_g(\tilde{X}_g^T\tilde{X}_g)^{-1} & -L_g(\tilde{X}_g^T\tilde{X}_g)^{-1}y_g^T\tilde{X}_g & -L_g(\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T \end{pmatrix}$$

We can rewrite this matrix using $\hat{B}_g^M$ as the unconstrained MLE of B_g and RSS as the residual sums of square. Then sweeping this matrix on bottom right corner

$-L_g(\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T$ and denoting $Q = (L_g(\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T)^{-1}$ which gives us

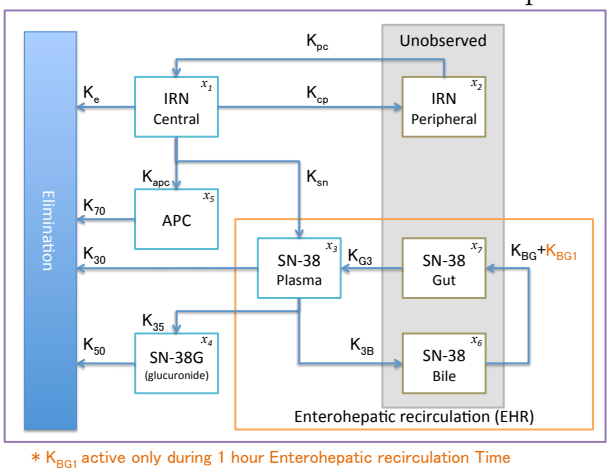
$$\begin{pmatrix} (\tilde{X}_g^T\tilde{X}_g)^{-1} & \hat{B}_g^M & (\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T \\ \hat{B}_g^M & RSS & -\hat{B}_g^M L_g^T \\ -L_g(\tilde{X}_g^T\tilde{X}_g)^{-1} & -L_g\hat{B}_g^M & -L(\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T \end{pmatrix} \overset{sweep}{\rightarrow}$$

$$\begin{pmatrix} (\tilde{X}_g^T\tilde{X}_g)^{-1} - (\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T Q L_g(\tilde{X}_g^T\tilde{X}_g)^{-1} & \hat{B}_g^M - (\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T Q L_g\hat{B}_g^M & (\tilde{X}_g^T\tilde{X}_g)^{-1}L_g^T Q \\ -\hat{B}_g^M + (\tilde{X}_g^T\tilde{X}_{-g})^{-1}L_g^T Q L_g\hat{B}_g^M & RSS + (\hat{B}_g^M)^T L_g^T Q L_g\hat{B}_g^M & -(\hat{B}_g^M)^T L_g^T Q \\ Q L_g(\tilde{X}_{-g}^T\tilde{X}_g)^{-1} & -Q L_g\hat{B}_g^M & -Q \end{pmatrix}$$

5.2 Appendix 2: Chapter 2 Supplements

5.2.1 Compartment model

Figure 5.3: Pharmacokinetic model with enterohepatic recirculation



The Compartment Model follows the following system of Ordinary Differential Equations:

$$
\begin{aligned}
x_1'(t) &= k_{pc}x_2 - (k_{cp} + k_{eapc} + k_{esn} + k_e)x_1 + R_1, \\
x_2'(t) &= k_{cp}x_1 - k_{pc}x_2, \\
x_3'(t) &= k_{esn}x_1 + k_{G3}x_7 - (k_{30} + k_{35} + k_{3B})x_3, \\
x_4'(t) &= k_{35}x_3 - k_{50}x_4, \\
x_5'(t) &= k_{eapc}x_1 - k_{70}x_5, \\
x_6'(t) &= \begin{cases} k_{3B}x_3 - (k_{BG} + k_{BG1})x_6 & \text{if} \quad \text{EHRT} \leq t \leq (\text{EHRT} + 1) \\ k_{3B}x_3 - k_{BG}x_6 & \text{o.w.} \end{cases} \\
x_7'(t) &= \begin{cases} (k_{BG} + k_{BG1})x_6 - k_{G3}x_7 & \text{if} \quad \text{EHRT} \leq t \leq (\text{EHRT} + 1) \\ k_{BG}x_6 - k_{G3}x_7 & \text{o.w.} \end{cases}
\end{aligned}
\tag{5.11}
$$

Note that R_1 is the injected dose and EHRT is the time when enterohepatic recirculation starts.

5.2.2 Computational details

5.2.2.1 Computation of PK parameter $\boldsymbol{\theta}$

To update $\boldsymbol{\theta}$ we will use Metropolis Hastings algorithm where a candidate values to update $\boldsymbol{\theta}$ are proposed from a chosen proposal distribution and move is accepted with certain probability. One may choose any proposal strategy in theory, however, once you factor in the computational toll associated with evaluation of the likelihood, a V variate proposal for $\boldsymbol{\theta}_i$ for $i = 1, \cdots, N$ is most cost efficient. The simplest of such choice is a multivariate Gaussian proposal centered at the current values of $\boldsymbol{\theta}_i$. The correlation amongst the $\boldsymbol{\theta}_i$ that might hinder the efficiency of the sampling can be dealt with using an adaptive Metropolis algorithm propose by Haario et al. (2001) where given current value of $\boldsymbol{\theta}_i$ at time ℓ, denoted as $\boldsymbol{\theta}_i^{(\ell)}$, proposal $\boldsymbol{\theta}_i^{(\ell')}$ is made from

$$\boldsymbol{\theta}_i^{(\ell')} \mid \boldsymbol{\theta}_i^{(\ell)}, \cdots, \boldsymbol{\theta}_i^{(0)} \sim N_V\left(\boldsymbol{\theta}_i^{(\ell)}, \frac{2.38^2}{V}\Sigma_i^{(\ell+1)}\right)$$

where $\Sigma_i^{(\ell)}$ is defined as $\frac{2.38^2}{V}\mathbf{C}_i^{(\ell)} + \frac{2.38^2\epsilon}{V}I_V$ where $\mathbf{C}_i^{(\ell)}$ is the empirical covariance matrix of $\left(\boldsymbol{\theta}_i^{(0)}, \cdots, \boldsymbol{\theta}_i^{(\ell)}\right)$ s.t.

$$\mathbf{C}_i^{(\ell)} = \frac{1}{\ell}\left(\sum_{j=0}^{\ell} \boldsymbol{\theta}_i^{(j)}\left(\boldsymbol{\theta}_i^{(j)}\right)^T - (\ell+1)\bar{\boldsymbol{\theta}}_i^{(\ell)}\left(\bar{\boldsymbol{\theta}}_i^{(\ell)}\right)^T\right)$$

where $\bar{\boldsymbol{\theta}}_i^{(\ell)}$ is the empirical mean defined as $\bar{\boldsymbol{\theta}}_i^{(\ell)} = \frac{1}{\ell+1}\sum_{j=0}^{\ell}\boldsymbol{\theta}_i^{(j)}$.

For the initial ℓ_0 steps, arbitrary initial covariance matrix Σ_0 should be used until enough samples have been gathered. With each iteration $\Sigma_i^{(\ell+1)}$ is updated

as

$$\Sigma_i^{(\ell+1)} = \frac{\ell-1}{\ell}\Sigma_i^{(\ell)} + \frac{2.38^2}{V}\frac{1}{\ell}\left((\ell)\bar{\boldsymbol{\theta}}_i^{(\ell-1)}\left(\bar{\boldsymbol{\theta}}_i^{(\ell-1)}\right)^T + (\ell+1)\bar{\boldsymbol{\theta}}_i^{(\ell)}\left(\bar{\boldsymbol{\theta}}_i^{(\ell)}\right)^T + \boldsymbol{\theta}_i^{(\ell)}\left(\boldsymbol{\theta}_i^{(\ell)}\right)^T + \epsilon I_V\right)$$

Given the proposed $\boldsymbol{\theta}_i^{(\ell')}$ the move is accepted with probability $min(1, A_t)$ where the acceptance probability A_t is calculated as

$$A_t = \frac{p(\boldsymbol{\theta}_i^{(\ell')} \mid Y_i, \tilde{Z}_i)p(\boldsymbol{\theta}_i^{(\ell')})q(\boldsymbol{\theta}_i^{(\ell')} \to \boldsymbol{\theta}_i^{(\ell)})}{p(\boldsymbol{\theta}_i^{(\ell)} \mid Y_i, \tilde{Z}_i)p(\boldsymbol{\theta}_i^{(\ell)})q(\boldsymbol{\theta}_i^{(\ell)} \to \boldsymbol{\theta}_i^{(\ell')})}$$

If accepted $\boldsymbol{\theta}_i^{(\ell+1)}$ is set to $\boldsymbol{\theta}_i^{(\ell')}$ otherwise it is kept as $\boldsymbol{\theta}_i^{(\ell)}$.

Algorithm

- Initialization
  - $\bar{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}$
  - $\Sigma_i^{(0)} = \mathbf{0}_{V\times V}$ for $i = 1, \cdots, N$
- For each $i \in \{1, \cdots, N\}$
  1. update $\bar{\boldsymbol{\theta}}^{(\ell)}$

$$\bar{\boldsymbol{\theta}}_i^{(\ell)} = \frac{\ell\bar{\boldsymbol{\theta}}_i^{(\ell-1)} + \boldsymbol{\theta}_i^{(\ell)}}{\ell+1}$$

  2. update $\Sigma_i^{(\ell+1)}$

$$\Sigma_i^{(\ell+1)} = \frac{\ell-1}{\ell}\Sigma_i^{(\ell)} + \frac{2.38^2}{V}\frac{1}{\ell}\left((\ell)\bar{\boldsymbol{\theta}}_i^{(\ell-1)}\left(\bar{\boldsymbol{\theta}}_i^{(\ell-1)}\right)^T + (\ell+1)\bar{\boldsymbol{\theta}}_i^{(\ell)}\left(\bar{\boldsymbol{\theta}}_i^{(\ell)}\right)^T + \boldsymbol{\theta}_i^{(\ell)}\left(\boldsymbol{\theta}_i^{(\ell)}\right)^T + \epsilon I_V\right)$$

3. propose $\boldsymbol{\theta}_i^{(\ell')}$

$$\boldsymbol{\theta}_i^{(\ell')} \sim N_V\left(\boldsymbol{\theta}_i^{(\ell)}, \frac{2.38^2}{V}\Sigma_i^{(\ell+1)}\right)$$

4. calculate the acceptance probability

$$A_t = \frac{p(\boldsymbol{\theta}_i^{(\ell')} \mid Y_i, \tilde{Z}_i)p(\boldsymbol{\theta}_i^{(\ell')})q(\boldsymbol{\theta}_i^{(\ell')} \to \boldsymbol{\theta}_i^{(\ell)})}{p(\boldsymbol{\theta}_i^{(\ell)} \mid Y_i, \tilde{Z}_i)p(\boldsymbol{\theta}_i^{(\ell)})q(\boldsymbol{\theta}_i^{(\ell)} \to \boldsymbol{\theta}_i^{(\ell')})}$$

5. decide to reject or accept Draw a random number u from uniform $(0, 1)$,

$$\boldsymbol{\theta}_i^{(\ell+1)} \Leftarrow \begin{cases} \boldsymbol{\theta}_i^{(\ell')} & \text{if } u < min(1, A_t) \\ \boldsymbol{\theta}_i^{(\ell)} & \text{o.w.} \end{cases}$$

5.2.3 Computation of latent probit score $\boldsymbol{Z}$

The full conditional distribution of $\boldsymbol{Z}_i$ is a truncated Q-variate Gaussian density $N_Q(\tilde{m}_{z_i}, \tilde{S}_{z_i}^{-1})I\{\boldsymbol{Z}_i \in \mathcal{A}_i\}$ where truncation $\mathcal{A}_i$ given $\tilde{Z}_i$ is defined as

$$\mathcal{A}_{iq} = \begin{cases} (-\infty, 0) & \tilde{Z}_{iq} = 0 \\ [0, 1) & \tilde{Z}_{iq} = 1 \\ [1, \infty) & \tilde{Z}_{iq} = 2 \end{cases}$$

for each $i = 1, \cdots, N$, and $q = 1, \cdots, Q$. Parameters $\tilde{m}_{z_i}$ and $\tilde{S}^{-1}_{z_i}$ are the full conditional mean and variance defined as

$$\begin{aligned} \tilde{S}_{z_i} &= \left(\Omega_Z + \tau_i \boldsymbol{\rho}^T \Omega_\theta \boldsymbol{\rho}\right) \\ \tilde{m}_{z_i} &= \tilde{S}^{-1}_{z_i} \left[\Omega_Z \boldsymbol{\gamma} \boldsymbol{U}_i + \tau_i \boldsymbol{\rho}^T \Omega_\theta \left(\log(\boldsymbol{\theta}_i) - \boldsymbol{\beta} \boldsymbol{X}_i\right)\right] \end{aligned}$$

For a general multivariate Gaussian variable $\mathfrak{Z} \sim N(m, S^{-1})$, conditional distribution of $\mathfrak{Z}_a \mid \mathfrak{Z}_b$ follows normal distribution. Using the canonical parametrization this is represented as $\mathfrak{Z}_a \mid \mathfrak{Z}_b \sim Nc(S_{ab}(\mathfrak{Z}_b - m_b), S_{aa})$ (Rue and Held, 2005). Hence $\boldsymbol{Z}_{i,q} \mid \boldsymbol{Z}_{i,-q}$ for $q = 1, \cdots, Q$ can be sampled recursively using the algorithm 2.5 of Rue and Held (2005) by sampling from a truncated Gaussian distribution.

5.2.3.1 Computation of σ^2

For subject and PK parameter level variability parameter σ^2_{ik} we exploit the conjugacy and define independent inverse gamma prior over time as.

$$p(\sigma^2_{ik}) \sim \textit{Inv-Ga}\left(\frac{r_1}{2}, \frac{r_2}{2}\right) \text{ for } i = 1, \cdots, N \text{ and } k = 1, \cdots, K$$

If the temporal correlation that is not accounted for by Σ_θ becomes a problem, one can consider expanding this to an AR type prior. Given the prior, the full conditional for σ^2_{ik} is again inverse gamma distributed with parameters.

$$p(\sigma^2_{ik} \mid \cdot) \sim \textit{Inv-Ga}\left(\frac{t_n + r_1}{2}, \frac{1}{2}\left(\left[\log(y_{ik}(t)) - \log\left(g_{ik}(t, \boldsymbol{\theta}_{i.}, D_i)\right)\right]^T \left[\log(y_{ik}(t)) - \log\left(g_{ik}(t, \boldsymbol{\theta}_{i.}, D_i)\right)\right] + r_2\right)\right)$$

5.2.3.2 Computation of τ

For parameter τ_i to account for the subject level extra variability, we define a conjugate Gamma distribution defined as.

$$p(\tau_i) \sim Ga\left(\frac{u_1}{2}, \frac{u_2}{2}\right) \text{ for } i = 1, \cdots, N$$

Then the full conditional distribution is a Gamma distribution s.t.

$$p(\tau_i \mid \cdot) \sim Ga\left(\frac{V + u_1}{2}, \frac{1}{2}\left(u_2 + \left[\log(\boldsymbol{\theta}_i) - \boldsymbol{\beta}\boldsymbol{X}_i - \boldsymbol{\rho}\boldsymbol{Z}_i\right]^T \Omega_\theta \left[\log(\boldsymbol{\theta}_i) - \boldsymbol{\beta}\boldsymbol{X}_i - \boldsymbol{\rho}\boldsymbol{Z}_i\right]\right)\right)$$

5.2.3.3 Computation of $\boldsymbol{\beta}$

We define conjugate matrix normal distribution for the PK level regression coefficients $\boldsymbol{\beta}$ as.

$$\boldsymbol{\beta} - \boldsymbol{b}_0 \sim \mathcal{MN}\left(\boldsymbol{B}_0^{-1}, \Omega_\theta^{-1}\right)$$

The full conditional distribution is a matrix normally distributed s.t.

$$\begin{aligned}
\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} &\sim \mathcal{MN}\left(\boldsymbol{S}_b^{-1}, \Omega_\theta^{-1}\right) \\
\boldsymbol{S}_b &= \left(\boldsymbol{B}_0 + (\boldsymbol{X}^T \boldsymbol{D}_\tau \boldsymbol{X})\right) \\
\tilde{\boldsymbol{\beta}} &= \boldsymbol{S}_b^{-1}\left(\boldsymbol{B}_0\boldsymbol{b}_0 + (\boldsymbol{X}^T \boldsymbol{D}_\tau \boldsymbol{X})\hat{\boldsymbol{\beta}}\right) \\
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^T \boldsymbol{D}_\tau \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{D}_\tau \left(\log(\boldsymbol{\theta}) - \boldsymbol{Z}\boldsymbol{\rho}\right)
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau)$.

5.2.3.4 Computation of γ

For PG level regression coefficients we define conjugate Zellner's G-prior

$$\gamma - \gamma_0 \sim \mathcal{MN}\left(I_Q, c(U^TU)^{-1}\right)$$

where c is hyperparameter to be specified. The posterior is again matrix normal defined as

$$\begin{aligned}
\gamma - \tilde{\gamma} &\sim \mathcal{MN}\left(\left(\frac{1}{c}I_Q + \Omega_z\right)^{-1}, \left(U^TU\right)^{-1}\right) \\
\tilde{\gamma} &= \left(\frac{1}{c}I_Q + \Omega_z\right)^{-1}\left(\frac{1}{c}\gamma_0^T + \Omega_z\hat{\gamma}^T\right) \\
\hat{\gamma} &= \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{U}^TZ
\end{aligned}$$

5.2.3.5 Computation of ρ

We first rewrite the conditional distribution of $\boldsymbol{\theta}_i$, $i = 1, \ldots, N$ as

$$\begin{aligned}
log(\boldsymbol{\theta}_i) &\sim N_V\left(\boldsymbol{\beta X}_i + \boldsymbol{\rho Z}_i, (\tau_i\Omega_\theta)^{-1} = \Sigma_{\theta_i}\right) \\
&\Rightarrow (\Sigma_{\theta_i})^{-1/2}log(\boldsymbol{\theta}_i) - \boldsymbol{\beta X}_i \sim N_V\left(\boldsymbol{\rho Z}_i, I_V\right)
\end{aligned}$$

where $(\Sigma_{\theta_i})^{-1/2}$ is the Cholesky decomposition of the covariance matrix Σ_{θ_i}. Defining the prior distribution for $\boldsymbol{\rho}$ as (2.15), we can follow the algorithm suggested by Sun et al. (2010) to sample from the full conditional distribution of $\boldsymbol{\rho}$ by letting $y_i = (\Sigma_{\theta_i})^{-1/2}log(\boldsymbol{\theta}_i) - \boldsymbol{\beta X}_i$, $x_i = \boldsymbol{Z}_i$, and $\sigma^2 = 1$. We leave the details of the sampling algorithm to Sun et al. (2010).

5.3 Appendix 3: Chapter 3 Supplements

5.3.1 Acceptance probability for the birth and the death moves.

The RJMCMC on $\mathcal{G}_0$ is complicated by the fact that move on $\mathcal{G}_0$ may also alter $\mathcal{G}_1$ as the edge $(l \to j) \in \mathcal{E}_1$ is defined in terms of $\beta_{lj} + \gamma_{lj}$.

While it is possible to propose a joint move on $\mathcal{G}_1$ along with $\mathcal{G}_0$, we prefer local moves and propose to "insulate" $\mathcal{G}_1$ from the move on $\mathcal{G}_0$ by proposing changes in γ_{lj} and z_{lj} in accordance with changes in β_{lj}.

The joint move on $\mathcal{G}_0$ and $\boldsymbol{z}$ is interpretable as an expansion on the RJMCMC on DAG algorithm proposed by Fronk and Giudici (2004). Birth and the death moves on $\mathcal{G}_0$ are engineered to achieve

$$\frac{q\left(z \Rightarrow z' \mid \mathcal{G}_0'\right)}{q\left(z' \Rightarrow z \mid \mathcal{G}_0\right)} = \frac{q\left(z' \Rightarrow z \mid \mathcal{G}_0\right)}{q\left(z \Rightarrow z' \mid \mathcal{G}_0'\right)}.$$

Thus the acceptance probability of the birth move on the edge $(l \to j)$ for $\mathcal{G}_0$ is defined as

$$A_B = min\left\{1, \frac{p\left(\boldsymbol{\beta}_j', \boldsymbol{\gamma}_j', \boldsymbol{z}_j' \mid y\right)}{p\left(\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{z}_j \mid y\right) q_b(\beta_{lj}')} \frac{q\left(z' \Rightarrow z \mid \mathcal{G}_0\right)}{q\left(z \Rightarrow z' \mid \mathcal{G}_0'\right)}\right\}, \tag{5.12}$$

and the acceptance probability of the corresponding death move on the edge $(l \to j)$ for $\mathcal{G}_0$ is defined as

$$A_D = min\left\{1, \frac{p\left(\boldsymbol{\beta}_j', \boldsymbol{\gamma}_j', \boldsymbol{z}_j' \mid y\right) q_b(\beta_{lj})}{p\left(\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{z}_j \mid y\right)} \frac{q\left(z \Rightarrow z' \mid \mathcal{G}_0\right)}{q\left(z' \Rightarrow z \mid \mathcal{G}_0'\right)}\right\}, \tag{5.13}$$

where $q_b(\cdot)$ is the proposal distribution for β'_{lj}, $\boldsymbol{\beta}'_j$ refers to $\boldsymbol{\beta}_j$ with the lth element set to β'_{lj}. We will get to the details on $\boldsymbol{\gamma}'_j$ and $\boldsymbol{z}'_j$ in the following section, but for now they are the proposed values for $\boldsymbol{\gamma}_j$ and $\boldsymbol{z}_j$ if there is to be any change and we will denote the proposal distribution of γ_{lg} as $q_g(\cdot)$. Note that the Jacobian term does not come into play when we change $\boldsymbol{\gamma}$ since it is 1 similar to the proposal for the $\boldsymbol{\beta}$ in Fronk and Giudici (2004).

In the following section we will consider how to define $q\left(z \Rightarrow z' \mid \mathcal{G}'_0\right)/q\left(z' \Rightarrow z \mid \mathcal{G}_0\right)$ conditioned on $\mathcal{G}_1$ so that the above symmetry is preserved.

5.3.1.1 When $(l \rightarrow j) \notin \mathcal{E}_1$

There are two senarios where an edge $(l \rightarrow j)$ does not exist in the differential graph $\mathcal{G}_1$.

C_{00}**:** $(l \rightarrow j) \notin \mathcal{E}_0$ so $\beta_{lj} = 0$ and $\gamma_{lj} = 0$ so that $z_{lj} = 0$ or

C_{11}**:** $(l \rightarrow j) \in \mathcal{E}_0$ so $\beta_{lj} \neq 0$ and $\gamma_{lj} = -\beta_{lj}$ so that $z_{lj} = 1$

Hence conditioned on $(l \rightarrow j) \notin \mathcal{E}_1$ a legal move will be to move between these two conditions. If we make this move deterministic, because $\gamma_{lj} = 0 \Rightarrow \gamma_{lj} = -\beta_{lj}$ does not alter the dimension of $\boldsymbol{\gamma}$ and hence $q\left(z \Rightarrow z' \mid \mathcal{G}'_0\right)/q\left(z' \Rightarrow z \mid \mathcal{G}_0\right) = 1$. Therefore the move $C_{00} \Rightarrow C_{11}$ is accepted with probability

$$A_{B_0} = min\left\{1, \frac{p\left(\boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \boldsymbol{z}'_j \mid y\right)}{p\left(\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{z}_j \mid y\right) q_b(\beta'_{lj})} \frac{q\left(z' \Rightarrow z \mid \mathcal{G}_0\right)}{q\left(z \Rightarrow z' \mid \mathcal{G}'_0\right)}\right\}, \tag{5.14}$$

and the reverse move $C_{11} \Rightarrow C_{00}$ is accepted with probability

$$A_{D_0} = min\left\{1, \frac{p\left(\boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \boldsymbol{z}'_j \mid y\right) q_b(\beta_{lj})}{p\left(\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{z}_j \mid y\right)} \frac{q\left(z \Rightarrow z' \mid \mathcal{G}_0\right)}{q\left(z' \Rightarrow z \mid \mathcal{G}'_0\right)}\right\}. \tag{5.15}$$

5.3.1.2 When $(l \to j) \in \mathcal{E}_1$

The situation is slightly complicated when $(l \to j) \in \mathcal{E}_1$ do to the restriction imposed by the conditional prior on the γ_{lj} that does not allow $(l \to j) \notin \mathcal{E}_0$ so $\beta_{lj} = 0$ and $\gamma_{lj} = -\beta_{lj}$ so that $z_{lj} = 1$ do to the lack of identifiability with the C_{00} case. Therefore only allowed combination of the parameters are the following.

C_{02}**:** $(l \to j) \notin \mathcal{E}_0$ so $\beta_{lj} = 0$ and $\gamma_{lj} \neq 0$ so that $z_{lj} = 2$,

C_{10}**:** $(l \to j) \in \mathcal{E}_0$ so $\beta_{lj} \neq 0$ and $\gamma_{lj} = 0$ so that $z_{lj} = 0$, or

C_{12}**:** $(l \to j) \in \mathcal{E}_0$ so $\beta_{lj} \neq 0$ and $\gamma_{lj} \neq 0$ so that $z_{lj} = 2$

If we first consider the death move on $\mathcal{G}_0$ there are two possibilities $C_{10} \Rightarrow C_{02}$ or $C_{12} \Rightarrow C_{02}$ and both moves will not alter $\mathcal{G}_1$. To conserve the symmetry with the death move, when proposing a birth move on $\mathcal{G}_0$ we need to allow both of the reverse moves $C_{02} \Rightarrow C_{10}$ and $C_{02} \Rightarrow C_{12}$ to be possible. We can do this by choosing either of the revers moves with equal probability. Another thing to keep in mind is that although $C_{10} \Rightarrow C_{02}$ is a death move on $\boldsymbol{\beta}$, in terms of the $\boldsymbol{\gamma}$ it is a birth move. Hence the proposal ratio are defined as

- for $C_{10} \Rightarrow C_{02}$ is
$$\frac{p(C_{02} \Rightarrow C_{10}) q_b(\beta_{lj})}{p(C_{10} \Rightarrow C_{02}) q_g(\gamma'_{lj})} = \frac{\left(\frac{1}{2}\right)}{q_g(\gamma'_{lj})}$$

- and for $C_{12} \Rightarrow C_{02}$ is

$$\frac{p(C_{02} \Rightarrow C_{12}) q_b(\beta_{lj})}{p(C_{12} \Rightarrow C_{02})} = \frac{\left(\frac{1}{2}\right)}{1}$$

and it is not hard to see that the proposal ratio for $C_{02} \Rightarrow C_{10}$ and $C_{02} \Rightarrow C_{12}$ are just their inverses.

As a result the reversible jump ratio of a death move is defined as

- For $C_{10} \Rightarrow C_{02}$

$$R_{D_1} = \frac{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\boldsymbol{\gamma}'_{lj} \mid z'_{lj}, \beta'_{lj}, \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right) \left(\frac{1}{2}\right) q_b\left(\beta_{lj}\right)}{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\beta_{lj} \mid \sigma_j^2, \mathcal{G}_0\right) p\left(\mathcal{G}_0\right) q_g\left(\gamma'_{lj}\right) \left(\pi_j\right)}, \quad (5.16)$$

- and for $C_{12} \Rightarrow C_{02}$

$$R_{D_2} = \frac{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\mathcal{G}'_0\right) \left(\frac{1}{2}\right) q_b\left(\beta_{lj}\right)}{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\beta_{lj} \mid \sigma_j^2, \mathcal{G}_0\right) p\left(\mathcal{G}_0\right) \left(1 - \pi_j\right)}, \quad (5.17)$$

Similarly the reversible jump ratio for birth moves are defined as

- for $C_{02} \Rightarrow C_{10}$

$$R_{B_1} = \frac{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\beta'_{lj} \mid \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right) q_g\left(\gamma_{lj}\right) \left(\pi_j\right)}{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\boldsymbol{\gamma}_{lj} \mid z_{lj}, \beta_{lj}, \sigma_j^2, \mathcal{G}_0\right) p\left(\mathcal{G}_0\right) \left(\frac{1}{2}\right) q_b\left(\beta'_{lj}\right)}, \quad (5.18)$$

- and for $C_{02} \Rightarrow C_{12}$

$$R_{B_2} = \frac{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}'_j, \boldsymbol{\gamma}'_j, \sigma_j^2\right) p\left(\beta'_{lj} \mid \sigma_j^2, \mathcal{G}'_0\right) p\left(\mathcal{G}'_0\right) \left(1 - \pi_j\right)}{p\left(\boldsymbol{y}_j \mid \tilde{\boldsymbol{X}}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \sigma_j^2\right) p\left(\mathcal{G}_0\right) \left(\frac{1}{2}\right) q_b\left(\beta'_{lj}\right)}, \quad (5.19)$$

and each of the move is accepted with probability

$$A_{D_i} = min\left\{1, R_{D_i}\right\} \text{ or } A_{B_i} = min\left\{1, R_{B_i}\right\} \tag{5.20}$$

5.3.2 Conditional posterior distribution of σ_j^2

The conditional posterior distribution for σ_j^2 is inverse gamma distribution with

$$\begin{aligned}
p(\sigma_j^2 \mid Y, \alpha, \beta, \gamma, z, \mathcal{G}, \psi) \quad \propto \quad & \left(\sigma_j^2\right)^{-\frac{\delta_j}{2}-1} \exp\left(-\frac{\tau_j}{2\sigma_j^2}\right)\left(\sigma_j^2\right)^{-\frac{n}{2}} \\
& \exp\left\{-\frac{1}{2\sigma_j^2}\left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right)^T\left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right)\right\}\left(\sigma_j^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\omega_j}{2\sigma_j^2}\left(\alpha_j - a_j\right)^2\right\} \\
& \prod_{l=1}^{p}\left(\left(\sigma_j^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\omega_j}{2\sigma_j^2}\left(\beta_{lj} - b_{lj}\right)^2\right\}\right)^{I\{l \in pa_0(j)\}} \\
& \prod_{l=1}^{p}\left(\left(\sigma_j^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\omega_j}{2\sigma_j^2}\left(\gamma_{lj} - \nu_{lj}\right)^2\right\}\right)^{I\{z_{lj}=2\}} \\
\propto \quad & \left(\sigma_j^2\right)^{-\frac{1}{2}\left(\delta_j + n + 1 + \sum_l I\{\mathcal{G}_{lj}=1\} + \sum_l I\{z_{lj}=2\}\right)-1} \\
& \exp\left(-\frac{1}{2\sigma_j^2}\left(\tau_j + \left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right)^T\left(y_j - \tilde{\boldsymbol{X}}_j \boldsymbol{B}_j\right)\right)\right) \\
& \exp\left(-\frac{1}{2\sigma_j^2}\left(\omega_j\left(\left(\alpha_j - a_j\right)^2 + \sum_l \left(\beta_{lj} - b_{lj}\right)^2 I\{l \in pa_0(j)\} + \sum_l \left(\gamma_{lj} - \nu_{lj}\right)^2 I\{z_{lj}=2\}\right)\right)\right)
\end{aligned}$$

List of Notations

Acronyms

ADME absorption, distribution, metabolization, and elimination

AML Acute Myeloid Leukemia

AR auto regressive

bp basepair

CES-2 carboxylesterase-2

CG chain graph

CI credible interval

CRAN Comprehensive R Archive Network

DAG directed acyclic graph

DLT dose limiting toxicities

DNA deoxyribonucleic acid

EHRT enterohepatic recirculation time

FD false discovery

FDR false discovery rates

FN false negative

FNR false negative rates

FP false positive

FWE family wise errors

GDAG Gaussian directed acyclic graph

GGM Gaussian graphical models

GO Gene Oncology

GWAS genome-wide association studies

HIW hyper inverse wishart

HPD highest posterior probability

MAF minor allele frequency

MCMC Markov Chain Monte Carlo

MDR Missed Detection Rate

MHGJ Metropolis-Hastings-Green with Jacobians

MLE maximum likelihood estimator

MTD maximum tolerated dose

NCBI National Center for Biotechnology Information

PDAG partial directed acyclic graph

PD pharmacodynamic

PKGm Pharmacogenomics

PKGx Pharmacogenetics

PK pharmacokinetics

PopPK population pharmacokinetics

RJMCMC reversible jumps Markov Chain Monte Carlo

RPPA reverse phase protein array

RPTD recommended phase two dose

SNP single nucleotide polymorphism

SN-38 7-ethyl-10-hydroxycamptothecin

TN true negative

TP true positive

Probability Distribution

$Beta(\alpha, \beta)$ Beta distribution with parameters $\alpha > 0$ and $\beta > 0$.

$Bin(n, p)$ Binomial distribution with sample size $n \in \mathbb{N} \setminus \{0\}$ and success probability $p \in [0, 1]$.

$Exp(\lambda)$ Exponential distribution with rate parameter $\lambda > 0$.

$Ga(\alpha, \beta)$ Gamma distribution with shape $\alpha > 0$ and inverse scale $\beta > 0$.

$Inv\text{-}Ga(\alpha, \beta)$ Inverse-Gamma distribution with shape $\alpha > 0$ and scale $\beta > 0$.

$HIW_g(\alpha, \Phi)$ Hyper inverse Wishart distribution with parameters α and Φ.

$Inv\text{-}N(\mu, \lambda)$ Inverse Gaussian distribution with mean $\mu > 0$ and shape parameter $\lambda > 0$.

$\mathcal{MN}_{n\times p}(\Phi, \Sigma)$ $n\times p$ Matrix variate Gaussian distributio with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_n^T)^T$, row covariance $\Phi = [\phi_{ij}]_{n\times n}$, and column covariance $\Sigma = [\sigma_{ij}]_{p\times p}$.

$\mathcal{MT}(\upsilon, \Phi, \Sigma)$ Central matrix Student T distribution with parameters υ,Φ, and Σ.

$N_p(\boldsymbol{\mu}, \Sigma)$ p variate Multivariate Gaussian distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and covariance $\Sigma = [\sigma_{ij}]_{p\times p}$.

$Wishart_\nu(S)$ Wishart distribution with degrees of freedom $\nu > 0$ and symmetric positive definite $k \times k$ matrix S.

$Inv\text{-}Wishart(\nu, S)$ Inverse Wishart distribution with degrees of freedom $\nu > 0$ and symmetric positive definite $k \times k$ matrix S.

Symbols

$=_d$ Equality in distribution.

$\perp\!\!\!\perp$ Independence.

$\propto$ Proportional Relation.

$\det(x)$ Determinant of a matrix x.

$diag(x)$ Diagonal matrix with x as the diagonal entries.

E Expectation.

$\text{trace}(x)$ Trace of a matrix x.

var Variance.

$an(v)$ A set of ancestor vertices for a vertex v.

$bd(v)$ A set of boundary vertices for a vertex v.

$ch(v)$	A set of child vertices for a vertex v.
$cl(v)$	A set of closure vertices for a vertex v.
$de(v)$	A set of decendant vertices for a vertex v.
$nd(v)$	A set of non-decendant vertices for a vertex v.
$pa(v)$	A set of parent vertices for a vertex v.
$\mathcal{G}$	Graph.
$\mathcal{E}$	Edge set of a graph.
$\mathcal{G}^M$	Moralization of a graph $\mathcal{G}$.
$\mathcal{V}$	Vertice set of a graph.
$\mathbb{I}$	Integer.
$\mathbb{N}$	Natural number.
$\mathbb{R}$	Real number.
$p(\cdot)$	Probability density function.
$P(\omega)$	Probability of an event ω.

Index

Bibliography

Adler SL (1981). Over-relaxation method for the monte carlo evaluation of the partition function for multiquadratic actions. *Phys. Rev. D*, 23:2901–2904.

Albert JH and Chib S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Albuquerque P and Bronnenberg BJ (2012). Measuring the impact of negative demand shocks on car dealer networks. *Marketing Science*, 31(1):4–23.

Atay-Kayis A and Massam H (2005). A monte carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335.

Barbieri M and Berger J (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Barker DJ, Hill SM, and Mukherjee S (2010). Mc4: a tempering algorithm for large-sample network inference. In *Proceedings of the 5th IAPR international conference on Pattern recognition in bioinformatics*, PRIB'10, pages 431–442. Springer-Verlag, Berlin, Heidelberg.

Barnard J, McCulloch R, and Meng XL (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312.

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Berger JO (1985). *Statistical decision theory and Bayesian analysis*. Springer Verlag.

Berger JO and Delampady M (1987). Testing precise hypotheses. *Statistical Science*, pages 317–335.

Berry DA and Hochberg Y (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1-2):215 – 227.

Bertrand J and Balding DJ (2013). Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenetics and genomics*, 23(3):167–174.

Besag J (1974). Spatial interections and the statistical analysis of lattice systems. *JRSS B*, pages 302–339.

Brooks S, Gelman A, Jones G, and Meng X (2011). Handbook of markov chain monte carlo.

Brown B, Fearn T, and Vannucci M (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, 86(3):635–648.

Brown PJ, Vannucci M, and Fearn T (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(3):pp. 627–641.

Cai B and Dunson DB (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62:446–457.

Cai TT and Sun W (2009). Simultaneous testing of grouped hypotheses: Finding

needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488).

Carvalho C and Scott J (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497.

Celeux G, El Anbari M, Marin JM, and Robert CP (2012). Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502.

Chen M and Dey D (2000). Bayesian analysis for correlated ordinal data models. In DK Dey, WS Ghosh, and B Mallick, editors, *Generalized Linear Models: A Bayesian Perspective*, pages 135–162. Marcel Dekker, New York.

Cheng Y and Lenkoski A (2012). A multivariate graphical stochastic volatility model. *arXiv.org*.

Chib S and Greenberg E (1996). Bayesian analysis of multivariate probit models.

Chib S and Greenberg E (1998). Analysis of multivariate probit models. *Biometrika*, 85:347–361.

Chiu TYM, Leonard T, and Tsui KW (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):pp. 198–210.

Cox DR and Wermuth N (1996). *Multivariate dependencies: Models, analysis and interpretation*, volume 67. Chapman & Hall/CRC.

Danaher P, Wang P, and Witten D (2011). The joint graphical lasso for inverse covariance estimation across multiple classes. *arXiv preprint arXiv:1111.0324*.

Daniels MJ and Kass RE (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263.

Davidian M and Giltinan DM (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387–419.

Dawid AP and Lauritzen SL (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

de Bragança Pereira CA and Stern JM (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110.

Dempster AP (1972). Covariance selection. *Biometrics*, 28(1):pp. 157–175.

Dobra A, Hans C, Jones B, Nevins JR, Yao G, and West M (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196 – 212. Special Issue on Multivariate Methods in Genomic Data Analysis.

Dobra A and Lenkoski A (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.

Duncan DB (1965). A Bayesian approach to multiple comparisons. *Technometrics*, 7(2):pp. 171–222.

Edwards D (2000). *Introduction to graphical modelling*. Springer Verlag.

Efron B (2007). Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377.

Efron B and Tibshirani R (2006). On testing the significance of sets of genes. *ArXiv Mathematics e-prints*.

Friedman J, Hastie T, and Tibshirani R (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Fronk EM (2002). Model selection for dags via rjmcmc for the discrete and mixed case.

Fronk EM and Giudici P (2004). Markov chain monte carlo model selection for dag models. *Statistical Methods & Applications*, 13:259–273. 10.1007/s10260-004-0097-z.

Gelman A (2014). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*.

Gelman A, Bois F, and Jiang J (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91(436):1400–1412.

Gelman A, Hill J, and Yajima M (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211.

Gelman A and Rubin DB (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, 25:165–173.

Genovese C and Wasserman L (2003). Bayesian and frequentist multiple testing. In JMBernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, and AFM Smith, editors, *Bayesian Statistics 7*, volume 7. Oxford University Press.

George EI and McCulloch RE (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):pp. 881–889.

Geyer CJ (1991). Markov chain monte carlo maximum likelihood. In EM Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,*, pages 156–163. American Statistical Association, New York.

Geyer CJ (2010a). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo*, chapter 11, pages 295–311. CRC Press.

Geyer CJ (2010b). Introduction to Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*. CRC Press.

Geyer CJ and Thompson EA (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):pp. 909–920.

Gilks WR (1992). Derivative-free adaptive rejection sampling for gibbs sampling. In J Bernardo, J Berger, AP Dawid, and AFM Smith, editors, *Bayesian Statistics 4*, page 641?649. Oxford University Press.

Giudici P and Green P (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801.

Golub G (1965). Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7(3):206–216.

Green PJ (1995). Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green PJ and Mira A (2001). Delayed rejection in reversible jump metropolis-hastings. *Biometrika*, 88(4):pp. 1035–1053.

Guo J, Levina E, Michailidis G, and Zhu J (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Guo W (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128.

Haario H, Saksman E, and Tamminen J (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):pp. 223–242.

Hans C (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

He Q and Lin DY (2011). A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8.

He Y, Jia J, and Yu B (2012). Reversible mcmc on markov equivalence classes of sparse directed acyclic graphs. *Annals fo Statistics*.

Hegemann RA, Smith LM, Barbaro AB, Bertozzi AL, Reid SE, and Tita GE (2011). Geographical influences of an emerging network of gang rivalries. *Physica A: Statistical Mechanics and its Applications*, 390(21?22):3894 – 3914.

Hoff PD and Niu X (2012). A covariance regression model. *Statistica Sinica*, 22:729–753.

Hoffman MD and Gelman A (2011). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *ArXiv e-prints*.

Ickstadt K, Bornkamp B, Grzegorczyk M, Wieczorek J, Sheriff M, Grecco H, and Zamir E (2010). Nonparametric Bayesian networks. *Bayesian Statistics*, 9:283–316.

Innocenti F, Undevia SD, Iyer L, Xian Chen P, Das S, Kocherginsky M, Karrison T, Janisch L, Ramírez J, Rudin CM, Vokes EE, and Ratain MJ (2004a). Genetic variants in the udp-glucuronosyltransferase 1a1 gene predict the risk of severe neutropenia of irinotecan. *Journal of Clinical Oncology*, 22(8):1382–1388.

Innocenti F, Undevia SD, Ramírez J, Mani S, Schilsky RL, Vogelzang NJ, Prado M, and Ratain MJ (2004b). A phase i trial of pharmacologic modulation of irinotecan with cyclosporine and phenobarbital&ast. *Clinical Pharmacology & Therapeutics*, 76(5):490–502.

Iyer L, Das S, Janisch L, Wen M, Ramirez J, Karrison T, Fleming G, Vokes E, Schilsky R, and Ratain M (2002). Ugt1a1* 28 polymorphism as a determinant of irinotecan disposition and toxicity. *The pharmacogenomics journal*, 2(1):43–47.

Jasra A, Stephens DA, and Holmes CC (2007). Population-based reversible jump markov chain monte carlo. *Biometrika*, 94(4):787–807.

Johnson JA, Klein TE, and Relling MV (2013). Clinical implementation of pharmacogenetics: More than one gene at a time. *Clin Pharmacol Ther*, 93(5):384–385.

Jones B, Carvalho C, Dobra A, Hans C, Carter C, and West M (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Kornblau S, Covey T, Putta S, Cohen A, Woronicz J, Fantl W, Gayko U, and Cesano A (2011). Signaling changes in the stem cell factor–AKT-S6 pathway in diagnostic AML samples are associated with disease relapse. *Blood Cancer Journal*, 1(2):e3.

Koster JTA (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5):pp. 2148–2177.

Kuo L and Mallick B (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lauritzen S (1996). *Graphical models*, volume 17. Oxford University Press, USA.

Lauritzen SL and Richardson TS (2002). Chain graph models and their causal interpretation. *Journal of the Royal Statistical Society (B)*, 64(3):321–361.

Lauritzen SL and Wermuth N (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57.

Leonard T and Hsu JS (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, pages 1669–1696.

Liang F (2010). A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Liechty JC, Liechty MW, and Muller P (2004). Bayesian correlation estimation. *Biometrika*, 91(1):1–14.

Madigan D, York J, and Allard D (1995). Bayesian graphical models for discrete

data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215–232.

Meinshausen N and Büllmann P (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462.

Mohler G and Short M (2012). Geographic profiling from kinetic models of criminal behavior. *SIAM Journal on Applied Mathematics*, 72(1):163–180.

Moon JY, Neto EC, Deng X, and Yandell BS (2013). *Bayesian Causal Phenotype Network Incorporating Genetic Variation and Biological Knowledge*, chapter 4, pages 75–. Oxford Univ Press.

Mukherjee S and Speed TP (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318.

Müller P, Parmigiani G, and Rice K (2006). FDR and Bayesian multiple comparisons rules. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, page 115.

Müller P, Parmigiani G, Robert C, and Rousseau J (2004). Optimal sample size for multiple testing. *Journal of the American Statistical Association*, 99(468):990–1001.

Müller P and Quintana F (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801 – 2808. Interdisciplinary Mathematical and Statistical Techniques, International Conference on Interdisciplinary Mathematical and Statistical Techniques.

Murray I and Ghahramani Z (2004). Bayesian learning in undirected graphical models: Approximate mcmc algorithms.

Neal R (1995). Suppressing random walks in markov chain monte carlo using ordered overrelaxation. In *eprint arXiv:bayes-an/9506004*, pages 6004–+.

Neal RM (2012). MCMC using Hamiltonian dynamics. *ArXiv e-prints*.

Newton MA (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics (Oxford)*, 5(2):155–176.

Neytchev PN (1995). Sweep operator for least-squares subject to linear constraints. *Computational Statistics & Data Analysis*, 20(6):599 – 609.

O'Dwyer PJ and Catalano RB (October 1, 2006). Uridine diphosphate glucuronosyltransferase (ugt) 1a1 and irinotecan: Practical pharmacogenomics arrives in cancer therapy. *Journal of Clinical Oncology*, 24(28):4534–4538.

Ozawa Y, Williams A, Estes M, Matsushita N, Boschelli F, Jove R, and List A (2008). Src family kinases promote aml cell survival through activation of signal transducers and activators of transcription (stat). *Leukemia research*, 32(6):893–903.

Park T and Casella G (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Pearl J (1985). A constraint-propagation approach to probabilistic reasoning. In LM Kanal and J Lemmer, editors, *Uncertainty in artificial intelligence*, pages 357–370. North-Holland, North-Holland, Amsterdam.

Pearl J (2000). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Peng J, Wang P, Zhou N, and Zhu J (2009). Partial correlation estimation by

joint sparse regression models. *Journal of the American Statistical Association*, 104(486).

Pereira CAdB, Stern JM, and Wechsler S (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1):79–100.

Phillips C (2007). Online resources for snp analysis: A review and route map. *Mol Biotechnol*, 35:65–97.

Rice K (2010). A decision-theoretic formulation of fisher?s approach to testing. *The American Statistician*, 64(4).

Ring HZ and Kroetz DL (2002). Candidate gene approach for pharmacogenetic studies. *Pharmacogenomics*, 3(1):47–56.

Roberts GO and Rosenthal JS (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Roden DM, Altman RB, Benowitz NL, Flockhart DA, Giacomini KM, Johnson JA, Krauss RM, McLeod HL, Ratain MJ, Relling MV, Ring HZ, Shuldiner AR, Weinshilboum RM, and Weiss ST (2006). Pharmacogenomics: Challenges and opportunities. *Annals of Internal Medicine*, 145(10):749–757.

Rodriguez A, Lenkoski A, and Dobra A (2011). Sparse covariance estimation in heterogeneous samples. *Electronic Journal of Statistics*, 5:981–1014.

Ronning G and Kukuk M (1996). Efficient estimation of ordered probit models. *Journal of The American Statistical Association*, 97:1122–1140.

Roses AD (2000). Pharmacogenetics and the practice of medicine. *Nature*, 405:857–865.

Rosner GL and Müller P (1997). Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. *Journal of pharmacokinetics and biopharmaceutics*, 25(2):209–233.

Rosner GL, Panetta J, Innocenti F, and Ratain M (2008). Pharmacogenetic pathway analysis of irinotecan. *Clinical Pharmacology & Therapeutics*, 84(3):393–402.

Rothman AJ, Bickel PJ, Levina E, and Zhu J (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Roverato A (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.

Rue H and Held L (2005). *Gaussian Markov Random Fields Theory and Applications*. Chapman & Hall/CRC.

Rue H, Martino S, and Chopin N (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Saatchi SS, Harris NL, Brown S, Lefsky M, Mitchard ETA, Salas W, Zutta BR, Buermann W, Lewis SL, Hagen S, Petrova S, White L, Silman M, and Morel A (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences*, 108(24):9899–9904.

Sallas W (1988). Remark as r75: Some remarks on algorithm as 164: Least

squares subject to linear constraints. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(3):484–489.

Scott G and Carvalho CM (2007). Feature-inclusion stochastic search for gaussian graphical models. Technical report, Duke University.

Scott J and Berger J (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Scott JG and Berger JO (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144 – 2162.

Senn SS (2008). *Statistical Issues in Drug Development*. John Wiley & Sons.

Shastry B (2005). Pharmacogenetics and the concept of individualized medicine. *The Pharmacogenomics Journal*, 6(1):16–21.

Sheiner LB and Steimer JL (2000). Pharmacokinetic/pharmacodynamic modeling in drug development. *Annu Rev Pharmacol Toxicolor*, 40:67–95.

Sherry S, Ward MH, and Kholodov Mea (2001). dbsnp: The ncbi database of genetic variation. *Nucleic Acids Res*, 29:308–311.

Smith M and Kohn R (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153.

Soetaert K SR Petzoldt T (2010). Solving differential equations in r: Package desolve. *Journal of Statistical Software*, 33(9):1–25.

Srinivasan S, Vanhuele M, and Pauwels K (2010). Mind-set metrics in market response models: An integrative approach. *Journal of Marketing Research*, 47(4):672–684.

Stan Development Team (2013). Stan: A C++ library for probability and sampling, version 1.3.

Steenburgh TJ, Ainslie A, and Engebretson PH (2003). Massively categorical variables: Revealing the information in zip codes. *Marketing Science*, 22(1):40–57.

Stirling W (1981). Algorithm as 164: Least squares subject to linear constraints. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(2):204–212.

Sun W, Ibrahim J, and Zou F (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics*, 185(1):349–359.

Telesca D, Inoue L, Neira M, Etzioni R, Gleave M, and Nelson C (2009). Differential expression and network inferences through functional data modeling. *Biometrics*, 65(3):793–804.

Telesca D, Müller P, Kornblau SM, Suchard MA, and Ji Y (2012a). Modeling protein expression and protein signaling pathways. *Journal of the American Statistical Association*, 107(500):1372–1384.

Telesca D, Parmigiani G, Müller P, and Freedman RS (2012b). Modeling dependent gene expression. *Annals of Applied Statistics*.

Thiesson B, Meek C, Chickering D, Chickering DM, and Heckerman D (1997). Learning mixtures of dag models. In *In Proc. of the Conf. on Uncertainty in AI*, pages 504–513. Morgan Kaufmann, Inc.

Thulin M (2012). Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. *arXiv:1210.1066 [math.ST]*.

Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, and Kornblau SM (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*, 5(10):2512–2521.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288.

Uher R, Perroud N, Ng MY, Hauser J, Henigsberg N, Maier W, Mors O, Placentino A, Rietschel M, Souery D, et al. (2010). Genome-wide pharmacogenetics of antidepressant response in the gendep project. *American Journal of Psychiatry*, 167(5):555–564.

Valcárcel B, Würtz P, al Basatena N, Tukiainen T, Kangas A, Soininen P, Järvelin M, Ala-Korpela M, Ebbels T, and de Iorio M (2011). A differential network approach to exploring differences between biological states: An application to prediabetes. *PLoS ONE*, 6(9).

Wacheck V (2010). Biomarkers. In M MÃijller, editor, *Clinical Pharmacology: Current Topics and Case Studies*, pages 225–239. Springer Vienna.

Wakefield J (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association*, 91(433):pp. 62–75.

Wakefield J, Aarons L, and Racine-Poon A (1999). The Bayesian approach to population pharmacokinetic/pharmacodynamic modeling. In C Gatsonis, R Kass, B Carlin, A Carriquiry, A Gelman, I Verdinelli, and M West, editors, *Case Studies in Bayesian Statistics*, volume 140 of *Lecture Notes in Statistics*, pages 205–265. Springer New York.

Wang H (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(2):771–790.

Wang H and Li S (2012). Efficient Gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198.

Wang H and West M (2009). Bayesian analysis of matrix normal graphical models. *Biometrika*, 96(4):821–834.

Wermuth N and Lauritzen SL (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 21–50.

West M (1984). Outlier models and prior distributions in bayesial linear regression. *Journal of the Royal Statistical Society (B)*, 46(3):431–439.

Whittaker J (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Witten DM, Friedman JH, and Simon N (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Wittkowski KM, Sonakya V, Song T, Seybold MP, Keddache M, and Durner M (2013). From single-snp to wide-locus: genome-wide association studies identifying functionally related genes and intragenic regions in small sample studies. *Pharmacogenomics*, 14(4):391–401.

Wong F, Carter CK, and Kohn R (2003). Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830.

Wu R and Lin M (2010). *Statistical and computational pharmacogenomics*. Chapman and Hall/CRC.

Yang Q, Khoury MJ, Friedman J, Little J, and Flanders WD (2005). How many genes underlie the occurrence of common complex diseases in the population? *International Journal of Epidemiology*, 34(5):1129–1137.

Yang R and Berger JO (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, pages 1195–1211.

Yiannakopoulou EC (2013). Pharmacogenomics of phase ii metabolizing enzymes and drug transporters: clinical implications. *Pharmacogenomics J*, 13(2):105–109.

Yuan M and Lin Y (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.