UNIVERSITY OF CALIFORNIA SAN DIEGO

# Detection of Sparse Heterogeneous Mixtures: Theory, Methods and Algorithms

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

in

Mathematics (with a specialization in Statistics)

by

Rong Huang

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Loki Natarajan
Professor Ronghui (Lily) Xu
Professor Danna Zhang
Professor Wenxin Zhou

2020

The dissertation of Rong Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Ery Arias-Castro, for his countless support and guidance throughout my Ph.D. study at UC San Diego. Professor Arias-Castro introduced me to this research field, and inspired me with his deep insight in mathematics and statistics. It was a great privilege and honor to study under his supervision.

I would like to thank my co-advisor Professor Ronghui Xu for offering me an excellent project and invaluable guidance through the study. It was a great journey in a totally different field. I would like to thank Professor Ian Abramson and Professor Dimitris Politis for providing a systematic and solid training in mathematical statistics and encouraging me pursing a Ph.D. degree. I would like to thank the rest of my doctoral committee: Professor Loki Natarajan, Professor Danna Zhang and Professor Wenxin Zhou for insightful advice.

I would like to thank my whole family for their support, endless love and understanding of my PhD study. My parents have always been supportive of every decision I made. My husband, Yunjiang Qiu, has been the most important part of my life. He has always been there for me as we have gone through ups and downs. My life during the nine years wouldn't be such an incredible journey without him.

I would like to thank my friends inside and outside our Math department. We have spent a lot of festivals together and their friendship made my life outside the research fun and memorable.

Chapter 2, in full, is a version of the paper " The Sparse Variance Contamination Model", Arias-Castro, Ery; Huang, Rong. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a version of the paper "Detection of Sparse Positive Dependence", *Electronic Journal of Statistics*, Arias-Castro, Ery; Huang, Rong; Verzelen, Nicolas, Volume 14, Number 1 (2020). The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is a version of the paper “Detecting Sparse Heterogeneous Mixtures in a Two-sample Problem”, Huang, Rong. The manuscript is currently being prepared for submission for publication. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, is a version of the paper “Sensitivity Analysis of Treatment Effect to Unmeasured Confounding in Observational Studies with Survival and Competing Risks Outcomes”, Huang, Rong; Xu, Ronghui; Dulai, Parambir S. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

# VITA

| | |
|---|---|
| 2013 | B. S. in Biology, Peking University, Beijing China |
| 2015 | M. S. in Molecular Genetics, University of Toronto, Toronto Canada |
| 2017-2020 | Graduate Teaching Assistant, University of California San Diego |
| 2020 | Ph. D. in Mathematics, University of California San Diego |

# PUBLICATIONS

Ery Arias-Castro, Rong Huang and Nicolas Verzelen. “Detection of Sparse Positive Dependence”, *Electronic Journal of Statistics*, Volume 14, Number 1 (2020).

Ery Arias-Castro and Rong Huang. “The Sparse Variance Contamination Model”, *Submitted*, 2020.

Rong Huang. “Detecting Sparse Heterogeneous Mixtures in a Two-sample Problem”, *manuscript in preparation*, 2020.

Rong Huang, Ronghui Xu and Parambir S. Dulai. “Sensitivity Analysis of Treatment Effect to Unmeasured Confounding in Observational Studies with Survival and Competing Risks Outcomes”, *Submitted*, 2020.

ABSTRACT OF THE DISSERTATION

## Detection of Sparse Heterogeneous Mixtures: Theory, Methods and Algorithms

by

Rong Huang

Doctor of Philosophy in Mathematics (with a specialization in Statistics)

University of California San Diego, 2020

Professor Ery Arias-Castro, Chair

The detection of sparse heterogeneous mixtures becomes important in settings where a small proportion of a population may be affected by a given treatment, for example. The situation is typically formalized as a contamination model. We consider such models in asymptotic regimes where the contamination proportion tends to zero at various rates. We study the following three settings: the contamination manifests itself as a change in variance, the contamination manifests itself as a positive dependence between the variables in the bivariate setting, and the effect is a shift in mean without knowing the null distribution. In each setting, we study how large the effect needs to be in order to reliably distinguish the null hypothesis and the alternative hypothesis. We show that the corresponding higher criticism test is first-order comparable to the likelihood ratio test, while other classical tests are suboptimal. In particular, we make connections between the first two settings. We consider the dependence problem from both parametric and nonparametric perspectives. In the last chapter, we consider a different problem, that is to examine the extent to which the causal inference resulting estimate is sensitive to the unmeasured confounders for survival and competing risks data.

# Chapter 1

# Introduction

The detection of rare effects becomes important in settings where a small proportion of a population may be affected by a given treatment, for example. The situation is typically formalized as a contamination model. Although such models have a long history (e.g., in the theory of robust statistics), we adopt the perspective of Ingster [28] and Donoho and Jin [20]. Ingster [28] has studied the normal mixture model, that is, considering the following contamination model:

$$(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\mathcal{N}(\mu,1), \tag{1.1}$$

where $\varepsilon \in [0,1/2)$ is the contamination proportion and $\mu \geq 0$ is the shift in mean of the contaminated component. The following hypothesis testing problem is considered: based on $X_1,\ldots,X_n$ drawn iid from (1.1), decide

$$\mathcal{H}_0 : \varepsilon = 0 \quad \text{versus} \quad \mathcal{H}_1 : \varepsilon > 0, \mu > 0.$$

The problem is investigated in various asymptotic regimes where the contamination proportion tends to zero at various rates. The detection boundary of the likelihood ratio test (LRT) (then any other tests) is derived. Donoho and Jin [20] further derived the detection boundary with the generalized Gaussian mixture model:

$$f(x) \propto \exp\Big(-\frac{|x|^\gamma}{\gamma}\Big),$$

where $\gamma > 0$. Note that $\gamma = 2$ corresponds to the normal distribution and $\gamma = 1$ corresponds to the double-exponential distribution. They parameterized $\varepsilon = \varepsilon_n$ as

$$\varepsilon_n = n^{-\beta}, \quad 0 < \beta < 1 \text{ fixed.}$$

In the sparse setting where $1/2 < \beta < 1$, let

$$\mu_n = (\gamma r \log n)^{1/\gamma}, \quad 0 < r < 1 \text{ fixed,}$$

then the detection boundary when $\gamma > 1$ is

$$\rho^*_\gamma(\beta) = \begin{cases} (2^{1/(\gamma-1)} - 1)^{\gamma-1}(\beta - \frac{1}{2}), & \frac{1}{2} < \beta < 1 - 2^{-\gamma/(\gamma-1)}; \\ (1-(1-\beta)^{1/\gamma})^\gamma, & 1 - 2^{-\gamma/(\gamma-1)} < \beta < 1. \end{cases}$$

and for the case $\gamma \leq 1$

$$\rho^*_\gamma(\beta) = 2\beta - 1.$$

That means, if $r > \rho^*(\beta)$, $\mathcal{H}_0$ and $\mathcal{H}_1$ separate asymptotically, while if $r < \rho^*(\beta)$, $\mathcal{H}_0$ and $\mathcal{H}_1$ merge asymptotically.

The detection boundary in the dense regime where $0 < \beta < 1/2$ is given in [3]. Let

$$\mu_n = n^{s-1/2}, \quad 0 < s < 1/2 \text{ fixed,}$$

then the hypotheses merge asymptotically when $s < \beta$ if $\gamma \geq 1/2$ and $s < \frac{1}{2} - \frac{1-2\beta}{1+2\gamma}$ if $\gamma < 1/2$.

This line of work has mostly focused on models where the effect is a shift in mean, with some rare exceptions [12, 11]. In Chapter 2, we consider a Gaussian contamination model where the contamination manifests itself as a change in variance. We show that the higher criticism test is (first-order) comparable to the likelihood ratio test in all sparsity regimes, while the chi-squared test and the extremes test are suboptimal.

In Chapter 3, in a bivariate setting, we consider the problem of detecting a sparse contamination or mixture component, where the effect manifests itself as a positive dependence between the variables, which are otherwise independent in the main component. We first look at this problem in the context of a normal mixture model. In essence, the situation reduces to a univariate setting where the effect is a decrease in variance. In particular, a higher criticism test based on the pairwise differences is shown to achieve the detection boundary defined by the (oracle) likelihood ratio test. We then turn to a Gaussian copula model where the marginal distributions are unknown. Standard invariance considerations lead us to consider rank tests. In fact, a higher criticism test based on the pairwise rank differences achieves the detection boundary in the normal mixture model, although not in the very sparse regime. We do not know of any rank test that has any power in that regime.

In Chapter 4, we consider the problem of detecting sparse heterogeneous mixtures in a two-sample setting from a nonparametric perspective, where the effect manifests itself as a positive shift. We suggest a two-sample higher criticism test, and show that it is first-order comparable to the likelihood ratio test for the normal mixture models in all sparsity regimes.

In Chapter 5, we turn to a causal inference problem in observational studies with survival and competing risks outcomes. No unmeasured confounding is often assumed in estimating treatment effects in observational data, whether using classical regression models or approaches such as propensity scores and inverse probability weighting. However, in many such studies, collection of confounders cannot possibly be exhaustive in practice, and it is crucial to examine the extent to which the resulting estimate is sensitive to the unmeasured confounders. We consider this problem for survival and competing risks data. Due to the complexity of models for such data, we adapt the simulated potential confounders approach of Carnegie et al. [14], which provides a general tool for sensitivity analysis due to unmeasured confounding. More specifically, we specify one sensitivity parameter to quantify the association between an unmeasured confounder and the exposure or treatment received, and another set of parameters to quantify the association between the confounder and the time-to-event outcomes. By varying the magnitudes of the sensitivity parameters, we estimate the treatment effect of interest using the stochastic EM and the EM algorithms. We demonstrate the performance of our methods on simulated data, and apply them to a comparative effectiveness study in inflammatory bowel disease (IBD).

# Chapter 2

# The Sparse Variance Contamination Model

## 2.1 Introduction

The detection of rare effects becomes important in settings where a small proportion of a population may be affected by a given treatment, for example. The situation is typically formalized as a contamination model. Although such models have a long history (e.g., in the theory of robust statistics), we adopt the perspective of Ingster [28] and Donoho and Jin [20], who consider such models in asymptotic regimes where the contamination proportion tends to zero at various rates. This line of work has mostly focused on models where the effect is a shift in mean, with some rare exceptions [12, 11]. In this paper, instead, we model the effect as a change in variance.

We consider the following contamination model:

$$(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\mathcal{N}(0,\sigma^2), \tag{2.1}$$

where $\varepsilon \in [0,1/2)$ is the contamination proportion and $\sigma > 0$ is the standard deviation of the contaminated component. (Note that this is a Gaussian mixture model with two components.) Following [28, 20], we consider the following hypothesis testing problem: based on $X_1,\ldots,X_n$

drawn iid from (2.1), decide

$$\mathcal{H}_0 : \varepsilon = 0 \quad \text{versus} \quad \mathcal{H}_1 : \varepsilon > 0, \sigma \neq 1. \tag{2.2}$$

As usual, we study the behavior of the likelihood ratio test, which is optimal in this simple versus simple hypothesis testing problem if we assume that the model parameters $(\varepsilon, \sigma)$ are known. We also study some testing procedures that, unlike the likelihood ratio test, do not require knowledge of $(\varepsilon, \sigma)$:

- The *chi-squared test* rejects for large values of $\left|\sum_i X_i^2 - n\right|$. This is the typical variance test when the sample is known to be zero mean.
- The *extremes test* combines the test that rejects for small values of $\min_i |X_i|$ and the test that rejects for large values of $\max_i |X_i|$ using Bonferroni's method.
- The *higher criticism test* [20] amounts to applying one of the tests proposed by Anderson and Darling [1] for normality. One variant is based on rejecting for large values of

  $$\text{HC} = \sup_{x \geq 0} \frac{\sqrt{n}|F_n(x) - \Psi(x)|}{\sqrt{\Psi(x)(1 - \Psi(x))}}, \tag{2.3}$$

  where $\Psi(x) := 2\Phi(x) - 1$, where $\Phi$ denotes the standard normal distribution, and $F_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{|X_i| \leq x\}$, is the empirical distribution of $|X_i|$.

The testing problem (2.2) was partially addressed by Cai, Jeng, and Jin [11], who consider a contamination model where the effect manifests itself as a shift in mean and a change in variance. However, in their setting the variance is fixed, while we let the variance change with the sample size in an asymptotic analysis that is now standard in this literature.

In the tradition of Ingster [28], we set

$$\varepsilon = n^{-\beta}, \quad \beta \in (0,1) \text{ fixed.} \tag{2.4}$$

The setting where $\beta \le 1/2$ is often called the dense regime while the setting where $\beta > 1/2$ is often called the sparse regime. (Note that the setting where $\beta > 1$ is uninteresting since in that case there is no contamination with probability tending to 1.)

Our analysis reveals three distinct situations:

(a) *Near zero* ($\sigma \to 0$): In the sparse regime, the higher criticism test is as optimal as the likelihood ratio test, while the chi-squared test is powerless and the extremes test is suboptimal.

(b) *Near one* ($\sigma \to 1$): In the dense regime, the chi-squared test and the higher criticism test are as optimal as the likelihood ratio test, while the extremes test has no power.

(c) *Away from zero and one ($\sigma$ fixed)*: In the sparse regime, the extremes test and the higher criticism test are as optimal as the likelihood ratio test, while the chi-squared test is asymptotically powerless if $\sigma$ is bounded.

These results are summarized in Table 2.1 and Figure 2.1.

**Table 2.1**: The detection boundary for the likelihood ratio test, the chi-squared test, the extremes test, and the higher criticism test.

| | dense regime ($\beta < 1/2$) | | sparse regime ($\beta > 1/2$) | |
|---|---|---|---|---|
| | $\lvert\sigma_n - 1\rvert = n^{-\gamma}, \gamma > 0$ | $\sigma$ fixed | $\sigma_n = n^{-\gamma}, \gamma > 0$ | $\sigma$ fixed |
| likelihood ratio | $\gamma < 1/2 - \beta$ | $\sigma \neq 1$ | $\gamma > 2\beta - 1$ | $\sigma > 1/\sqrt{1-\beta}$ |
| chi-squared | $\gamma < 1/2 - \beta$ | $\sigma \neq 1$ | no power | no power |
| extremes | no power | $\sigma > 1/\sqrt{1-\beta}$ | $\gamma > \beta$ | $\sigma > 1/\sqrt{1-\beta}$ |
| higher criticism | $\gamma < 1/2 - \beta$ | $\sigma \neq 1$ | $\gamma > 2\beta - 1$ | $\sigma > 1/\sqrt{1-\beta}$ |

## 2.2 The likelihood ratio test

We start with bounding the performance of the likelihood ratio test. As this is the most powerful test by the Neyman–Pearson Lemma, this bound also applies to any other test. In the

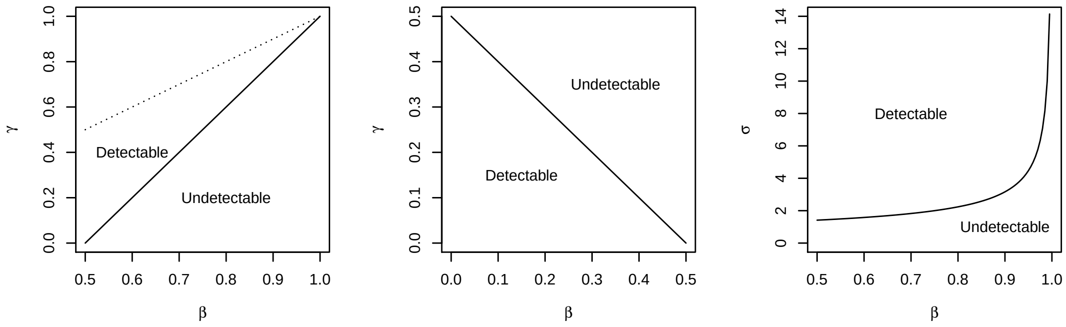


**Figure 2.1**: The detection boundary for the Gaussian mixture model (2.1). From left to right, the plots correspond to (2.5), (2.8), and σ fixed. In each case, the detection boundary (defined by the first-order performance of the likelihood ratio test) is drawn as a solid line, and is shown to also apply to the higher criticism test. The dotted line corresponds to the detection boundary for the extremes test. See Table 2.1 for more details.

present setting, the likelihood ratio is given by

$$L := \prod_{i=1}^{n} L_i,$$

where $L_i$ is the likelihood ratio for observation $X_i$, which in this case is

$$\begin{aligned} L_i &= \frac{\frac{1-\varepsilon}{\sqrt{2\pi}}\exp(-\frac{1}{2}X_i^2) + \frac{\varepsilon}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}X_i^2)}{\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}X_i^2)} \\ &= 1 - \varepsilon + \frac{\varepsilon}{\sigma}\exp\left(\frac{\sigma^2 - 1}{2\sigma^2}X_i^2\right), \end{aligned}$$

so that

$$L = \prod_{i=1}^{n}\left[1 - \varepsilon + \frac{\varepsilon}{\sigma}\exp\left(\frac{\sigma^2 - 1}{2\sigma^2}X_i^2\right)\right].$$

We say that a testing procedure is asymptotically powerless if the sum of its probabilities of Type I and Type II errors (its risk) has limit inferior at least 1 in the large sample limit. The likelihood ratio test has optimum risk when applied with critical value equal to 1, meaning with rejection region $\{L > 1\}$, and its risk is then equal to

$$\text{risk}(L) := 1 - \frac{1}{2}\mathbb{E}_0|L-1|.$$

(We refer the reader to [34, Problem 3.10].)

### 2.2.1 Near zero

Consider the testing problem (2.2) in the regime where $\sigma = \sigma_n \to 0$ as $n \to \infty$. More specifically, we adopt the following parameterization as it brings into focus the first-order asymptotics:

$$\sigma = n^{-\gamma}, \quad \gamma > 0 \text{ fixed.} \tag{2.5}$$

**Theorem 1.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4) *and* (2.5)*, the likelihood ratio test (and then any other test procedure) is asymptotically powerless when*

$$\gamma < 2\beta - 1. \tag{2.6}$$

*Proof.* Our goal is to show that $\text{risk}(L) = 1 + o(1)$ under the stated conditions. When $\sigma$ is below and bounded away from $\sqrt{2}$, it turns out that a crude method, the so-called 2nd moment method which relies on the Cauchy-Schwarz Inequality, is enough to lower bound the risk. Indeed, by the Cauchy-Schwarz Inequality,

$$\text{risk}(L) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0[L^2] - 1},$$

and we are left with the task of finding conditions under which $\mathbb{E}_0[L^2] \leq 1 + o(1)$.

We have

$$\mathbb{E}_0[L^2] = \prod_{i=1}^{n} \mathbb{E}_0[L_i^2] = (\mathbb{E}_0[L_1^2])^n,$$

where

$$\begin{aligned}\mathbb{E}_0[L_1^2] &= \mathbb{E}_0\left[\left(1-\varepsilon+\frac{\varepsilon}{\sigma}\exp\left(\frac{\sigma^2-1}{2\sigma^2}X_1^2\right)\right)^2\right]\\ &= 1-\varepsilon^2+\frac{\varepsilon^2}{\sigma^2}\mathbb{E}_0\left[\exp\left(\frac{\sigma^2-1}{\sigma^2}X_1^2\right)\right]\\ &= 1-\varepsilon^2+\varepsilon^2\left[\sigma^2(2-\sigma^2)\right]^{-1/2}\\ &= 1+\varepsilon^2\left(\left[\sigma^2(2-\sigma^2)\right]^{-1/2}-1\right).\end{aligned}$$

Therefore,

$$\mathbb{E}_0[L^2] = \left[1+\varepsilon^2\left(\left[\sigma^2(2-\sigma^2)\right]^{-1/2}-1\right)\right]^n \leq \exp\left[n\varepsilon^2\left(\left[\sigma^2(2-\sigma^2)\right]^{-1/2}-1\right)\right],$$

so that $\mathbb{E}_0[L^2] \leq 1+o(1)$ when

$$n\varepsilon^2\left(\left[\sigma^2(2-\sigma^2)\right]^{-1/2}-1\right) \to 0. \tag{2.7}$$

Plugging in the parameterization (2.4) and (2.5), we immediately see that this condition is fulfilled when (2.6) holds, and this concludes the proof. □

### 2.2.2 Near one

Consider the testing problem (2.2) in the regime where $\sigma^2 \to 1$. More specifically, we adopt the following parameterization:

$$|\sigma-1| = n^{-\gamma}, \quad \gamma > 0 \text{ fixed.} \tag{2.8}$$

**Theorem 2.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4) *and* (2.8)*, the likelihood ratio test (and then any other test procedure) is asymptotically powerless* *when*

$$\gamma > 1/2 - \beta. \tag{2.9}$$

*Proof.* Restarting the proof of Theorem 1 at (2.7), and plugging in the parameterization (2.4) and (2.8), we immediately see that $\mathbb{E}_0[L^2] \le 1 + o(1)$ when (2.9) holds. □

### 2.2.3 Away from zero and one

Consider the testing problem (2.2) in the regime where $\sigma$ is fixed away from 0 and 1. (Some of the results developed in this section are special cases of results in [11].)

**Theorem 3.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4) *and* $\sigma > 0$ *is fixed, the likelihood ratio test (and therefore any other test) is asymptotically powerless when* $\beta > 1/2$ *and*

$$\sigma < 1/\sqrt{1-\beta}. \tag{2.10}$$

*Proof.* We use a refinement of the second moment method, sometimes called the truncated second moment method, which is based on bounding the moments of a thresholded version of the likelihood ratio. Define the indicator variable $D_i = \mathbb{I}\{|X_i| \le \sqrt{2\log n}\}$ and the corresponding truncated likelihood ratio

$$\bar{L} = \prod_{i=1}^{n} \bar{L}_i, \quad \bar{L}_i := L_i D_i.$$

Using the triangle inequality, the fact that $\bar{L} \le L$, and the Cauchy-Schwarz Inequality, we have the following upper bound:

$$\begin{aligned} \mathbb{E}_0[|L-1|] &\le \mathbb{E}_0[|\bar{L}-1|] + \mathbb{E}_0[L-\bar{L}] \\ &\le \left[\mathbb{E}_0[\bar{L}^2] - 1 + 2(1-\mathbb{E}_0[\bar{L}])\right]^{1/2} + (1-\mathbb{E}_0[\bar{L}]) \ , \end{aligned}$$

so that $\text{risk}(L) \ge 1 + o(1)$ when $\mathbb{E}_0[\bar{L}^2] \le 1 + o(1)$ and $\mathbb{E}_0[\bar{L}] \ge 1 + o(1)$.

For the first moment, we have

$$\mathbb{E}_0[\bar{L}] = \prod_{i=1}^{n} \mathbb{E}_0[\bar{L}_i] = \mathbb{E}_0[\bar{L}_1]^n,$$

so that it suffices to prove that $\mathbb{E}_0[\bar{L}_1] \geq 1 - o(1/n)$. We develop

$$\begin{aligned}
\mathbb{E}_0[\bar{L}_1] &= \mathbb{E}_0\left[\left(1 - \varepsilon + \frac{\varepsilon}{\sigma}\exp\left(\frac{\sigma^2 - 1}{2\sigma^2} X_1^2\right)\right) D_1\right] \\
&= (1-\varepsilon)(1 - 2\bar{\Phi}(\sqrt{2\log n})) + \varepsilon(1 - 2\bar{\Phi}(\sqrt{2\log n}/\sigma)) \\
&= (1-\varepsilon)(1 - O(n^{-1}/\sqrt{\log n})) + \varepsilon(1 - O(n^{-1/\sigma^2}/\sqrt{\log n})) \\
&= 1 - o(1/n) - o(\varepsilon n^{-1/\sigma^2}),
\end{aligned}$$

where $\bar{\Phi}$ is the standard normal survival function. We used the well-known fact that $\bar{\Phi}(t) \sim e^{-t^2/2}/\sqrt{2\pi}t$ as $t \to \infty$. Since $\varepsilon = n^{-\beta}$ with $\beta > 1/2$, and (2.10) holds, we have $\varepsilon n^{-1/\sigma^2} = o(1/n)$, so that $\mathbb{E}_0[\bar{L}_1] \geq 1 - o(1/n)$.

For the second moment, we have

$$\mathbb{E}_0[\bar{L}^2] = \prod_{i=1}^{n} \mathbb{E}_0[\bar{L}_i^2] = \mathbb{E}_0[\bar{L}_1^2]^n,$$

so that it suffices to prove that $\mathbb{E}_0[\bar{L}_1^2] \leq 1 + o(1/n)$. We develop

$$\begin{aligned}
\mathbb{E}_0[\bar{L}_1^2] &= \mathbb{E}_0\left[\left(1 - \varepsilon + \frac{\varepsilon}{\sigma}\exp\left(\frac{\sigma^2-1}{2\sigma^2}X_1^2\right)\right)^2 D_1\right] \\
&= (1-\varepsilon)^2(1-2\bar{\Phi}(\sqrt{2\log n})) + 2(1-\varepsilon)\varepsilon(1-2\bar{\Phi}(\sqrt{2\log n}/\sigma)) \\
&\quad + \frac{\varepsilon^2}{\sigma^2}\mathbb{E}_0\left[\exp\left(\frac{\sigma^2-1}{\sigma^2}X_1^2\right)D_1\right] \\
&\leq 1 - \varepsilon^2 + \frac{\varepsilon^2}{\sqrt{2\pi}\sigma^2}\int_{-\sqrt{2\log n}}^{\sqrt{2\log n}} \exp\left(\left(\frac{\sigma^2-1}{\sigma^2} - \frac{1}{2}\right)x^2\right)\mathrm{d}x \\
&\leq 1 + O\left(\varepsilon^2 \exp\left(\frac{(\sigma^2-2)_+}{\sigma^2}\log n\right)\sqrt{\log n}\right).
\end{aligned}$$

Hence, it suffices that $-2\beta + (\sigma^2-2)_+/\sigma^2 < -1$, which is equivalent to (2.10). □

Though we only provide lower bounds on what can be achieved, they turn out to be sharp once we analyze the performance of other tests, especially the higher criticism test, which is shown to achieves these lower bounds to firs-order accuracy.

## 2.3 Other tests

Having studied the performance of the likelihood ratio test, we now turn to studying the performance of the chi-squared test, the extremes test, and the higher criticism test. These tests are more practical in that they do not require knowledge of the parameters driving the alternative, $(\varepsilon, \sigma)$, to be implemented.

### 2.3.1 The chi-squared test

The chi-squared test is the classical variance test. It happens to only be asymptotically optimal in the dense regime.

**Proposition 1.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4)*, the* *chi-squared test is asymptotically powerful when* $\beta < 1/2$ *and either* $\sigma$ *is bounded away from 1 or* (2.8) *holds with* $\gamma < 1/2 - \beta$. *The chi-squared test is asymptotically powerless when* $\beta > 1/2$ *and* $\sigma$ *is bounded.*

*Proof.* We divide the proof into the two regimes.

*Dense regime (*$\beta < 1/2$*).* We show that there is a chi-squared test that is asymptotically powerful when $\beta < 1/2$. Under $\mathcal{H}_0$, $W := \sum_{i=1}^n X_i^2$ has the chi-squared distribution with $n$ degrees of freedom. But using only the fact that $\mathbb{E}_0(W) = n$ and $\mathrm{Var}_0(W) = 2n$, by Chebyshev's inequality, we have

$$\mathbb{P}_0(|W - n| \geq a_n\sqrt{n}) \to 0,$$

for any sequence $(a_n)$ diverging to infinity. Under $\mathcal{H}_1$, $\mathbb{E}_1(W) = n(1 - \varepsilon + \varepsilon\sigma^2)$. Let $I_i$ indicate whether $X_i$ comes from the contaminated component. Note that $I_i \sim \mathrm{Bernoulli}(\varepsilon)$. Then

$$\begin{aligned}\mathrm{Var}_1(W) = n\mathrm{Var}_1(X_1) &= n[\mathbb{E}_1(\mathrm{Var}_1(X_1|I_1)) + \mathrm{Var}_1(\mathbb{E}_1(X_1|I_1))] \\ &= n[2 - 2\varepsilon + 2\varepsilon\sigma^4 + (\sigma^2 - 1)^2\varepsilon(1 - \varepsilon)].\end{aligned}$$

Note that $\mathrm{Var}_1(W) \sim 2n$ eventually. By Chebyshev's inequality,

$$\mathbb{P}_1(|W - n(1 - \varepsilon + \varepsilon\sigma^2)| \geq a_n\sqrt{n}) \to 0.$$

We choose $a_n = \log n$ and consider the test with rejection region $\{|W - n| \geq a_n\sqrt{n}\}$. This test is asymptotically powerful when, eventually,

$$|n(1 - \varepsilon + \varepsilon\sigma^2) - n| \geq 2a_n\sqrt{n},$$

meaning,

$$|\sigma^2 - 1|\varepsilon\sqrt{n} \geq 2a_n.$$

This is the case when $\beta < 1/2$ with no condition on $\sigma$ other than remaining bounded away from 1, and also when (2.8) holds and $\gamma < 1/2 - \beta$.

*Sparse regime (*$\beta > 1/2$*).* To prove that the chi-squared procedure is asymptotically powerless when $\beta > 1/2$, we argue in terms of convergence in distribution rather than the simple bounding of moments. Under $\mathcal{H}_0$, the usual Central Limit Theorem implies that $(W-n)/\sqrt{2n}$ converges weakly to the standard normal distribution. Under $\mathcal{H}_1$, the same is true using the Lyapunov Central Limit Theorem for triangular arrays. Indeed, even though the distribution of $X_1, \dots, X_n$ depends on $(n, \varepsilon)$, uniformly

$$\frac{\sum_{i=1}^n \mathbb{E}_1\left[(X_i^2-1)^4\right]}{\left(\sum_{i=1}^n \mathbb{E}_1\left[(X_i^2-1)^2\right]\right)^2} = \frac{n\mathbb{E}_1\left[(X_1^2-1)^4\right]}{n^2\left(\mathbb{E}_1\left[(X_1^2-1)^2\right]\right)^2} \asymp 1/n \to 0,$$

so that $(W - \mathbb{E}_1(W))/\sqrt{\mathrm{Var}_1(W)}$ converges weakly to the standard normal distribution. Since

$$\frac{W-\mathbb{E}_1(W)}{\sqrt{\mathrm{Var}_1(W)}} = \left(\frac{W-n}{\sqrt{2n}} + \frac{n-\mathbb{E}_1(W)}{\sqrt{2n}}\right)\frac{\sqrt{2n}}{\sqrt{\mathrm{Var}_1(W)}},$$

with

$$\mathbb{E}_1(W) = n(1-\varepsilon+\varepsilon\sigma^2) = n + o(\sqrt{n}), \quad \text{(since } \beta > 1/2\text{)},$$

and

$$\mathrm{Var}_1(W) = \sum_{i=1}^n \mathbb{E}_1\left[(X_i^2-1)^2\right] = n[2-2\varepsilon+2\varepsilon\sigma^4+(\sigma^2-1)^2\varepsilon(1-\varepsilon)] \sim 2n, \quad \text{(since } \sigma \text{ is bounded)},$$

it is also the case that $(W-n)/\sqrt{2n}$ converges weakly to the standard normal distribution by Slutsky's theorem. Hence, there is no test based on $W$ that has any asymptotic power. □

### 2.3.2 The extremes test

The extremes test, as the name indicates, focuses on the extreme observations, disregarding the rest of the sample. It happens to be suboptimal in the setting where $\sigma \to 0$, while it achieves the detection boundary in the sparse regime in the setting where $\sigma$ is fixed.

**Proposition 2.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4) *and* (2.5)*, the extremes test is asymptotically powerful when* $\gamma > \beta$ *(and asymptotically powerless when* $\gamma < \beta$*). If instead* $\sigma > 0$ *is fixed, the extremes test is asymptotically powerful when* $\sigma > 1/\sqrt{1-\beta}$ *(and asymptotically powerless when* $\sigma < 1/\sqrt{1-\beta}$*).*

*Proof.* Under $\mathcal{H}_0$, for any $a_n \to \infty$, we have

$$\begin{aligned}\mathbb{P}_0\left(\min_i |X_i| \geq 1/na_n\right) &= \left[\mathbb{P}_0\left(|X_1| \geq 1/na_n\right)\right]^n \\ &= \left[2\bar{\Phi}(1/na_n)\right]^n \\ &= \left[1 - O(1/na_n)\right]^n \to 1.\end{aligned}$$

Similarly, as is well-known,

$$\mathbb{P}_0\left(\max_i |X_i| \leq \sqrt{2\log n}\right) \to 1.$$

We thus consider the test with rejection region $\{\min_i |X_i| \leq 1/n\log n\} \cup \{\max_i |X_i| \geq \sqrt{2\log n}\}$.

We now consider the alternative. We first consider the case where (2.5) holds. We focus on the main sub-case where, in addition, $\gamma < 1$. Let $I \subset \{1,\ldots,n\}$ index the contaminated observations, meaning those sampled from $\mathcal{N}(0,\sigma^2)$. In our mixture model, $|I|$ is binomial with parameters

$(n,\varepsilon)$. Let $Z_1,\dots,Z_n$ be iid standard normal variables and set $b_n = \sigma n \log n$. We have

$$\begin{aligned}\mathbb{P}_1\left(\min_i |X_i| \le 1/n\log n\right) &\ge \mathbb{P}_1\left(\min_{i\in I} |X_i| \le 1/n\log n\right)\\ &= 1 - \mathbb{E}\left[\mathbb{P}\left(\min_{i\in I}|Z_i| \ge 1/b_n \mid I\right)\right]\\ &= 1-\mathbb{E}\left[(2\bar{\Phi}(1/b_n))^{|I|}\right]\\ &= 1-\left[1-\varepsilon+\varepsilon 2\bar{\Phi}(1/b_n)\right]^n.\end{aligned}$$

Since we have assumed that $\gamma < 1$ in (2.5), we have $1/b_n \to 0$, and therefore

$$2\bar{\Phi}(1/b_n) = 1 - \frac{2+o(1)}{\sqrt{2\pi}b_n}.$$

This in turn implies that

$$\left[1-\varepsilon+\varepsilon 2\bar{\Phi}(1/b_n)\right]^n = \left[1-\frac{(2+o(1))\varepsilon}{\sqrt{2\pi}b_n}\right]^n \to 0$$

when $n\varepsilon/b_n \to \infty$, which is the case when $\gamma > \beta$.

Assume instead that $\gamma < \beta$. Fix a level $\alpha \in (0,1)$ and consider the extremes test at that level. Based on the same calculations, this test has rejection region $\{\min_i |X_i| \le c_n\} \cup \{\max_i |X_i| \ge d_n\}$, where $c_n$ and $d_n$ are defined by $[2\bar{\Phi}(c_n)]^n = 1-\alpha/2$ and $[2\Phi(d_n)-1]^n = 1-\alpha/2$, respectively. Note that

$$c_n \sim -\sqrt{\pi/2}\log(1-\alpha/2)/n, \quad d_n \sim \sqrt{2\log n}.$$

For the minimum, we have

$$\mathbb{P}_1\left(\min_i |X_i| \le c_n\right) \le \mathbb{P}_1\left(\min_{i\notin I} |X_i| \le c_n\right) + \mathbb{P}_1\left(\min_{i\in I} |X_i| \le c_n\right).$$

Let $Z_1, \dots, Z_n$ be iid standard normal variables. Clearly,

$$\mathbb{P}_1\left(\min_{i \notin I} |X_i| \le c_n\right) \le \mathbb{P}\left(\min_i |Z_i| \le c_n\right) = \alpha/2,$$

and, as was derived above,

$$\begin{aligned}\mathbb{P}_1\left(\min_{i \in I} |X_i| \le c_n\right) &= \mathbb{P}_1\left(\min_{i \in I} |Z_i| \le c_n/\sigma\right) \\ &= 1 - \left[1 - \varepsilon + \varepsilon 2\bar{\Phi}(c_n/\sigma)\right]^n,\end{aligned}$$

with

$$\left[1 - \varepsilon + \varepsilon 2\bar{\Phi}(c_n/\sigma)\right]^n = \left[1 - \frac{(2+o(1))\varepsilon c_n}{\sqrt{2\pi}\sigma}\right]^n \to 1,$$

since $\varepsilon c_n/\sigma \asymp n^{-1-\beta+\gamma} = o(1/n)$. Thus, $\mathbb{P}_1(\min_i |X_i| \le c_n) \to 0$. And since $\max_i |X_i|$ under the alternative is stochastically bounded from above by its distribution under the null (since $\sigma < 1$), we also have $\mathbb{P}_1(\max_i |X_i| \ge d_n) \le \alpha/2$. Hence, the extremes test (at level $\alpha$ arbitrary) has asymptotic power $\alpha$, meaning it is asymptotically powerless. (It is no better than random guessing.)

Next, we consider the case where $\sigma$ is fixed. Following similar arguments, now with $b_n = \sigma^{-1}\sqrt{2\log n}$, we have

$$\begin{aligned}\mathbb{P}_1\left(\max_i |X_i| \ge \sqrt{2\log n}\right) &\ge \mathbb{P}_1\left(\max_{i \in I} |X_i| \ge \sqrt{2\log n}\right) \\ &= 1 - \mathbb{E}\left[\mathbb{P}\left(\max_{i \in I} |Z_i| \le b_n \mid I\right)\right] \\ &= 1 - \mathbb{E}\left[(2\Phi(b_n) - 1)^{|I|}\right] \\ &= 1 - \left[1 - \varepsilon + \varepsilon(2\Phi(b_n) - 1)\right]^n.\end{aligned}$$

We have

$$2\Phi(b_n) - 1 \asymp 1 - o(n^{-1/\sigma^2}),$$

so that

$$\left[1-\varepsilon+\varepsilon(2\Phi(b_n)-1)\right]^n \asymp \left[1-o(\varepsilon n^{-1/\sigma^2})\right]^n \to 0$$

when $n\varepsilon n^{-1/\sigma^2} \to \infty$, which is the case when $\sigma > 1/\sqrt{1-\beta}$.

Using a similar line of arguments, it can also be shown that the test is asymptotically powerless when $\sigma < 1/\sqrt{1-\beta}$ is fixed. □

### 2.3.3 The higher criticism test

The higher criticism, which looks at the entire sample via excursions of its empirical process, happens to achieve the detection boundary in all regimes, and is thus (first-order) comparable to the likelihood ratio test while being adaptive to the model parameters.

**Proposition 3.** *For the model* (2.1) *and the testing problem* (2.2) *with parameterization* (2.4)*, the higher criticism test is asymptotically powerful when either* (2.5) *holds with* $\gamma > 2\beta - 1$*, or* (2.8) *holds with* $\gamma < 1/2 - \beta$*, or* $\sigma > 1/\sqrt{1-\beta}$ *is fixed, or* $\beta < 1/2$ *and* $\sigma \neq 1$ *is fixed.*

*Proof.* Jaeschke [29] derived the asymptotic distribution of HC defined in (2.3) under the null, and this weak convergence result in particular implies that

$$\mathbb{P}_0\left(\mathrm{HC} \geq \sqrt{3\log\log n}\right) \to 0.$$

For simplicity, because it is enough for our purposes, we consider the test with rejection region $\{\mathrm{HC} \geq \log n\}$. Note that the test is asymptotically powerful if, under the alternative, there is $t_n \geq 0$ such that

$$\frac{\sqrt{n}|F_n(t_n) - \Psi(t_n)|}{\sqrt{\Psi(t_n)(1-\Psi(t_n))}} \geq \log n$$

with probability tending to 1. To establish this, we will apply Chebyshev's inequality. Indeed,

$nF_n(t)$ is binomial with parameters $n$ and $\Lambda(t) := (1-\varepsilon)\Psi(t) + \varepsilon\Psi(t/\sigma)$, so that

$$\frac{\sqrt{n}|F_n(t_n) - \Lambda(t_n)|}{\sqrt{\Lambda(t_n)(1-\Lambda(t_n))}} \leq \log n$$

with probability tending to 1. When this is the case, we have

$$\frac{\sqrt{n}|F_n(t_n) - \Psi(t_n)|}{\sqrt{\Psi(t_n)(1-\Psi(t_n))}} \geq u_n - (\log n)\sqrt{v_n},$$

where

$$u_n := \frac{\sqrt{n}\varepsilon|\Psi(t_n/\sigma) - \Psi(t_n)|}{\sqrt{\Psi(t_n)(1-\Psi(t_n))}}, \quad v_n := \frac{\Lambda(t_n)(1-\Lambda(t_n))}{\Psi(t_n)(1-\Psi(t_n))},$$

and only need to prove that

$$u_n \geq (\sqrt{v_n} + 1)\log n. \tag{2.11}$$

First, assume that (2.5) holds with $\gamma > 2\beta - 1$. We focus on the interesting sub-case where $\gamma < \beta$. Fix $q$ such that $q > \gamma$ and $1/2 - \beta - q/2 + \gamma > 0$ and set $t_n = n^{-q}$. Then, using the fact that $\varepsilon/\sigma = n^{\gamma-\beta} = o(1)$, we have

$$\Psi(t_n) \asymp t_n, \quad \Psi(t_n/\sigma) \asymp t_n/\sigma, \quad \Lambda(t_n) \asymp t_n + \varepsilon t_n/\sigma \asymp t_n,$$

so that

$$u_n \asymp \sqrt{n}\varepsilon(t_n/\sigma)/\sqrt{t_n} = n^{1/2-\beta-q/2+\gamma} \gg \log n, \quad v_n \asymp 1,$$

and therefore (2.11) is fulfilled, eventually.

Next, we assume that (2.8) holds with $\gamma < 1/2 - \beta$. Here we set $t_n = 1$, and get $0 < \Psi(t_n) = \Psi(1) < 1$, and

$$|\Psi(t_n/\sigma) - \Psi(t_n)| \sim |(1/\sigma - 1)\Psi'(1)| \asymp |\sigma - 1|, \quad \Lambda(t_n) \asymp 1,$$

so that

$$u_n \asymp \sqrt{n}\varepsilon|\sigma - 1| = n^{1/2-\beta-\gamma} \gg \log n, \quad v_n \asymp 1,$$

and therefore (2.11) is fulfilled, eventually.

The same arguments apply to the case where $\beta < 1/2$ and $\sigma \neq 1$ is fixed. (It essentially corresponds to the previous case with $\gamma = 0$.)

The remaining case is where $\sigma > 1/\sqrt{1-\beta}$ is fixed, with $\beta > 1/2$ (for otherwise it is included in the previous case). We choose $t_n = \sqrt{2q\log n}$, with $q := \beta/(1-1/\sigma^2)$, and get

$$t_n\bar{\Psi}(t_n) \asymp e^{-t_n^2/2} = n^{-q}, \quad t_n\bar{\Psi}(t_n/\sigma) \asymp e^{-t_n^2/2\sigma^2} = n^{-q/\sigma^2},$$

and

$$t_n\bar{\Lambda}(t_n) \asymp e^{-t_n^2/2} + \varepsilon e^{-t_n^2/2\sigma^2} = n^{-q} + n^{-\beta-q/\sigma^2} = 2n^{-q},$$

so that

$$u_n \asymp \frac{\sqrt{n}\varepsilon e^{-t_n^2/2\sigma^2}/t_n}{\sqrt{e^{-t_n^2/2}/t_n}} \asymp n^{1/2-\beta-q/\sigma^2+q/2}/(\log n)^{1/4} \gg \log n, \quad v_n \asymp 1,$$

and therefore (2.11) is fulfilled, eventually. □

## 2.4 Numerical experiments

We performed some numerical experiments to investigate the finite sample performance of the tests considered here: the likelihood ratio test, the chi-squared test, the extremes test, the higher criticism test. The sample size $n$ was set large to $10^5$ in order to capture the large-sample behavior of these tests. We tried four scenarios with different combinations of $(\beta, \sigma)$. The p-values for each test are calibrated as follows:

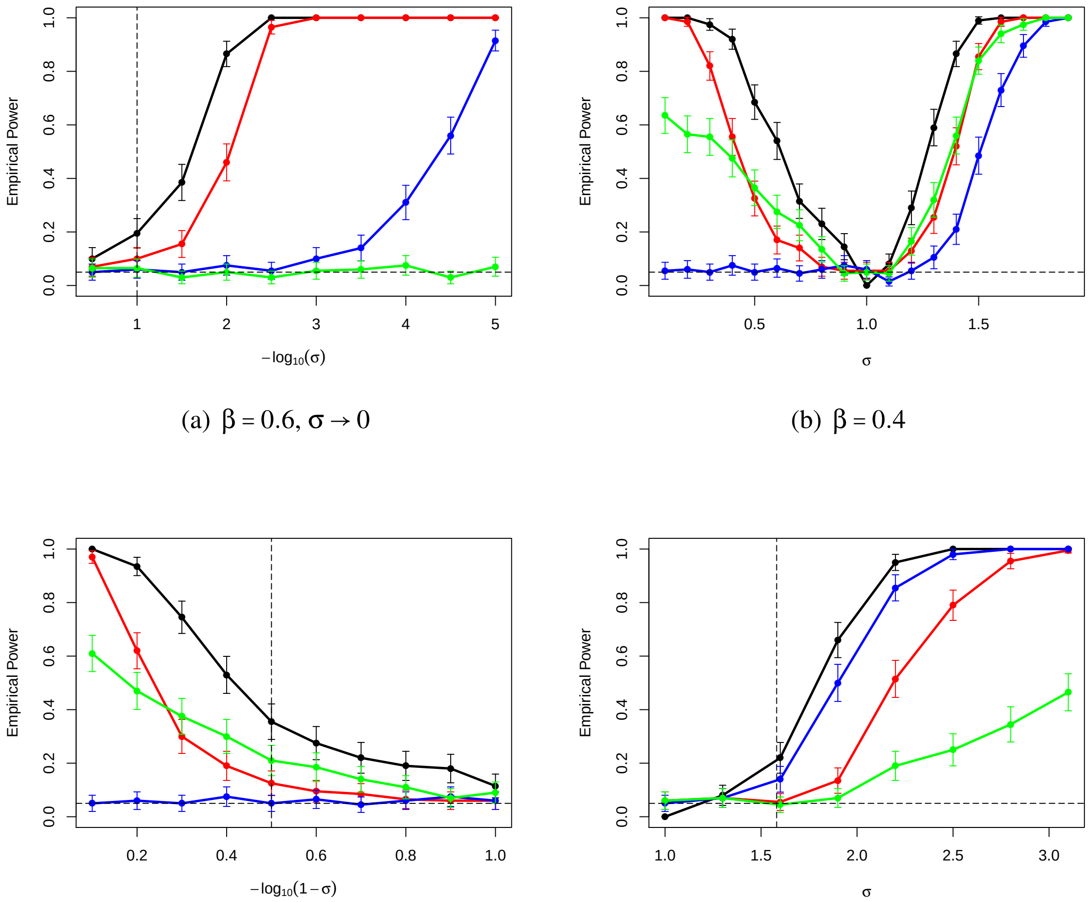
(a) For the *likelihood ratio test* and the *higher criticism test*, we simulated the null distribution based on $10^4$ Monte Carlo replicates.

(b) For the *extremes test* and the *chi-squared test*, we used the exact null distribution, which in each case is available in closed form.

For each combination of $(\beta, \sigma)$, we repeated the whole process 200 times and recorded the fraction of p-values smaller than 0.05, representing the empirical power at the 0.05 level. The result of this experiment is reported in Figure 2.2 and is largely congruent with the theory developed earlier in the paper.

## 2.5 Acknowledgements

This work was partially supported by the US National Science Foundation (DMS 1223137). Chapter 2, in full, is a version of the paper " The Sparse Variance Contamination Model", Arias-Castro, Ery; Huang, Rong. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.



(a) $\beta = 0.6$, $\sigma \to 0$

(b) $\beta = 0.4$

(c) $\beta = 0.4$, $\sigma \to 1$

(d) $\beta = 0.6$

**Figure 2.2**: Empirical power comparison with 95% error bars for the likelihood ratio test (black), the higher criticism test (red), the extremes test (blue) and the chi-squared test (green). (a) Sparse regime where $\beta = 0.6$ and $\sigma \to 0$. (b) Dense regime where $\beta = 0.4$ and $\sigma$ fixed. Note that the LR test is here asymptotically powerful at any $\sigma \neq 1$. (c) Dense regime where $\beta = 0.4$ and $\sigma \to 1$. (d) Sparse regime where $\beta = 0.6$ and $\sigma > 1$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $n = 10^5$ and the power curves and error bars are based on 200 replications.

# Chapter 3

# Detection of Sparse Positive Dependence

## 3.1 Introduction

The detection of rare effects has been an important problem for years in settings, and may be particularly relevant today, for example, with the search for personalized care in the health industry, where a small fraction of a population may respond particularly well, or particularly poorly, to some given treatment [47].

Following a theoretical investigation initiated in large part by Ingster [28] and broadened by Donoho and Jin [20], we are interested in studying two-component mixture models, also known as contamination models, in various asymptotic regimes defined by how the small mixture weight converges to zero. Most of the existing work in the setting of univariate data has focused on models where the contamination manifests itself as a shift in mean [22, 21, 26, 12, 43] with a few exceptions where the effect is a change in variance [2], or a change in both mean and variance [11].

In the present paper, we are interested in bivariate data, instead, and more specifically in a situation where the effect felt in the dependence between the two variables being measured. This setting has been recently considered in the literature in the context of assessing the reproducibility of studies. For example, [36] aims to identify significant features from separate studies using an expectation-maximization (EM) algorithm. They applied a copula mixture model and assumed that changes in the mean and covariance matrix differentiate the contaminated component from the null component. [57] studies another model where variables from the contamination are stochastically larger marginally. In both models, the marginal distributions have some non-null effects. Similar settings have been considered within a multiple testing framework [9, 56].

While existing work has focused on models motivated by questions of reproducibility, in the present work we come back to basics and directly address the problem of detecting a bivariate mixture with a component where the variables are independent and a component where the variables are positively dependent.

### 3.1.1 Gaussian mixture model

Ingster [28] and Donoho and Jin [20] started with a mixture of Gaussians, and we do the same, and in our setting, this means we consider the following mixture model

$$(X,Y) \sim (1-\varepsilon)\mathcal{N}(0,\mathrm{I}) + \varepsilon\mathcal{N}(0,\Sigma_\rho), \quad \Sigma_\rho := \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{3.1}$$

where $\varepsilon \in [0,1/2)$ is the contamination proportion and $0 \le \rho \le 1$ is the correlation between the two variables under contamination. We consider the following hypothesis testing problem: based on $(X_1,Y_1),\dots,(X_n,Y_n)$ drawn iid from (3.1), decide

$$\mathcal{H}_0 : \varepsilon = 0 \quad \text{versus} \quad \mathcal{H}_1 : \varepsilon > 0, \rho > 0. \tag{3.2}$$

Note that under the null hypothesis, $(X,Y)$ is from the bivariate standard normal. Under the alternative, $X$ and $Y$ remain standard normal marginally. Following the literature on the detection of sparse mixtures [28, 20], we are most interested in a situation, asymptotic as $n \to \infty$, where $\varepsilon = \varepsilon_n \to 0$, and the central question is how large $\rho = \rho_n$ needs to be in order to reliability distinguish these hypotheses.

The formulation (3.1) suggests that the alternative hypothesis is composite, but if we assume that $(\varepsilon, \rho)$ are known under the alternative, then the likelihood ratio test (LRT) is optimal by Neyman-Pearson lemma. We start with characterizing the behavior of the LRT, which provides a benchmark. We then study some other testing procedures that do not require knowledge of the model parameters:[1]

- The *covariance test* rejects for large values of $\sum_i X_i Y_i$, and coincides with Rao's score test in the present context. This is the classical test for independence, specifically designed for the case where $\varepsilon = 1$ and $\rho > 0$ under the alternative. We shall see that it is suboptimal in some regimes.

- The *extremes test* rejects for small values of $\min_i |X_i - Y_i|$. This test exploits the fact that, because $\rho$ is assumed positive, the variables in the contaminated component are closer to each other than in the null component.

- The *higher criticism test* was suggested by John Tukey and deployed by [20] for the testing of sparse mixtures. We propose a version of that test based on the pairwise differences, $U_i := (X_i - Y_i)/\sqrt{2}$. In detail, the test rejects for large values of

$$\sup_{u \geq 0} \frac{\sqrt{n}(\hat{F}(u) - \Psi(u))}{\sqrt{\Psi(u)(1 - \Psi(u))}}, \tag{3.3}$$

  where $\Psi(u) := 2\Phi(u) - 1$, with $\Phi$ denotes the standard normal distribution function, and $\hat{F}(u) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{|U_i| \leq u\}$, the empirical distribution function of $|U_1|, \dots, |U_n|$.

As is common practice in this line of work [28, 20], under $\mathcal{H}_1$ we set

$$\varepsilon = n^{-\beta}, \quad \beta \in (0,1) \text{ fixed.} \tag{3.4}$$

[1] Such procedures are said to be *adaptive*.

The setting where $\beta \le 1/2$ is often called the dense regime and the setting where $\beta > 1/2$ is often called the sparse regime. Our analysis reveals the following:

(a) *Dense regime.* The dense regime is most interesting when $\rho \to 0$. In that case, we find that the covariance test and the higher criticism test match the asymptotic performance of the likelihood ratio test to first-order, while the extremes test has no power.

(b) *Sparse regime.* The sparse regime is most interesting when $\rho \to 1$. In that case, we find that the higher criticism test still performs as well as the likelihood ratio test to first order, while the covariance test is powerless, and the extremes test is suboptimal.

### 3.1.2 Gaussian mixture copula model

From a practical point of view, the assumption that both $X$ and $Y$ are normally distributed is quite stringent. Hence, we would like to know if there are nonparametric procedures that do not require such a condition but can still achieve the same performance as the likelihood ratio test. In the univariate setting where the effect arises as a shift in mean, this was investigated in [3]. In the bivariate setting, in a model for reproducibility, [57] proposes a nonparametric test based on a weighted version of Hoeffding's test for independence.

Here, instead of model (3.1), we suppose $(X,Y)$ follows a Gaussian mixture copula model (GMCM) [8], meaning that there is a latent random vector $(Z^1, Z^2)$ such that

$$\begin{aligned} &F(X) = \Phi(Z^1), \quad G(Y) = \Phi(Z^2), \\ &(Z^1, Z^2) \sim (1-\varepsilon)\mathcal{N}(0,I) + \varepsilon\mathcal{N}(0,\Sigma_\rho), \quad \Sigma_\rho := \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \end{aligned} \tag{3.5}$$

where $F$ and $G$ are unknown distribution functions on the real line, and $\Phi$ is the standard normal distribution function, while $\varepsilon \in [0, 1/2)$ is the contamination proportion and $0 \le \rho \le 1$ is the correlation between $Z^1$ and $Z^2$ in the contaminated component, as before in model (3.1). [36]

also uses a copula mixture model, but they placed emphasis on the mean while we focus on the dependence.

We still consider the testing problem (3.2), but now in the context of Model (3.5). The setting is nonparametric in that both $F$ and $G$ are unknown. Model (3.5) is crafted in such a way that the marginal distributions of $X$ and $Y$ contain absolutely no information that is pertinent to the testing problem under consideration.

The model is also attractive because of an invariance under all increasing marginal transformations of the variables. This is the same invariance that leads to considering rank based methods such as the Spearman correlation test [34, Ch 6]. In fact, we analyze the Spearman correlation test, which is the nonparametric analog to the covariance test, showing that it is first-order asymptotically optimal in the dense regime. We also propose and analyze a nonparametric version of the higher criticism based on ranks which we show is first-order asymptotically optimal in the moderately sparse regime where $1/2 < \beta < 3/4$. In the very sparse regime, where $\beta > 3/4$, we do not know of any rank-based test that has any power.

## 3.2 Gaussian mixture model

In this section, we focus on the Gaussian mixture model (3.1). We start by deriving a lower bound on the performance of the likelihood ratio test, which provides a benchmark for the other (adaptive) tests, which we subsequently analyze.

We distinguish between the dense and sparse regimes:

$$\text{dense regime} \quad \rho = n^{-\gamma}, \quad \gamma > 0 \text{ fixed}; \tag{3.6}$$

$$\text{sparse regime} \quad \rho = 1 - n^{-\gamma}, \quad \gamma > 0 \text{ fixed}. \tag{3.7}$$

We say that a testing procedure is asymptotically powerful (resp. powerless) if the sum of its probabilities of Type I and Type II errors (its risk) has limit 0 (resp. limit inferior at least 1) in the large sample asymptote.

### 3.2.1 The likelihood ratio test

**Theorem 4.** *Consider the testing problem* (3.2) *with* $\varepsilon$ *parameterized as in* (3.4)*. In the dense regime, with* $\rho$ *parameterized as in* (3.6)*, the likelihood ratio test is asymptotically powerless when* $\gamma > 1/2 - \beta$*. In the sparse regime, with* $\rho$ *parameterized as in* (3.7)*, the likelihood ratio test is asymptotically powerless when* $\gamma < 4(\beta - 1/2)$.

This only provides a lower bound on what can be achieved, but it will turn out that to be sharp once we establish the performance of the higher criticism test in Proposition 5 below.

*Proof.* The proof techniques are standard and already present in [22, 28], and many of the subsequent works.

Defining $U := (X-Y)/\sqrt{2}$ and $V := (X+Y)/\sqrt{2}$, the model (3.1) is equivalently expressed in terms of $(U,V)$, which has distribution

$$(U,V) \sim (1-\varepsilon)\mathcal{N}(0,\mathrm{I}) + \varepsilon\mathcal{N}(0,\Delta_\rho), \quad \Delta_\rho := \mathrm{diag}(1-\rho, 1+\rho). \tag{3.8}$$

Note that $U$ and $V$ are independent only conditional on knowing what distribution they were sampled from. In terms of the $(U,V)$'s, the likelihood ratio is

$$L := \prod_{i=1}^{n} L_i,$$

where $L_i$ is the likelihood ratio for observation $(U_i, V_i)$, which in the present case takes the following expression

$$\begin{aligned} L_i &= \frac{\frac{1-\varepsilon}{2\pi}\exp(-\frac{1}{2}U_i^2 - \frac{1}{2}V_i^2) + \frac{\varepsilon}{2\pi\sqrt{1-\rho^2}}\exp(-\frac{1}{2(1-\rho)}U_i^2 - \frac{1}{2(1+\rho)}V_i^2)}{\frac{1}{2\pi}\exp(-\frac{1}{2}U_i^2 - \frac{1}{2}V_i^2)} \\ &= 1-\varepsilon+\varepsilon(1-\rho^2)^{-1/2}\exp(-\tfrac{\rho}{2(1-\rho)}U_i^2 + \tfrac{\rho}{2(1+\rho)}V_i^2). \end{aligned}$$

The risk of the likelihood ratio test is equal to [34, Problem 3.10]

$$\operatorname{risk}(L) := 1 - \frac{1}{2}\mathbb{E}_0[|L-1|].$$

We show that $\operatorname{risk}(L) = 1 + o(1)$ under each of the stated conditions. We consider each regime in turn.

*Dense regime.* It turns out that it suffices to bound the second moment. Indeed, using the Cauchy-Schwarz inequality, we have

$$\operatorname{risk}(L) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0[L^2]-1},$$

reducing the task to showing that $\mathbb{E}_0[L^2] \leq 1+o(1)$. We have

$$\mathbb{E}_0[L^2] = \prod_{i=1}^{n}\mathbb{E}_0[L_i^2] = (\mathbb{E}_0[L_1^2])^n$$

where

$$\begin{aligned} \mathbb{E}_0[L_1^2] &= \mathbb{E}_0\Big[\Big(1-\varepsilon+\varepsilon(1-\rho^2)^{-1/2}\exp(-\tfrac{\rho}{2(1-\rho)}U_1^2 + \tfrac{\rho}{2(1+\rho)}V_1^2)\Big)^2\Big] \\ &= (1-\varepsilon)^2 + 2(1-\varepsilon)\varepsilon \\ &\quad + \varepsilon^2(1-\rho^2)^{-1}\mathbb{E}_0\big[\exp(-\tfrac{\rho}{(1-\rho)}U_1^2)\big]\mathbb{E}_0\big[\exp(\tfrac{\rho}{(1+\rho)}V_1^2)\big] \\ &= 1-\varepsilon^2+\varepsilon^2(1-\rho^2)^{-1}\mathbb{E}_0\big[\exp(-\tfrac{\rho}{(1-\rho)}U_1^2)\big]\mathbb{E}_0\big[\exp(\tfrac{\rho}{(1+\rho)}V_1^2)\big]. \end{aligned}$$

For the third term, we have

$$\mathbb{E}_0\big[\exp(-\tfrac{\rho}{(1-\rho)}U_1^2)\big] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{\rho}{1-\rho}u^2-\frac{1}{2}u^2}\mathrm{d}u = \sqrt{\frac{1-\rho}{1+\rho}},$$

and

$$\mathbb{E}_0\big[\exp(\tfrac{\rho}{(1+\rho)}V_1^2)\big] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{\rho}{1+\rho}v^2-\frac{1}{2}v^2}\mathrm{d}v = \sqrt{\frac{1+\rho}{1-\rho}}.$$

Hence, we have

$$\mathbb{E}_0[L_1^2] = 1+\varepsilon^2\rho^2/(1-\rho^2),$$

and, therefore,

$$\mathbb{E}_0[L^2] = \big[1+\varepsilon^2\rho^2/(1-\rho^2)\big]^n \le \exp\big[n\varepsilon^2\rho^2/(1-\rho^2)\big],$$

so that $\mathbb{E}_0[L^2] \le 1+o(1)$ when

$$n\varepsilon^2\rho^2 = o(1),$$

since $\rho$ is assumed to be bounded away from 1. Under the specified parameterization, this happens exactly when $\gamma > 1/2-\beta$.

*Sparse regime.* It turns out that simply bounding the second moment, as we did above, does not suffice. Instead, we truncate the likelihood and study the behavior of its first two moments. Define the indicator variable $D_i = \mathbb{I}\{|V_i| \le \sqrt{2\log n}\}$ and the corresponding truncated likelihood ratio

$$\bar{L} = \prod_{i=1}^{n} \bar{L}_i, \quad \bar{L}_i := L_i D_i.$$

Using the triangle inequality, the fact that $\bar{L} \le L$, and the Cauchy-Schwarz inequality, we have the following upper bound:

$$\begin{aligned}\mathbb{E}_0|L-1| &\le \mathbb{E}_0|\bar{L}-1| + \mathbb{E}_0(L-\bar{L})\\ &\le \left[\mathbb{E}_0[\bar{L}^2]-1+2(1-\mathbb{E}_0[\bar{L}])\right]^{1/2} + (1-\mathbb{E}_0[\bar{L}])\ ,\end{aligned}$$

so that $\text{risk}(L) = 1+o(1)$ when $\mathbb{E}_0[\bar{L}^2] \le 1+o(1)$ and $\mathbb{E}_0[\bar{L}] \ge 1-o(1)$.

For the first moment, we have

$$\mathbb{E}_0[\bar{L}] = \prod_{i=1}^{n} \mathbb{E}_0[\bar{L}_i] = (\mathbb{E}_0[\bar{L}_1])^n$$

where, using the independence of $U_1$ and $V_1$, and taking the expectation with respect to $U_1$ first,

$$\begin{aligned}\mathbb{E}_0[\bar{L}_1] &= \mathbb{E}_0\left[\left(1-\varepsilon+\varepsilon(1+\rho)^{-1/2}\exp(\tfrac{\rho}{2(1+\rho)}V_1^2)\right)D_1\right]\\ &= (1-\varepsilon)\Psi(\sqrt{2\log n}) + \varepsilon\Psi(\sqrt{2\log n}/\sqrt{1+\rho})\\ &= (1-\varepsilon)(1-O(n^{-1}/\sqrt{\log n})) + \varepsilon(1-O(n^{-1/(1+\rho)}/\sqrt{\log n}))\\ &= 1-o(1/n)-o(\varepsilon n^{-1/(1+\rho)}),\end{aligned}$$

where, for $t \ge 0$,

$$\Psi(t) = \mathbb{P}(|\mathcal{N}(0,1)| \le t) = 2\Phi(t)-1 = \int_{-t}^{t} \frac{e^{-s^2/2}}{\sqrt{2\pi}}\mathrm{d}s,$$

and we used the fact that $1-\Psi(t) \asymp e^{-t^2/2}/t$ when $t \to \infty$. Since $\varepsilon = n^{-\beta}$ with $\beta > 1/2$ in the sparse regime, for $\rho$ sufficiently close to 1, $\varepsilon n^{-1/(1+\rho)} \le 1/n$, in which case $\mathbb{E}_0[\bar{L}_1] \ge 1-o(1/n)$. This yields

$$\mathbb{E}_0[\bar{L}] \ge (1-o(1/n))^n = 1-o(1).$$

For the second moment, we have

$$\mathbb{E}_0[\bar{L}^2] = \prod_{i=1}^{n} \mathbb{E}_0[\bar{L}_i^2] = \mathbb{E}_0[\bar{L}_1^2]^n,$$

where

$$\begin{aligned}\mathbb{E}_0[\bar{L}_1^2] &= \mathbb{E}_0\Big[\Big(1-\varepsilon+\varepsilon(1-\rho^2)^{-1/2}\exp(-\tfrac{\rho}{2(1-\rho)}U_1^2+\tfrac{\rho}{2(1+\rho)}V_1^2)\Big)^2 D_1\Big]\\ &= (1-\varepsilon)^2\Psi(\sqrt{2\log n})+2(1-\varepsilon)\varepsilon\Psi(\sqrt{2\log n}/\sqrt{1+\rho})\\ &\quad +\varepsilon^2(1-\rho^2)^{-1}\,\mathbb{E}_0[\exp(-\tfrac{\rho}{(1-\rho)}U_1^2)]\,\mathbb{E}_0[\exp(\tfrac{\rho}{(1+\rho)}V_1^2)D_1].\end{aligned}$$

The sum of first two terms is bounded from above by $(1-\varepsilon)^2+2(1-\varepsilon)\varepsilon = 1-\varepsilon^2$. For the third term, we have

$$\mathbb{E}_0[\exp(-\tfrac{\rho}{(1-\rho)}U_1^2)] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{\rho}{1-\rho}u^2-\frac{1}{2}u^2}\mathrm{d}u = \sqrt{\frac{1-\rho}{1+\rho}},$$

and

$$\mathbb{E}_0[\exp(\tfrac{\rho}{(1+\rho)}V_1^2)D_1] = \frac{1}{\sqrt{2\pi}}\int_{-\sqrt{2\log n}}^{\sqrt{2\log n}} e^{\frac{\rho}{1+\rho}v^2-\frac{1}{2}v^2}\mathrm{d}v \le \frac{1}{\sqrt{2\pi}}2\sqrt{2\log n},$$

using the fact that $\rho \le 1$. Hence,

$$\begin{aligned}\mathbb{E}_0[\bar{L}_1^2] &\le 1-\varepsilon^2+\varepsilon^2(1-\rho^2)^{-1}\sqrt{\frac{1-\rho}{1+\rho}}\frac{1}{\sqrt{2\pi}}2\sqrt{2\log n}\\ &\le 1+\varepsilon^2(1-\rho)^{-1/2}(\log n)^{1/2},\end{aligned}$$

when $\rho$ is sufficiently close to 1. This in turn yields the following bound

$$\mathbb{E}_0[\bar{L}^2] \le \big[1+\varepsilon^2(1-\rho)^{-1/2}(\log n)^{1/2}\big]^n \le \exp\big[n\varepsilon^2(1-\rho)^{-1/2}(\log n)^{1/2}\big],$$

so that $\mathbb{E}_0[\bar{L}^2] \leq 1 + o(1)$ when

$$n\varepsilon^2(1-\rho)^{-1/2}(\log n)^{1/2} = o(1).$$

Under the specified parameterization, this happens exactly when $\gamma < 4\beta - 2$. □

In the dense regime, with $\rho$ parameterized as in (3.6), we say that a test achieves the detection boundary if it is asymptotically powerful when $\gamma < 1/2 - \beta$, and in the sparse regime, with $\rho$ parameterized as in (3.7), we say that a test achieves the detection boundary if it is asymptotically powerful when $\gamma > 4(\beta - 1/2)$.

### 3.2.2 The covariance test

Recall that the covariance test rejects for large values of $T_n := \sum_{i=1}^n X_i Y_i$, calibrated under the null where $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are iid standard normal.

**Proposition 4.** *For the testing problem* (3.2)*, the covariance test achieves the detection boundary in the dense regime, while it is asymptotically powerless in the sparse regime.*

*Proof.* We divide the proof into the two regimes.

*Dense regime.* Under $\mathcal{H}_0$, we have

$$\mathbb{E}_0(T_n) = n\mathbb{E}_0(X_1Y_1) = n\mathbb{E}_0(X_1)\mathbb{E}_0(Y_1) = 0,$$
$$\operatorname{Var}_0(T_n) = n\operatorname{Var}_0(X_1Y_1) = n\mathbb{E}_0(X_1^2)\mathbb{E}_0(Y_1^2) = n,$$

so that, by Chebyshev's inequality,

$$\mathbb{P}_0(|T_n| \geq a_n\sqrt{n}) \to 0,$$

for any sequence $(a_n)$ diverging to infinity.

Under $\mathcal{H}_1$, we have

$$\mathbb{E}_1(T_n) = n\mathbb{E}_1(X_1Y_1) = n\varepsilon\rho,$$

$$\mathrm{Var}_1(T_n) = n\mathrm{Var}_1(X_1Y_1) = n(1+2\varepsilon\rho^2 - \varepsilon^2\rho^2) \leq 3n,$$

so that, by Chebyshev's inequality,

$$\mathbb{P}_1(|T_n - n\varepsilon\rho| \geq a_n\sqrt{n}) \to 0.$$

Thus the test with rejection region $\{T_n \geq a_n\sqrt{n}\}$ is asymptotically powerful when

$$\sqrt{n}\varepsilon\rho \geq 2a_n.$$

If we choose $a_n = \log n$, for example, and $\rho$ is parameterized as in (3.6), this happens for $n$ large enough when $\gamma < 1/2 - \beta$.

*Sparse regime.* To prove that the covariance test is asymptotically powerless when $\beta > 1/2$, we show that, under $\mathcal{H}_1$, $T_n$ converges to the same limiting distribution as under $\mathcal{H}_0$.

Under $\mathcal{H}_0$, by the central limit theorem,

$$\frac{T_n}{\sqrt{n}} \rightharpoonup \mathcal{N}(0,1).$$

Under $\mathcal{H}_1$ the distribution of the $(X_i, Y_i)$'s (which remain iid) depends on $n$, but the condition for applying Lyapunov's central limit theorem are satisfied since

$$\mathbb{E}_1[(X_iY_i - \varepsilon\rho)^4] \leq 8(\mathbb{E}_1[(X_iY_i)^4] + (\varepsilon\rho)^4),$$

with $(\varepsilon\rho)^4 \leq 1$ and

$$\mathbb{E}_1[(X_iY_i)^4] \leq \left[\mathbb{E}_1(X_i^8)\mathbb{E}_1(Y_i^8)\right]^{1/2} = \mathbb{E}(Z^8) = \text{const},$$

where $Z \sim \mathcal{N}(0,1)$ and the inequality is Cauchy-Schwarz's, while

$$\text{Var}_1(X_iY_i) = 1 + 2\varepsilon\rho^2 - \varepsilon^2\rho^2 \geq 1,$$

so that the test statistic still converges weakly to a normal distribution,

$$\frac{T_n - \mathbb{E}_1(T_n)}{\sqrt{\text{Var}_1(T_n)}} \rightharpoonup \mathcal{N}(0,1).$$

In the present regime, we have

$$\mathbb{E}_1(T_n) = n\varepsilon\rho, \quad \text{Var}_1(T_n) = n(1 + 2\varepsilon\rho^2 - \varepsilon^2\rho^2),$$

so that $\mathbb{E}_1(T_n)/\sqrt{\text{Var}_1(T_n)} \to 0$ and $\text{Var}_1(T_n) \sim n$, and thus we conclude by Slutsky's theorem that $T_n/\sqrt{n} \rightharpoonup \mathcal{N}(0,1)$. □

*Remark* 1. There are good reasons to consider the covariance test in this specific form since the means and variances are known. It is worth pointing out that the Pearson correlation test, which is more standard in practice since it does not require knowledge of the means or variances, has the same asymptotic power properties.

### 3.2.3 The higher criticism test and the extremes test

Define $U_i = (X_i - Y_i)/\sqrt{2}$, and note that

$$U_1, \ldots, U_n \overset{\text{iid}}{\sim} (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(0, 1-\rho).$$

Seen through the $U_i$'s, the problem becomes that of detecting a sparse contamination where the effect is in the variance. We recently studied this problem in detail [2], extending previous work by Cai et al [11], who considered a setting where the effect is both in the mean and variance. Borrowing from our prior work, we consider a higher criticism test, already defined in (3.3), and an extremes test, which rejects for small values of $\min_i |U_i|$.

**Proposition 5.** *For the testing problem* (3.2)*, the higher criticism test achieves the detection boundary in the dense and sparse regimes.*

*Proof.* Set $\sigma^2 = 1 - \rho$, which is the variance of the contaminated component. In our prior work [2, Prop 3], we showed that the higher criticism test as defined in (3.3) is asymptotically powerful when

(a) $\sigma^2 = n^{-\gamma}$ with $\gamma > 0$ fixed such that $\gamma > 4(\beta - 1/2)$;

(b) $|\sigma^2 - 1| = n^{-\gamma}$ with $\gamma > 0$ fixed such that $\gamma < 1/2 - \beta$.

This can be directly translated into the present setting, yielding the stated result. □

**Proposition 6.** *For the testing problem* (3.2)*, the extremes test is asymptotically powerless when* $\rho$ *is bounded away from 1, while when* $\varepsilon$ *parameterized as in* (3.4) *and* $\rho$ *parameterized as in* (3.7)*, it is asymptotically powerful when* $\gamma > 2\beta$*, and asymptotically powerless when* $\gamma < 2\beta$*.*

*Proof.* This is also a direct corollary from our prior work our prior work [2, Prop 2]. □

Thus the extremes test is grossly suboptimal in the dense regime, while it is suboptimal in the sparse regime due to the fact that $2\beta - 4(\beta - 1/2) = 2 - 2\beta > 0$.

*Remark* 2. The higher criticism and extremes tests are both based on the $U_i$'s. This was convenient as it reduced the problem of testing for independence to the problem of testing for a change in variance (both in a contamination model). However, reducing the original data, meaning the $(X_i, Y_i)$'s, to the $U_i$'s implies a loss of information. Indeed, a lossless reduction would be from the

$(X_i, Y_i)$'s to the $(U_i, V_i)$'s, where $V_i := (X_i + Y_i)/\sqrt{2}$, with joint distribution given in (3.8). It just turns out that ignoring the $V_i$'s does not lead to any loss in first-order asymptotic power.

### 3.2.4 Numerical experiments

We performed some numerical experiments to investigate the finite sample performance of the tests considered here: the likelihood ratio test, the Pearson correlation test (instead of the covariance test from a practical point of view), the extremes test, the higher criticism test, and also a plug-in version of the higher criticism test where the parameters of the bivariate normal distribution (the two means and two variances) are estimated under the null. The sample size $n$ is set large to $n = 10^6$ in order to capture the large-sample behavior of these tests. We tried four sparsity levels, setting $\beta \in \{0.2, 0.4, 0.6, 0.8\}$. The p-values for each test are computed as follows:

(a) For the *likelihood ratio test*, the p-values are estimated based on $10^3$ permutations.

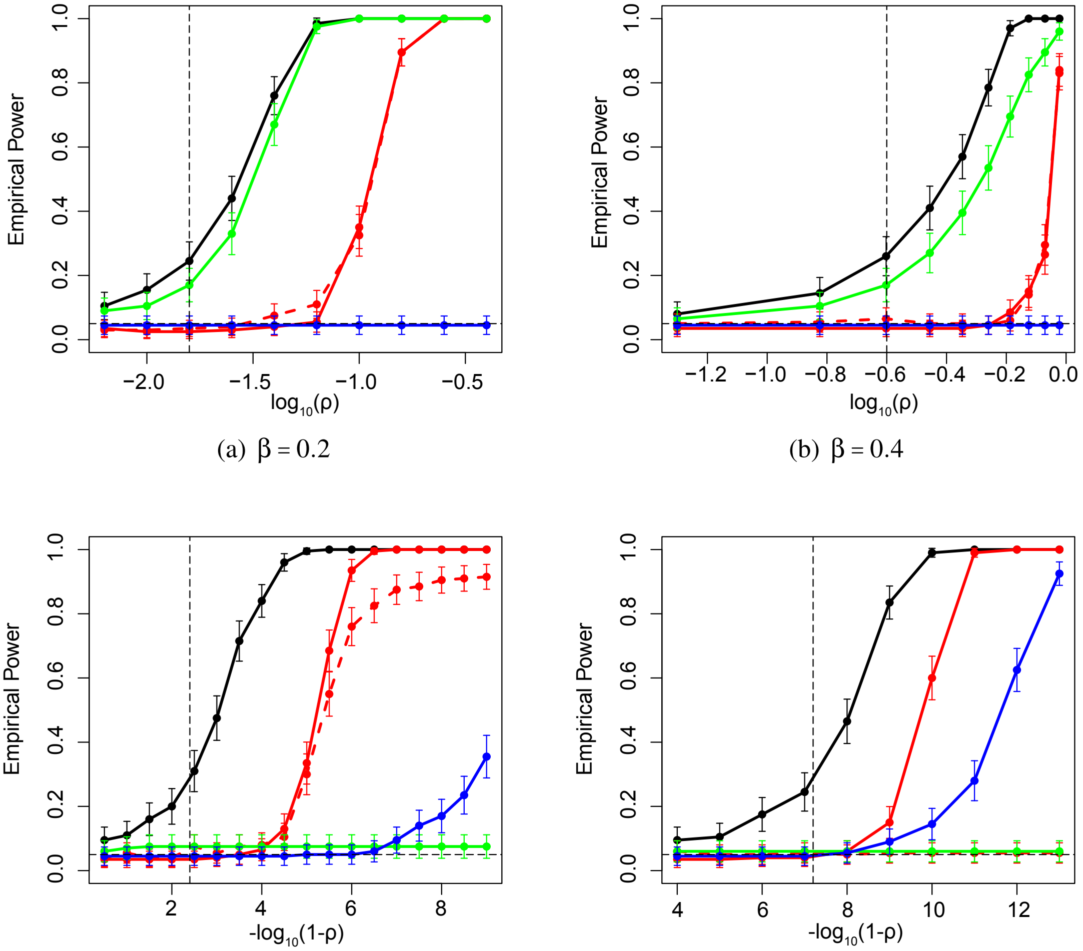
(b) For the *higher criticism test* and the *plug-in higher criticism test*, the p-values are estimated based on 200 permutations.

(c) For the *extremes test*, we used the exact null distribution, which is available in a closed form.

(d) For the *Pearson correlation test*, the p-values are from the limiting distribution.

For each scenario, we repeated the process 200 times and calculated the fraction of p-values smaller than 0.05, representing the empirical power at the 0.05 level.

The results of this experiment are reported in Figure 3.1 and are broadly consistent with the theory developed earlier in this section. Though we show that the higher criticism test is first-order comparable to the likelihood ratio test in the dense regime, even with a large sample, its power is much lower. The Pearson correlation test does better in that regime. The plug-in higher criticism test has a similar performance as the higher criticism test in the dense regime, while it loses some power in the moderately sparse regime, and is powerless in the very sparse regime.



(a) $\beta = 0.2$

(b) $\beta = 0.4$

(c) $\beta = 0.6$

(d) $\beta = 0.8$

**Figure 3.1**: Empirical power comparison with 95% error bars for the likelihood ratio test (black), the Pearson correlation test (green), the extremes test (blue), the higher criticism test (red, solid) and the plug-in higher criticism test (red, dashed). (a) Dense regime where $\beta = 0.2$. (b) Dense regime where $\beta = 0.4$. (c) Sparse regime where $\beta = 0.6$ and $\rho \to 1$. (d) Sparse regime where $\beta = 0.8$ and $\rho \to 1$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $n = 10^6$ and the power curves and error bars are based on 200 replications.

## 3.3 Gaussian mixture copula model

In this section we turn to the Gaussian mixture copula model introduced in (3.5). The setting is thus nonparametric, since the marginal distributions are completely unknown, and standard invariance considerations [34, Ch 6] lead us to consider test procedures that are based on the ranks. For this, we let $R_i$ denote the rank of $X_i$ among $\{X_1, \ldots, X_n\}$, and similarly, we let $S_i$ denote the rank of $Y_i$ among $\{Y_1, \ldots, Y_n\}$. (The ranks are in increasing order, say.)

Although not strictly necessary, we will assume that $F$ and $G$ in (3.5) are strictly increasing and continuous. In that case, the ranks are invariant with respect to transformations of the form $(x,y) \mapsto (p(x), q(y))$ with $p$ and $q$ strictly increasing on the real line. In particular, for the rank tests that follow, this allows us to reduce their analysis under (3.5) to their analysis under (3.1).

### 3.3.1 The covariance rank test

The covariance rank test is the analog of the covariance test of Section 3.2.2. It rejects for large values of $T_n := \sum_i R_i S_i$ (redefined). As is well-known, this is equivalent to rejecting for large values of the Spearman rank correlation.

**Proposition 7.** *For the testing problem* (3.2) *under the model* (3.5)*, the covariance rank test achieves the detection boundary in the dense regime, while it is asymptotically powerless in the sparse regime.*

*Proof.* We again divide the proof into the two regimes.

*Dense regime.* We start by considering the null hypothesis $\mathcal{H}_0$. From [25, Eq 3.11-3.12, Ch 11], we have

$$\begin{aligned} \mathbb{E}_0(T_n) &= n(n+1)^2/4 = n^3/4 + O(n^2), \\ \operatorname{Var}_0(T_n) &= n^2(n-1)(n+1)^2/144 \asymp n^5, \end{aligned} \tag{3.9}$$

so that, using Chebyshev's inequality,

$$\mathbb{P}_0(T_n \geq n^3/4 + a_n n^{5/2}) \to 0,$$

for any sequence $(a_n)$ diverging to infinity.

We now turn to the alternative hypothesis $\mathcal{H}_1$. For convenience, we assume that the ranks run from 0 to $n-1$. This does not change the test procedure since $T_n = -\frac{1}{2}\sum_i (R_i - S_i)^2 + \text{const}$, but makes the derivations somewhat less cumbersome. In particular, we have

$$R_i = \sum_{j=1}^{n} A_{ij}, \quad A_{ij} := \mathbb{I}\{X_i > X_j\},$$
$$S_i = \sum_{j=1}^{n} B_{ij}, \quad B_{ij} := \mathbb{I}\{Y_i > Y_j\},$$

so that

$$T_n = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{ij} B_{ik}.$$

For the expectation, we have

$$\begin{aligned} \mathbb{E}_1(T_n) &= n(n-1)(n-2)\,\mathbb{E}_1[A_{12}B_{13}] + O(n^2) \\ &= n^3\,\mathbb{E}_1[A_{12}B_{13}] + O(n^2). \end{aligned}$$

The expectation is with respect to $(X_1, Y_1), X_2, Y_3$ independent, with $(X_1, Y_1)$ drawn from the mixture (3.1), and $X_2$ and $Y_3$ standard normal. Let $U = (X_1 - X_2)/\sqrt{2}$ and $V = (Y_1 - Y_3)/\sqrt{2}$, so that $\mathbb{E}_1[A_{12}B_{13}] = \mathbb{P}_1(U > 0, V > 0)$. We note that $(U, V)$ is bivariate normal with standard marginals. Moreover, when $(X_1, Y_1)$ comes from the main component, $U$ and $V$ are uncorrelated, and therefore independent; while when $(X_1, Y_1)$ comes from the contaminated component, $U$ and

$V$ have correlation $\rho/2$. Therefore,

$$\mathbb{E}_1[A_{12}B_{13}] = (1-\varepsilon)\Lambda(0) + \varepsilon\Lambda(\rho/2),$$

where $\Lambda(\rho) = \mathbb{P}(U > 0, V > 0)$ under $(U,V) \sim \mathcal{N}(0,\Sigma_\rho)$. We immediately have $\Lambda(0) = 1/4$, and in general,[2]

$$\Lambda(\rho) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}(\rho).$$

We conclude that

$$\begin{aligned}\mathbb{E}_1(T_n) &= n^3\left[\tfrac{1}{4} + \tfrac{1}{2\pi}\varepsilon\sin^{-1}(\rho/2)\right] + O(n^2)\\ &\geq \tfrac{1}{4}n^3 + \tfrac{1}{4\pi}n^3\varepsilon\rho + O(n^2),\end{aligned}$$

using the fact that $\sin^{-1}(a) \geq a$ for all $a \geq 0$. For the variance, we start with the second moment

$$\begin{aligned}\mathbb{E}_1(T_n^2) &= n(n-1)\cdots(n-5)\,\mathbb{E}_1[A_{12}B_{13}A_{45}B_{46}] + O(n^5)\\ &= n^6\,\mathbb{E}_1[A_{12}B_{13}A_{45}B_{46}] + O(n^5),\end{aligned}$$

which then implies that

$$\begin{aligned}\mathrm{Var}_1(T_n) &= n^6\,\mathbb{E}_1[A_{12}B_{13}A_{45}B_{46}] + O(n^5) - \left[n^3\,\mathbb{E}_1[A_{12}B_{13}] + O(n^2)\right]^2\\ &= O(n^5),\end{aligned}$$

the same bound we had for $\mathrm{Var}_0(T_n)$. Thus, by Chebyshev's inequality, we have

$$\mathbb{P}_1\left(T_n \leq \tfrac{1}{4}n^3 + \tfrac{1}{4\pi}n^3\varepsilon\rho - a_n n^{5/2}\right) \to 0,$$

[2] This identity is well-known, and not hard to prove (https://math.stackexchange.com/questions/255368/getting-px0-y0-for-a-bivariate-distribution). It also appears, for example, in [55, Lem 1].

for any sequence $(a_n)$ diverging to infinity.

We consider the test with rejection region $\{T_n \geq n^3/4 + a_n n^{5/2}\}$. Our analysis implies that this test is asymptotically powerful when

$$n^3 \varepsilon \rho / 4\pi \geq 2 a_n n^{5/2},$$

If we choose $a_n = \log n$, for example, and $\rho$ is parameterized as in (3.6), this happens for $n$ large enough when $\gamma < 1/2 - \beta$.

*Sparse regime.* To prove that the covariance rank test is asymptotically powerless when $\beta > 1/2$, similarly as the covariance test, we show that, under $\mathcal{H}_1$, $T_n$ converges to the same limiting distribution as under $\mathcal{H}_0$. Under $\mathcal{H}_0$, we have [25, Ch 11],

$$\frac{T_n - \zeta_n}{\tau_n} \rightharpoonup \mathcal{N}(0,1), \quad n \to \infty, \tag{3.10}$$

where $\zeta_n := \mathbb{E}_0(T_n)$ and $\tau_n^2 := \mathrm{Var}_0(T_n)$. We place ourselves under $\mathcal{H}_1$, and show that (3.10) continues to hold. For this we use a simple coupling. We couple $T_n$ with a new statistic $T_n'$, defined just like $T_n$, except that, for each pair $(X_i, Y_i)$ drawn from the contaminated component, we replace $Y_i$ by $Y_i' \sim \mathcal{N}(0,1)$ independent of $X_i$ and any other variable. Let $M$ denote the number of pairs drawn from the contaminated component, and note that $M$ is random, having the binomial distribution with parameters $(n, \varepsilon)$. It's not hard to show that $|T_n - T_n'| \leq M n^2$, so that $|T_n - T_n'| = O_P(n^3 \varepsilon)$. And by construction, $T_n'$ has the same distribution as $T_n$ under $\mathcal{H}_0$. We use this in what follows

$$\frac{T_n - \zeta_n}{\tau_n} = \frac{T_n' - \zeta_n}{\tau_n} + \frac{T_n - T_n'}{\tau_n},$$

where, on the RHS, the first term converges weakly to the standard normal distribution, while the second term is $= O_P(n^3 \varepsilon / \tau_n) = o_P(1)$, since $\varepsilon = n^{1-\beta}$ with $\beta > 1/2$ and $\tau_n \asymp n^{5/2}$ by (3.9). We thus conclude that (3.10) with an application of Slutsky's theorem. □

### 3.3.2 The higher criticism rank test

The analog of the higher criticism test of (3.3) is a higher criticism based on the pairwise differences in ranks, $D_i := |R_i - S_i|$. To be specific, we define

$$\mathrm{HC}_{\mathrm{rank}} = \max_{0 \le t \le n/2} \frac{\sum_{i=1}^n \mathbb{I}\{D_i \le t\} - nu(t)}{\sqrt{nu(t)(1-u(t))}},$$

where $u(t)$ is the probability $\mathbb{P}_0(D_i \le t)$, which can be expressed in closed form as

$$u(t) = \frac{n^2 - (n-t)(n-t-1)}{n^2} = \frac{n(2t+1) - t(t+1)}{n^2}.$$

Note that in this definition the denominator is only an approximation to the standard deviation of the numerator. The standard deviation has a closed-form expression which can be derived from a more general result of Hoeffding [27, Th 2], but it is cumbersome and relatively costly to compute (although its computation is only done once for each $n$). Also, there is a fair amount of flexibility in the choice of range of thresholds $t$ considered. This particular choice seems to work well enough. As any other rank test, it is calibrated by permutation (or Monte Carlo if there are no ties in the data).

**Theorem 5.** *For the testing problem* (3.2) *under the model* (3.5)*, the higher criticism rank test achieves the detection boundary in the dense and in the moderately sparse regimes.*

*Proof.* As usual, we first control the test statistic under the null, and then analyze its behavior under the alternative.

*Under the null hypothesis*

We start with the situation under the null hypothesis $\mathcal{H}_0$, where we show that $\mathrm{HC}_{\mathrm{rank}}$ is of order at most $O(\log n)$ based on the concentration inequality for randomly permuted sums. Fixing critical value $t$, define

$$a_{i,j} = \mathbb{I}\{|i-j| \le t\}, \quad \text{for } 1 \le i, j \le n.$$

Since $X$ is independent of $Y$, as we are under the null, we have that

$$\Delta(t) := \sum_{i=1}^{n} \mathbb{I}\{D_i \le t\} \tag{3.11}$$

has the same distribution as $A_n := \sum_{i=1}^n a_{i,\pi_n(i)}$ when $\pi_n$ is a uniformly distributed random permutation of $[n] := \{1, \cdots, n\}$. Note that

$$\mathbb{E}(A_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} = \frac{n(2t+1) - t(t+1)}{n} = nu(t). \tag{3.12}$$

By [15, Prop 1.1],

$$\mathbb{P}(|A_n - \mathbb{E}(A_n)| \ge b) \le 2\exp\left(-\frac{b^2}{4\mathbb{E}(A_n) + 2b}\right). \tag{3.13}$$

This implies that, for $q \ge 1$,

$$\begin{aligned}
&\mathbb{P}_0\left(\Delta(t) \ge nu(t) + q\sqrt{nu(t)(1-u(t))}\right) \\
&\le 2\exp\left(-\frac{q^2 nu(t)(1-u(t))}{4nu(t) + 2q\sqrt{nu(t)(1-u(t))}}\right) \\
&\le 2\exp\left(-q/c_1\right),
\end{aligned}$$

for some other constant $c_1 > 0$, using the fact that $1/n \le u(t) \le 3/4 + 1/2n$ when $0 \le t \le n/2$, which is the range of $t$'s we are considering. Hence, choosing $q = 2c_1 \log n$ and using the union bound, we have

$$\begin{aligned}\mathbb{P}_0(\mathrm{HC}_{\mathrm{rank}} \geq q) &\leq \sum_{t \leq n/2} \mathbb{P}_0\left(\Delta(t) \geq nu(t) + q\sqrt{nu(t)(1-u(t))}\right) \\ &\leq 2(n/2+1)\exp\left(-q/c_1\right) \asymp 1/n \to 0.\end{aligned}$$

*Under the alternative hypothesis*

We now consider the alternative $\mathcal{H}_1$, and show that $\mathrm{HC}_{\mathrm{rank}} \gg \log n$ in probability under the stated condition. For this, it suffices to find some $t = t_n \leq n/2$ such that, for some $q = q_n \gg \log n$,

$$\Delta(t) \geq nu(t) + q\sqrt{nu(t)(1-u(t))}, \tag{3.14}$$

with probability tending to 1 (under $\mathcal{H}_1$).

Since rank-based methods are invariant with respect to increasing transformations, in the following analysis we simply assume that $F = G = \Phi$.

*Dense regime.* Define $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \leq x\}$ and $\hat{G}(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i \leq y\}$. These empirical distribution functions are useful because, by definition, $R_i = n\hat{F}(X_i)$ and $S_i = n\hat{G}(Y_i)$, so that

$$\begin{aligned}D_i/n &= |R_i - S_i|/n \\ &= |\hat{F}(X_i) - \hat{G}(Y_i)| \\ &\leq |\hat{F}(X_i) - \Phi(X_i)| + |\Phi(X_i) - \Phi(Y_i)| + |\Phi(Y_i) - \hat{G}(Y_i)| \\ &\leq \underbrace{|\Phi(X_i) - \Phi(Y_i)|}_{M_i} + \underbrace{\|\hat{F} - \Phi\|_\infty + \|\hat{G} - \Phi\|_\infty}_{K}.\end{aligned}$$

This gives

$$\Delta(t) \geq \mathbb{I}\{K \leq k/n\}\Lambda(t), \quad \Lambda(t) := \sum_{i=1}^n \mathbb{I}\{M_i \leq (t-k)/n\}. \tag{3.15}$$

By the Dvoretzky-Kiefer-Wolfowitz (DKW) concentration inequality, there is a universal constant $c_0$ such that, for any $b \geq 0$,

$$\mathbb{P}(K \geq b) \leq c_0 \exp(-nb^2/c_0).$$

We choose $k = (\log n)\sqrt{n}$, and with that choice we have that $\mathbb{I}\{K \leq k/n\} = 1 - Q_n$, where $Q_n$ is Bernoulli with parameter bounded by $\eta := c_0 \exp(-(\log n)^2/c_0)$ (so that $Q_n = O_P(\eta)$).

As for the sum, the $M_i$ are iid, and for an observation $(X_i, Y_i)$ that comes from the null component, $X_i, Y_i$ are iid standard normal, while when it comes from the contaminated component, $X_i, Y_i$ are still marginally standard normal but no longer independent: $Y_i = \sqrt{1-\rho^2}\,\tilde{Y}_i + \rho X_i$, where $\tilde{Y}_i$ is independent of $X_i$ and also standard normal. We thus have

$$\mathbb{P}_1(M_i \leq s) = (1-\varepsilon)v_s(0) + \varepsilon v_s(\rho),$$

where

$$v_s(\rho) := \mathbb{E}[f_s(Z, \sqrt{1-\rho^2}Z' + \rho Z)],$$

where in the expectation $Z, Z'$ are iid standard normal, and $f_s(z, z') := \mathbb{I}\{|\Phi(z) - \Phi(z')| \leq s\}$ is bounded and measurable. Elementary calculations show that $v_s(0) = 1 - (1-s)^2$, and an application of Lemma 1 shows that $v_s$ is infinitely differentiable, with derivative at 0 equal to $\mathbb{E}[f_s(Z, Z')ZZ']$, and second derivative uniformly bounded over $[-1/2, 1/2]$ by some numerical constant, say $c_2$, independently of $s$. Recalling that $\rho$ is small in the present regime, a Taylor development based on the above gives

$$v_s(\rho) \geq 1 - (1-s)^2 + v_s'(0)\rho - c_2\rho^2/2, \quad \rho \in [-1/2, 1/2].$$

In the dense regime, remember that $0 < \beta < 1/2$ and $\rho = n^{-\gamma}$. We place ourselves above the detection boundary, meaning that we fix $\gamma < 1/2 - \beta$. Here we choose $t = n/2$ (assumed to be an integer for convenience), let $s = (t-k)/n = 1/2 - k/n$. We note that $v_s'(0)$ is continuous in $s$ (by dominated convergence), and because $s \to 1/2$ in our setting, we have

$$v_s'(0) \to v_{1/2}'(0) = \mathbb{E}[\mathbb{I}\{|\Phi(Z) - \Phi(Z')| \leq 1/2\}ZZ'] =: c_1 > 0.$$

Indeed, using the fact that

$$|\Phi(z) - \Phi(z')| \leq 1/2 \;\Leftrightarrow\; (\Phi(z) - 1/2) \vee 0 \leq \Phi(z') \leq (\Phi(z) + 1/2) \wedge 1,$$

with $\Phi(z) \leq 1/2$ if and only if $z \leq 0$, we have

$$\begin{aligned} c_1 = &\int_0^\infty \underbrace{\int_{\Phi^{-1}(\Phi(z)-1/2)}^\infty \phi(z')\mathrm{d}z'}_{>0} \underbrace{\phi(z)z}_{>0}\mathrm{d}z \\ &+ \int_{-\infty}^0 \underbrace{\int_{-\infty}^{\Phi^{-1}(\Phi(z)+1/2)} \phi(z')\mathrm{d}z'}_{<0} \underbrace{\phi(z)z}_{<0}\mathrm{d}z, \end{aligned}$$

where the inner integrals are positive by the fact that $\phi$ is symmetric, and the inequalities are indeed strict except when $z = 0$.

Thus, eventually (as $n \to \infty$),

$$v_s(\rho) \geq 1 - (1-s)^2 + (c_1/2)\rho.$$

Thus, an application of Chebyshev's inequality gives

$$\Lambda(n/2) \geq n\big[(1-\varepsilon)v_s(0) + \varepsilon v_s(\rho)\big] + O_P(\sqrt{n}).$$

Putting everything together, we have

$$\begin{aligned}
&\Delta(t) - nu(t) \\
&= (1 + O_P(\eta))n\big[(1-\varepsilon)v_s(0) + \varepsilon v_s(\rho)\big] + O_P(\sqrt{n}) - nu(t) \\
&\ge n\big[1 - (1-(t-k)/n)^2 - u(t)\big] + n\varepsilon(c_1/2)\rho + O_P(n\eta) + O_P(\sqrt{n}) \\
&= n\varepsilon(c_1/2)\rho + O_P((\log n)\sqrt{n}),
\end{aligned}$$

using the fact that $\eta = o(1/n^2)$. For (3.14) to hold it thus suffices that $n\varepsilon\rho \gg (\log n)\sqrt{n}$, which is the case since $n\varepsilon\rho = n^{1-\beta-\gamma}$ with $1-\beta-\gamma > 1/2$.

*Moderately sparse regime.* Let $I_0$ and $I_1$ index the observations coming from the null and contaminated components, respectively. We have

$$\Delta(t) = \sum_{i \in I_0} \mathbb{I}\{D_i \le t\} + \sum_{i \in I_1} \mathbb{I}\{D_i \le t\} =: \Delta_0(t) + \Delta_1(t). \tag{3.16}$$

We lower bound both terms on the right-hand side, starting with $\Delta_0(t)$. To do this, we consider a slightly smaller threshold, specifically $t_0 = (1-\omega)t$ with $\omega = o(1)$ specified below, and compare $\Delta_0(t)$ with $\Delta^0(t_0) := \sum_{i \in I_0} \mathbb{I}\{D_i^0 \le t_0\}$, where $D_i^0 := |R_i^0 - S_i^0|$ with $R_i^0$ denoting the rank of $X_i$ among $\{X_j : j \in I_0\}$ and $S_i^0$ denoting the rank of $Y_i$ among $\{Y_j : j \in I_0\}$. Conditional on $|I_0| = n_0$, $\Delta^0(t_0)$ has the same distribution as $\Delta(t_0)$ in (3.11) under the null hypothesis but with $n$ replaced by $n_0$, so that from (3.12) we deduce that it has expectation

$$\mu := (n_0(2t_0+1) - t_0(t_0+1))/n_0,$$

and from (3.13) that

$$\Delta^0(t_0) \ge \mu - 8(\log n)\sqrt{\mu \vee \log n}$$

with probability at least $1 - 2/n$ when $n$ is large enough. (Again, this is conditional on $|I_0| = n_0$.)

Because $\varepsilon \ll n^{-1/2}$ in the present regime, we have $|I_0| \geq n - (\log n)\sqrt{n}$ with probability at least $1 - 1/n$ when $n$ is large enough. Also, we will choose $t$ below such that $\sqrt{n} \ll t \ll n$, and $\omega$ such that $\omega \ll 1$, so that $t_0 \sim t$. Together, this implies that

$$\Delta^0(t_0) \geq 2t_0 + 1 - \frac{t_0(t_0+1)}{n - (\log n)\sqrt{n}} - 8(\log n)\sqrt{2t_0+1} = 2t_0 - \frac{t_0^2}{n} - O((\log n)\sqrt{t}),$$

eventually, with probability at least $1 - 3/n$.

We now claim that, with probability tending to 1, $\Delta_0(t) \geq \Delta^0(t_0)$. Indeed, by definition of the ranks $R_i$ and modified ranks $R_i^0$, we have

$$R_i - R_i^0 = \sum_{j \in I_1} \mathbb{I}\{X_j \leq X_i\} = |I_1|\hat{F}_1(X_i),$$

where $\hat{F}_1(x) := \frac{1}{|I_1|}\sum_{j \in I_1} \mathbb{I}\{X_j \leq x\}$ is the empirical distribution function associated with the contaminated $X$ observations. In particular, when $|I_0| = n_0$, so that $|I_1| = n - n_0 =: n_1$, we have

$$\left|R_i - R_i^0 - n_1\Phi(X_i)\right| \leq n_1\|\hat{F}_1 - \Phi\|_\infty,$$

valid for all $i \in I_0$. At the same time, and with analogous notation, we also have

$$\left|S_i - S_i^0 - n_1\Phi(Y_i)\right| \leq n_1\|\hat{G}_1 - \Phi\|_\infty,$$

valid for all $i \in I_0$. Combining these, we obtain

$$\underbrace{|R_i - S_i|}_{D_i} \leq \underbrace{|R_i^0 - S_i^0|}_{D_i^0} + n_1|\Phi(X_i) - \Phi(Y_i)| + n_1\underbrace{\left(\|\hat{F}_1 - \Phi\|_\infty + \|\hat{G}_1 - \Phi\|_\infty\right)}_{=:K_1},$$

valid for all $i \in I_0$. Letting $\hat{F}_0$ denote the empirical distribution function of $\{X_i : i \in I_0\}$ and $\hat{G}_0$

denote that of $\{Y_i : i \in I_0\}$, we have

$$|\Phi(X_i) - \Phi(Y_i)| \le \underbrace{|\hat{F}_0(X_i) - \hat{G}_0(Y_i)|}_{D_i^0/n_0} + \underbrace{\|\hat{F}_0 - \Phi\|_\infty + \|\hat{G}_0 - \Phi\|_\infty}_{=:K_0}, \tag{3.17}$$

valid for all $i \in I_0$. Note that this is conditional on $|I_0| = n_0$ and that the distributions of $K_0$ and $K_1$ depend (implicitly) on $n_0$ (and $n_1$). We conclude that, conditional on $|I_0| = n_0$, for any $i \in I_0$,

$$D_i \le (n/n_0)D_i^0 + n_1(K_0 + K_1). \tag{3.18}$$

Applying the DKW inequality with the tight constant, we have that $K_0 \le (\log n)/\sqrt{n_0}$ and $K_1 \le (\log n)/\sqrt{n_1}$ with probability at least $1 - 2/n$ when $n$ is large enough, and when this is the case, $D_i \le (n/n_0)D_i^0 + 2(\log n)\sqrt{n_1}$, assuming that $n_0 \ge n_1$. This is given $|I_0| = n_0$ and (therefore) $|I_1| = n_1$, and we also know that $|I_0| \ge n - (\log n)\sqrt{n}$ and $|I_1| \le 2n\varepsilon$ with probability at least $1 - 1/n$ when $n$ is large enough. (We are using that $|I_1| \sim \text{Bin}(n, \varepsilon)$ with $n\varepsilon = n^{1-\beta}$ with $\beta < 1$.) Hence, with probability at least $1 - 3/n$,

$$D_i \le \frac{nD_i^0}{n - (\log n)\sqrt{n}} + 2(\log n)\sqrt{2n\varepsilon},$$

for any $i \in I_0$. In particular, if we choose $\omega = (\log n)^2 \max\left(1/\sqrt{n}, \sqrt{n\varepsilon}/t\right)$, then, with probability at least $1 - 2/n$ when $n$ is large enough, $D_i^0 \le t_0$ implies that $D_i \le t$ for any $i \in I_0$, implying that $\Delta_0(t) \ge \Delta^0(t_0)$.

We thus conclude that

$$\Delta_0(t) \ge 2t_0 - t_0^2/n - O_P((\log n)\sqrt{t}).$$

As for $\Delta_1(t)$, as in (3.15), we have

$$\Delta_1(t) \geq \mathbb{I}\{K \leq k/n\}\Lambda_1(t), \quad \Lambda_1(t) := \sum_{i \in I_1} \mathbb{I}\{M_i \leq (t-k)/n\}.$$

We choose $k = (\log n)\sqrt{n}$ as we did before, so that $\mathbb{I}\{K \leq k/n\} = 1 + O_P(\eta)$, with the same $\eta$ defined previously. As for the sum, $\Lambda_1(t)$ has the same distribution as $\sum_{i=1}^{B} \mathbb{I}\{\tilde{M}_i \leq (t-k)/n\}$, with $B$ binomial with parameters $(n, \varepsilon)$ and $\tilde{M}_i = |\Phi(\tilde{X}_i) - \Phi(\tilde{Y}_i)|$ with $(\tilde{X}_i, \tilde{Y}_i)$ iid normal with standard normal marginals and correlation $\rho$. In particular,

$$\tilde{M}_i \leq \frac{1}{\sqrt{2\pi}}|\tilde{X}_i - \tilde{Y}_i| =: \frac{1}{\sqrt{\pi}}|\tilde{U}_i|,$$

by the fact that $\Phi$ has derivative bounded by $1/\sqrt{2\pi}$ everywhere, and where $\tilde{U}_i \sim \mathcal{N}(0, 1-\rho)$, and simple calculations give

$$v(s) := \mathbb{P}(\tilde{M}_i \leq s) \geq \Psi\left(\frac{\sqrt{\pi}s}{\sqrt{1-\rho}}\right) =: \lambda(s), \quad s \in [0,1].$$

We thus have

$$\mathbb{E}_1(\Lambda_1(t)) = n\varepsilon v((t-k)/n),$$

and

$$\mathrm{Var}_1(\Lambda_1(t)) = \mathrm{Var}(B)v((t-k)/n)^2 + \mathbb{E}(B)v((t-k)/n) \leq 2n\varepsilon v((t-k)/n),$$

and applying Chebyshev's inequality, we thus have

$$\begin{aligned}\Lambda_1(t) &= n\varepsilon v((t-k)/n) + O(\sqrt{n\varepsilon v((t-k)/n)}) \\ &\geq (1 + o_P(1))n\varepsilon\lambda((t-k)/n)),\end{aligned}$$

as long as the right-hand side diverges.

In the moderately sparse regime, remember that $1/2 < \beta < 3/4$ and $\rho = 1 - n^{-\gamma}$. We place ourselves just above the detection boundary, meaning that we fix $\gamma > 4(\beta - 1/2)$. We focus on the harder sub-case where, in addition, $\gamma < 2\beta$. In that case, we can fix $a$ such that $1/2 > a > \gamma/2$ and $1/2 - \beta + \gamma/2 - a/2 > 0$, and set $t = \lfloor n^{1-a} \rfloor$. Note that such a real number $a$ exists, and that $t \le n/2$ with $t \gg k$. We also have $n\varepsilon = n^{1-\beta}$ and $u(t) \asymp t/n \asymp n^{-a}$, as well as

$$\lambda((t-k)/n) = \Psi\left(\frac{\sqrt{\pi}(t-k)}{n\sqrt{1-\rho}}\right) \asymp n^{\gamma/2-a}, \quad \text{since } \frac{\sqrt{\pi}(t-k)}{n\sqrt{1-\rho}} \asymp n^{\gamma/2-a} \to 0,$$

and $\Psi$ is differentiable at 0 with positive derivative. In particular, $n\varepsilon\lambda((t-k)/n) \asymp n^{1-\beta+\gamma/2-a} \to \infty$. Putting everything together, we have

$$\begin{aligned}\Delta(t) - nu(t) &\ge 2t_0 - t_0^2/n - O_P((\log n)\sqrt{t}) + (1 + o_P(1))n\varepsilon\lambda((t-k)/n) \\ &\quad - \big(2t + 1 - t(t+1)/n\big) \\ &= -O_P((\log n)\sqrt{t}) + n\varepsilon(1 + o_P(1))\lambda((t-k)/n),\end{aligned}$$

after some simplifications, using the definition of $\omega$ above and the fact that $\sqrt{n} \ll t \ll n$. For (3.14) to hold, it is thus enough to have $n\varepsilon\lambda((t-k)/n) \gg (\log n)\sqrt{t}$, which is the case since

$$\frac{n\varepsilon\lambda((t-k)/n)}{\sqrt{t}} \asymp \frac{n^{1-\beta+\gamma/2-a}}{n^{1/2-a/2}} = n^{1/2-\beta+\gamma/2-a/2},$$

with $1/2 - \beta + \gamma/2 - a/2 > 0$ by our choice of $a$. □

**Lemma 1.** *Let $A, B$ be iid standard normal, and for $f : \mathbb{R}^2 \to [0,1]$ measurable and $r \in [-1,1]$, define $\Gamma_f(r) = \mathbb{E}[f(A, \sqrt{1-r^2}B + rA)]$. Then $\Gamma_f$ is infinitely differentiable, with $\Gamma_f'(0) = \mathbb{E}[f(A,B)AB]$, and with $\sup_{|r|\le 1/2} |\Gamma_f''(r)|$ bounded by some numerical constant (independent of $f$).*

*Proof.* We have

$$\Gamma_f(r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(a,b)\phi(a,b;r)\mathrm{d}a\mathrm{d}b,$$

where

$$\phi(a,b;r) := \frac{\exp\big[-(a^2-2rab+b^2)/(2-2r^2)\big]}{2\pi\sqrt{1-r^2}}.$$

An application of the dominated convergence theorem allows us to differentiate under the integral at will. In particular,

$$\Gamma_f^{(k)}(r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(a,b)\partial_r^k\phi(a,b;r)\mathrm{d}a\mathrm{d}b,$$

Elementary calculations show that $\partial_r\phi(a,b;0) = (2\pi)^{-1}ab\exp[-(a^2+b^2)/2]$. We also obtain

$$|\Gamma_f''(r)| \le \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\partial_r^2\phi(a,b;r)|\mathrm{d}a\mathrm{d}b,$$

which is easily seen to uniformly bounded for $|r| \le 1/2$. □

It is natural to wonder whether the higher criticism rank test has some power in the very sparse regime. The following indicates that it is powerless in that regime.

**Proposition 8.** *Consider the very sparse regime in the most extreme case where* $\rho = 1$. *In that setting, any test that rejects for large values of* $\Delta(t) := \sum_{i=1}^n \mathbb{I}\{D_i \le t\}$ *(where the threshold* $t$ *is allowed to vary with* $n$*) is asymptotically powerless.*

*Proof.* By a compactness argument, we may assume that either $t \to \infty$ or $t$ is constant (as $n$ varies). We start with the former and address the latter at the end. We focus on the case where $t \ll n$, as the case where $t \asymp n$ can be dealt with in a very similar fashion.

*Under the null hypothesis*

We first consider the behavior of $\Delta(t)$ under the null hypothesis, and argue that $\Delta(t)$ is asymptotically normally distributed. This is based on an application of a combinatorial central limit theorem due to Hoeffding [27]. Remember that under $\mathcal{H}_0$, $\Delta(t)$ has the distribution of $A_n = \sum_{i=1}^n a_{i,\pi_n(i)}$

when $\pi_n$ is a uniformly distributed random permutation of $[n]$ and $a_{i,j} = \mathbb{I}\{|i-j| \le t\}$. We saw that

$$\mathbb{E}(A_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} = \frac{n(2t+1)-t(t+1)}{n} = nu(t),$$

and, as derived in [27], we also have

$$\operatorname{Var}(A_n) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,j}^2,$$

where

$$d_{i,j} = a_{i,j} - \frac{1}{n}\sum_{g=1}^{n} a_{g,j} - \frac{1}{n}\sum_{h=1}^{n} a_{i,h} + \frac{1}{n^2}\sum_{g=1}^{n}\sum_{h=1}^{n} a_{g,h}.$$

[27, Th 3] implies that $A_n$ is asymptotically normal when

$$\frac{\max_{i,j\in[n]} d_{i,j}^2}{\frac{1}{n^2}\sum_{i\in[n]}\sum_{j\in[n]} d_{i,j}^2} \to 0.$$

Elementary but somewhat tedious calculations yield that this is the case if and only if $t \to \infty$, which we assume. Further elementary calculations, in part similar to some appearing in the proof of Theorem 5, yield that

$$\frac{\operatorname{Var}(A_n)}{nu(t)(1-u(t))} \to 1,$$

We thus have, under the null hypothesis,

$$\frac{\Delta(t) - nu(t)}{\sqrt{nu(t)(1-u(t))}} \rightharpoonup \mathcal{N}(0,1),$$

and therefore, together with the fact that $1 \ll t \ll n$, we conclude that

$$\frac{\Delta(t) - 2t + t^2/n}{\sqrt{2t}} \rightharpoonup \mathcal{N}(0,1), \tag{3.19}$$

again under the null hypothesis.

*Under the alternative hypothesis*

We now consider the alternative, again in the very sparse regime and in the most advantageous case where $\rho = 1$, and show that the same weak limit holds. For this, we follow the arguments of the proof of Theorem 5 in the moderately sparse regime, although in the reverse direction so-to-speak. We use the same notation.

Starting from the decomposition (3.16), we have

$$\frac{\Delta(t) - 2t + t^2/n}{\sqrt{2t}} = \frac{\Delta_0(t) - 2t + t^2/n}{\sqrt{2t}} + \frac{\Delta_1(t)}{\sqrt{2t}}. \tag{3.20}$$

In what follows, we first show that the first term on the RHS is asymptotically standard normal, and then we show that the second term converges to 0 in probability.

*First term in* (3.20). For $i \in I_0$, as in (3.18) but in reverse, we have

$$\begin{aligned} D_i^0 &\le (1 + |I_1|/n) D_i + |I_1|(K_0 + K_1) \\ &\le \big(1 + \varepsilon + (\log n)\sqrt{\varepsilon/n}\big) D_i + (\log n)\sqrt{n\varepsilon}, \end{aligned}$$

with probability tending to 1 uniformly over $i \in I_0$. Assuming this is true, then $D_i \le t$ implies that

$$\begin{aligned} D_i^0 &\le \big(1 + \varepsilon + (\log n)\sqrt{\varepsilon/n}\big) t + (\log n)\sqrt{n\varepsilon} \\ &\le t_0 := (1 + \varepsilon) t + 2(\log n)\sqrt{n\varepsilon}. \end{aligned}$$

Hence, with probability tending to 1,

$$\Delta_0(t) \le \Delta^0(t_0).$$

As before, conditional on $|I_0| = n_0$, $\Delta^0(t_0)$ has the same distribution as $\Delta(t_0)$ in (3.11) under the null hypothesis but with $n$ replaced by $n_0$. This, the fact that $|I_0| \ge n - O_P(\sqrt{n})$, and (3.19), implies that

$$\frac{\Delta^0(t_0) - 2t_0 + t_0^2/n}{\sqrt{2t_0}} \rightharpoonup \mathcal{N}(0,1).$$

We used the fact that $t_0^2/n \le t_0^2/|I_0| \le t_0^2/(n - O(\sqrt{n}))$, which implies that

$$\frac{t_0^2/|I_0|}{\sqrt{t_0}} = \frac{t_0^2/n}{\sqrt{t_0}} + \underbrace{O(t_0\sqrt{t_0}/n\sqrt{n})}_{o(1)},$$

where the $O$ term is $o(1)$ by the fact that $t_0/n = o(1)$. Continuing, with probability tending to 1, we have

$$\begin{aligned}
\frac{\Delta_0(t) - 2t + t^2/n}{\sqrt{2t}} &\le \frac{\Delta^0(t_0) - 2t + t^2/n}{\sqrt{2t}} \\
&= \sqrt{t_0/t}\,\frac{\Delta^0(t_0) - 2t_0 + t_0^2/n}{\sqrt{2t_0}} + \frac{2t_0 - t_0^2/n - 2t + t^2/n}{\sqrt{2t}} \\
&\rightharpoonup \mathcal{N}(0,1), \qquad (3.21)
\end{aligned}$$

whenever $t_0/t \to 1$ and $(t_0 - t)/\sqrt{t} \to 0$ (using the fact that $t \le t_0 \ll n$). This is the case exactly when $t \gg (\log n)^2 n\varepsilon$.

We now consider the complementary case. In fact, what follows applies when $t \le \sqrt{n}$. We use a slightly different strategy. Recall that, for $i \in I_0$,

$$R_i - R_i^0 = \sum_{j \in I_1} \mathbb{I}\{X_j \le X_i\},$$

$$S_i - S_i^0 = \sum_{j \in I_1} \mathbb{I}\{Y_j \le Y_i\},$$

and combined with the triangle inequality, and recalling that $X_j = Y_j$ when $j \in I_1$, we have

$$|D_i - D_i^0| \le W_i := \sum_{j \in I_1} \mathbb{I}\{X_i \wedge Y_i \le X_j \le X_i \vee Y_i\}.$$

Consider the event

$$\Omega = \left\{|I_1| \le 2n\varepsilon, K_0 \le (\log n)/\sqrt{n}\right\},$$

which happens with probability tending to one. Given $\Omega$, we have

$$\begin{aligned}\{D_i \le t\} &= \{D_i \le t, D_i^0 \le t\} \cup \{D_i \le t, D_i^0 > t\}\\ &\subset \{D_i^0 \le t\} \cup \{W_i \ge D_i^0 - t, 2n\varepsilon + t \ge D_i^0 > t\},\end{aligned}$$

using the fact that $D_i^0 \le D_i + |I_1|$, so that

$$\Delta_0(t) \le \Delta^0(t) + \sum_{i \in I_0} \mathbb{I}\{W_i \ge D_i^0 - t, 2n\varepsilon + t \ge D_i^0 > t\}. \tag{3.22}$$

Given $\{(X_k, Y_k) : k \in I_0\}$, and conditional on $(|I_0|, |I_1|) = (n_0, n_1)$, $W_i$ is binomial with parameters $n_1$ and $P_i := |\Phi(X_i) - \Phi(Y_i)|$. As in (3.17), the latter is bounded by $D_i^0/n + K_0$, which itself is bounded (eventually) by $2(\log n)/\sqrt{n}$ under $\Omega$ when $D_i^0 = d$ with $d \le t + 2n\varepsilon$ (since we work under the assumption that $t \le \sqrt{n}$). Thus, for such a $d$, eventually,

$$\begin{aligned}\mathbb{P}(W_i \ge w \mid \Omega, D_i^0 = d) &\le \mathbb{E}\Big[\mathbb{P}\big(W_i \ge w \mid \Omega, D_i^0 = d, (X_k, Y_k)_{k \in I_0}\big)\Big]\\ &\le 2\,\mathbb{P}\big(W_i \ge w \mid P_i \le 2(\log n)/\sqrt{n}\big)\\ &\le 2\,\mathrm{Prob}\big(\mathrm{Bin}(2n\varepsilon, 2(\log n)/\sqrt{n}) \ge w\big)\\ &\le c_0(n\varepsilon \times (\log n)/\sqrt{n})^w,\end{aligned}$$

where $c_0$ is a universal constant. The factor of 2 in the second inequality comes from deconditioning from $\{K_0 \le (\log n)/\sqrt{n}$. In the last line we used the fact that $\mathrm{Prob}(\mathrm{Bin}(m,q) \ge k) \le \binom{m}{k} q^k$, referred to as the Giné–Zinn inequality in [18]. We also have

$$\mathbb{P}(D_i^0 = d \mid \Omega) \le 2\,\mathbb{P}(D_i^0 = d \mid |I_1| \le 2n\varepsilon) \le 2\frac{2}{n - 2n\varepsilon} \le \frac{5}{n},$$

eventually, using the fact that $\mathbb{P}(D_i^0 = d \mid |I_0| = n_0) \le 2/n_0$. Together, this yields

$$\begin{aligned}
&\mathbb{P}\big(W_i \ge D_i^0 - t, 2n\varepsilon + t \ge D_i^0 > t \mid \Omega\big) \\
&\le \sum_{d \ge t+1}^{t+2\lfloor n\varepsilon\rfloor} \mathbb{P}\big(W_i \ge d - t \mid \Omega, D_i^0 = d\big) \times \mathbb{P}(D_i^0 = d \mid \Omega) \\
&\le c_1 \sum_{d \ge t+1}^{t+2\lfloor n\varepsilon\rfloor} ((\log n)\sqrt{n}\varepsilon)^{d-t} \times \frac{1}{n}, \\
&\le c_1 \times \frac{2}{n} \times ((\log n)\sqrt{n}\varepsilon).
\end{aligned}$$

Hence, the second term on the RHS of (3.22) has expectation of order at most $n$ times the last term in our last derivations, which is of order at most $(\log n)\sqrt{n}\varepsilon = o(1)$. Since that term is integer-valued, this implies that $\Delta_0(t) \le \Delta^0(t)$ with probability tending to one. In particular, (3.21) applies.

*Second term in* (3.20). Consider $i \in I_1$. Because $\rho = 1$, we have $X_i = Y_i$, and conditional on $X_i = z$, $R_i - 1$ and $S_i - 1$ are iid with distribution $\text{Bin}(n-1, p)$ where $p := \Phi(z)$. In particular, $D_i$ has the distribution of $|U - V|$ where $U$ and $V$ are iid with distribution $\text{Bin}(n-1, P)$ and $P \sim \text{Unif}(0,1)$. Let $u_2(t)$ denote the probability that $D_i \le t$. We want to bound $u_2(t)$ from above.

For $p \in [0,1]$, define $g(p)$ as the probability that $|U - V| \le t$ when $U$ and $V$ are iid $\text{Bin}(n-1, p)$, and note that $u_2(t) = \int_0^1 g(p)\mathrm{d}p$. Define $\sigma^2 = 2(n-1)p(1-p)$, which is the variance of $U - V$, and also $h(a) = \mathbb{P}((U-V)/\sigma \le a)$. Using the fact that $U - V$ is integer valued, we have

$$g(p) = h(t/\sigma) - h(-(t+1)/\sigma) \le \Phi(t/\sigma) - \Phi(-(t+1)/\sigma) + 2\|h - \Phi\|_\infty.$$

Where $\Phi$ is the standard normal distribution function. Because $\Phi$ has derivative bounded by $1/\sqrt{2\pi}$ everywhere, the first term on the RHS is $= O(t/\sigma)$. For the second term, we use the Berry–Esseen inequality (seeing $U$ and $V$, each, as the sum of $n-1$ iid $\text{Ber}(p)$ random variables), to get that it is $= O(1/\sigma)$. Therefore, since $t \ge 1$, there is a universal constant $c_0$ such that $g(p) \le c_0 t/\sigma$.

Of course, being a probability, we also have $g(p) \leq 1$. Hence,

$$u_2(t) = \int_0^1 g(p)\mathrm{d}p \leq \int_0^1 \left(1 \wedge \frac{c_0 t}{2(n-1)p(1-p)}\right)\mathrm{d}p \asymp 1 \wedge t/\sqrt{n}.$$

Now, by Markov's inequality, and the fact that $|I_1|$ is binomial with parameters $(n, \varepsilon)$, the second term in (3.20) is

$$= \frac{O_P(n\varepsilon)O_P(u_2(t))}{\sqrt{nu(t)(1-u(t))}} \asymp \frac{n^{1-\beta}(1 \wedge t/\sqrt{n})}{\sqrt{t}} \asymp n^{1-\beta}t^{-1/2} \wedge n^{1/2-\beta}t^{1/2} \to 0,$$

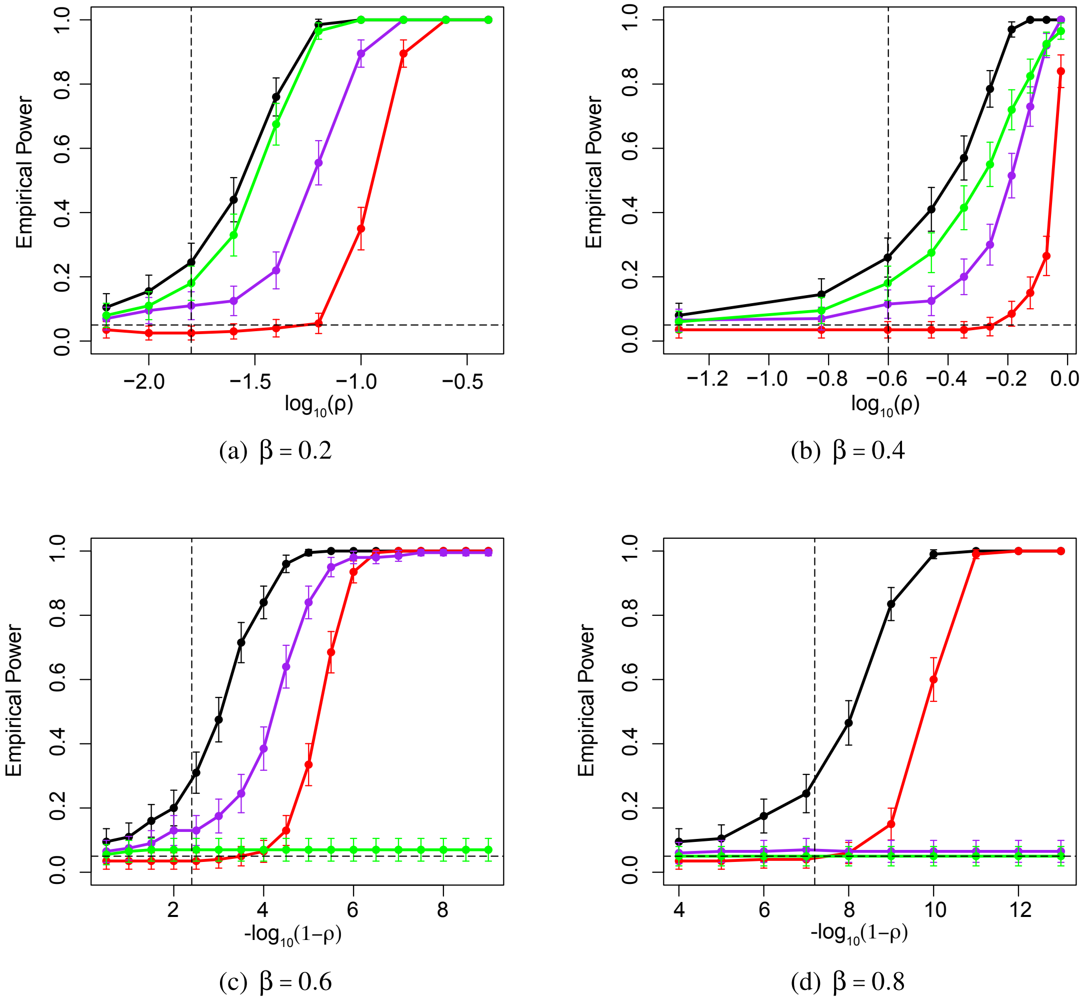
for any choice of $t$ when $\beta > 3/4$ (very sparse regime).

*Special case: t constant.* When $t$ is constant, the null distribution of $\Delta(t)$ is known to converge to the Poisson distribution of mean $2t+1$. (See [5, Exa 1.3], which is only cosmetically different.) The control under the alternative can be secured in exactly the same way. In particular, it holds that $\Delta_0(t) \leq \Delta^0(t)$ with probability tending to one, with $\Delta^0(t)$ having the same asymptotic distribution (Poisson with mean $2t+1$). □

### 3.3.3 Numerical experiments

We consider the same setting as in Section 3.2 and compare the two nonparametric tests, the covariance rank test and the higher criticism rank test, to the parametric tests. The p-values for the higher criticism rank test are obtained based on $10^5$ permutations, while the p-values for the covariance rank test are taken from the limiting distribution based on its correspondence with the Spearman rank correlation.

The results are presented in Figure 3.2. In finite samples, the higher criticism rank test exhibits substantially more power than the higher criticism in the dense and moderately sparse regime. We have no good explanation for this rather surprising phenomenon. However, the higher criticism rank test has no power in the very sparse regime, and neither does the covariance rank test.



(a) $\beta = 0.2$

(b) $\beta = 0.4$

(c) $\beta = 0.6$

(d) $\beta = 0.8$

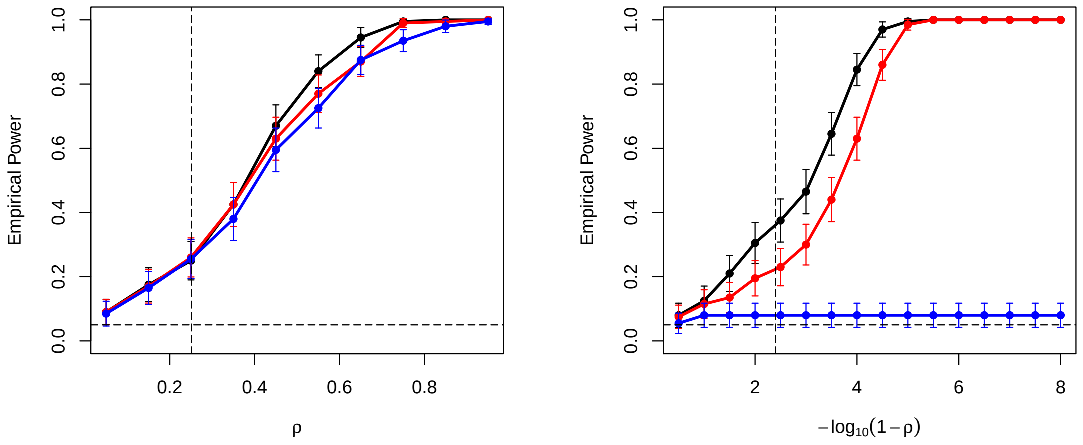
**Figure 3.2**: Empirical power comparison with 95% error bars for the likelihood ratio test (black), the covariance rank test (green), the higher criticism test (red) and the higher criticism rank test (purple). (a) Dense regime where $\beta = 0.2$. (b) Dense regime where $\beta = 0.4$. (c) Sparse regime where $\beta = 0.6$ and $\rho \to 1$. (d) Sparse regime where $\beta = 0.8$ and $\rho \to 1$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $n = 10^6$ and the power curves and error bars are based on 200 replications.

## 3.4 Discussion

**The power residing in the $V_i$** In Proposition 5 we established that the higher criticism test based on $U_1,\dots,U_n$ achieves the detection boundary in the Gaussian mixture model. It is natural, however, to ask whether one could do better in finite samples by also utilizing $V_1,\dots,V_n$. We performed some side experiments to quantify this by comparing the full LRT, meaning the LRT based on $(U_1,V_1),\dots,(U_n,V_n)$, the LRT based on $U_1,\dots,U_n$ only, and the LRT based on $V_1,\dots,V_n$ only. We did so in the same parametric setting of Section 3.2.4. The results are reported in Figure 3.3, and can be to some extent anticipated from our previous work [2]. In a nutshell, in the dense regime, what matters is the deviation of the variance from 1, and this is felt by all tests, so that the $U$-LRT and the $V$-LRT are seen to be also as powerful as the full LRT. In the sparse regime, however, we can see that the $V$-LRT has essentially no power. This is due to the fact that the $V_i$'s in that case have variance $1+\rho$, which is bounded from above by 2, so that no test depending on the $V_i$'s can have any power as we show in [2]. The $U$-LRT, which we know to be asymptotically optimal to first order, remains competitive, although now clearly less powerful than the full LRT.

**The power of rank tests in the very sparse regime** In Proposition 8 we argued, we hope convincingly, that no test that resembles the higher criticism rank test has any power in the very sparse regime ($\beta > 3/4$). This seems clear from the experiments reported in Figure 3.2. This begs the question of whether there are any rank tests that have any (asymptotic) power in the very sparse regime. We do not know the answer to that question, but are willing to conjecture that there are no such tests.

**The two-sided problem** We focused on the one-sided setting (3.1), effectively testing $\rho = 0$ versus $\rho > 0$. Knowing the sign of $\rho$ is not crucial, as one can apply a one-sided test for $\rho > 0$ to the transformed data $(X_1,-Y_1),\dots,(X_n,-Y_n)$. Less trivial is the case where there are three



(a) $\beta = 0.4$

(b) $\beta = 0.6$

**Figure 3.3**: Empirical power comparison with 95% error bars for the full LRT (black), the $U$-LRT (red) and the $V$-LRT (blue). (a) Dense regime where $\beta = 0.4$. (b) Sparse regime where $\beta = 0.6$ and $\rho \to 1$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $n = 10^6$ and the power curves and error bars are based on 200 replications.

components

$$(X,Y) \sim (1-\varepsilon)\mathcal{N}(0,\mathrm{I}) + \frac{\varepsilon}{2}\mathcal{N}(0,\Sigma_\rho) + \frac{\varepsilon}{2}\mathcal{N}(0,\Sigma_{-\rho}).$$

We did not look at this model, in part because we wanted to test against a monotonic association (in the contamination component), which is perhaps the most popular alternative in a nonparametric context.

## 3.5 Acknowledgements

Chapter 3, in full, is a version of the paper “Detection of Sparse Positive Dependence”, *Electronic Journal of Statistics*, Arias-Castro, Ery; Huang, Rong; Verzelen, Nicolas, Volume 14, Number 1 (2020). The dissertation author was the primary investigator and author of this material.

# Chapter 4

# Detecting Sparse Heterogeneous Mixtures in a Two-sample Problem

## 4.1 Introduction

The detection of sparse mixtures has been studied for decades [28, 20]. Most work has focused on detecting deviation of data from the known distribution. However, in practice, it's more common that we do not have access to the null distribution and have to estimate it from a control group. For example, in a clinical trial, patients were assigned to two groups randomly, given either the placebo or the treatment. For some reasons, the treatment could only affect a small proportion of the patients treated, while the remaining patients reacted the same as patients in the control group. Similar settings were investigated in Conover and Salsburg [16] and they modeled the shift in the distribution as a Lehmann alternative and focused on the locally most powerful tests.

We study this situation in parallel with the work of Ingster [28] and Donoho and Jin [20]. Let $F$ and $G$ be two continuous (unknown) distribution functions on the real line. We consider the following hypothesis testing problem: based on a random sample $X_1, \cdots, X_m$ drawn iid from $F$

and another independent random sample $Y_1, \cdots, Y_n$ drawn iid from $G$, decide

$$\mathcal{H}_0 : G = F \quad \text{versus} \quad \mathcal{H}_1 : G = (1-\varepsilon)F + \varepsilon F(\cdot - \mu), \quad \varepsilon > 0, \quad \mu > 0. \tag{4.1}$$

where $\varepsilon \in (0, 1/2)$ is the fraction of non-null effect and $\mu$ is the size of the location shift. Hence, under the alternative, $G$ is stochastic larger than $F$. We assume that

$$\lim_{m,n\to\infty} \frac{n}{m+n} = \eta \in (0, 1/2] \tag{4.2}$$

at a sufficient fast rate.

The optimum of rank tests was mainly investigated as locally most powerful [33]. Following the line of our work in the one-sample setting, we still focus on the asymptotically optimum here. As usual, a testing procedure is asymptotically powerful (resp. powerless) if the sum of its probabilities of Type I and Type II errors (its risk) has limit 0 (resp. inferior at least 1) in the large sample asymptote.

### 4.1.1 A benchmark: generalized Guassian mixture model

The normal mixture model has been studied in Ingster [28] considering the one-sample setting, that is, $F$ is known to be standard normal and only the $Y$-sample is collected. The problem is investigated in various asymptotic regimes defined by how fast $\varepsilon$ goes to zero. The detection boundary of the likelihood ratio test (LRT) (then any other tests) is derived. Donoho and Jin [20] further derived the detection boundary when $F$ is generalized Gaussian:

$$f(x) \propto \exp\Big(-\frac{|x|^\gamma}{\gamma}\Big),$$

where $\gamma > 0$. Note that $\gamma = 2$ corresponds to the normal distribution and $\gamma = 1$ corresponds to the double-exponential distribution. They parameterized $\varepsilon = \varepsilon_n$ as

$$\varepsilon_n = n^{-\beta}, \quad 0 < \beta < 1 \text{ fixed.} \tag{4.3}$$

In the sparse setting where $1/2 < \beta < 1$, let

$$\mu_n = (\gamma r \log n)^{1/\gamma}, \quad 0 < r < 1 \text{ fixed,} \tag{4.4}$$

then the detection boundary when $\gamma > 1$ is

$$\rho^*_\gamma(\beta) = \begin{cases} (2^{1/(\gamma-1)} - 1)^{\gamma-1}(\beta - \frac{1}{2}), & \frac{1}{2} < \beta < 1 - 2^{-\gamma/(\gamma-1)}; \\ (1 - (1-\beta)^{1/\gamma})^\gamma, & 1 - 2^{-\gamma/(\gamma-1)} < \beta < 1. \end{cases}$$

and for the case $\gamma \leq 1$

$$\rho^*_\gamma(\beta) = 2\beta - 1.$$

That means, if $r > \rho^*(\beta)$, $\mathcal{H}_0$ and $\mathcal{H}_1$ separate asymptotically, while if $r < \rho^*(\beta)$, $\mathcal{H}_0$ and $\mathcal{H}_1$ merge asymptotically.

The detection boundary in the dense regime where $0 < \beta < 1/2$ is given in [3]. Let

$$\mu_n = n^{s-1/2}, \quad 0 < s < 1/2 \text{ fixed,} \tag{4.5}$$

then the hypotheses merge asymptotically when $s < \beta$ if $\gamma \geq 1/2$ and $s < \frac{1}{2} - \frac{1-2\beta}{1+2\gamma}$ if $\gamma < 1/2$.

### 4.1.2 The two-sample higher criticism test

The one-sample higher criticism test was suggested in Donoho and Jin [20] and proved to be first-order asymptotically comparable to the LRT in the normal mixture model. In the two-sample setting, an analogous higher criticism statistic was proposed as [13, 24]:

$$\mathrm{HC} = \sup_{t\in\mathbb{R}} \sqrt{\frac{mn}{m+n}} \frac{[F_m(t) - G_n(t)]}{\sqrt{H_{m+n}(t)(1 - H_{m+n}(t))}}, \tag{4.6}$$

where $F_m$ and $G_n$ are empirical distributions of the $X$-sample and the $Y$-sample, respectively, and $H_{m+n}(t) = \frac{1}{m+n}(mF_m(t) + nG_n(t))$ is the empirical distribution of the combined sample. The test rejects for large values of (4.6). This is to the two-sample Kolmogorov-Smirnov test [51] what the Anderson-Darling test (aka the higher criticism test) is to the Kolmogorov-Smirnov test.

Pettitt [46] proposed the integral version of the two-sample Anderson-Darling statistic and gave an approximation of the distribution. Finner and Gontscharuk [24] studied the supremum version (4.6) in terms of local levels and focused on the Type I error. It is also connected to the work of Zhao et al [57], which normalizes the Hoeffding test for independence in an analogous way. Note that all these tests are based on ranks only, so they are nonparametric tests. Distribution-free tests for sparse heterogeneous mixtures in the one-sample setting was investigated in [3] where they assumed that $F$ is symmetric about zero and the true effects have positive median. In our work, we do not pose any assumptions on $F$ except it is continuous, and we use the $X$-sample to estimate $F$. In addition, the two-sample higher criticism is parallel to the CUSUM sign test in [3] as the Wilcoxon test to the Wilcoxon signed-rank test [53].

## 4.2 Lower bound

The formulation (4.1) indicates that both the null and the alternative hypotheses are composite. If we assume model parameters $(F, \varepsilon, \mu)$ are known, then the likelihood ratio test (LRT) is the most powerful test by Neyman-Pearson lemma. In particular, if $F$ is given, we don't need the $X$-sample, then the question is reduced to the one-sample situation [28, 20, 12]. The general detection boundary was given in [4, Lemma A.1] as follows in our context: let $f$ denote the density of $F$, then the hypotheses (4.1) merge asymptotically when there is a sequence $(x_n)$

such that

$$n\bar{F}(x_n) \to 0, \quad n\varepsilon_n \bar{F}(x_n - \mu_n) \to 0,$$

and

$$n\varepsilon_n^2 \Big[ \int_{-\infty}^{x_n} \frac{f(x-\mu_n)^2}{f(x)} dx - 1 \Big]_+ \to 0.$$

## 4.3 The two-sample higher criticism test

For a distribution $F$, $\bar{F}(x) = 1 - F(x)$ will denote its survival function.

**Theorem 6.** *For the testing problem* (4.1) *and under* (4.2)*, the two-sample higher criticism test is asymptotically powerful if either there is a sequence* $(t_n)$ *such that* $t_n \to \infty$*,*

$$n(\bar{F}(t_n) \vee \varepsilon \bar{F}(t_n - \mu)) \gg \log^2 n, \tag{4.7}$$

*and*

$$\frac{\sqrt{n}\varepsilon[\bar{F}(t_n - \mu) - \bar{F}(t_n)]}{\sqrt{\bar{F}(t_n) + \varepsilon\eta \bar{F}(t_n - \mu)}} \gg \log n; \tag{4.8}$$

*or* $t$ *is the median of* $F$ *and*

$$\sqrt{n}\varepsilon[\bar{F}(t-\mu) - \frac{1}{2}] \gg \log n. \tag{4.9}$$

*Proof.* Finner and Gontscharuk [24] showed that the two-sample HC statistic (4.6) is almost surely equal to

$$\mathrm{HC}^* = \sup_{s \in I_{m,n}} \sqrt{\frac{m+n}{m+n-1}} \frac{[V_{m,s} - \mathbb{E}_0[V_{m,s}]]}{\sqrt{\mathrm{Var}_0[V_{m,s}]}},$$

where $V_{m,s}$ denotes the number of ranks related to the $X$-sample being not larger than $s$ and $I_{m,n} := \{1, \cdots, m+n-1\}$. They also showed that $\mathrm{HC}^*$ coincides asymptotically in distribution with the one-sample HC statistic with sample size $n$ under the null as we assume that $n \le m$, which was derived in [29]. Thus, we have

$$\mathbb{P}_0\left(\mathrm{HC} \geq \sqrt{3\log\log n}\right) \to 0.$$

For simplicity, we consider the test with rejection region $\{\mathrm{HC} \geq \log n\}$. Hence, the test is asymptotically powerful if, under the alternative, there is $t_n$ (or $t$) $\in \mathbb{R}$ such that

$$\mathbb{P}_1\left(\sqrt{\frac{mn}{m+n}}\frac{[F_m(t_n) - G_n(t_n)]}{\sqrt{H_{m+n}(t_n)(1-H_{m+n}(t_n))}} \geq \log n\right) \to 1.$$

Indeed, $mF_m(t)$ is binomial with parameters $m$ and $F(t)$, $nG_n(t)$ is binomial with parameters $n$ and $G(t) = (1-\varepsilon)F(t) + \varepsilon F(t-\mu)$, and

$$H_{m+n}(t) = \frac{m}{m+n}F_m(t) + \frac{n}{m+n}G_n(t).$$

We also define $H(\cdot)$ as

$$H(t) = (1-\eta)F(t) + \eta G(t) = (1-\varepsilon\eta)F(t) + \varepsilon\eta F(t-\mu).$$

Hence, by Chebyshev's inequality, we have

$$\frac{\sqrt{m}|F_m(t_n) - F(t_n)|}{\sqrt{F(t_n)(1-F(t_n))}} \leq \log m$$

with probability tending to 1, and

$$\frac{\sqrt{n}|G_n(t_n) - G(t_n)|}{\sqrt{G(t_n)(1-G(t_n))}} \leq \log n$$

with probability tending to 1. By triangular inequality, we have

$$\begin{aligned}|F(t_n)-G(t_n)| &= |F(t_n)-F_m(t_n)+F_m(t_n)-G_n(t_n)+G_n(t_n)-G(t_n)| \\ &\leq |F(t_n)-F_m(t_n)|+|F_m(t_n)-G_n(t_n)|+|G_n(t_n)-G(t_n)| \\ &\leq |F_m(t_n)-G_n(t_n)|+\log m\sqrt{\frac{F(t_n)(1-F(t_n))}{m}}+\log n\sqrt{\frac{G(t_n)(1-G(t_n))}{n}}\end{aligned}$$

Hence, we have

$$\sqrt{\frac{mn}{m+n}}\frac{[F_m(t_n)-G_n(t_n)]}{\sqrt{H_{m+n}(t_n)(1-H_{m+n}(t_n))}} \geq a_n - b_n,$$

where

$$\begin{aligned}a_n &:= \sqrt{\frac{mn}{m+n}}\frac{\varepsilon[\bar{F}(t_n-\mu)-\bar{F}(t_n)]}{\sqrt{H_{m+n}(t_n)(1-H_{m+n}(t_n))}}, \\ b_n &:= \sqrt{\frac{n}{m+n}}\log m\sqrt{\frac{F(t_n)(1-F(t_n))}{H_{m+n}(t_n)(1-H_{m+n}(t_n))}}+\sqrt{\frac{m}{m+n}}\log n\sqrt{\frac{G(t_n)(1-G(t_n))}{H_{m+n}(t_n)(1-H_{m+n}(t_n))}},\end{aligned}$$

as we know that $F > G$ under the alternative, and it suffices to show that

$$a_n \geq b_n + \log n \tag{4.10}$$

with probability tending to 1.

$$\begin{aligned}
|\bar{H}_{m+n}(t_n) - \bar{H}(t_n)| &= \Big|\frac{m}{m+n}\bar{F}_m(t_n) + \frac{n}{m+n}\bar{G}_n(t_n) - (1-\eta)\bar{F}(t_n) - \eta\bar{G}(t_n)\Big| \\
&= \Big|\frac{m}{m+n}(\bar{F}_m(t_n) - \bar{F}(t_n)) + (\frac{m}{m+n} - (1-\eta))\bar{F}(t_n) \\
&\quad + \frac{n}{m+n}(\bar{G}_n(t_n) - \bar{G}(t_n)) + (\frac{n}{m+n} - \eta)\bar{G}(t_n)\Big| \\
&\le \Big|\frac{m}{m+n}(\bar{F}_m(t_n) - \bar{F}(t_n))\Big| + \Big|\frac{n}{m+n}(\bar{G}_n(t_n) - \bar{G}(t_n))\Big| \\
&\quad + O\Big(\frac{\log n}{\sqrt{n}}\Big)(\bar{F}(t_n) + \bar{G}(t_n)) \\
&\le \frac{m}{m+n}\log m\sqrt{\frac{\bar{F}(t_n)(1-\bar{F}(t_n))}{m}} + \frac{n}{m+n}\log n\sqrt{\frac{\bar{G}(t_n)(1-\bar{G}(t_n))}{n}} \\
&\quad + O\Big(\frac{\log n}{\sqrt{n}}\Big)(\bar{F}(t_n) + \bar{G}(t_n)) \\
&\asymp O\Big(\frac{\log m}{\sqrt{m}}\sqrt{\bar{F}(t_n)}\Big) + O\Big(\frac{\log n}{\sqrt{n}}\sqrt{\bar{G}(t_n)}\Big) + O\Big(\frac{\log n}{\sqrt{n}}(\bar{F}(t_n) + \bar{G}(t_n))\Big) \\
&\asymp O\Big(\frac{\log n}{\sqrt{n}}\sqrt{(1-\varepsilon)\bar{F}(t_n) + \varepsilon\bar{F}(t_n - \mu)}\Big),
\end{aligned}$$

as we assume that

$$|\frac{n}{m+n} - \eta| = O(\log n/\sqrt{n}).$$

Thus, under (4.7) or (4.9), we have

$$|\bar{H}_{m+n}(t_n) - \bar{H}(t_n)| \ll \bar{H}(t_n).$$

Then if $t_n \to \infty$, we have $H(t_n) \to 1$, and

$$\begin{aligned} a_n &= \sqrt{\frac{mn}{m+n}} \frac{\varepsilon[\bar{F}(t_n - \mu) - \bar{F}(t_n)]}{\sqrt{H_{m+n}(t_n)(1 - H_{m+n}(t_n))}} \\ &\asymp \frac{\sqrt{n(1-\eta)}\varepsilon[\bar{F}(t_n - \mu) - \bar{F}(t_n)]}{\sqrt{(1-\varepsilon\eta)\bar{F}(t_n) + \varepsilon\eta\bar{F}(t_n - \mu)}} \cdot \sqrt{\frac{\bar{H}(t_n)}{\bar{H}_{m+n}(t_n)}} \\ &\asymp \frac{\sqrt{n}\varepsilon[\bar{F}(t_n - \mu) - \bar{F}(t_n)]}{\sqrt{\bar{F}(t_n) + \varepsilon\eta\bar{F}(t_n - \mu)}} \gg \log n, \end{aligned}$$

under (4.8).

If $t$ is the median of $F$ such that $F(t) = 1/2$, then $H(t)$ is bounded away from 0 and 1, and under condition (4.9),

$$\begin{aligned} a_n &= \sqrt{\frac{mn}{m+n}} \frac{\varepsilon[\bar{F}(t - \mu) - \bar{F}(t)]}{\sqrt{H_{m+n}(t)(1 - H_{m+n}(t))}} \\ &= \sqrt{n}\varepsilon[\bar{F}(t - \mu) - \frac{1}{2}] \gg \log n. \end{aligned}$$

In addition, we have

$$b_n \asymp \log n.$$

Therefore, (4.10) is fulfilled eventually. □

In the generalized Gaussian mixture model, with parameterization (4.3) and (4.4), we choose $t_n = (\gamma q \log n)^{1/\gamma}$, $r < q \le 1$ fixed. By the tail behavior of $F$, we have

$$\bar{F}(t_n) = L_n n^{-q}, \quad \bar{F}(t_n - \mu_n) = L_n n^{-(q^{1/\gamma} - r^{1/\gamma})^\gamma},$$

where $L_n$ denotes any factor logarithmic in $n$.

If $\gamma > 1$, we define $r_\gamma = (1 - 2^{-1/(\gamma-1)})^\gamma$. If $r < r_\gamma$, we set $q = r/r_\gamma$. Then the LHS in (4.7) is

$$n(L_n n^{-r/r_\gamma} \vee \varepsilon L_n n^{-r(r_\gamma^{-1/\gamma} - 1)^\gamma}) \asymp L_n(n^{1-r/r_\gamma} \vee n^{1-\beta-r(r_\gamma^{-1/\gamma}-1)^\gamma}) \gg \log^2 n.$$

The LHS in (4.8) is

$$L_n \frac{n^{\frac{1}{2}-\beta}(n^{-r(r_\gamma^{-1/\gamma}-1)^\gamma} - n^{-r/r_\gamma})}{\sqrt{n^{-r/r_\gamma} + \varepsilon\eta n^{-r(r_\gamma^{-1/\gamma}-1)^\gamma}}} \asymp L_n(n^{\frac{1+r/r_\gamma}{2}-\beta-r(r_\gamma^{-1/\gamma}-1)^\gamma} \wedge n^{\frac{1}{2}(1-\beta-r(r_\gamma^{-1/\gamma}-1)^\gamma)}),$$

where both exponents are positive when $r > (2^{1/(\gamma-1)} - 1)^{\gamma-1}(\beta - \frac{1}{2})$.

If $r \geq r_\gamma$, we set $q = 1$. Then the LHS in (4.7) is

$$n(L_n n^{-1} \vee \varepsilon L_n n^{-(1-r^{1/\gamma})^\gamma}) \asymp L_n(1 \vee n^{1-\beta-(1-r^{1/\gamma})^\gamma}) \gg \log^2 n,$$

if $1 - \beta - (1 - r^{1/\gamma})^\gamma > 0$. And the LHS in (4.8) is

$$L_n \frac{n^{\frac{1}{2}-\beta}(n^{-(1-r^{1/\gamma})^\gamma} - n^{-1})}{\sqrt{n^{-1} + \varepsilon\eta n^{-(1-r^{1/\gamma})^\gamma}}} \asymp L_n(n^{1-\beta-(1-r^{1/\gamma})^\gamma} \wedge n^{\frac{1}{2}(1-\beta-(1-r^{1/\gamma})^\gamma)}),$$

where both exponents are positive when if $1 - \beta - (1 - r^{1/\gamma})^\gamma > 0$.

If $\gamma \leq 1$, we set $q = r$, so that $t_n = \mu_n$. Then the LHS in (4.7) is

$$n(L_n n^{-r} \vee L_n) \gg \log^2 n,$$

and the LHS in (4.8) is

$$L_n \frac{n^{\frac{1}{2}-\beta}(1 - n^{-r})}{\sqrt{n^{-r} + n^{-\beta}}} \asymp L_n n^{\frac{1}{2}-\beta+\frac{r}{2}},$$

where the exponent is positive when $r > 2\beta - 1$. Comparing with the detection boundary, we see that the two-sample higher criticism test achieves the detection boundary in the generalized Gaussian model in the sparse regimes for any $\gamma > 0$.

In the dense regime, with parameterization (4.3) and (4.5), $t = F^{-1}(1/2) = 0$, hence,

$$\sqrt{n}\varepsilon[\bar{F}(-\mu) - \frac{1}{2}] \asymp \sqrt{n}\varepsilon\mu = n^{s-\beta},$$

and the exponent is positive when $s < \beta$. So that the two-sample HC test achieves the detection boundary when $\gamma \geq 1/2$.

## 4.4 Other tests

It is well-known that in the more classical setting, the two-sample situation is closely related to the one-sample tests for symmetry. Essentially, the main two-sample tests have the same relatively efficiency between them as the corresponding one-sample tests. We analyzed some classical tests in this section.

### 4.4.1 The Wilcoxon test

The Wilcoxon test is the classical nonparametric test for location shift between two samples [53, 41]. In particular, in this case, it rejects for large values of the Wilcoxon statistic $U$ which counts the number of pairs $X_i$, $Y_j$ with $X_i < Y_j$.

**Proposition 9.** *For the testing problem* (4.1) *and under* (4.2)*, the Wilcoxon test is asymptotically powerful (resp. powerless) when*

$$\sqrt{n}\varepsilon\Big[\frac{1}{2} - \int F(\cdot - \mu)dF\Big] \gg \log n \quad (reps. \to 0). \tag{4.11}$$

*Proof.* Mann and Whitney [41] proved that under the null $F = G$, for large samples,

$$\frac{\frac{U}{mn} - \mathbb{E}_0(\frac{U}{mn})}{\sigma_0(\frac{U}{mn})}$$

is approximately normally distributed. In particular, the first two moments of $U$ are [41, 33]

$$\mathbb{E}\Big(\frac{U}{mn}\Big) = \int F dG,$$

$$mn\operatorname{Var}\Big(\frac{U}{mn}\Big) = \Big[\frac{m+n+1}{12} + (m-1)(\lambda-\varepsilon_1) + (n-1)(\lambda-\varepsilon_2) - \lambda^2(m+n-1)\Big],$$

where

$$\lambda = \frac{1}{2} - \int F dG, \quad \varepsilon_1 = \frac{1}{3} - \int F^2 dG, \quad \varepsilon_2 = \frac{1}{3} - \int (1-G)^2 dF.$$

Hence, we have

$$\mathbb{E}_0\Big(\frac{U}{mn}\Big) = \int F dF = \frac{1}{2},$$

$$\operatorname{Var}_0\Big(\frac{U}{mn}\Big) = \frac{m+n+1}{12mn}.$$

By Chebyshev's inequality, we have

$$\mathbb{P}_0\Big(|U - \frac{mn}{2}| \geq a_n\sqrt{\frac{(m+n+1)mn}{12}}\Big) \to 0,$$

for any sequence $a_n$ diverging to infinity. Under $\mathcal{H}_1$, $G = (1-\varepsilon)F + \varepsilon F(\cdot - \mu)$, we have

$$\mathbb{E}_1\Big(\frac{U}{mn}\Big) = \int F dG = \frac{1}{2} + \frac{\varepsilon}{2} - \varepsilon \int F(\cdot - \mu) dF,$$

and

$$\operatorname{Var}_1\Big(\frac{U}{mn}\Big) = O\Big(\frac{m+n+1}{12mn}\Big),$$

as $0 \leq \int F^k dG \leq 1$, $k = 1,2$ and $0 \leq \int (1-G)^2 dF \leq 1$. Then by Chebyshev's inequality, we have

$$\mathbb{P}_1\Big(|U - mn(\frac{1}{2} + \frac{\varepsilon}{2} - \varepsilon \int F(\cdot - \mu) dF)| \geq a_n\sqrt{\frac{(m+n+1)mn}{12}}\Big) \to 0,$$

for any sequence $a_n$ diverging to infinity. We choose $a_n = \log n$ and consider the test with rejection region $\{U - \frac{mn}{2} \geq \log n\sqrt{\frac{(m+n+1)mn}{12}}\}$. The test is asymptotically powerful when, eventually,

$$\varepsilon\big[\frac{1}{2} - \int F(\cdot - \mu) dF\big] \geq \log n\sqrt{\frac{m+n+1}{12mn}},$$

which is satisfied under (4.2) and (4.11).

Next we show that the Wilcoxon test is asymptotically powerless when (4.11) converges to zero. Lehmann [32] showed that asymptotic normality still holds for $U$ under the alternative and (4.2), that is

$$\frac{\frac{U}{mn} - \mathbb{E}_1(\frac{U}{mn})}{\sigma_1(\frac{U}{mn})} \to \mathcal{N}(0,1).$$

Hence, under $\mathcal{H}_1$, we have

$$\frac{U-\mathbb{E}_0(U)}{\sigma_0(U)} = \Big(\frac{U-\mathbb{E}_1(U)}{\sigma_1(U)} + \frac{\mathbb{E}_1(U)-\mathbb{E}_0(U)}{\sigma_1(U)}\Big)\cdot\frac{\sigma_1(U)}{\sigma_0(U)},$$

where $\frac{\mathbb{E}_1(U)-\mathbb{E}_0(U)}{\sigma_1(U)} \asymp \sqrt{n}\varepsilon[\frac{1}{2} - \int F(\cdot - \mu)dF] \to 0$ and $\sigma_1(U)/\sigma_0(U) \asymp 1$. Therefore, by Slutsky's theorem, $(U-\mathbb{E}_0(U))/\sigma_0(U)$ also converges to $\mathcal{N}(0,1)$ as under the null. No test based on $U$ would have any power. □

Note that the Wilcoxon test is asymptotically powerless when $\sqrt{n}\varepsilon_n \to 0$. In the generalized Gaussian mixture model, in the dense regime with parameterization (4.3) and (4.5), we have

$$\sqrt{n}\varepsilon_n[\frac{1}{2} - \int F(\cdot-\mu_n)dF] \approx \sqrt{n}\varepsilon_n[\frac{1}{2} - \int (F-\mu_n f)dF] = \sqrt{n}\varepsilon_n\mu_n \int f dF \asymp n^{s-\beta},$$

where $f$ is the density function. Hence, the Wilcoxon test is asymptotically powerful when $s > \beta$, and it achieves the detection boundary when $\gamma > 1/2$.

### 4.4.2 The two-sample Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov test [51] rejects for larges values of

$$D_{m,n} = \sup_{t\in\mathbb{R}}[F_m(t) - G_n(t)].$$

**Proposition 10.** *For the testing problem* (4.1) *and under* (4.2)*, the two-sample Kolmogorov-*

*Smirnov test is asymptotically powerful (resp. powerless) when*

$$\sqrt{n}\varepsilon \sup_{t\in\mathbb{R}}[\bar{F}(t-\mu)-\bar{F}(t)] \to \infty \quad (\textit{resp.} \to 0). \tag{4.12}$$

*Proof.* We already know the limiting distribution of $\sqrt{mn/(m+n)}D_{m,n}$ under the null hypothesis [51]. Under $\mathcal{H}_1$, by triangle inequality,

$$\begin{aligned}\sqrt{n}\sup_{t\in\mathbb{R}}[F_m(t)-G_n(t)] &\geq \sqrt{n}\sup_{t\in\mathbb{R}}[F(t)-G(t)]-\sqrt{n}\sup_{t\in\mathbb{R}}|F_m(t)-F(t)|-\sqrt{n}\sup_{t\in\mathbb{R}}|G_n(t)-G(t)|\\ &= \sqrt{n}\varepsilon\sup_{t\in\mathbb{R}}[F(t)-F(t-\mu)]-O_p(1)\to\infty,\end{aligned}$$

when the limit in (4.12) is infinity.

When the limit in (4.12) is 0, let $I_0$ and $I_1$ index the observations in the $Y$- sample coming from the null and contaminated components, respectively. Let $G_n^j(t)=\frac{1}{|I_j|}\sum_{i\in I_j}\mathbb{I}\{y_i\leq t\}$, $j=0,1$. We have

$$G_n(t)=\frac{|I_0|}{n}G_n^0(t)+\frac{|I_1|}{n}G_n^1(t).$$

By triangle inequality,

$$\begin{aligned}&|\sqrt{n}\sup_{t\in\mathbb{R}}[F_m(t)-G_n(t)]-\sqrt{|I_0|}\sup_{t\in\mathbb{R}}[F_m(t)-G_n^0(t)]|\\ &\leq \Big|\sqrt{\frac{|I_0|}{n}}-1\Big|\Big|\sqrt{|I_0|}\sup_{t\in\mathbb{R}}[F_m(t)-G_n^0(t)]\Big|+\sqrt{\frac{|I_1|}{n}}\Big|\sqrt{|I_1|}\sup_{t\in\mathbb{R}}[F_m(t)-G_n^1(t)]\Big|\\ &\leq \Big|\sqrt{\frac{|I_0|}{n}}-1\Big|O_p(1)+\sqrt{\frac{|I_1|}{n}}\Big|\sqrt{|I_1|}\sup_{t\in\mathbb{R}}[F(t)-F(t-\mu)]+O_p(1)\Big|=o_p(1),\end{aligned}$$

by the fact that $|I_0|\sim_p n$, $|I_1|\sim_p n\varepsilon$ and (4.12) converges to 0. Hence, $\sqrt{n}D_{m,n}\sim\sqrt{|I_0|}D_{m,|I_0|}$ under $\mathcal{H}_1$, which has the same limiting distribution as under $\mathcal{H}_0$. □

Note that the two-sample Kolmogorov-Smirnov test has no power in the sparse regime. In the generalized Gaussian mixture model, in the dense regime with parameterization (4.3) and

(4.5), we have

$$\sqrt{n}\varepsilon \sup_{t\in\mathbb{R}}[\bar{F}(t-\mu)-\bar{F}(t)] \geq \sqrt{n}\varepsilon[\bar{F}(-\mu)-\bar{F}(0)] \asymp n^{s-\beta} \to \infty,$$

when $s > \beta$. Same as the Wilcoxon test, it only achieves the detection boundary with $\gamma > 1/2$.

### 4.4.3 The tail-run test

We now consider the tail-run test. Let $\zeta_{(j)} = 0$ or 1, according to whether the $j$th largest observation is from the $X$-sample or the $Y$-sample, $j = 1,\cdots,m+n$. Then the tail-run test rejects for large values of

$$L^* = \max\{l \geq 0 : \zeta_{(1)} = \cdots = \zeta_{(l)} = 1\}.$$

The one-sample tail-run test for sparse mixtures is investigated in [3]. It is also analogous to the extremes tests in the normal mixture model.

**Proposition 11.** *For the testing problem* (4.1) *and under* (4.2)*, and let* $(l_n)$ *be a divergent sequence of positive integers. The tail-run test is asymptotically powerful when there exits a sequence* $(t_n)$ *such that*

$$m\bar{F}(t_n) \to 0, \quad n\varepsilon\bar{F}(t_n-\mu) \geq 2l_n. \tag{4.13}$$

*Proof.* We consider the tail-run test with rejection region $\{L^* \geq l_n\}$. Note that $L^*$ is the number of the $Y$-samples until the first $X$-sample is encountered. Under $\mathcal{H}_0$, $L^*$ is following negative hypergeometric distribution with the population size $m+n$, and we have

$$\mathbb{E}_0(L^*) = \frac{n}{m+1}, \quad \mathrm{Var}_0(L^*) = \frac{(m+n+1)n}{(m+1)(m+2)}[1-\frac{1}{m+1}].$$

Hence, $L^* = O_p(1)$ and $l_n \to \infty$, we have $\mathbb{P}_0(L^* \geq l_n) \to 0$ as $n \to \infty$.

Under $\mathcal{H}_1$, note that

$$\mathbb{P}_X(\max X_i \le t_n) = (1 - \bar{F}(t_n))^m \to 1,$$

under the condition $m\bar{F}(x_n) \to 0$. Therefore, $L^* \ge N := \#\{j, Y_j > t_n\}$ with high probability. And $N \sim \text{Bin}(n, p_y)$ where $p_y = (1-\varepsilon)\bar{F}(t_n) + \varepsilon\bar{F}(t_n - \mu)$. Eventually, under (4.13) , we have $N = (1 + o_p(1))np_y \ge l_n$. □

In the generalized Gaussian mixture model, with parameterization (4.3) and (4.4), we choose $t_n = (\gamma(1+q)\log n)^{1/\gamma}$, $q > 0$ fixed. By the tail behavior of $F$, we have

$$m\bar{F}(t_n) = L_n n^{-q}, \quad n\varepsilon\bar{F}(t_n - \mu_n) = L_n n^{1-\beta-((1+q)^{1/\gamma}-r^{1/\gamma})^\gamma},$$

where $L_n$ denotes any factor logarithmic in $n$. When $r > (1-(1-\beta)^{1/\gamma})^\gamma$ is fixed, we can choose $q > 0$ small enough that $1 - \beta - ((1+q)^{1/\gamma} - r^{1/\gamma})^\gamma > 0$. Hence, the tail-run test is suboptimal in the moderately sparse regime and is optimal in the very sparse regime.

## 4.5 Numerical experiments

We performed some numerical experiments to investigate the finite sample performance of the likelihood ratio test (LRT), the two-sample higher criticism (HC) test, the Wilcoxon test, the two-sample Kolmogorov-Smirnov (KS) test and the tail-run test. We set sample sizes $m = n = 10^5$ in order to capture the large-sample behavior of these tests. The p-values for each test are calibrated as follows:

(a) For the *likelihood ratio test* and the *two-sample higher criticism test*, we simulated the null distribution based on $4,000$ Monte Carlo replicates.

(b) For the *Wilcoxon test* and the *two-sample Kolmogorov-Smirnov test*, the p-values are from the limiting distributions.

(c) For the *tail-run test*, we used the exact null distribution, that is the negative hypergeometric distribution.

For each scenario, we repeated the whole process 200 times and recorded the fraction of p-values smaller than 0.05, representing the empirical power at the 0.05 level.

## Normal mixture model

In this model, $F$ is standard normal. The results are reported in Figure 4.1 and are largely congruent with the theory developed earlier.

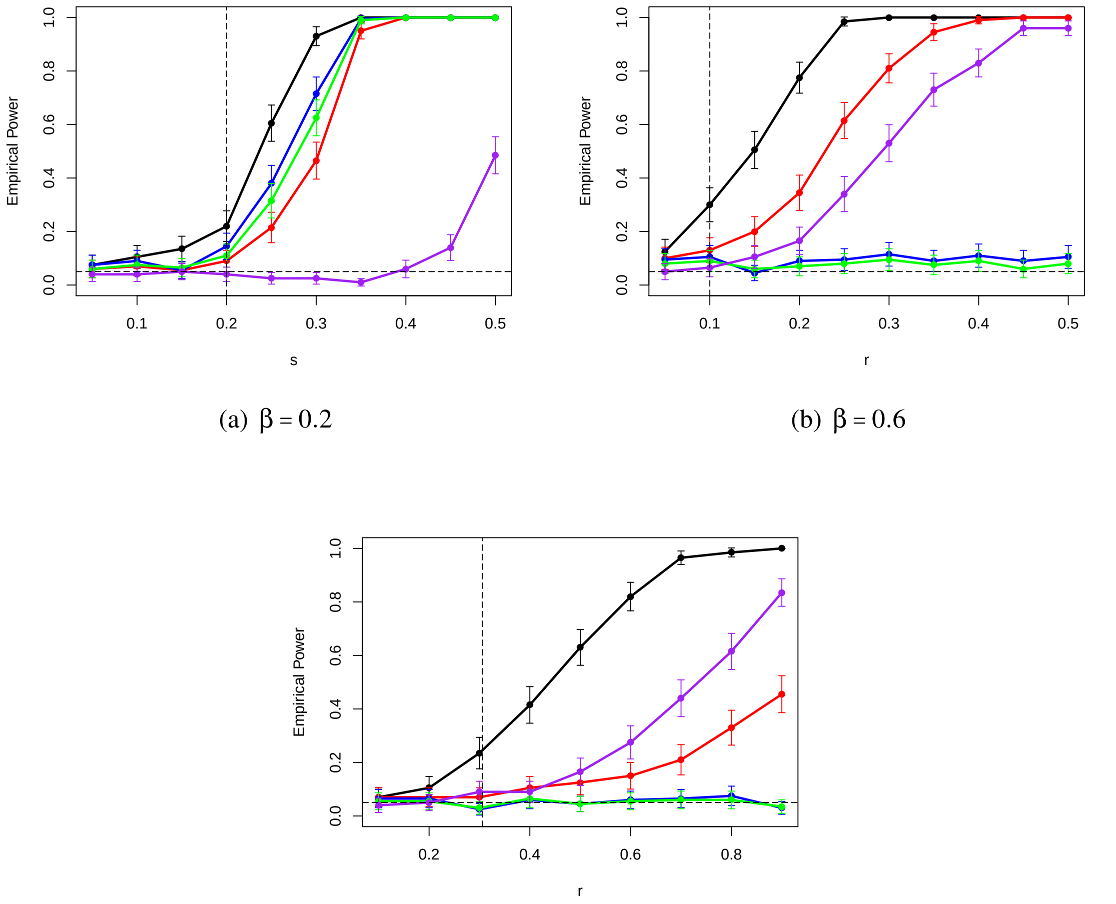
*Dense regime.* We set $\beta = 0.2$ and $\mu_n = n^{s-1/2}$ with $s$ ranging from 0.05 to 0.5 with increments of 0.05. The two-sample HC test, the Wilcoxon test and the two-sample KS test perform comparable to the LRT, while the tail-run test is obviously suboptimal.

*Moderately sparse regime.* We set $\beta = 0.6$ and $\mu_n = \sqrt{2r \log n}$ with $r$ ranging from 0.05 to 0.5 with increments of 0.05. The two-sample HC performs slightly worse than the LRT but better than the tail-run test, while the Wilcoxon test and the KS test are powerless.

*Very sparse regime.* We set $\beta = 0.8$ and $\mu_n = \sqrt{2r \log n}$ with $r$ ranging from 0.1 to 0.9 with increments of 0.1. Though our theory show that both the two-sample HC test and the tail-run test achieve the detection boundary, they both perform significantly below the LRT. The tail-run test is more powerful than the two-sample HC test, which is consistent with the observation in the one-sample setting [2].

## Double-exponential mixture model

In this model, $F$ is double-exponential with variance 1. The simulation results are reported in Figure 4.2. The results are largely congruent with our theory.



(a) $\beta = 0.2$

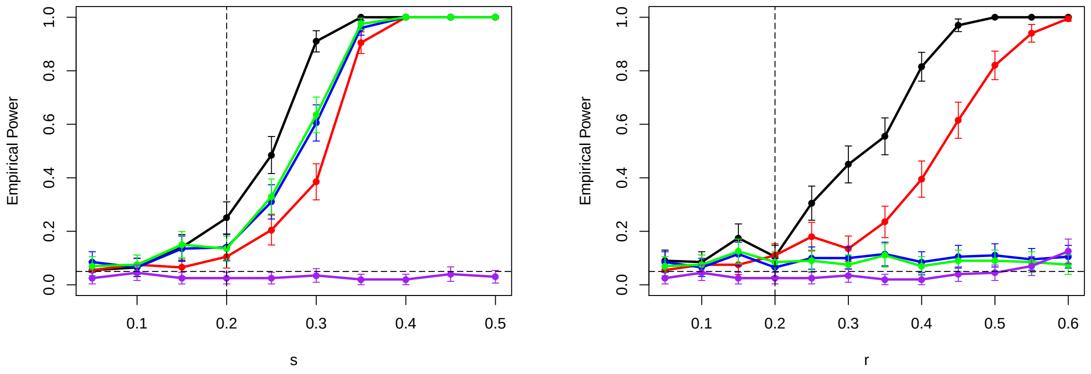
(b) $\beta = 0.6$

(c) $\beta = 0.8$

**Figure 4.1**: Empirical power comparison with 95% error bars for the likelihood ratio test (black), the two-sample higher criticism test (red), the Wilcoxon test (blue), the two-sample Kolmogorov-Smirnov test (green) and the tail-run test (purple). (a) Dense regime where $\beta = 0.2$. (b) Moderately sparse regime where $\beta = 0.6$. (c) Very sparse regime where $\beta = 0.8$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $m = n = 10^5$ and the power curves and error bars are based on 200 replications.

## 4.6 Acknowledgements

Chapter 4, in full, is a version of the paper "Detecting Sparse Heterogeneous Mixtures in a Two-sample Problem", Huang, Rong. The manuscript is currently being prepared for submission for publication. The dissertation author was the primary investigator and author of this material.



(a) $\beta = 0.2$

(b) $\beta = 0.6$

**Figure 4.2**: Empirical power comparison with 95% error bars for the likelihood ratio test (black), the two-sample higher criticism test (red), the Wilcoxon test (blue), the two-sample Kolmogorov-Smirnov test (green) and the tail-run test (purple). (a) Dense regime where $\beta = 0.2$. (b) Moderately sparse regime where $\beta = 0.6$. (c) Very sparse regime where $\beta = 0.8$. The horizontal line marks the level (set at 0.05) and the vertical line marks the asymptotic detection boundary derived earlier. The sample size is $m = n = 10^5$ and the power curves and error bars are based on 200 replications.

# Chapter 5

# Sensitivity Analysis of Treatment Effect to Unmeasured Confounding in Observational Studies with Survival and Competing Risks Outcomes

## 5.1 Introduction

One widely used yet untestable assumption when analyzing data from observational studies is that there is no unobserved confounding, which means that the treatment received and the potential outcomes are independent conditional on the observed pre-treatment covariates. Sensitivity analysis offers an approach to assess the extent to which the inference is robust to violation of this assumption. Rosenbaum [49] contains a nice introduction describing the idea based on association between the unobserved confounder and the treatment, and between the unobserved confounder and the outcome. Analytical approaches have been developed for simpler outcomes such as binary [38], as well as for survival outcomes under the assumption that the event is rare or the effect of the unmeasured confounder on the survival time is small [37]. Li *et al.* [35] and Shen *et al.* [50] considered sensitivity analysis methods for inverse probability weighted (IPW) estimators using propensity scores, an approach that was gaining popularity in practice [6].

Our motivation came from studies in inflammatory bowel disease (IBD). IBD is an umbrella term for two conditions, ulcerative colitis (UC) and Crohn's disease (CD), that are characterized by chronic inflammation of the gastrointestinal tract [31]. With rapid growth in treatment options, head-to-head comparisons are entirely lacking due to difficulty in performing randomized clinical trials (RCT). In order to compare the effectiveness between Vedolizumab and tumor necrosis factor (TNF)-antagonist therapies for UC and CD patients, data were collected between May 2014 and December 2017 from a North American based consortium registry [44, 23], which is a multi-center collaborative research group where outcomes are pooled for consecutive UC and CD patients treated with biologics. Our primary endpoint is time to clinical remission since treatment initiation. Although data collection was rather extensive and accounted for most known measurable confounders, treatment selection for IBD is known to be preference sensitive and influenced by patient and provider perceptions, experiences, and understandings of potential benefit and risk based on the data available to them, all of which are unmeasurable. We aim to assess to what extent our inference from the data is affected by potentially unmeasured confounding.

Time to clinical remission since treatment initiation is a survival endpoint; however, patients need time to achieve this endpoint. Wide variability exists across centers, patients, and providers, for their preference to proceed with surgery while awaiting response to therapy. Therefore, surgery presents a competing risk to clinical remission, in that surgery prevents the event of achieving clinical remission. In Lukin *et al.* [40] and Bohm *et al.* [10] the authors considered propensity score methods with IPW as the primary approach to account for the observed covariates. However, it is possible that there might be confounders not captured by the observed covariates. To carry out sensitivity analysis for this type of complex outcomes, we found the simulated unobserved confounder approach [14] to be useful, in particular since the analytical approaches seem difficult to derive for competing risks.

The paper is organized as follows. We describe our models in Section 5.2, including both the survival models and the competing risks models. We consider estimation in Section 5.3, using both the Expectation-Maximization (EM) algorithms and a stochastic EM algorithm. In Section 5.4, we demonstrate the performance of our algorithms via simulations. We apply our methods to the IBD data in Section 5.5. Finally, we conclude with discussion in Section 5.6.

## 5.2 Models

### 5.2.1 Survival outcome

Denote $T^0$ a time-to-event outcome, $Z$ a binary treatment indicator, and $\boldsymbol{X}$ a vector of observed covariates. Due to possible right censoring, we observe $T = \min(T^0, C)$ and $\delta = I(T^0 \leq C)$, where $C$ is the censoring time, and $I(\cdot)$ the indicator function. We consider $U$ which represents the portion of unmeasured confounder(s) that is independent of $\boldsymbol{X}$, and will simply refer to $U$ as the unmeasured confounder for the rest of the paper. We assume $U$ to be binary for ease of implementation, although other distributions are possible and will be discussed later. Given $Z$, $\boldsymbol{X}$ and $U$, the hazard rate of $T^0$ is modeled using the Cox proportional hazards (PH) regression [17]:

$$\lambda(t|Z, \boldsymbol{X}, U) = \lambda_0(t)\exp(\tau Z + \boldsymbol{X}'\boldsymbol{\beta} + \zeta U), \tag{5.1}$$

where $\lambda_0(\cdot)$ is the baseline hazard function, and $\tau$, $\boldsymbol{\beta}$ and $\zeta$ are the regression coefficients. In addition, we assume that given $\boldsymbol{X}$ and $U$, $Z$ follows a generalized linear model; for illustration purposes we assume a probit link below, although logistic would be an obvious alternative:

$$\mathbb{P}(Z=1|\boldsymbol{X},U)=\Phi(\boldsymbol{X}'\boldsymbol{\beta}_z+\zeta_z U), \tag{5.2}$$

where $\Phi$ is the standard normal cumulative distribution function (CDF), and $\boldsymbol{\beta}_z$ and $\zeta_z$ are the regression coefficients. In the above $\zeta_z$ and $\zeta$ are sensitivity parameters, which quantify the relationships between the unobserved confounder and the treatment received and the outcome, respectively. Finally, we assume that $U\sim\text{Bernoulli}(\pi)$, and we set $\pi=0.5$.

Our goal is to simulate $U$ given the observed $T$, $\delta$, $Z$ and $\boldsymbol{X}$. We note that if the parameters in the above models are known, then

$$U|T,\delta,Z,\boldsymbol{X}\sim\text{Bernoulli}\left(\frac{\mathbb{P}(T,\delta,Z,U=1|\boldsymbol{X})}{\mathbb{P}(T,\delta,Z|\boldsymbol{X})}\right), \tag{5.3}$$

where $\mathbb{P}(T,\delta,Z,U=u|\boldsymbol{X})$ is the joint probability of $(T,\delta,Z,U=u)$ given $\boldsymbol{X}$ for $u=0,1$, and $\mathbb{P}(T,\delta,Z|\boldsymbol{X})=\mathbb{P}(T,\delta,Z,U=1|\boldsymbol{X})+\mathbb{P}(T,\delta,Z,U=0|\boldsymbol{X})$. In particular,

$$\begin{aligned}\mathbb{P}(T,\delta,Z,U|\boldsymbol{X})=&\mathbb{P}(U|\boldsymbol{X})\cdot\mathbb{P}(Z|\boldsymbol{X},U)\cdot\mathbb{P}(T,\delta|Z,\boldsymbol{X},U)\\ =&\pi^U(1-\pi)^{1-U}\left\{\Phi(\boldsymbol{X}'\boldsymbol{\beta}_z+\zeta_z U)\right\}^Z\left\{1-\Phi(\boldsymbol{X}'\boldsymbol{\beta}_z+\zeta_z U)\right\}^{1-Z}\\ &\cdot\left\{\lambda_0(T)e^{\tau Z+\boldsymbol{X}'\boldsymbol{\beta}+\zeta U}\right\}^{\delta}\exp\left\{-\Lambda_0(T)\cdot e^{\tau Z+\boldsymbol{X}'\boldsymbol{\beta}+\zeta U}\right\}.\end{aligned} \tag{5.4}$$

Expression (5.3) will be used to simulate $U_i$ given the observed $T_i$, $\delta_i$, $Z_i$ and $\boldsymbol{X}_i$, where $(T_i,\delta_i,Z_i,\boldsymbol{X}_i)$ for $i=1,\dots,n$ are independent and identically distributed (i.i.d.) from the distribution of $(T,\delta,Z,\boldsymbol{X})$.

### 5.2.2 Competing risks

In the presence of competing risks, when an event occurs it may be one of $m$ distinct types of failures indexed by $j=1,2,\cdots,m$. Again denote $T^0$ the time-to-event, $Z$ a binary treatment indicator, and $\boldsymbol{X}$ a vector of observed covariates. We observe $T = \min(T^0, C)$, $\delta = I(T^0 \le C)$, $J$ the type of failure if $\delta = 1$ and $J = 0$ otherwise for notational purpose. We again consider an unmeasured binary confounder $U$ that is independent of $\boldsymbol{X}$. The cause-specific hazard function [30] for the $j$-th failure type is $\lambda_j(t|Z,\boldsymbol{X},U) = \lim_{\Delta t \to 0} \mathbb{P}(t \le T^0 < t + \Delta t, J = j | T^0 \ge t, Z, \boldsymbol{X}, U)/\Delta t$. Using the proportional hazards modeling of the cause-specific hazard function, we have:

$$\lambda_j(t|Z,\boldsymbol{X},U) = \lambda_{j0}(t)\exp(\tau_j Z + \boldsymbol{X}'\boldsymbol{\beta_j} + \zeta_j U), \quad j = 1,2,\cdots,m. \tag{5.5}$$

where $\lambda_{j0}(\cdot)$ is the baseline hazard function for type $j$, and ($\tau_j$, $\boldsymbol{\beta_j}$, $\zeta_j$) are the regression coefficients. The interpretation of treatment effect in the presence of competing risks needs caution, and is further discussed in the data analysis section. As before we also assume that given $\boldsymbol{X}$ and $U$, $Z$ follows a generalized linear model (5.2) with a probit link. Then parallel to (5.4) we have the joint probability of $(T,\delta,J,Z,U)$ given $\boldsymbol{X}$ as [30]

$$\begin{aligned}\mathbb{P}(T,\delta,J,Z,U|\boldsymbol{X}) = \pi^U(1-\pi)^{1-U}&\left\{\Phi(\boldsymbol{X}'\boldsymbol{\beta_z}+\zeta_z U)\right\}^Z\left\{1-\Phi(\boldsymbol{X}'\boldsymbol{\beta_z}+\zeta_z U)\right\}^{1-Z}\\ &\cdot\prod_{j=1}^{m}\left\{\lambda_{j0}(T)e^{\tau_j Z+\boldsymbol{X}'\beta_j+\zeta_j U}\right\}^{I(\delta=1,J=j)}\exp\left\{-\Lambda_{j0}(T)\cdot e^{\tau_j Z+\boldsymbol{X}'\beta_j+\zeta_j U}\right\},\end{aligned}$$

where $I(\delta = 1, J = j)$ indicates whether subject had the event $j$. The posterior probability of $U$ is then obtained similar to (5.3). In general, if there are $m$ distinct types of failures, then there would be $m+1$ sensitivity parameters, $\zeta_z, \zeta_1, \cdots, \zeta_m$.

## 5.3 Estimation

In order to simulate $U$ given the observed data, we first need to estimate the unknown parameters. As before denote $n$ the number of subjects. Conditional on the unobserved $U$ as well as $Z$ and $\boldsymbol{X}$, the likelihood function of the survival outcome without competing risks is

$$\begin{aligned}L_1(\tau,\boldsymbol{\beta},\zeta;T,\delta|Z,\boldsymbol{X},U) &= \prod_{i=1}^{n}\lambda(t_i|z_i,\boldsymbol{x_i},u_i)^{\delta_i}\exp\{-\Lambda(t_i|z_i,\boldsymbol{x_i},u_i)\}\\ &= \prod_{i=1}^{n}\left\{\lambda_0(t_i)e^{\tau z_i+\boldsymbol{x}_i'\boldsymbol{\beta}+\zeta u_i}\right\}^{\delta_i}\exp\left\{-\Lambda_0(t_i)e^{\tau z_i+\boldsymbol{x}_i'\boldsymbol{\beta}+\zeta u_i}\right\}.\end{aligned}$$

Similarly, the likelihood function of the competing risks outcome is

$$\begin{aligned}L_1(\tau,\boldsymbol{\beta_j},\zeta_j;T,\delta,J|Z,\boldsymbol{X},U) &= \prod_{i=1}^{n}\prod_{j=1}^{m}\lambda_j(t_i|z_i,\boldsymbol{x_i},u_i)^{\delta_{ij}}\exp\{-\Lambda_j(t_i|z_i,\boldsymbol{x_i},u_i)\}\\ &= \prod_{i=1}^{n}\prod_{j=1}^{m}\left\{\lambda_{j0}(t_i)e^{\tau_j z_i+\boldsymbol{x}_i'\boldsymbol{\beta}_j+\zeta_j u_i}\right\}^{\delta_{ij}}\exp\left\{-\Lambda_{j0}(t_i)e^{\tau_j z_i+\boldsymbol{x}_i'\boldsymbol{\beta}_j+\zeta_j u_i}\right\},\end{aligned} \tag{5.6}$$

where $\delta_{ij} := I(\delta_i = 1, J_i = j)$ indicates whether subject $i$ had event $j$.

### 5.3.1 The EM algorithm

The EM algorithm [19] is a commonly used approach to handle missing data, in this case $U$, in the likelihood function. Let $\boldsymbol{\theta}$ denote the unknown parameters, and $y_i$ the survival outcome for subject $i$. The EM algorithm iterates between the E-steps and the M-steps that are described below, where in the notation the covariate $\boldsymbol{x_i}$ is suppressed which is always being conditioned upon. The initial values can be set using the parameter estimates from the regression models ignoring $U$. We note that the sensitivity parameters, as well as $\pi = 0.5$, are known.

#### E-step

In the E-step we compute the conditional expectation of the log-likelihood of the complete data $(y_i, z_i, u_i)$ given the observed data and the current parameter value $\tilde{\boldsymbol{\theta}}$. For the survival outcome without competing risks, let

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}) &= \mathbb{E}[l(\boldsymbol{\theta};\mathbf{y},\mathbf{z},\mathbf{u})|\mathbf{y},\mathbf{z},\tilde{\boldsymbol{\theta}}] \\
&= \mathbb{E}[l_1(\boldsymbol{\beta},\tau,\lambda_0;\mathbf{y}|\mathbf{z},\mathbf{u})|\mathbf{y},\mathbf{z},\tilde{\boldsymbol{\theta}}] + \mathbb{E}[l_2(\boldsymbol{\beta_z};\mathbf{z}|\mathbf{u})|\mathbf{y},\mathbf{z},\tilde{\boldsymbol{\theta}}] + \mathbb{E}[l_3(\mathbf{u})|\mathbf{y},\mathbf{z},\tilde{\boldsymbol{\theta}}] \\
&:= \mathcal{Q}_1(\boldsymbol{\beta},\tau,\lambda_0) + \mathcal{Q}_2(\boldsymbol{\beta_z}) + \mathcal{Q}_3, \qquad (5.7)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{Q}_1(\boldsymbol{\beta},\tau,\lambda_0) = \sum_{i=1}^{n} \Big[ & \delta_i\{\log\lambda_0(t_i) + \boldsymbol{x_i'}\boldsymbol{\beta} + \zeta\mathbb{E}[u_i|y_i,z_i,\tilde{\boldsymbol{\theta}}] + \tau z_i\} \qquad (5.8) \\
& -\Lambda_0(t_i)\exp\left\{\boldsymbol{x_i'}\boldsymbol{\beta} + \log\mathbb{E}[e^{\zeta u_i}|y_i,z_i,\tilde{\boldsymbol{\theta}}] + \tau z_i\right\}\Big],
\end{aligned}
$$

$$
\mathcal{Q}_2(\boldsymbol{\beta_z}) = \sum_{i=1}^{n}\{z_i\mathbb{E}[\log(\Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i))|y_i,z_i,\tilde{\boldsymbol{\theta}}] + (1-z_i)\mathbb{E}[\log(1-\Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i))|y_i,z_i,\tilde{\boldsymbol{\theta}}]\},
$$

$$
(5.9)
$$

$$
\mathcal{Q}_3 = \sum_{i=1}^{n}\{\log\pi\mathbb{E}[u_i|y_i,z_i,\tilde{\boldsymbol{\theta}}] + \log(1-\pi)\mathbb{E}[1-u_i|y_i,z_i,\tilde{\boldsymbol{\theta}}]\}.
$$

We note that $\mathcal{Q}_3$ is in fact not used in the M-step since it does not involve unknown parameters. As described earlier, given the observed data, $U$ follows Bernoulli $(\tilde{\pi}_i)$ as in (5.3) where $\tilde{\pi}_i$ is calculated based on the current parameter value $\tilde{\boldsymbol{\theta}}$. So for any function $h(u_i)$ in (5.8) and (5.9), we have $\mathbb{E}[h(u_i)|y_i,z_i,\tilde{\boldsymbol{\theta}}] = h(1)\tilde{\pi}_i + h(0)(1-\tilde{\pi}_i)$.

For competing risks outcome, from (5.6) we see that the likelihood function is a product of $m$ likelihoods, one for each type of event with its own type specific parameters. The corresponding $\mathcal{Q}_1$ function is then a sum of $\mathcal{Q}_{1j}(\boldsymbol{\beta_j},\tau_j,\lambda_{j0})$'s, each having the same form as $\mathcal{Q}_1(\boldsymbol{\beta},\tau,\lambda_0)$ above but with parameters $\boldsymbol{\beta_j},\tau_j,\lambda_{j0}$ and data for the event type $j$ instead.

### M-step

From (5.7) it is clear that in the M-step we can update $(\boldsymbol{\beta}, \tau, \lambda_0)$ and $\boldsymbol{\beta}_z$ separately. In order to maximize $\mathcal{Q}_1$, we note that it has the same form as the log-likelihood in a Cox regression model with known offset $\log \mathbb{E}[e^{\zeta u_i} | y_i, z_i, \tilde{\boldsymbol{\theta}}]$, just like the Cox model with random effects [52]. For competing risks again because $\mathcal{Q}_1$ is a sum of $\mathcal{Q}_{1j}(\boldsymbol{\beta}_{\boldsymbol{j}}, \tau_j, \lambda_{j0})$'s for $j = 1, \ldots, m$, each set of parameters $\boldsymbol{\beta}_{\boldsymbol{j}}, \tau_j, \lambda_{j0}$ is updated separately using the Cox model software with offsets, the same way as a single survival outcome.

To maximize $\mathcal{Q}_2$, we have

$$\mathcal{Q}_2(\boldsymbol{\beta}_z) = \sum_{i=1}^{n} \Big( z_i \left[ \log\{\Phi(\boldsymbol{x}_i' \boldsymbol{\beta}_z + \zeta_z)\} \tilde{\pi}_i + \log\{\Phi(\boldsymbol{x}_i' \boldsymbol{\beta}_z)\}(1 - \tilde{\pi}_i) \right]$$

$$+ (1 - z_i) \left[ \log\{1 - \Phi(\boldsymbol{x}_i' \boldsymbol{\beta}_z + \zeta_z)\} \tilde{\pi}_i + \log\{1 - \Phi(\boldsymbol{x}_i' \boldsymbol{\beta}_z)\}(1 - \tilde{\pi}_i) \right] \Big).$$

This function can be maximized using the R function ‘optim’.

### Variance estimation

As in typical nonparametric maximum likelihood inference under semiparametric models, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is estimated by the inverse of a discrete observed information matrix $I(\hat{\boldsymbol{\theta}})$ following the EM algorithm, which is given by Louis' formula [39] based on missing information principle:

$$I(\boldsymbol{\theta}) = \mathbb{E}[-\ddot{l}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}, \mathbf{u}) | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}] - \mathbb{E}[s(\theta; \mathbf{y}, \mathbf{z}, \mathbf{u}) s(\theta; \mathbf{y}, \mathbf{z}, \mathbf{u})' | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}], \tag{5.10}$$

where $\ddot{l}$ and $s$ denote the second and first derivatives of $l$ with respect to $\boldsymbol{\theta}$. The components of $\ddot{l}$ and $s$ are given in the Appendix.

### 5.3.2 The stochastic EM algorithm

Instead of the EM algorithm described above, the stochastic EM algorithm was used in Carnegie *et al.* [14], we think primarily due to its ease of implementation for practitioners as well as intuitive appeal. It is similar to a Monte Carlo EM (MCEM) but in the E-steps only a single $U$ is drawn from the conditional distribution of $U$ given the observed data, so that in the M-steps the parameters are updated using that single sample of $U$ as if it were observed. A typical MCEM would otherwise draw many samples of $U$ in order to approximate the conditional expectations in the E-steps. The E- and M-steps are as described above for the models that we consider in this paper, for both survival and competing risks outcomes.

In order to obtain a more accurate estimate, the whole procedure is repeated $K$ times, and the final estimate of the treatment effect on the survival outcome is $\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k / K$, with the corresponding standard error

$$\hat{\sigma}_{\hat{\tau}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}^2_{\hat{\tau}_k} + \left(1+\frac{1}{K}\right)\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\tau}_k - \hat{\tau})^2}, \tag{5.11}$$

where $\hat{\sigma}^2_{\hat{\tau}_k}$ is the estimated variance of $\hat{\tau}_k$ pretending that the singly sampled $U_k$ is observed. For competing risks we have similarly for type $j$ event $\hat{\tau}_j = \sum_{k=1}^{K} \hat{\tau}_{jk} / K$, and the corresponding standard error is obtained using (5.11) with $\hat{\tau}_k$ replaced by $\hat{\tau}_{jk}$ and $\hat{\tau}$ replaced by $\hat{\tau}_j$.

Nielsen *et al.* [45] studied the asymptotic behavior of the stochastic EM algorithm, and showed that under certain assumptions it is root-$n$ consistent but not fully efficient. We show in our data analysis that it can be naturally adapted to the IPW approach and obtain inferential results in sensitivity analysis.

The implementation of R scripts used in the simulation and data analysis below are available on Github: https://github.com/Rong0707/sensitivity_survival

## 5.4 Simulations

We conducted simulation studies to investigate the performance of the EM as well as the stochastic EM algorithms, as compared to the estimation of the treatment effect using the true confounder $U$ with the given sensitivity parameters. For both survival and competing risks outcomes, we set sample size $n = 1{,}000$, $U \sim$ Bernoulli(0.5), and two independent covariates $X_1 \sim N(0,1), X_2 \sim N(1,1)$ with $\boldsymbol{\beta}_z = (0.25, -0.25)'$ in (5.2). The number of EM or stochastic EM steps was set to 20 (see Figure 5.1 and related discussion below), and true sensitivity parameter values were used in fitting the models. The final estimates from the stochastic EM were obtained by averaging over $K = 40$ estimates to reduce the variability. For each case we show the results of 200 simulation runs.

### 5.4.1 Survival outcome

To simulate survival outcomes under model (5.1), we set $\lambda_0(t) = 1$, $\boldsymbol{\beta} = (0.5, -1)'$ and $\tau = 1$. In addition, we set censoring times $C \sim$ Uniform(1, 2) which led to between 25~60% censoring, depending on the combinations of the parameter values.

We run simulations over each combination of $\zeta_z \in \{0, 1, 2\}$ and $\zeta \in \{-2, -1, 0, 1, 2\}$. The results of the simulation are reported in Table 5.1 and Supplement Figure A.1. From the table and figure it is clear that ignoring $U$ led to bias in the estimated treatment effect as long as $\zeta \neq 0$; this bias also increases with the magnitude of $\zeta$ as well as the magnitude of $\zeta_z$. On the other hand, both the stochastic EM and the EM algorithm gave good estimates of the treatment effect compared with the estimates using the true $U$'s. Closer comparison of the results in Table 5.1 shows that the EM algorithm gave more accurate estimates than the stochastic EM algorithm, both in terms of generally less bias and smaller variances.

**Table 5.1**: Estimated treatment effect (standard deviation) for the simulated survival data with $\tau = 1$.

| | Method | $\zeta_z = 0$ | $\zeta_z = 1$ | $\zeta_z = 2$ |
|---|---|---|---|---|
| $\zeta = -2$ | True $U$ | 1.0171 (0.1244) | 1.0108 (0.1171) | 1.0033 (0.1166) |
| | EM | 1.0216 (0.1496) | 1.0101 (0.1469) | 1.0015 (0.1297) |
| | Sto EM | 1.0257 (0.1502) | 1.0132 (0.1468) | 1.0066 (0.1296) |
| | No $U$ | 0.7873 (0.1219) | 0.1512 (0.1257) | -0.2052 (0.1141) |
| $\zeta = -1$ | True $U$ | 1.0206 (0.1015) | 1.0144 (0.1067) | 1.0129 (0.1071) |
| | EM | 1.0203 (0.1153) | 1.0121 (0.1173) | 1.0104 (0.1103) |
| | Sto EM | 1.0220 (0.1157) | 1.0125 (0.1172) | 1.0118 (0.1105) |
| | No $U$ | 0.9310 (0.1068) | 0.5664 (0.1109) | 0.3524 (0.1068) |
| $\zeta = 0$ | True $U$ | 1.0159 (0.0868) | 1.0124 (0.0996) | 1.0095 (0.1035) |
| | EM | 1.0159 (0.0868) | 1.0124 (0.0996) | 1.0095 (0.1035) |
| | Sto EM | 1.0159 (0.0868) | 1.0124 (0.0996) | 1.0095 (0.1035) |
| | No $U$ | 1.0159 (0.0868) | 1.0124 (0.0996) | 1.0095 (0.1035) |
| $\zeta = 1$ | True $U$ | 1.0148 (0.0797) | 1.0134 (0.0896) | 1.0110 (0.1004) |
| | EM | 1.0188 (0.0891) | 1.0167 (0.0977) | 1.0139 (0.1072) |
| | Sto EM | 1.0195 (0.0894) | 1.0183 (0.0977) | 1.0164 (0.1068) |
| | No $U$ | 0.9059 (0.0802) | 1.2601 (0.0878) | 1.4993 (0.0971) |
| $\zeta = 2$ | True $U$ | 1.0133 (0.0768) | 1.0164 (0.0875) | 1.0154 (0.1031) |
| | EM | 1.0226 (0.1047) | 1.0260 (0.1122) | 1.0263 (0.1218) |
| | Sto EM | 1.0225 (0.1052) | 1.0271 (0.1127) | 1.0303 (0.1228) |
| | No $U$ | 0.6946 (0.0783) | 1.2618 (0.0835) | 1.6734 (0.0942) |

### 5.4.2 Competing risks outcomes

To simulate competing risks outcomes, we followed the approach designed in Beyersmann *et al.* [7]. We assumed that $m = 2$, the baseline hazard functions for type 1 and type 2 failures to be $\lambda_{10}(t) = \lambda_{20}(t) = 1$, and $\boldsymbol{\beta_1} = (0.5, -1)'$, $\tau_1 = 1$, $\boldsymbol{\beta_2} = (-0.5, 0.2)'$, $\tau_2 = -1$ in model (5.5). We then simulated the survival times with all-causes hazard $\lambda = \lambda_1 + \lambda_2$, and the cause $J$ was generated from Bernoulli trials with $P(J = 1|Z, \boldsymbol{X}, U) = \lambda_1/(\lambda_1 + \lambda_2)$. We also set censoring times $C \sim$ Uniform(0.3, 0.7).

Similarly as the survival model, we first ran simulations over each combination of $\zeta_z \in \{0, 1, 2\}$ and $\zeta_1 = \zeta_2 \in \{-2, -1, 0, 1, 2\}$. This gave about 20~60% censoring, depending on the combinations of the parameter values, and about equal numbers of type 1 and type 2 events. In a second scenario, we fixed $\zeta_1 = 1$ and $\zeta_2 \in \{-2, -1, 0, 1, 2\}$ as before, which gave about 20~40% censoring, and type 1/2 event rates between 40/20% and 30/50%, again depending on the combinations of the parameter values. The results of experiments are reported in Table 5.2 - Table 5.5 and Figure A.2 - Figure A.5. All results show that for each type of failure, the estimated treatment effect by either the stochastic EM or the EM recovered the true treatment effect quite well, while ignoring $U$ induced a substantial bias. In particular, Table 5.4 and Supplement Figure A.4. show that varying $\zeta_2$ had a noticeable impact on the estimation of $\tau_1$, i.e. unobserved confounding for type 2 failure had a noticeable impact on the estimation of the treatment effect on type 1 failure.

Finally, we take a closer look at the EM and the stochastic EM algorithm in a single run. Figure 5.1 plots the values of the corresponding $\hat{\tau}_j$'s during the first 50 EM or stochastic EM steps. Such plots are often used to examine the behavior and convergence of EM type algorithms for a given data set. It is seen that the EM sequence displays a much smoother line than the stochastic EM sequence; and even at convergence, the stochastic EM sequence has quite some fluctuation compared to the EM sequence.

**Table 5.2**: Treatment effect estimate (standard deviation) on type 1 failures for the simulated competing risks data with $\tau_1 = 1$.

| | method | $\zeta_z = 0$ | $\zeta_z = 1$ | $\zeta_z = 2$ |
|---|---|---|---|---|
| $\zeta_1 = \zeta_2 = -2$ | True *U* | 1.0150 (0.1428) | 1.0293 (0.1603) | 1.0305 (0.1644) |
| | EM | 1.0154 (0.1767) | 1.0357 (0.1801) | 1.0354 (0.1831) |
| | StoEM | 1.0180 (0.1788) | 1.0390 (0.1825) | 1.0428 (0.1844) |
| | No *U* | 0.9312 (0.1583) | 0.3192 (0.1670) | -0.0215 (0.1678) |
| $\zeta_1 = \zeta_2 = -1$ | True *U* | 1.0141 (0.1327) | 1.0185 (0.1542) | 1.0269 (0.1613) |
| | EM | 1.0141 (0.1390) | 1.0186 (0.1585) | 1.0280 (0.1677) |
| | StoEM | 1.0150 (0.1388) | 1.0191 (0.1584) | 1.0304 (0.1675) |
| | No *U* | 0.9817 (0.1339) | 0.6243 (0.1536) | 0.4260 (0.1635) |
| $\zeta_1 = \zeta_2 = 0$ | True *U* | 1.0153 (0.1212) | 1.0258 (0.1329) | 1.0317 (0.1593) |
| | EM | 1.0153 (0.1212) | 1.0258 (0.1329) | 1.0317 (0.1593) |
| | StoEM | 1.0153 (0.1212) | 1.0258 (0.1329) | 1.0317 (0.1593) |
| | No *U* | 1.0153 (0.1212) | 1.0258 (0.1329) | 1.0317 (0.1593) |
| $\zeta_1 = \zeta_2 = 1$ | True *U* | 1.0078 (0.1115) | 1.0276 (0.1259) | 1.0348 (0.1524) |
| | EM | 1.0088 (0.1226) | 1.0251 (0.1415) | 1.0320 (0.1549) |
| | StoEM | 1.0095 (0.1230) | 1.0272 (0.1422) | 1.0346 (0.1549) |
| | No *U* | 0.9745 (0.1130) | 1.3518 (0.1319) | 1.5750 (0.1470) |
| $\zeta_1 = \zeta_2 = 2$ | True *U* | 1.0064 (0.1094) | 1.0263 (0.1238) | 1.0345 (0.1570) |
| | EM | 1.0022 (0.1370) | 1.0174 (0.1600) | 1.0266 (0.1645) |
| | StoEM | 1.0031 (0.1369) | 1.0230 (0.1615) | 1.0345 (0.1652) |
| | No *U* | 0.9238 (0.1141) | 1.5064 (0.1300) | 1.8424 (0.1411) |

**Table 5.3**: Treatment effect estimate (standard deviation) on type 2 failures for the simulated competing risks data with $\tau_2 = -1$.

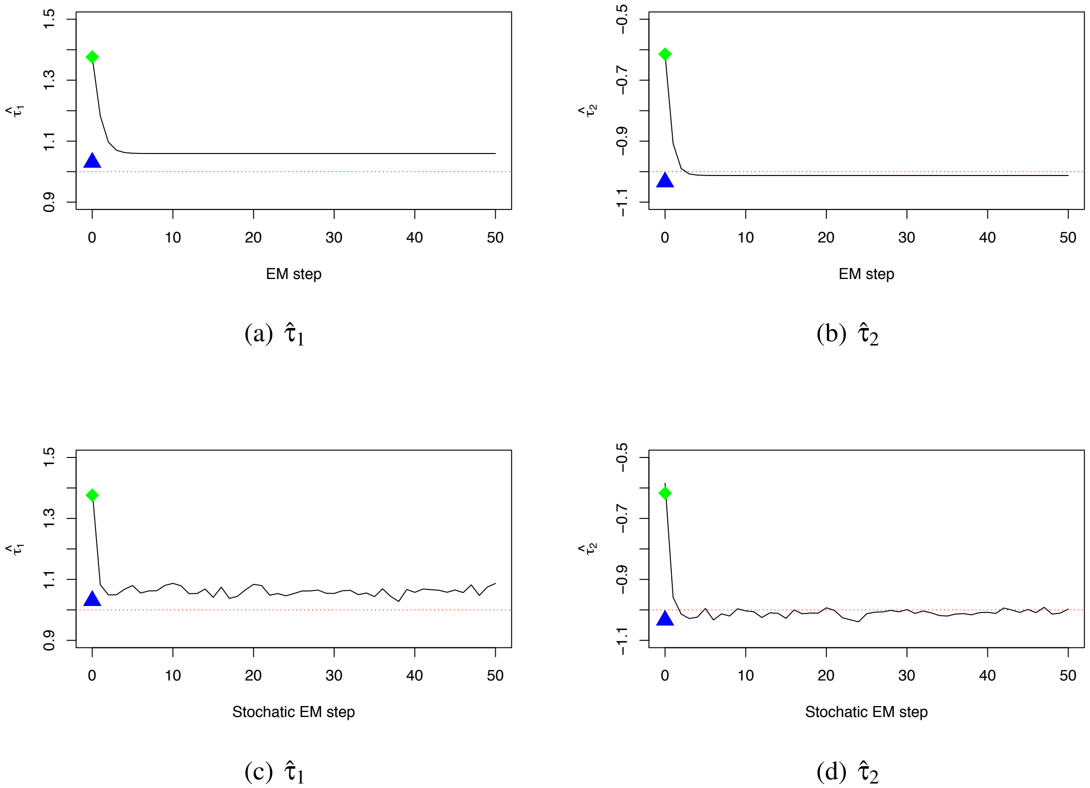
| | method | $\zeta_z = 0$ | $\zeta_z = 1$ | $\zeta_z = 2$ |
|---|---|---|---|---|
| $\zeta_1 = \zeta_2 = -2$ | true $U$ | -1.0170 (0.1974) | -0.9969 (0.1829) | -0.9994 (0.1707) |
| | EM | -1.0170 (0.2143) | -0.9892 (0.1837) | -0.9937 (0.1826) |
| | stoEM | -1.0176 (0.2163) | -0.9892 (0.1845) | -0.9911 (0.1822) |
| | no $U$ | -0.9893 (0.2002) | -1.6139 (0.1713) | -2.0409 (0.1709) |
| $\zeta_1 = \zeta_2 = -1$ | true $U$ | -1.0236 (0.1702) | -0.9996 (0.1484) | -0.9964 (0.1469) |
| | EM | -1.0231 (0.1772) | -0.9945 (0.1518) | -0.9940 (0.1517) |
| | stoEM | -1.0233 (0.1774) | -0.9951 (0.1518) | -0.9936 (0.1515) |
| | no $U$ | -1.0109 (0.1728) | -1.3505 (0.1482) | -1.5975 (0.1485) |
| $\zeta_1 = \zeta_2 = 0$ | true $U$ | -1.0152 (0.1472) | -0.9851 (0.1189) | -0.9797 (0.1238) |
| | EM | -1.0152 (0.1472) | -0.9851 (0.1189) | -0.9797 (0.1238) |
| | stoEM | -1.0152 (0.1472) | -0.9851 (0.1189) | -0.9797 (0.1238) |
| | no $U$ | -1.0152 (0.1472) | -0.9851 (0.1189) | -0.9797 (0.1238) |
| $\zeta_1 = \zeta_2 = 1$ | true $U$ | -1.0088 (0.1316) | -0.9886 (0.1019) | -0.9836 (0.1097) |
| | EM | -1.0108 (0.1423) | -0.9950 (0.1109) | -0.9887 (0.1115) |
| | stoEM | -1.0110 (0.1434) | -0.9953 (0.1117) | -0.9883 (0.1114) |
| | no $U$ | -0.9776 (0.1328) | -0.5956 (0.1039) | -0.3335 (0.1074) |
| $\zeta_1 = \zeta_2 = 2$ | true $U$ | -1.0069 (0.1281) | -0.9950 (0.0976) | -0.9910 (0.1069) |
| | EM | -1.0180 (0.1647) | -1.0093 (0.1322) | -1.0029 (0.1275) |
| | stoEM | -1.0184 (0.1662) | -1.0070 (0.1338) | -0.9983 (0.1282) |
| | no $U$ | -0.9075 (0.1313) | -0.2871 (0.1038) | 0.1522 (0.1050) |

**Table 5.4**: Treatment effect estimate (standard deviation) on type 1 failures for the simulated competing risks data with $\tau_1 = 1$ and $\zeta_1 = 1$ fixed.

| | method | $\zeta_z = 0$ | $\zeta_z = 1$ | $\zeta_z = 2$ |
|---|---|---|---|---|
| $\zeta_2 = -2$ | true *U* | 0.9971 (0.1011) | 1.0187 (0.1163) | 1.0271 (0.1505) |
| | EM | 0.9963 (0.1093) | 1.0174 (0.1320) | 1.0263 (0.1545) |
| | stoEM | 0.9971 (0.1101) | 1.0189 (0.1320) | 1.0282 (0.1549) |
| | no *U* | 0.8893 (0.0998) | 1.2872 (0.1214) | 1.5816 (0.1468) |
| $\zeta_2 = -1$ | true *U* | 1.0020 (0.1008) | 1.0212 (0.1227) | 1.0292 (0.1508) |
| | EM | 1.0028 (0.1080) | 1.0203 (0.1369) | 1.0280 (0.1544) |
| | stoEM | 1.0034 (0.1091) | 1.0214 (0.1368) | 1.0299 (0.1548) |
| | no *U* | 0.9023 (0.0992) | 1.2980 (0.1264) | 1.5844 (0.1465) |
| $\zeta_2 = 0$ | true *U* | 1.0033 (0.0966) | 1.0271 (0.1183) | 1.0326 (0.1481) |
| | EM | 1.0040 (0.1045) | 1.0258 (0.1338) | 1.0305 (0.1507) |
| | stoEM | 1.0050 (0.1045) | 1.0275 (0.1338) | 1.0325 (0.1510) |
| | no *U* | 0.9243 (0.0965) | 1.3220 (0.1244) | 1.5863 (0.1431) |
| $\zeta_2 = 1$ | true *U* | 1.0078 (0.1115) | 1.0276 (0.1259) | 1.0348 (0.1524) |
| | EM | 1.0088 (0.1226) | 1.0251 (0.1415) | 1.0320 (0.1549) |
| | stoEM | 1.0095 (0.1230) | 1.0272 (0.1422) | 1.0346 (0.1549) |
| | no U | 0.9745 (0.1130) | 1.3518 (0.1319) | 1.5750 (0.1470) |
| $\zeta_2 = 2$ | true *U* | 1.0148 (0.1231) | 1.0245 (0.1402) | 1.0301 (0.1592) |
| | EM | 1.0135 (0.1353) | 1.0195 (0.1506) | 1.0255 (0.1609) |
| | stoEM | 1.0146 (0.1356) | 1.0223 (0.1511) | 1.0295 (0.1615) |
| | no *U* | 1.0452 (0.1271) | 1.3591 (0.1415) | 1.5158 (0.1520) |

**Table 5.5**: Treatment effect estimate (standard deviation) on type 2 failures for the simulated competing risks data with $\tau_2 = -1$ and $\zeta_1 = 1$ fixed.

| | method | $\zeta_z = 0$ | $\zeta_z = 1$ | $\zeta_z = 2$ |
|---|---|---|---|---|
| $\zeta_2 = -2$ | true *U* | -1.0113 (0.1926) | -1.0100 (0.1818) | -1.0107 (0.1747) |
| | EM | -1.0118 (0.2107) | -0.9985 (0.1886) | -1.0031 (0.1832) |
| | stoEM | -1.0142 (0.2111) | -1.0001 (0.1898) | -1.0002 (0.1831) |
| | no *U* | -0.8715 (0.1996) | -1.4541 (0.1786) | -1.8513 (0.1766) |
| $\zeta_2 = -1$ | true *U* | -1.0224 (0.1803) | -1.0066 (0.1556) | -1.0082 (0.1537) |
| | EM | -1.0222 (0.1887) | -0.9995 (0.1546) | -1.0038 (0.1567) |
| | stoEM | -1.0227 (0.1888) | -1.0001 (0.1549) | -1.0035 (0.1568) |
| | no *U* | -0.9510 (0.1841) | -1.2829 (0.1511) | -1.5228 (0.1547) |
| $\zeta_2 = 0$ | true *U* | -1.0170 (0.1693) | -1.0010 (0.1383) | -0.9957 (0.1329) |
| | EM | -1.0170 (0.1693) | -1.0010 (0.1383) | -0.9957 (0.1329) |
| | stoEM | -1.0170 (0.1693) | -1.0010 (0.1383) | -0.9957 (0.1329) |
| | no *U* | -1.0170 (0.1693) | -1.0010 (0.1383) | -0.9957 (0.1329) |
| $\zeta_2 = 1$ | true *U* | -1.0088 (0.1316) | -0.9886 (0.1019) | -0.9836 (0.1097) |
| | EM | -1.0108 (0.1423) | -0.9950 (0.1109) | -0.9887 (0.1115) |
| | stoEM | -1.0110 (0.1434) | -0.9953 (0.1117) | -0.9883 (0.1114) |
| | no *U* | -0.9776 (0.1328) | -0.5956 (0.1039) | -0.3335 (0.1074) |
| $\zeta_2 = 2$ | true *U* | -1.0054 (0.1153) | -0.9918 (0.0933) | -0.9876 (0.0969) |
| | EM | -1.0143 (0.1513) | -1.0037 (0.1272) | -0.9976 (0.1175) |
| | stoEM | -1.0139 (0.1522) | -1.0026 (0.1276) | -0.9927 (0.1174) |
| | no *U* | -0.7921 (0.1136) | -0.1505 (0.0981) | 0.3065 (0.0977) |



(a) $\hat{\tau}_1$

(b) $\hat{\tau}_2$

(c) $\hat{\tau}_1$

(d) $\hat{\tau}_2$

**Figure 5.1**: Convergence of the EM (top) and stochastic EM (bottom) algorithms in a single simulation run. The red horizontal lines indicate the true values of $\tau_j$'s, and $\zeta_1 = \zeta_2 = \zeta_z = 1$. The blue triangles correspond to the estimated treatment effects with true $U$, the green diamonds correspond to the estimates without $U$, and the black lines show the values of $\hat{\tau}_j$ during the first 50 steps. All sequences met the convergence criterion in 50 steps.

## 5.5 Sensitivity analysis of the IBD data

### 5.5.1 Ulcerative colitis data

Ulcerative colitis (UC) is one type of IBD that occurs in the large intestine (colon) and the rectum, which is characterized clinically by bloody diarrhea and urgency. We are interested in comparing the effectiveness between Vedolizumab and TNF-antagonist therapy for UC patients. The data were collected between May 2014 and December 2017 from the North American based VICTORY consortium registry [44]. In brief, a total of 719 (453 treated with Vedolizumab, 266 with TNF-antagonist) UC patients with a median follow-up of 12 months were included. We focus on the treatment effect of Vedolizumab ($Z = 1$) versus TNF-antagonist ($Z = 0$) on clinical remission, which is defined as resolution of diarrhea, rectal bleeding and urgency. In the Vedolizumab group, 187 patients had clinical remission and no one had surgery, while in the TNF-antagonist group, 100 patients had clinical remission and 3 patients had surgery. Since there were only 3 competing events of surgery, too few to fit any model, we had to simply treat surgery as independent censoring and applied our approach under the survival models (i.e. without competing risks) to approximate the treatment effect of Vedolizumab.

In Lukin *et al.* [40] the propensity score for each subject $i$, denoted $\text{PS}_i$, was calculated using the R package 'twang' [42] based on pre-treatment variables, including age, disease extent, clinical disease severity, UC related hospitalization within the preceding 1-year, prior TNF-antagonist exposure, baseline steroid dependency or refractoriness, concomitant steroid use, and concomitant immunomodulator use.

To be consistent with Lukin *et al.* [40], here we consider a single covariate $X_i = \Phi^{-1}(\text{PS}_i)$ in our models, as this quantity is more likely to be normally distributed than $\text{PS}_i$. In the models without unmeasured confounding ($\zeta_z = \zeta = 0$), the estimates were $\hat{\beta}_z$ (SE) = 1.1002 (0.0926), $\hat{\beta}$ (SE) = -0.3250 (0.0994), and $\hat{\tau}$ (SE) = 0.5756 (0.1423), where 'SE' stands for standard error. We note that $\hat{\beta}_z$ would have been exactly one if, instead of 'twang', probit regression had been used to fit the propensity score model. In addition, the estimated treatment effect $\hat{\tau}$ here was obtained by regression adjustment, compared to the IPW estimate of Lukin *et al.* [40] (see sensitivity analysis for IPW below also).

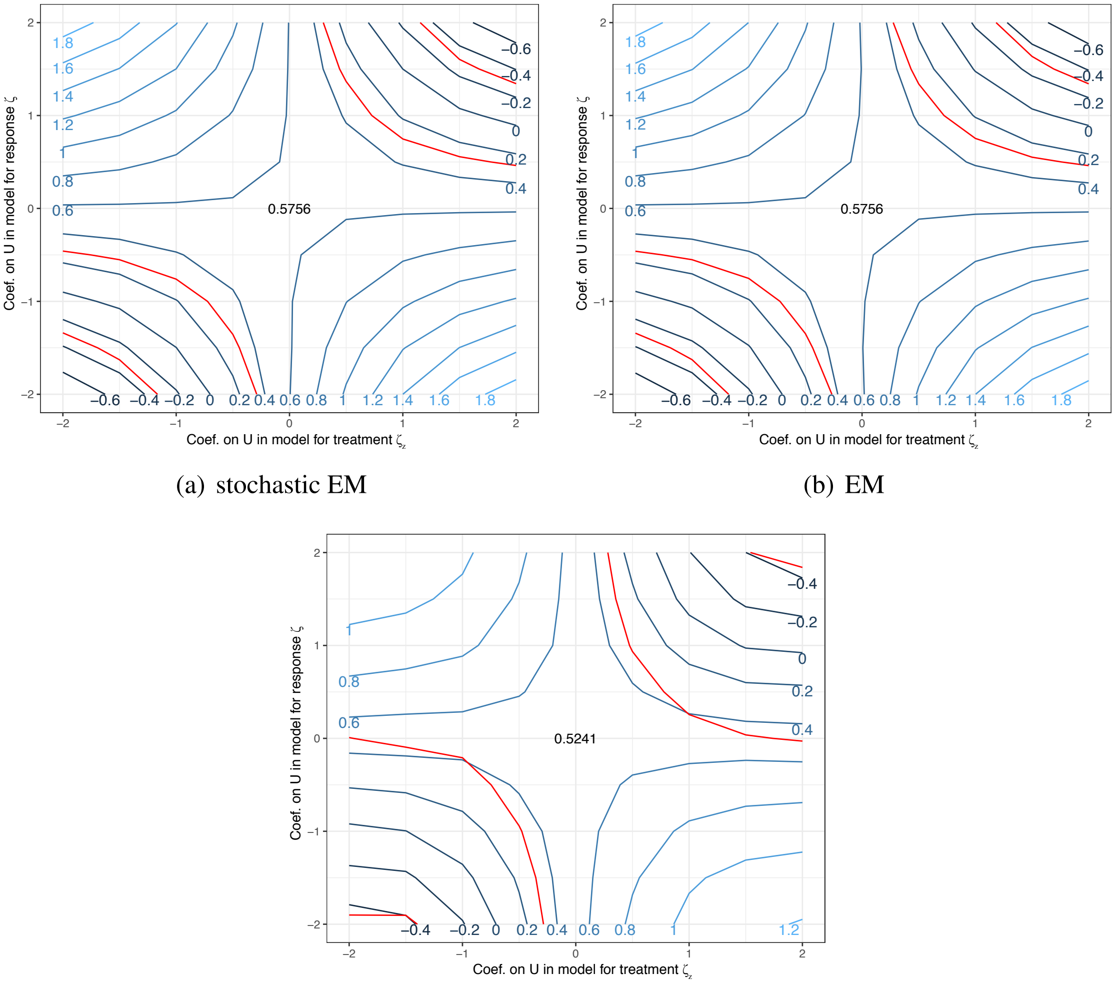
We then assume that there is an unmeasured confounder $U \sim \text{Bernoulli}(0.5)$. To determine the range for the sensitivity parameters, we take into consideration the observed association between a measured confounder and the treatment received or the outcome, in this case all less than one in absolute value in terms of log odds ratio (OR) or log hazard ratio (HR). In addition, a probit coefficient on a binary variable ($U$) is likely to lie in $[-2,2]$ in practice as suggested by Carnegie *et al.* [14]. Similarly under the Cox PH model, the log hazard ratio of ±2 is very substantial for a binary variable. Therefore, we focused on $\zeta_z \in [-2,2]$ and $\zeta \in [-2,2]$.

The EM and stochastic EM algorithms were then applied as described in Section 5.3. The estimates from the stochastic EM were obtained by averaging over $K = 100$ estimates. The sensitivity analysis results are reported in Figure 5.2 panels (a) and (b) and Supplement Table A.1. Figure 5.2 (a) and (b) show that over a wide range of sensitivity parameters, the EM and the stochastic EM gave very similar results. Note that except for very small random fluctuation in the stochastic EM results, the contours and curves are symmetric about the origin $(\zeta_z, \zeta) = (0,0)$, where the estimated $\hat{\tau} = 0.5756$ is marked.

A main usage of these sensitivity plots is to identify the magnitude of the unmeasured confounding, i.e. the sensitivity parameters $(\zeta_z, \zeta)$, needed to alter a conclusion on the treatment effect. This can be reflected in two ways: 1) to drive the estimated treatment effect to zero, or 2) to lead to a non-significant estimated treatment effect in this case. From the plots we see that $(\zeta_z, \zeta)$ will need to be close to (1.5, 1) or (1, 1.5), for example, in order to drive the estimated treatment effect to zero. To understand whether such a magnitude is likely in practice, we may again compare them to the observed association between a measured confounder and the treatment or the outcome, which were all less than one in absolute value in terms of log OR or log HR as we noted earlier (the largest log HR being just under 0.6 in absolute value). We may also compare them to the fitted values of $\hat{\beta}_z = 1.1002$ and $\hat{\beta} = -0.3250$ above. We see that such a very strong association between $U$ and the survival outcome, in particular, seems unlikely.

In Figure 5.2 (a) and (b) any combination of $(\zeta_z, \zeta)$ in the region between two red curves in the upper right or lower left quadrant leads to a non-significant estimated treatment effect at 0.05 level two-sided. For example, $(\zeta_z, \zeta)$ will need to be close to (1, 0.8), in order to drive the estimate to be non-significant. Similar to the discussion above, such a magnitude of unmeasured confounding seems unlikely in practice.

Finally, as IPW with $\text{PS}_i$ was the main statistical approach used in Lukin *et al.* [40] to estimate the treatment effect, we also carried out sensitivity analysis for this approach. We implemented this by combining the stochastic EM with IPW as follows. At convergence of the algorithm we simulated $U_i$ and estimated the propensity score $\mathbb{P}(Z = 1|X, U)$ by regressing $Z_i$ on $X_i = \Phi^{-1}(\text{PS}_i)$ and the simulated $U_i$, $i = 1, \ldots n$. Stabilized weights were obtained and further trimmed to be within (0.1, 10) if necessary. The IPW approach was then applied. The final estimates were also obtained by averaging over $K = 100$ estimates, with the corresponding standard errors obtained using (5.11) where $\hat{\sigma}^2_{\hat{\tau}_k}$ was the sandwich variance estimator following the IPW. The results are reported in Figure 5.2 (c) and Supplement Table A.1. It is seen that unlike the regression adjustment results above, where the estimated treatment effect remained the same as long as $\zeta = 0$, here instead the estimated treatment effect remained the same as long as $\zeta_z = 0$. We also note much larger standard errors when $|\zeta_z|$ is large, perhaps understandable as the treatment groups become more imbalanced. However, similar to the regression adjustment results above, in order to drive the estimated treatment effect to zero, $(\zeta_z, \zeta)$ will need to be close to (1.5, 1) or (1, 1.5). On the other hand, the estimated treatment effect may become non-significant at 0.05 level if $(\zeta_z, \zeta) = (0.5, 1)$ or $(\zeta_z, \zeta) = (0.8, 0.5)$.



(a) stochastic EM

(b) EM

(c) IPW

**Figure 5.2**: Sensitivity analysis results for UC patients data for outcome clinical remission. In all plots, the blue contours show the sensitivity parameter values corresponding to the estimated treatment effect $\hat{\tau}$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}/\hat{\sigma}_{\hat{\tau}}| = 1.96$.

### 5.5.2 Crohn's disease data

Crohn's disease (CD) is another type of IBD that can cause inflammation along anywhere of the digestive tract. We are again interested in comparing the effectiveness between Vedolizumab and TNF-antagonist therapy for CD patients. The data were collected between May 2014 and December 2017 from the North American based consortium registry [23]. A total of 1,242 patients were included (655 treated with Vedolizumab, 587 with TNF-antagonist therapy). The primary interest is the treatment effect of Vedolizumab ($Z = 1$) versus TNF-antagonist ($Z = 0$) on clinical remission, which is defined as complete resolution of CD-related symptoms. In the Vedolizumab group, 196 patients had clinical remission and 9 had surgery, while in the TNF-antagonist group, 255 patients had clinical remission and 18 patients had surgery. Supplement Figure A.6 shows the cumulative incidence curves for time to clinical remission and time to surgery in these patients. Due to the presence of competing events, we applied our approach under the competing risks models to estimate the treatment effect of Vedolizumab.

In Bohm *et al.* [10], the propensity score for each subject $i$, denoted $\mathrm{PS}_i$, was calculated using the R package 'twang' [42] based on pre-treatment variables, including prior TNF-antagonist exposure and number of prior TNF-antagonists exposed, disease extent, history of fistulizing disease, prior bowel surgery, disease phentyope, clinical disease severity, CD related hospitalization within the preceding 1-year, baseline steroid dependency or refractoriness, concomitant steroid use, or concomitant immunomodulator use.

To be consistent with Bohm *et al.* [10], we consider a single covariate $X_i = \Phi^{-1}(\mathrm{PS}_i)$ in our models. In the models without unmeasured confounding ($\zeta_z = \zeta_1 = \zeta_2 = 0$), the estimate of $\beta_z$ as defined in model (5.2) is $\hat{\beta}_z$ (SE) = 1.0631 (0.0513), the estimates of $\beta_j$ $(j = 1,2)$ as defined in model (5.5) are $\hat{\beta}_1$ (SE) = −0.1664 (0.0562) and $\hat{\beta}_2$ (SE) = −0.2601 (0.2401), and the estimates of $\tau_j$ $(j = 1,2)$ are $\hat{\tau}_1$ (SE) = 0.0605 (0.1318) and $\hat{\tau}_2$ (SE) = −0.0537 (0.5705).

We then assume an unmeasured confounder $U \sim$ Bernoulli(0.5). The range for the sensitivity parameters is determined similarly as the UC data. We focus on $\zeta_z \in [-2,2]$ and

$\zeta_1 \in [-2,2]$, and $\zeta_2 \in \{-2,0,2\}$. The EM and stochastic EM algorithms were then applied as described in Section 5.3. The estimates from the stochastic EM were obtained by averaging over $K = 100$ estimates. The sensitivity analysis results are reported in Figure 5.3 panels (a) and (b) and Supplement Table A.2 - Table A.4.

Note that by our algorithms, $\zeta_2$ affects $\hat{\tau}_1$ only through the conditional probability of $U$ as shown in (5.3). In these data, as the number of surgery is relatively small compared to the number of clinical remission, the effect of $\zeta_2$ on $\hat{\tau}_1$ is subtle (Supplement Table A.2 - Table A.4.). This is, of course, not necessarily true if the number of the competing risk events is comparable to the number of events of interest. We further discuss the impact of the competing risk towards the end of this analysis.

Figure 5.3 (a) and (b) show that when $\zeta_2 = 0$, over a wide range of $(\zeta_z, \zeta_1)$, the EM and the stochastic EM gave similar results. In the plots, the blue contours show the values of $(\zeta_z, \zeta_1)$ corresponding to the estimated treatment effect $\hat{\tau}_1$, and the red curves correspond to where the absolute value of the $t-$statistic $|t| = |\hat{\tau}_1/\hat{\sigma}_{\hat{\tau}_1}| = 1.96$. Hence, any combination of $(\zeta_z, \zeta_1)$ in the region surrounded by four red curves leads to a non-significant estimated treatment effect at level 0.05 two-sided. Except for very small random fluctuation in the stochastic EM results, the contours and curves are symmetric about the origin $(\zeta_z, \zeta_1) = (0,0)$, where the estimated $\hat{\tau}_1 = 0.0605$ is marked. We see that in order to drive the estimated treatment effect to being significant, given $\zeta_2 = 0$, $(\zeta_z, \zeta_1)$ will need to be close to (1,1) or (−0.8, 0.8), for example. Compared to $\hat{\beta}_z = 1.0631$ and $\hat{\beta}_1 = -0.1664$ above, such a strong association between $U$ and the outcome seems unlikely in practice.

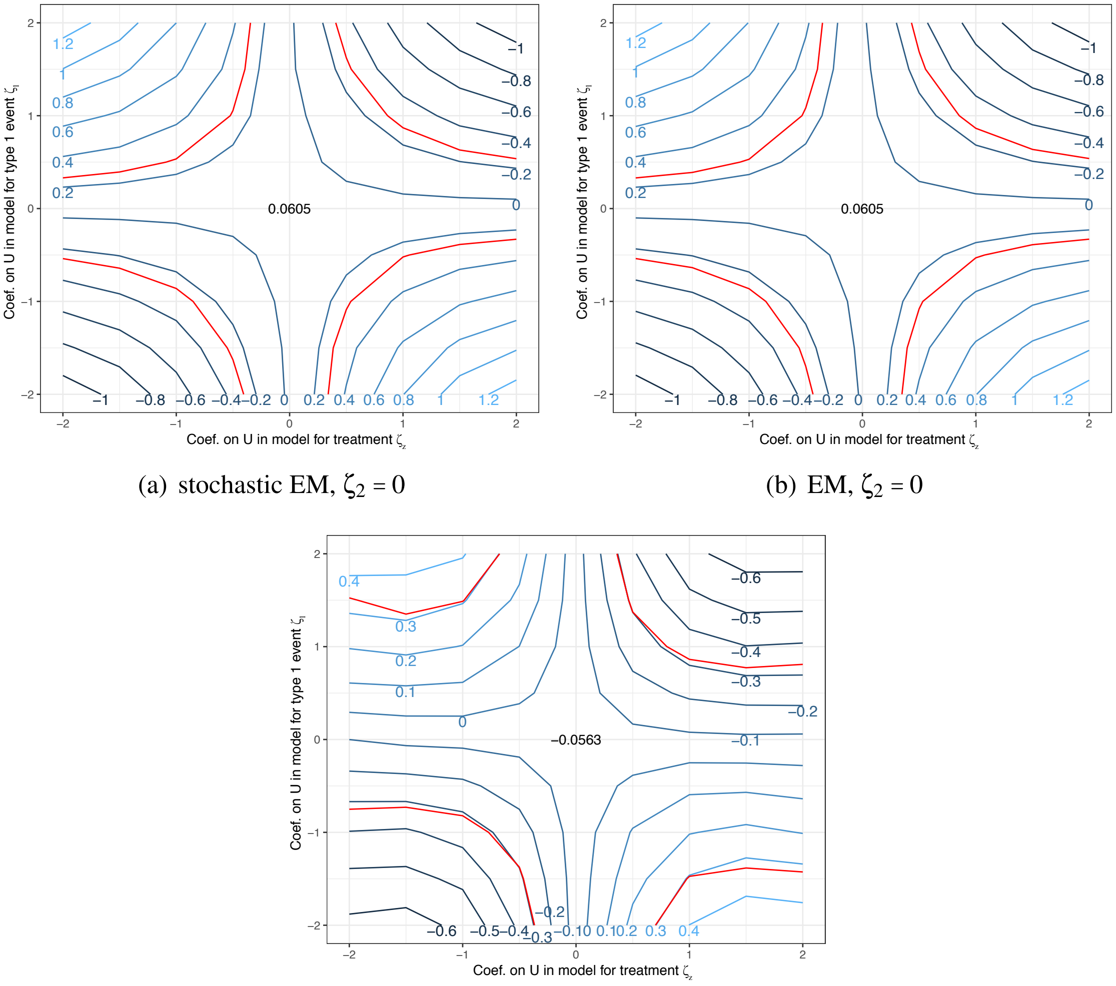
As IPW with PS$_i$ was the main statistical approach used in Bohm *et al.* [10] to estimate the treatment effect, we also carried out sensitivity analysis for this approach by combining the stochastic EM with IPW as under the survival models. The final estimates were also obtained by averaging over $K = 100$ estimates. The results are reported in Figure 5.3 (c) and Supplement Table A.2 - Table A.4. Similar to the regression adjustment results, in order to drive the estimated treatment effect to being significant, given $\zeta_2 = 0$, $(\zeta_z, \zeta_1)$ will need to be close to (1, 1) or (-1, 1.5), which seems unlikely in practice.

We emphasize that the interpretation of treatment effect in the presence of competing risks needs caution in general. Here the comparison of the two treatment groups as reflected in the effect $\tau_1$ for clinical remission is among those without surgery. In this case, the effect of treatment on the competing risk, i.e. time to surgery, is not significant for a broad range of sensitivity parameter values (data not shown). Therefore we are not in a situation where a treatment appears to increase the risk of one type of events while reducing the risk of another type of events, which could otherwise happen in practice as the probabilities from different types of events must sum up to one as time goes to infinity.

Finally, as suggested by a reviewer, we explore the sensitivity analyses when the distribution of $U$, instead of being symmetric, has $\pi = 0.7$ or 0.3. The results are in the Supplement Figure A.7 - Figure A.10. It is seen that for the same values of $\zeta$ or $\zeta_z$, the change in the estimated treatment effect is not as large. A possible explanation is that the magnitude of unmeasured confounding, as reflected in the variance of $U$, is reduced when $\pi = 0.7$ or 0.3. We also note that in this case the contours are no longer symmetric about the origin $(\zeta_z, \zeta_1) = (0, 0)$.

## 5.6 Discussion

In this paper we developed approaches to perform sensitivity analysis of the estimated treatment effect with regard to unobserved confounding in observational studies with survival or competing risks outcomes. The approaches we developed are based on models for survival or competing risks outcomes, which allow simulating the unobserved confounder given the observed data. The sensitivity parameters reflect the association between the unobserved confounder and the outcomes, as well as the association between the unobserved confounder and the treatment assignments. The interpretation of these sensitivity parameters is straightforward, which leads to



(a) stochastic EM, $\zeta_2 = 0$

(b) EM, $\zeta_2 = 0$

(c) IPW, $\zeta_2 = 0$

**Figure 5.3**: Sensitivity analysis results for CD patients data for outcome clinical remission. In all plots, the blue contours show the values of $(\zeta_z, \zeta_1)$ corresponding to the estimated treatment effect $\hat{\tau}_1$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}_1 / \hat{\sigma}_{\hat{\tau}_1}| = 1.96$.

relative ease in choosing plausible ranges for them. Simulation studies show that both the EM and the stochastic EM algorithm are able to recover the true treatment effect if the correct sensitivity parameter values are used. The EM algorithm is clearly optimal in theory [45], although the stochastic EM allows easy incorporation of IPW approaches for estimating treatment effects, which are commonly used in practice and as we have illustrated in our data analysis.

Lin *et al.* [37] developed an analytic approach with closed-form formulas for assessing the sensitivity of regression results to unmeasured confounders in observational studies with either binary or survival outcomes. Under certain conditions, they derived simple algebraic relationships between the true treatment effect and the apparent treatment effect ignoring the unmeasured confounder $U$. For survival data, they assumed that the event was rare or the effect of $U$ on the survival outcome was small. Their parameterization is different from ours here: they parametrized the conditional distribution of $U$ given the treatment $Z$, and assumed that $U$ was independent of the observed covariates $X$ given $Z$; in contrast we have modeled the distribution of $Z$ given $X$ and $U$. The Lin *et al.* approach does not apply to competing risks.

For the distribution of the unobserved confounder we used binary 0, 1 with probability 0.5 each, which were recommended and used throughout the book by Rosenbaum [49]. It is also possible to incorporate normally distributed $U$, such as in Shen *et al.* [50] and Xu *et al.* [54], in which case the probit link in model (5.2) allows closed-form marginal propensity scores given $X$ after integrating out $U$. The $\mathcal{Q}_1$ part of the EM algorithm would be similar to that under the proportional hazards mixed-effects model (PHMM) and Monte Carlo approximation would be needed in the E-steps [52].

Carnegie *et al.* [14] discussed the advantages and disadvantages of using parametric versus nonparametric approaches in sensitivity analysis. Parametric approaches are typically needed in order to simulate the unobserved confounder; in survival analysis however, the outcome models are often semiparametric, allow flexibility in modeling in particular the nuisance parameters. On the other hand, nonparametric bounds might be considered under minimal assumptions in place of sensitivity analysis [48]. However, such bounds can be very difficult to derive for complex outcomes like what we consider here in the presence of right censoring, which is unlike in Shen *et al.* [50] where it is possible to derive these bounds for binary or continuous outcomes without censoring. Also evident in Shen *et al.* [50] is that parametric settings are often needed in order to aid in the interpretation of the sensitivity parameters in the corresponding nonparametric settings, and extensive simulations have to be conducted in order to determine sensible ranges for these sensitivity parameters [54].

## 5.7 Acknowledgement

This work was partially supported by National Institutes of Health, Grant UL1TR001442 of CTSA funding. Chapter 5, in full, is a version of the paper “Sensitivity Analysis of Treatment Effect to Unmeasured Confounding in Observational Studies with Survival and Competing Risks Outcomes”, Huang, Rong; Xu, Ronghui; Dulai, Parambir S. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

## 5.8 Appendix

In the following we write out the components of $\ddot{l}$ and $s$ for competing risks with $j = 1,\dots,m$. For a single survival outcome without competing risks, we should simply take $m = 1$ and the corresponding parameters are the same as without the subscript $j$.

The components of $s$ are:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta_j}} &= \sum_{i=1}^{n} \boldsymbol{x_i}\{\delta_{ij} - \Lambda_{j0}(t_i)\exp(\tau_j z_i + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \zeta_j u_i)\} \\
\frac{\partial l}{\partial \tau_j} &= \sum_{i=1}^{n} z_i\{\delta_{ij} - \Lambda_{j0}(t_i)\exp(\tau_j z_i + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \zeta_j u_i)\} \\
\frac{\partial l}{\partial \lambda_{j0}(t_i)} &= \frac{1}{\lambda_{j0}(t_i)} - \sum_{t_k \ge t_i} \exp(\tau_j z_k + \boldsymbol{x_k'}\boldsymbol{\beta_j} + \zeta_j u_k) \\
\frac{\partial l}{\partial \boldsymbol{\beta_z}} &= \sum_{i=1}^{n} \Big\{ z_i \frac{\phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)}{\Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)} - (1 - z_i)\frac{\phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)}{1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)} \Big\} \boldsymbol{x_i}
\end{aligned}$$

for $j = 1, \cdots, m$. For the second derivatives,

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta_j}^2} &= -\sum_{i=1}^{n} \boldsymbol{x_i}^{\otimes 2} \Lambda_{j0}(t_i)\exp(\tau_j z_i + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \zeta_j u_i) \\
\frac{\partial^2 l}{\partial \tau_j^2} &= -\sum_{i=1}^{n} z_i \Lambda_{j0}(t_i)\exp(\tau_j z_i + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \zeta_j u_i) \\
\frac{\partial^2 l}{\partial \lambda_{j0}(t_i)^2} &= -\frac{1}{\lambda_{j0}(t_i)^2} \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta_j} \partial \tau_j} &= -\sum_{i=1}^{n} z_i \boldsymbol{x_i} \Lambda_{j0}(t_i)\exp(\tau_j z_i + \boldsymbol{x_i'}\boldsymbol{\beta_j} + \zeta_j u_i) \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta_j} \partial \lambda_{j0}(t_i)} &= -\sum_{t_k \ge t_i} \boldsymbol{x_k} \exp(\tau_j z_k + \boldsymbol{x_k'}\boldsymbol{\beta_j} + \zeta_j u_k) \\
\frac{\partial^2 l}{\partial \tau_j \partial \lambda_{j0}(t_i)} &= -\sum_{t_k \ge t_i} z_k \exp(\tau_j z_k + \boldsymbol{x_k'}\boldsymbol{\beta_j} + \zeta_j u_k) \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta_z}^2} &= -\sum_{i=1}^{n} \phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i) \Big\{ z_i \frac{\phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i) + (\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)\Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)}{\Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)^2} \\
&\quad + (1 - z_i) \frac{\phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i) - (\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i)(1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i))}{(1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta_z} + \zeta_z u_i))^2} \Big\} \boldsymbol{x_i}^{\otimes 2}
\end{aligned}$$

where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{aa'}$ for a vector $\boldsymbol{a}$, $\phi$ is the probability density function (pdf) of the standard normal distribution, and all other off-diagonal elements are zeros. The computation of the first term in (5.10) is similar to the computation in the E-step for different functions $h(u_i)$. To calculate the second term in (5.10), we sample $U$ from Bernoulli$(\tilde{\pi})$ for 1,000 times after convergence of the

EM, and take the average of $s(\boldsymbol{\theta};\mathbf{y},\mathbf{z},\mathbf{u})s(\boldsymbol{\theta};\mathbf{y},\mathbf{z},\mathbf{u})'$ over the sampled $U$'s.

# Appendix A

# Supplemental Materials

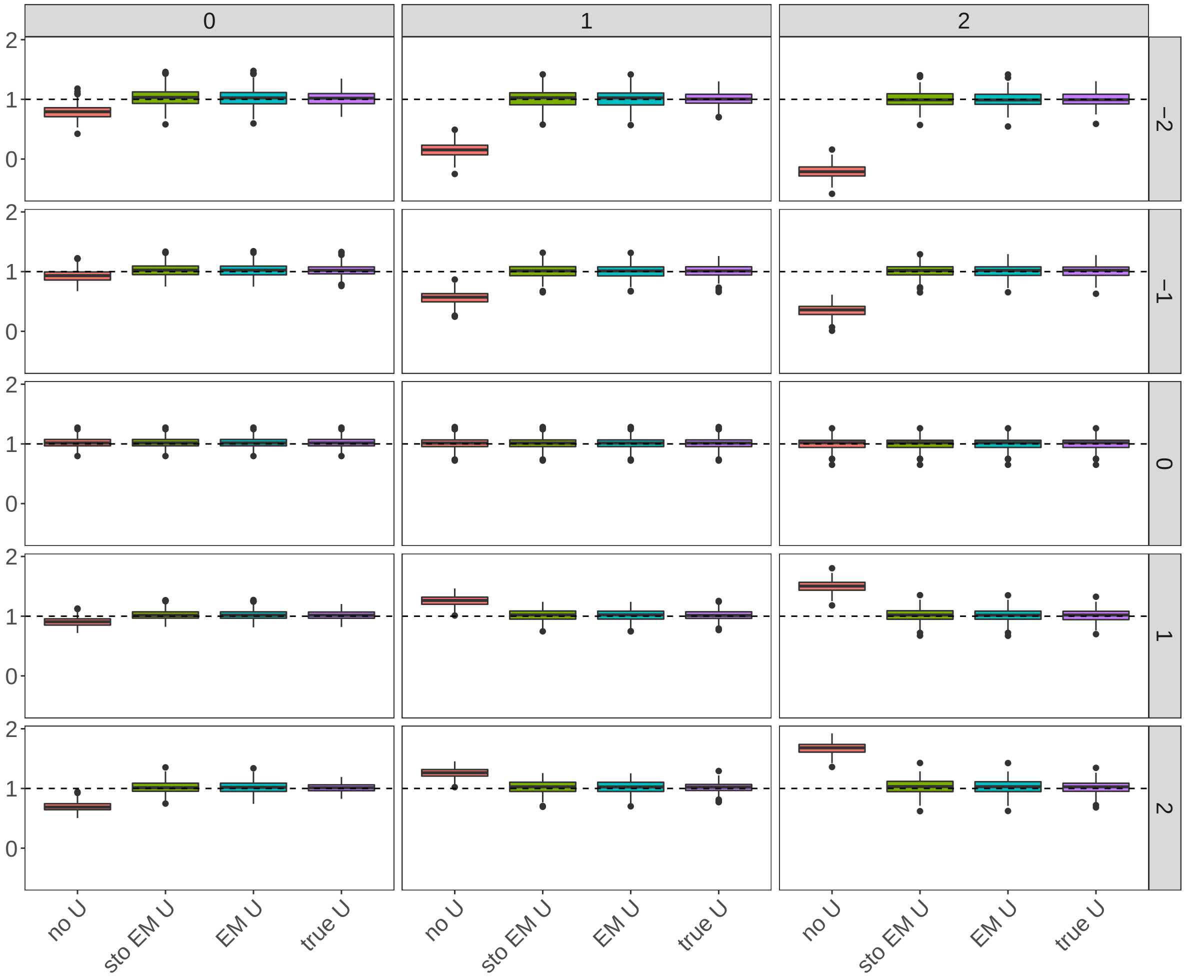


**Figure A.1**: Distributions of the estimated treatment effect ($\hat{\tau}$) for the simulated survival data. $\zeta_z \in \{0,1,2\}$ on the horizontal label and $\zeta \in \{-2,-1,0,1,2\}$ on the vertical label. Each boxplot displays $\hat{\tau}$ from 200 simulations.

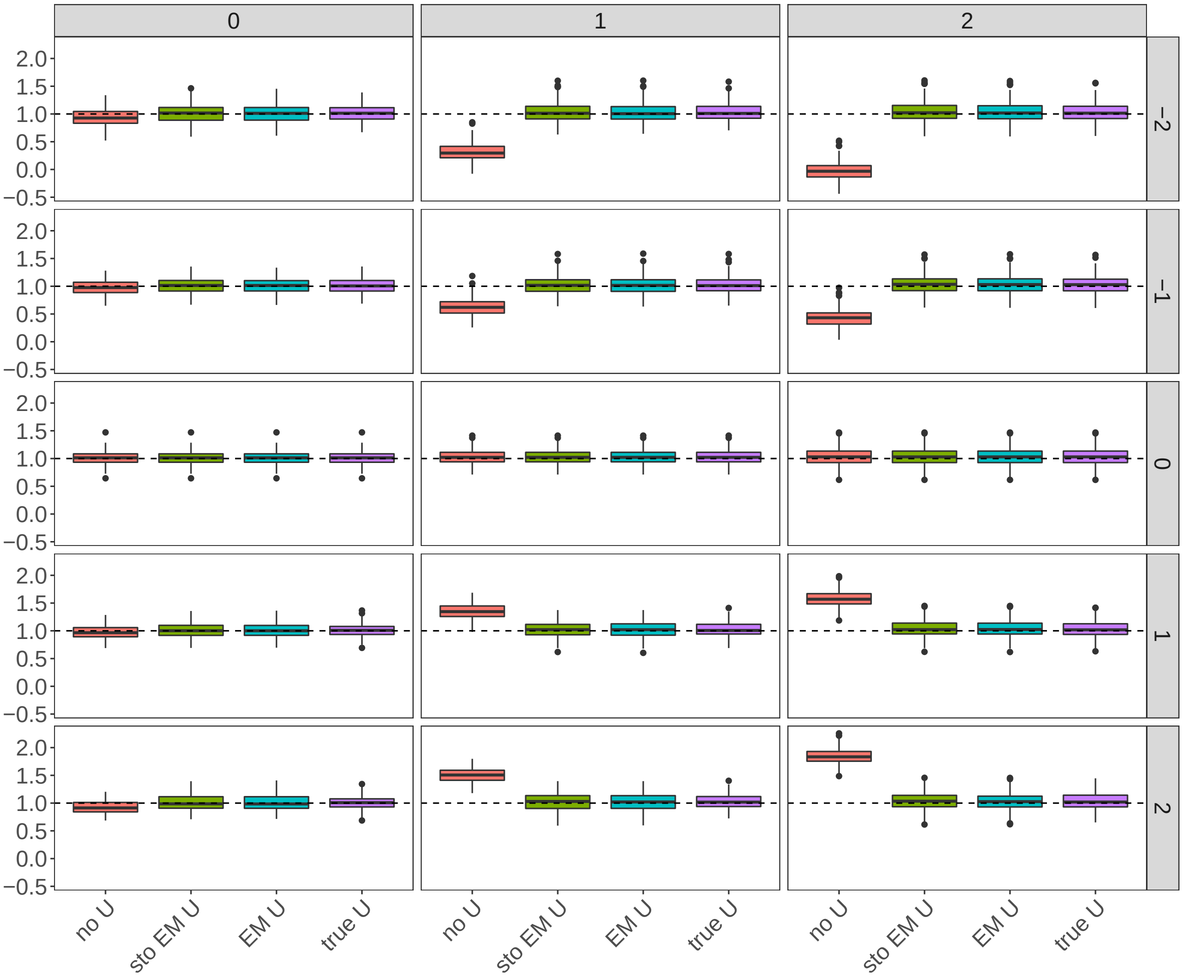


**Figure A.2**: Distributions of the estimated treatment effect on type 1 failures for the simulated competing risks data. $\zeta_z \in \{0, 1, 2\}$ on the horizontal label and $\zeta_1 = \zeta_2 \in \{-2, -1, 0, 1, 2\}$ on the vertical label. Each boxplot displays $\hat{\tau}_1$ from 200 simulation runs.

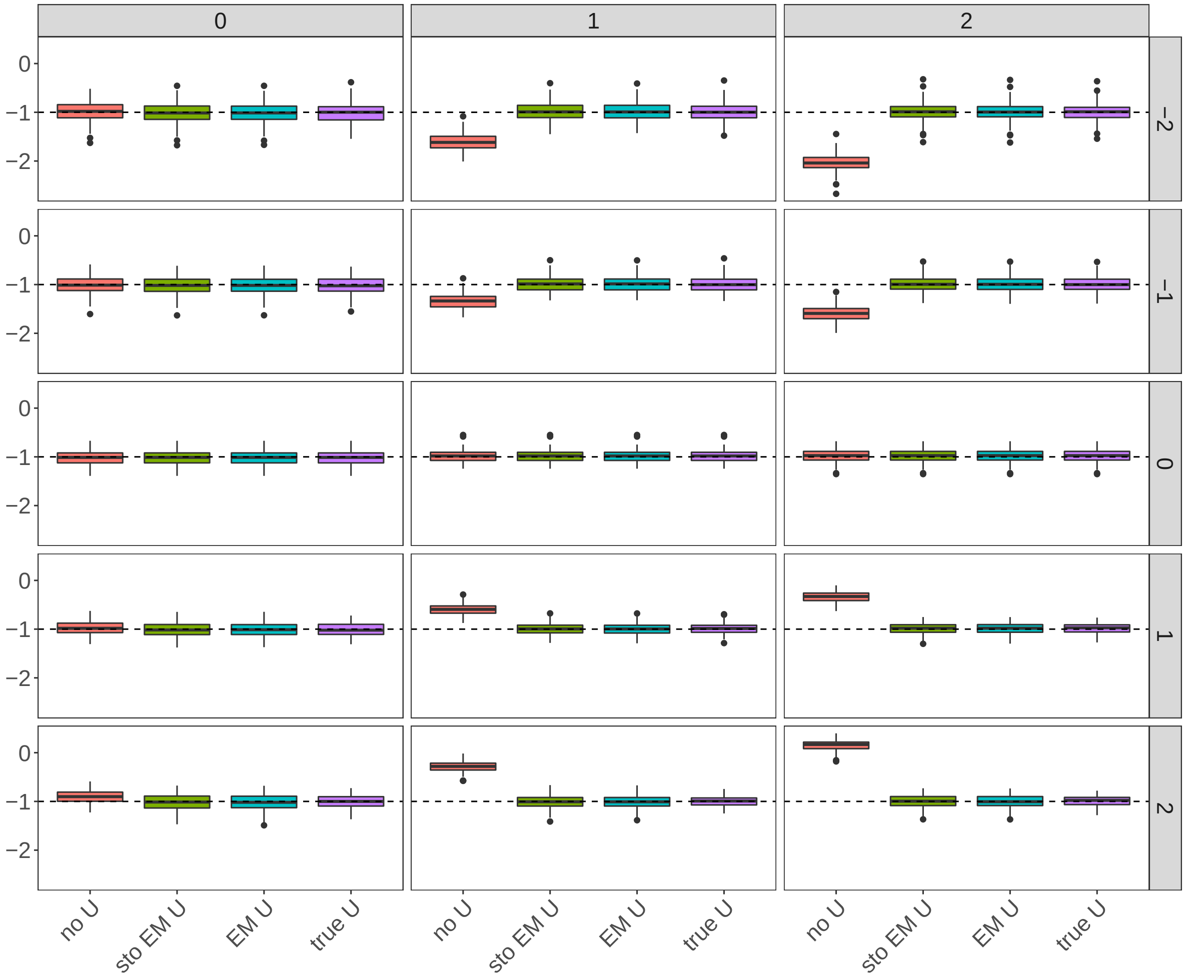


**Figure A.3**: Distributions of the estimated treatment effect on type 2 failures for the simulated competing risks data. $\zeta_z \in \{0,1,2\}$ on the horizontal label and $\zeta_1 = \zeta_2 \in \{-2,-1,0,1,2\}$ on the vertical label. Each boxplot displays $\hat{\tau}_2$ from 200 simulations.

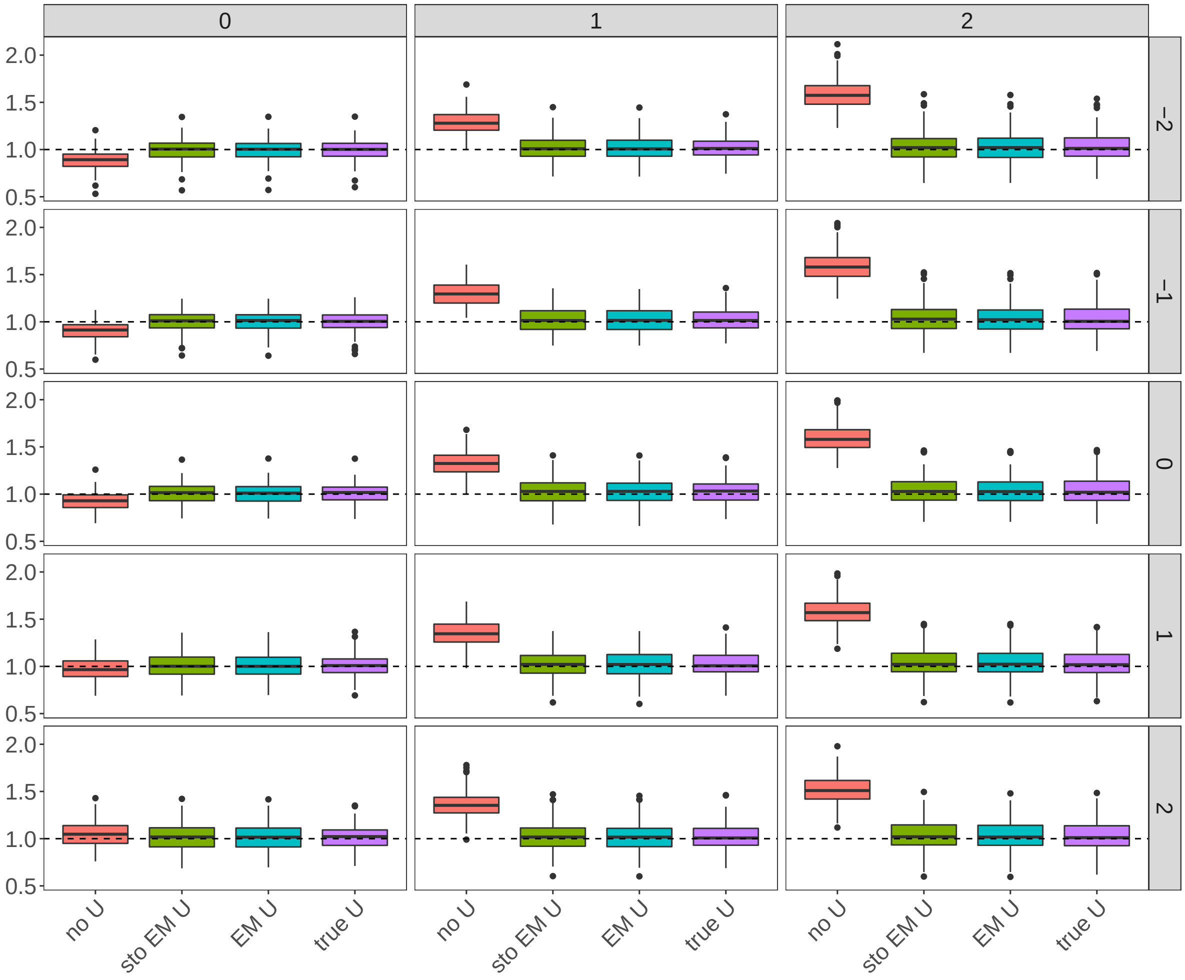


**Figure A.4**: Distributions of the estimated treatment effect on type 1 failures for the simulated competing risks data. $\zeta_z \in \{0,1,2\}$ on the horizontal label, $\zeta_1 = 1$ and $\zeta_2 \in \{-2,-1,0,1,2\}$ on the vertical label. Each boxplot displays $\hat{\tau}_1$ from 200 simulation runs.

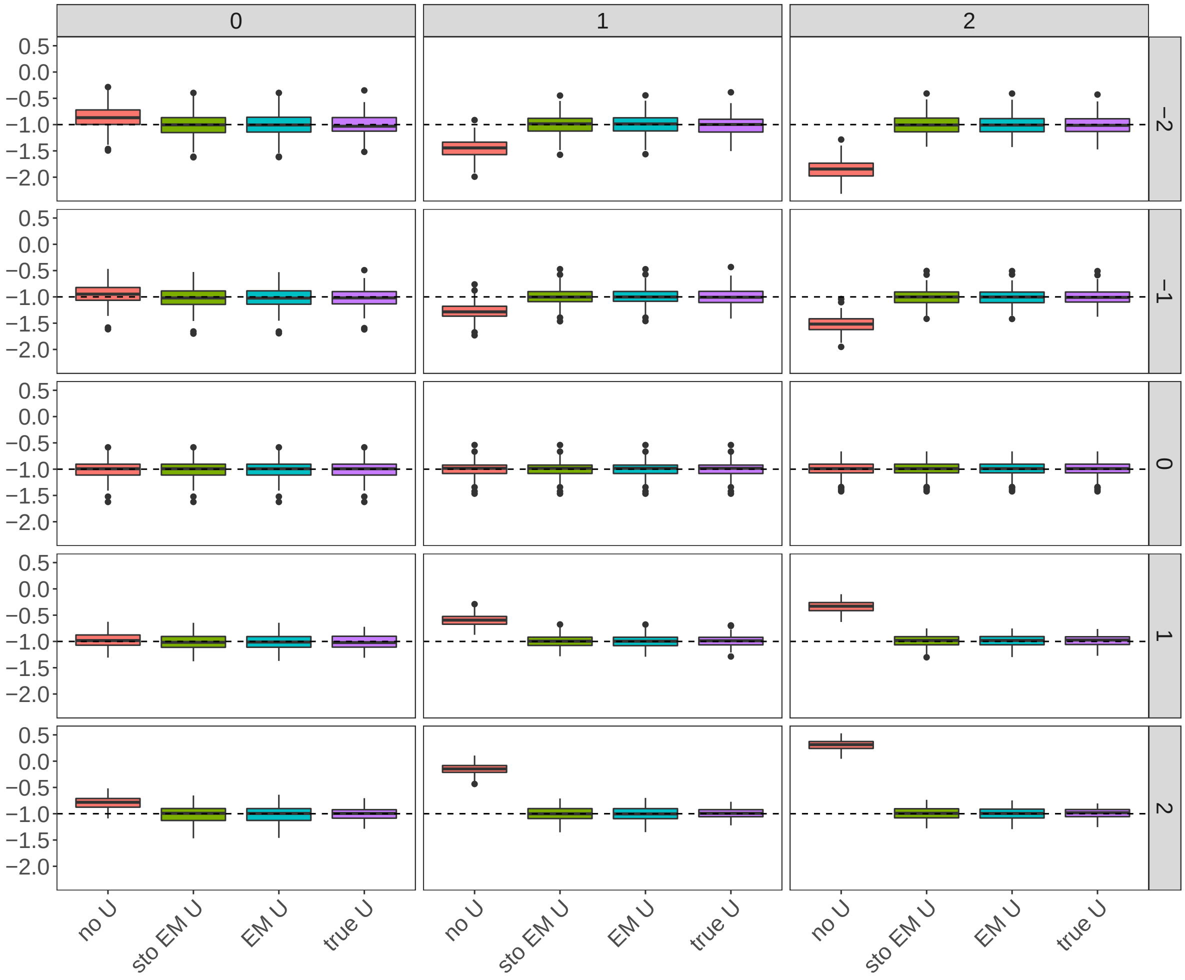


**Figure A.5**: Distributions of the estimated treatment effect on type 2 failures for the simulated competing risks data. $\zeta_z \in \{0, 1, 2\}$ on the horizontal label, $\zeta_1 = 1$ and $\zeta_2 \in \{-2, -1, 0, 1, 2\}$ on the vertical label. Each boxplot displays $\hat{\tau}_2$ from 200 simulations.

**Table A.1**: Sensitivity analysis results on the IBD for UC patients data for outcome clinical remission.

| $\zeta$ | method | $\zeta_z = 0$ | $\zeta_z = 0.5$ | $\zeta_z = 1$ | $\zeta_z = 1.5$ | $\zeta_z = 2$ |
|---|---|---|---|---|---|---|
| -2 | EM | 0.5833 (0.1825) | 1.0176 (0.1780) | 1.3987 (0.1731) | 1.6922 (0.1690) | 1.9005 (0.1677) |
| | stoEM | 0.5948 (0.1718) | 1.0238 (0.1702) | 1.4000 (0.1637) | 1.7116 (0.1658) | 1.9064 (0.1629) |
| | IPW | 0.5241 (0.1563) | 0.8406 (0.1691) | 1.0589 (0.1866) | 1.1615 (0.2019) | 1.2121 (0.2125) |
| -1.5 | EM | 0.5954 (0.1640) | 0.9119 (0.1626) | 1.1886 (0.1616) | 1.4033 (0.1582) | 1.5570 (0.1560) |
| | stoEM | 0.5866 (0.1601) | 0.9148 (0.1605) | 1.1917 (0.1570) | 1.4016 (0.1558) | 1.5672 (0.1535) |
| | IPW | 0.5241 (0.1563) | 0.7763 (0.1703) | 0.9701 (0.1981) | 1.0716 (0.2072) | 1.0961 (0.2183) |
| -1 | EM | 0.5895 (0.1523) | 0.7950 (0.1520) | 0.9759 (0.1506) | 1.1185 (0.1495) | 1.2217 (0.1485) |
| | stoEM | 0.5894 (0.1532) | 0.7964 (0.1507) | 0.9691 (0.1488) | 1.1197 (0.1491) | 1.2212 (0.1473) |
| | IPW | 0.5241 (0.1563) | 0.7126 (0.1704) | 0.8327 (0.1985) | 0.8835 (0.2207) | 0.9221 (0.2387) |
| -0.5 | EM | 0.5798 (0.1449) | 0.6811 (0.1448) | 0.7711 (0.1445) | 0.8431 (0.1442) | 0.8956 (0.1440) |
| | stoEM | 0.5799 (0.1450) | 0.6782 (0.1451) | 0.7732 (0.1443) | 0.8416 (0.1444) | 0.8987 (0.1439) |
| | IPW | 0.5241 (0.1563) | 0.6210 (0.1693) | 0.6862 (0.2079) | 0.7287 (0.2315) | 0.7246 (0.2309) |
| 0 | EM | 0.5756 (0.1424) | 0.5756 (0.1424) | 0.5756 (0.1424) | 0.5756 (0.1424) | 0.5756 (0.1424) |
| | stoEM | 0.5756 (0.1423) | 0.5756 (0.1423) | 0.5756 (0.1423) | 0.5756 (0.1423) | 0.5756 (0.1423) |
| | IPW | 0.5241 (0.1563) | 0.5216 (0.1700) | 0.4979 (0.2088) | 0.4847 (0.2391) | 0.4730 (0.2506) |
| 0.5 | EM | 0.5798 (0.1449) | 0.4777 (0.1451) | 0.3852 (0.1445) | 0.3105 (0.1441) | 0.2561 (0.1436) |
| | stoEM | 0.5829 (0.1447) | 0.4774 (0.1443) | 0.3875 (0.1448) | 0.3135 (0.1442) | 0.2573 (0.1438) |
| | IPW | 0.5241 (0.1563) | 0.4196 (0.1701) | 0.3132 (0.2012) | 0.2540 (0.2243) | 0.2408 (0.2470) |
| 1 | EM | 0.5895 (0.1524) | 0.3811 (0.1516) | 0.1923 (0.1508) | 0.0400 (0.1490) | -0.0701 (0.1477) |
| | stoEM | 0.5888 (0.1493) | 0.3850 (0.1502) | 0.1844 (0.1498) | 0.0442 (0.1506) | -0.0688 (0.1468) |
| | IPW | 0.5241 (0.1563) | 0.3156 (0.1675) | 0.1231 (0.1971) | -0.0152 (0.2264) | -0.0443 (0.2486) |
| 1.5 | EM | 0.5954 (0.1641) | 0.2767 (0.1629) | -0.0105 (0.1581) | -0.2419 (0.1587) | -0.4084 (0.1541) |
| | stoEM | 0.5901 (0.1618) | 0.2699 (0.1551) | -1e-04 (0.1574) | -0.2469 (0.1549) | -0.4040 (0.1528) |
| | IPW | 0.5241 (0.1563) | 0.2280 (0.1652) | -0.0649 (0.1924) | -0.2371 (0.2213) | -0.2932 (0.2340) |
| 2 | EM | 0.5833 (0.1784) | 0.1560 (0.1756) | -0.2250 (0.1743) | -0.5365 (0.1720) | -0.7611 (0.1615) |
| | stoEM | 0.5927 (0.1677) | 0.1459 (0.1631) | -0.2268 (0.1685) | -0.5524 (0.1650) | -0.7519 (0.1614) |
| | IPW | 0.5241 (0.1563) | 0.1434 (0.1610) | -0.1950 (0.1855) | -0.3996 (0.2070) | -0.5294 (0.2309) |

**Table A.2**: Sensitivity analysis results on the IBD for CD patients data for outcome clinical remission with $\zeta_2 = -2$.

| $\zeta_1$ | method | $\zeta_z = 0$ | $\zeta_z = 0.5$ | $\zeta_z = 1$ | $\zeta_z = 1.5$ | $\zeta_z = 2$ |
|---|---|---|---|---|---|---|
| -2 | EM | 0.0229 (0.1535) | 0.4237 (0.1546) | 0.8001 (0.1588) | 1.0966 (0.1551) | 1.2968 (0.1509) |
| | stoEM | 0.0336 (0.1486) | 0.4195 (0.1485) | 0.7958 (0.1509) | 1.1031 (0.1490) | 1.2963 (0.1480) |
| | IPW | -0.0563 (0.1496) | 0.2309 (0.1501) | 0.4044 (0.154) | 0.4651 (0.1586) | 0.4516 (0.1642) |
| -1.5 | EM | 0.0403 (0.1471) | 0.3505 (0.1461) | 0.6294 (0.1463) | 0.8422 (0.1438) | 0.9854 (0.1433) |
| | stoEM | 0.0394 (0.1421) | 0.3488 (0.1441) | 0.6279 (0.1433) | 0.8475 (0.1422) | 0.9856 (0.1427) |
| | IPW | -0.0563 (0.1496) | 0.1644 (0.1506) | 0.3140 (0.1547) | 0.3662 (0.1671) | 0.3498 (0.1666) |
| -1 | EM | 0.0519 (0.1392) | 0.2600 (0.1381) | 0.4426 (0.1388) | 0.5801 (0.1372) | 0.6728 (0.1365) |
| | stoEM | 0.0531 (0.1371) | 0.2566 (0.1379) | 0.4397 (0.1375) | 0.5835 (0.1365) | 0.6730 (0.1363) |
| | IPW | -0.0563 (0.1496) | 0.1068 (0.1530) | 0.1991 (0.1591) | 0.2324 (0.1634) | 0.1995 (0.1677) |
| -0.5 | EM | 0.0583 (0.1337) | 0.1619 (0.1336) | 0.2516 (0.1335) | 0.3188 (0.1333) | 0.3641 (0.1330) |
| | stoEM | 0.0567 (0.1334) | 0.1580 (0.1336) | 0.2528 (0.1332) | 0.3202 (0.1331) | 0.3636 (0.1330) |
| | IPW | -0.0563 (0.1496) | 0.0196 (0.1548) | 0.0820 (0.1626) | 0.0844 (0.1648) | 0.0685 (0.1687) |
| 0 | EM | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) |
| | stoEM | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) |
| | IPW | -0.0563 (0.1496) | -0.0703 (0.1540) | -0.0751 (0.1609) | -0.0792 (0.1676) | -0.0744 (0.1719) |
| 0.5 | EM | 0.0591 (0.1336) | -0.0432 (0.1335) | -0.1306 (0.1333) | -0.1957 (0.1332) | -0.2396 (0.1332) |
| | stoEM | 0.0561 (0.1331) | -0.0447 (0.1334) | -0.1338 (0.1332) | -0.1964 (0.1336) | -0.2407 (0.1332) |
| | IPW | -0.0563 (0.1496) | -0.1584 (0.1556) | -0.2141 (0.1625) | -0.2314 (0.1772) | -0.2357 (0.1711) |
| 1 | EM | 0.0538 (0.1382) | -0.1489 (0.1385) | -0.3215 (0.1377) | -0.4504 (0.1373) | -0.5375 (0.1375) |
| | stoEM | 0.0584 (0.1376) | -0.1464 (0.1365) | -0.3160 (0.1371) | -0.4584 (0.1370) | -0.5373 (0.1364) |
| | IPW | -0.0563 (0.1496) | -0.2453 (0.1529) | -0.3511 (0.1625) | -0.3908 (0.1612) | -0.3827 (0.1678) |
| 1.5 | EM | 0.0432 (0.1453) | -0.2542 (0.1444) | -0.5084 (0.1439) | -0.7011 (0.1448) | -0.8322 (0.1430) |
| | stoEM | 0.0322 (0.1441) | -0.2530 (0.1421) | -0.5164 (0.1417) | -0.6963 (0.1414) | -0.8396 (0.1421) |
| | IPW | -0.0563 (0.1496) | -0.3192 (0.1530) | -0.4691 (0.1584) | -0.5321 (0.1607) | -0.5289 (0.1652) |
| 2 | EM | 0.0268 (0.1527) | -0.3520 (0.1519) | -0.6806 (0.1535) | -0.9394 (0.1527) | -1.1181 (0.1508) |
| | stoEM | 0.0280 (0.1523) | -0.3612 (0.1481) | -0.6909 (0.1454) | -0.9496 (0.1481) | -1.1195 (0.1466) |
| | IPW | -0.0563 (0.1496) | -0.3878 (0.1508) | -0.5729 (0.1532) | -0.6322 (0.1574) | -0.6322 (0.1601) |

**Table A.3**: Sensitivity analysis results on the IBD for CD patients data for outcome clinical remission with $\zeta_2 = 0$.

| $\zeta_1$ | method | $\zeta_z = 0$ | $\zeta_z = 0.5$ | $\zeta_z = 1$ | $\zeta_z = 1.5$ | $\zeta_z = 2$ |
|---|---|---|---|---|---|---|
| -2 | EM | 0.0263 (0.1538) | 0.4227 (0.1544) | 0.7952 (0.1568) | 1.0920 (0.1551) | 1.2941 (0.1513) |
| | stoEM | 0.0363 (0.1506) | 0.4126 (0.1491) | 0.7955 (0.1508) | 1.1013 (0.1487) | 1.2955 (0.1488) |
| | IPW | -0.0563 (0.1496) | 0.2300 (0.1507) | 0.4010 (0.1537) | 0.4628 (0.1579) | 0.4488 (0.1638) |
| -1.5 | EM | 0.0427 (0.1469) | 0.3492 (0.1451) | 0.6257 (0.1462) | 0.8388 (0.1437) | 0.9833 (0.1429) |
| | stoEM | 0.0406 (0.1420) | 0.3494 (0.1419) | 0.6260 (0.1422) | 0.8478 (0.1427) | 0.9840 (0.1426) |
| | IPW | -0.0563 (0.1496) | 0.1641 (0.1507) | 0.3087 (0.1543) | 0.3622 (0.1695) | 0.3483 (0.1661) |
| -1 | EM | 0.0533 (0.1387) | 0.2591 (0.1381) | 0.4404 (0.1385) | 0.5781 (0.1372) | 0.6714 (0.1365) |
| | stoEM | 0.0534 (0.1368) | 0.2558 (0.1380) | 0.4382 (0.1371) | 0.5801 (0.1366) | 0.6717 (0.1360) |
| | IPW | -0.0563 (0.1496) | 0.1078 (0.1526) | 0.1956 (0.1602) | 0.2241 (0.1630) | 0.1965 (0.1678) |
| -0.5 | EM | 0.0588 (0.1336) | 0.1615 (0.1335) | 0.2507 (0.1333) | 0.3180 (0.1332) | 0.3636 (0.1330) |
| | stoEM | 0.0581 (0.1332) | 0.1574 (0.1334) | 0.2536 (0.1332) | 0.3194 (0.1332) | 0.3627 (0.1329) |
| | IPW | -0.0563 (0.1496) | 0.0210 (0.1554) | 0.0778 (0.1623) | 0.0802 (0.1643) | 0.0631 (0.1679) |
| 0 | EM | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) |
| | stoEM | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) |
| | IPW | -0.0563 (0.1496) | -0.0703 (0.1545) | -0.0782 (0.1612) | -0.0826 (0.1692) | -0.0807 (0.1718) |
| 0.5 | EM | 0.0588 (0.1336) | -0.0429 (0.1335) | -0.1301 (0.1333) | -0.1952 (0.1333) | -0.2393 (0.1332) |
| | stoEM | 0.0561 (0.1330) | -0.0425 (0.1334) | -0.1327 (0.1334) | -0.1966 (0.1336) | -0.2400 (0.1332) |
| | IPW | -0.0563 (0.1496) | -0.1597 (0.1559) | -0.2179 (0.1625) | -0.2411 (0.1769) | -0.2436 (0.1709) |
| 1 | EM | 0.0533 (0.1383) | -0.1485 (0.1386) | -0.3207 (0.1378) | -0.4498 (0.1375) | -0.5371 (0.1374) |
| | stoEM | 0.0576 (0.1377) | -0.1437 (0.1364) | -0.3150 (0.1370) | -0.4572 (0.1369) | -0.5379 (0.1363) |
| | IPW | -0.0563 (0.1496) | -0.2455 (0.1530) | -0.3554 (0.1620) | -0.3976 (0.1618) | -0.3887 (0.1697) |
| 1.5 | EM | 0.0427 (0.1453) | -0.2536 (0.1442) | -0.5075 (0.1440) | -0.7005 (0.1448) | -0.8319 (0.1430) |
| | stoEM | 0.0301 (0.1447) | -0.2555 (0.1420) | -0.5189 (0.1418) | -0.6974 (0.1418) | -0.8379 (0.1414) |
| | IPW | -0.0563 (0.1496) | -0.3192 (0.1534) | -0.4751 (0.1599) | -0.5378 (0.1609) | -0.5351 (0.1646) |
| 2 | EM | 0.0263 (0.1527) | -0.3514 (0.1534) | -0.6800 (0.1540) | -0.9390 (0.1528) | -1.1179 (0.1508) |
| | stoEM | 0.0286 (0.1512) | -0.3634 (0.1482) | -0.6890 (0.1467) | -0.9512 (0.1484) | -1.1178 (0.1472) |
| | IPW | -0.0563 (0.1496) | -0.3878 (0.1509) | -0.5791 (0.1541) | -0.6408 (0.1574) | -0.6410 (0.1597) |

**Table A.4**: Sensitivity analysis results on the IBD for CD patients data for outcome clinical remission with $\zeta_2 = 2$.

| $\zeta_1$ | method | $\zeta_z = 0$ | $\zeta_z = 0.5$ | $\zeta_z = 1$ | $\zeta_z = 1.5$ | $\zeta_z = 2$ |
|---|---|---|---|---|---|---|
| -2 | EM | 0.0268 (0.1533) | 0.4254 (0.1545) | 0.8006 (0.1568) | 1.0991 (0.1547) | 1.3010 (0.1509) |
| | stoEM | 0.0349 (0.1488) | 0.4152 (0.1487) | 0.8002 (0.1506) | 1.1081 (0.1482) | 1.3077 (0.1490) |
| | IPW | -0.0563 (0.1496) | 0.2313 (0.1508) | 0.3976 (0.1543) | 0.4547 (0.1568) | 0.4415 (0.1637) |
| -1.5 | EM | 0.0432 (0.1466) | 0.3516 (0.1457) | 0.6304 (0.1463) | 0.8450 (0.1435) | 0.9896 (0.1427) |
| | stoEM | 0.0423 (0.1416) | 0.3551 (0.1425) | 0.6304 (0.1425) | 0.8537 (0.1426) | 0.9907 (0.1428) |
| | IPW | -0.0563 (0.1496) | 0.1681 (0.1512) | 0.3073 (0.1551) | 0.3536 (0.1686) | 0.3372 (0.1649) |
| -1 | EM | 0.0538 (0.1387) | 0.2610 (0.1380) | 0.4440 (0.1386) | 0.5828 (0.1371) | 0.6764 (0.1365) |
| | stoEM | 0.0537 (0.1366) | 0.2574 (0.1380) | 0.4411 (0.1373) | 0.5855 (0.1369) | 0.6780 (0.1358) |
| | IPW | -0.0563 (0.1496) | 0.1106 (0.1534) | 0.1943 (0.1596) | 0.2151 (0.1626) | 0.1871 (0.1662) |
| -0.5 | EM | 0.0591 (0.1335) | 0.1626 (0.1334) | 0.2528 (0.1333) | 0.3207 (0.1332) | 0.3664 (0.1330) |
| | stoEM | 0.0582 (0.1331) | 0.1595 (0.1332) | 0.2553 (0.1331) | 0.3225 (0.1330) | 0.3657 (0.1328) |
| | IPW | -0.0563 (0.1496) | 0.0202 (0.1559) | 0.0758 (0.1618) | 0.0735 (0.1647) | 0.0567 (0.1664) |
| 0 | EM | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) | 0.0605 (0.1319) |
| | stoEM | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) | 0.0605 (0.1318) |
| | IPW | -0.0563 (0.1496) | -0.0688 (0.1546) | -0.0797 (0.1611) | -0.0907 (0.1691) | -0.0893 (0.1717) |
| 0.5 | EM | 0.0583 (0.1337) | -0.0446 (0.1337) | -0.1328 (0.1333) | -0.1986 (0.1333) | -0.2427 (0.1332) |
| | stoEM | 0.0560 (0.1332) | -0.0435 (0.1333) | -0.1354 (0.1334) | -0.2005 (0.1335) | -0.2437 (0.1332) |
| | IPW | -0.0563 (0.1496) | -0.1567 (0.1567) | -0.2186 (0.1616) | -0.2461 (0.1781) | -0.2511 (0.1700) |
| 1 | EM | 0.0519 (0.1384) | -0.1523 (0.1387) | -0.3270 (0.1381) | -0.4571 (0.1373) | -0.5440 (0.1375) |
| | stoEM | 0.0566 (0.1380) | -0.1491 (0.1363) | -0.3217 (0.1368) | -0.4649 (0.1372) | -0.5443 (0.1361) |
| | IPW | -0.0563 (0.1496) | -0.2458 (0.1544) | -0.3561 (0.1625) | -0.4040 (0.1614) | -0.3972 (0.1708) |
| 1.5 | EM | 0.0403 (0.1464) | -0.2600 (0.1451) | -0.5176 (0.1445) | -0.7115 (0.1441) | -0.8414 (0.1429) |
| | stoEM | 0.0296 (0.1454) | -0.2664 (0.1410) | -0.5270 (0.1423) | -0.7087 (0.1425) | -0.8466 (0.1408) |
| | IPW | -0.0563 (0.1496) | -0.3153 (0.1538) | -0.4771 (0.1599) | -0.5403 (0.1616) | -0.5395 (0.1634) |
| 2 | EM | 0.0229 (0.1531) | -0.3591 (0.1548) | -0.6929 (0.1540) | -0.9521 (0.1528) | -1.1279 (0.1507) |
| | stoEM | 0.0294 (0.1544) | -0.3707 (0.1471) | -0.6994 (0.1474) | -0.9558 (0.1466) | -1.1310 (0.1484) |
| | IPW | -0.0563 (0.1496) | -0.3869 (0.1511) | -0.5796 (0.1543) | -0.6431 (0.1585) | -0.6405 (0.1590) |

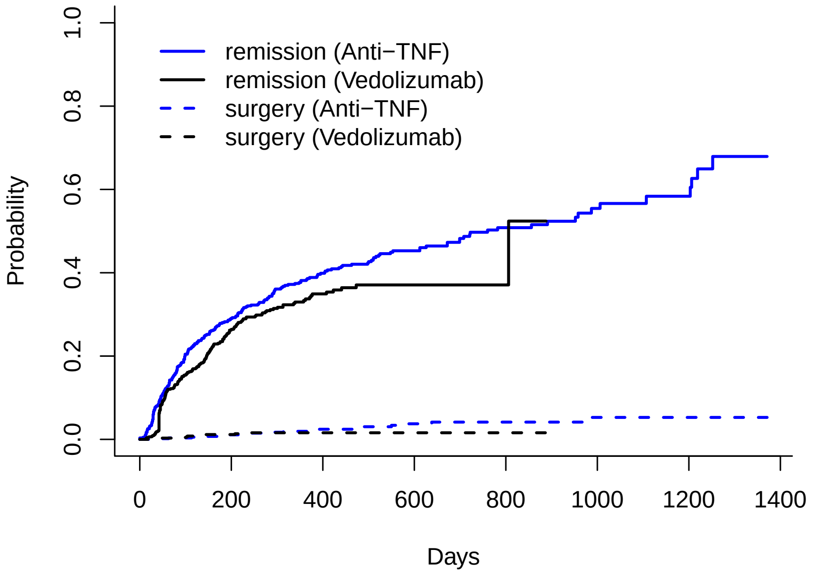
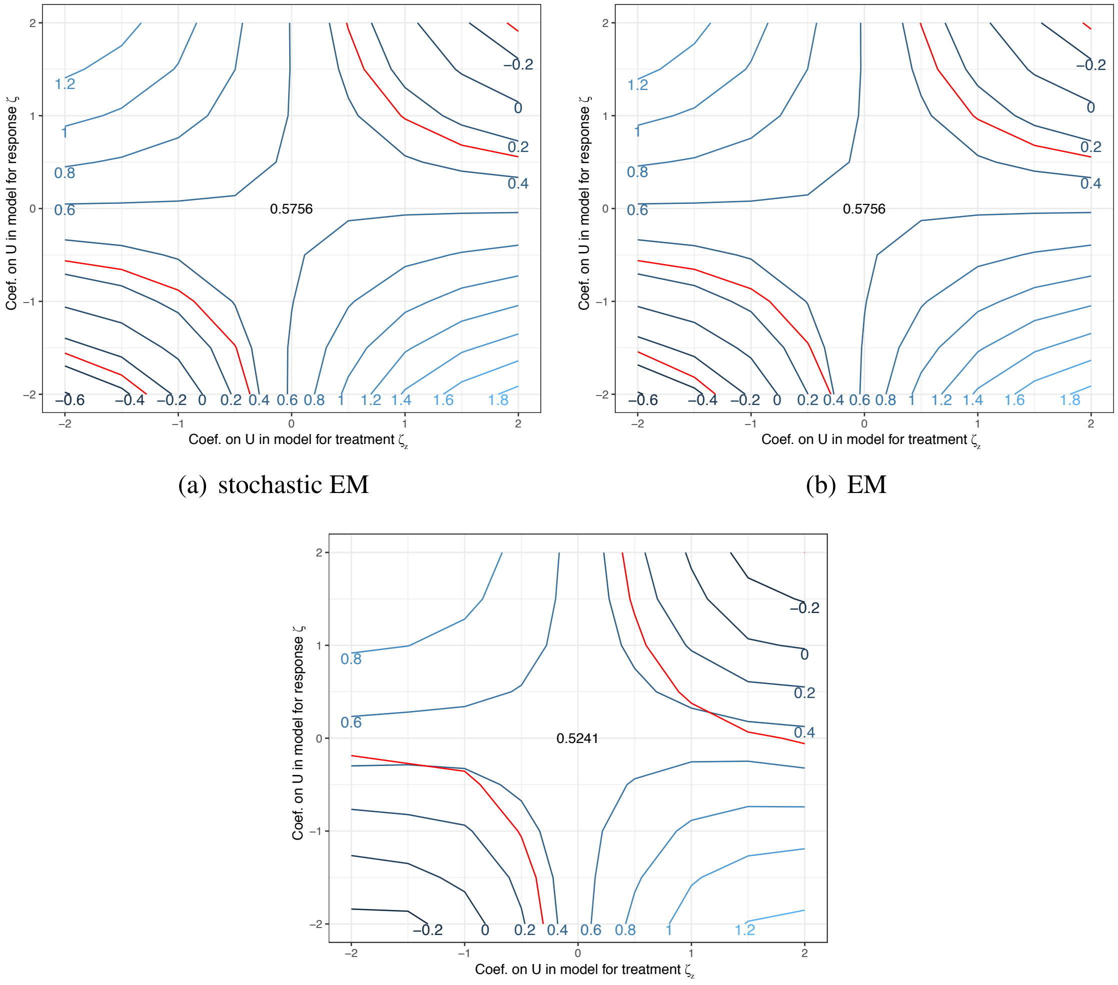


**Figure A.6**: Estimated cumulative incidence functions for time to clinical remission and time to surgery in CD patients.

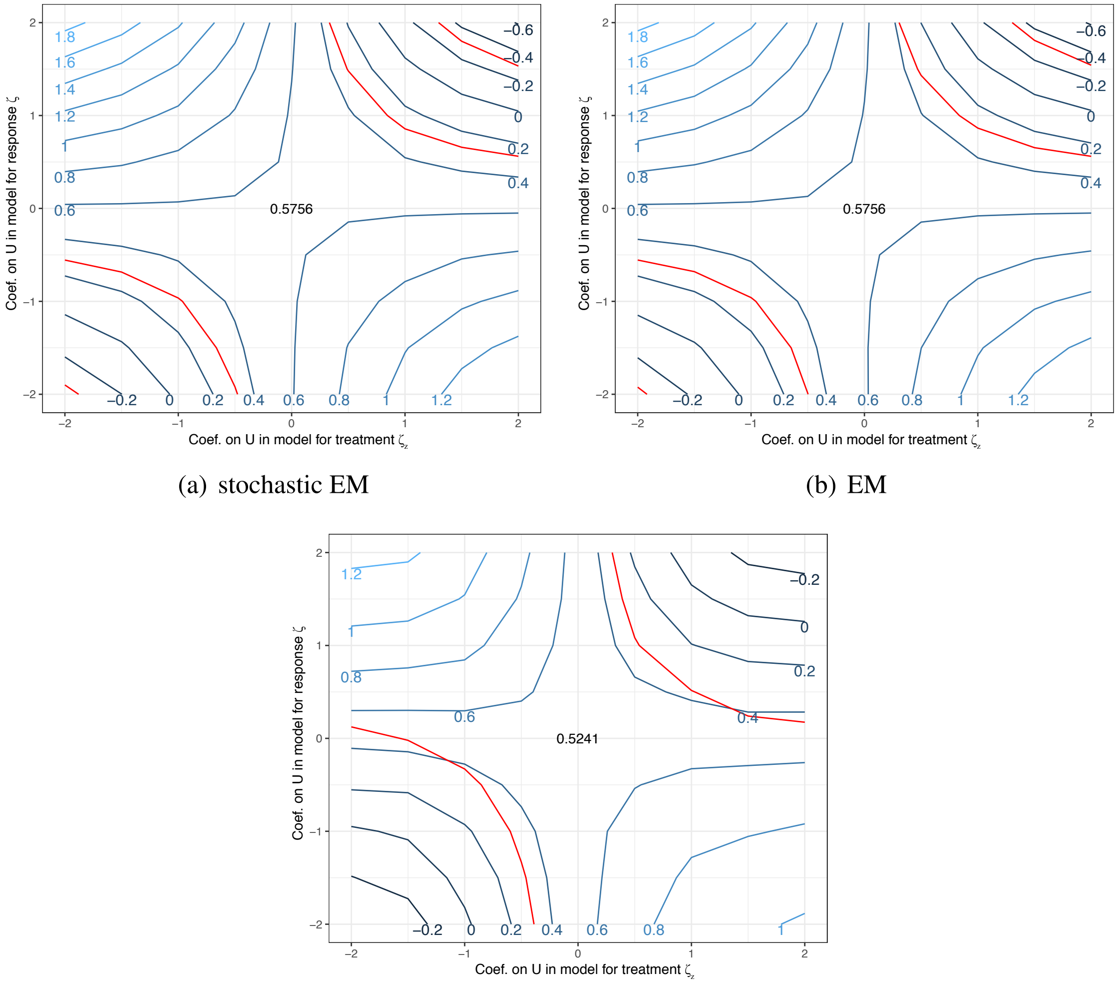


(a) stochastic EM

(b) EM

(c) IPW

**Figure A.7**: Sensitivity analysis results for UC patients data for outcome clinical remission. In all plots, the blue contours show the sensitivity parameter values corresponding to the estimated treatment effect $\hat{\tau}$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}/\hat{\sigma}_{\hat{\tau}}| = 1.96$. $U \sim$ Bernoulli(0.7).

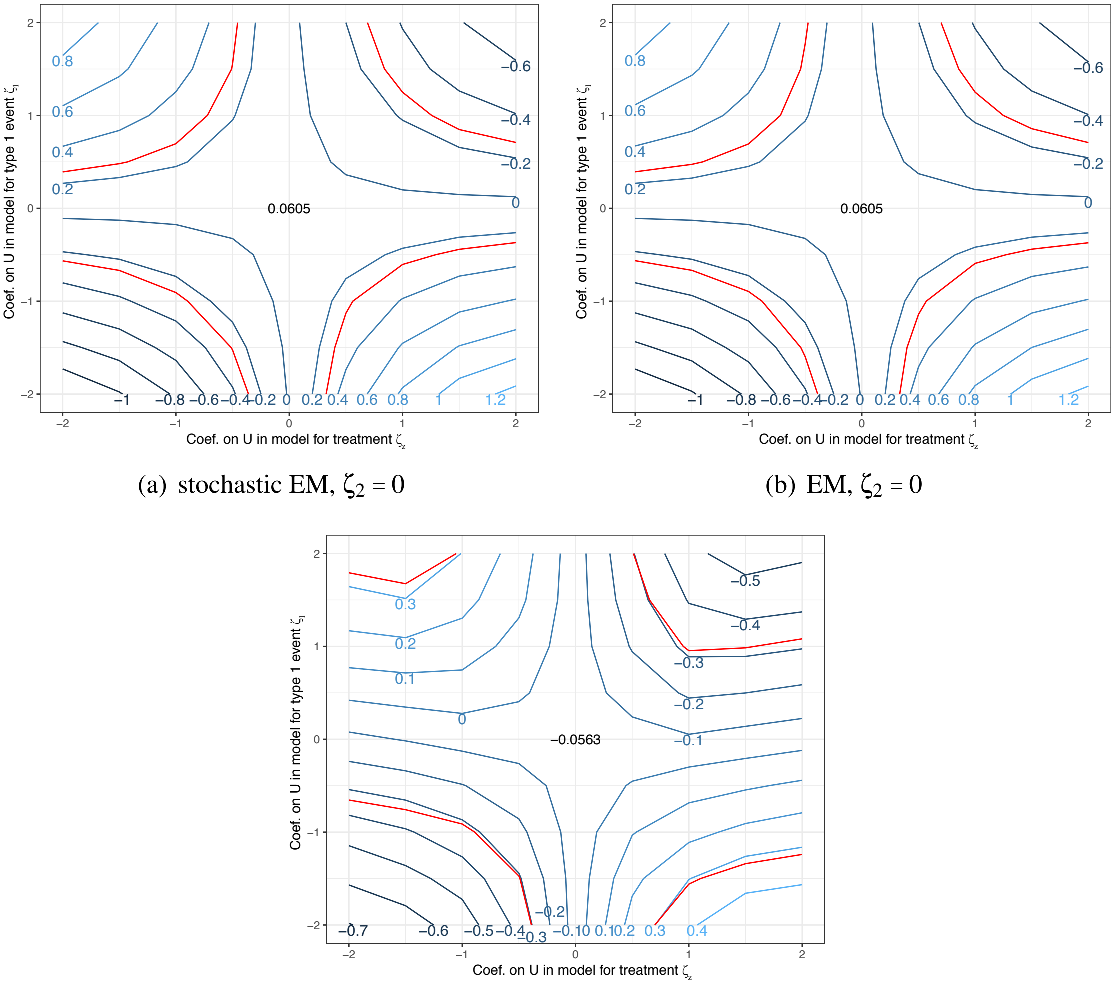


(a) stochastic EM

(b) EM

(c) IPW

**Figure A.8**: Sensitivity analysis results for UC patients data for outcome clinical remission. In all plots, the blue contours show the sensitivity parameter values corresponding to the estimated treatment effect $\hat{\tau}$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}/\hat{\sigma}_{\hat{\tau}}| = 1.96$. $U \sim$ Bernoulli(0.3).

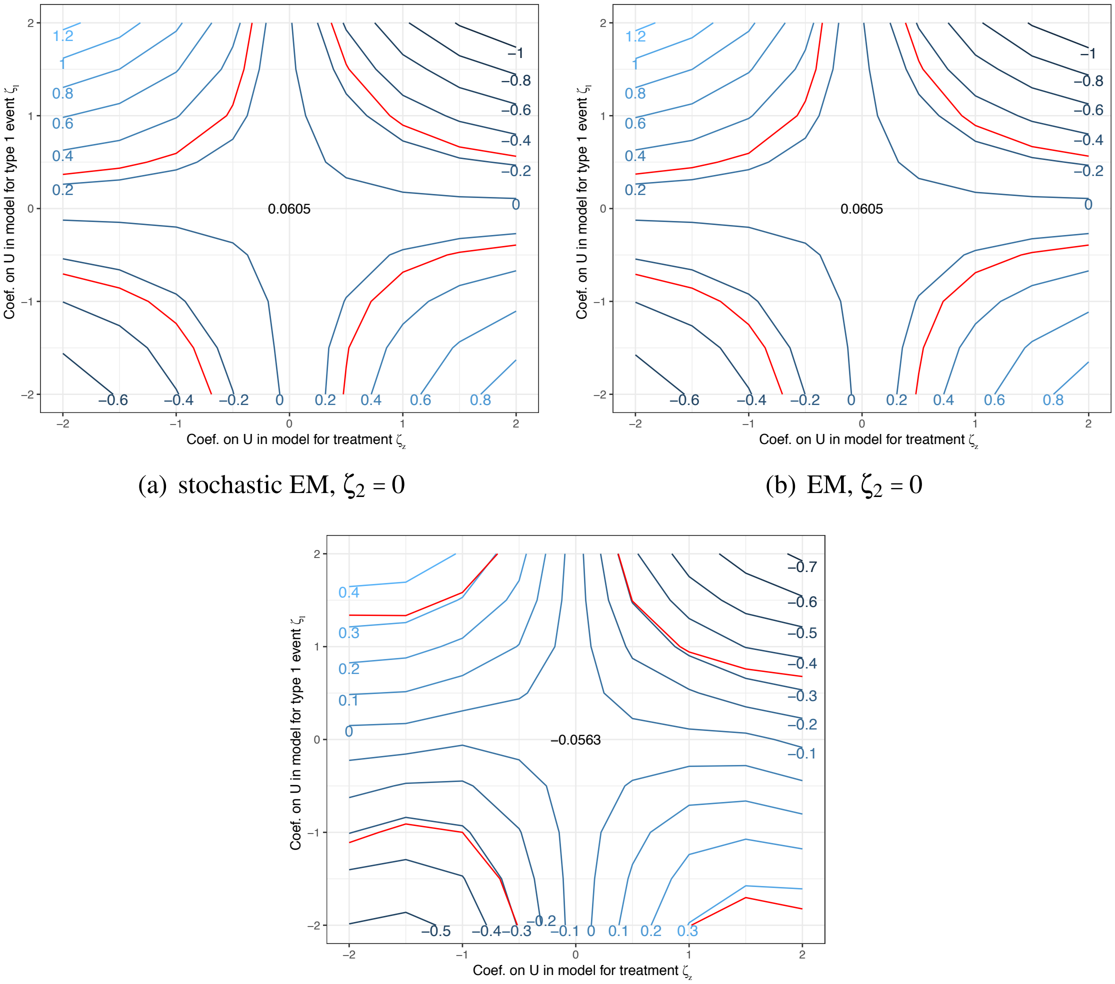


(a) stochastic EM, $\zeta_2 = 0$

(b) EM, $\zeta_2 = 0$

(c) IPW, $\zeta_2 = 0$

**Figure A.9**: Sensitivity analysis results for CD patients data for outcome clinical remission. In all plots, the blue contours show the sensitivity parameter values corresponding to the estimated treatment effect $\hat{\tau}_1$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}_1/\hat{\sigma}_{\hat{\tau}_1}| = 1.96$. $U \sim$ Bernoulli(0.7).



(a) stochastic EM, $\zeta_2 = 0$

(b) EM, $\zeta_2 = 0$

(c) IPW, $\zeta_2 = 0$

**Figure A.10**: Sensitivity analysis results for CD patients data for outcome clinical remission. In all plots, the blue contours show the sensitivity parameter values corresponding to the estimated treatment effect $\hat{\tau}_1$, and the red curves correspond to where the absolute value of the $t$-statistic $|t| = |\hat{\tau}_1/\hat{\sigma}_{\hat{\tau}_1}| = 1.96$. $U \sim \text{Bernoulli}(0.3)$.

# Bibliography

[1] T W Anderson and D A Darling. Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[2] Ery Arias-Castro and Rong Huang. The sparse variance contamination model. *arXiv preprint arXiv:1807.10785*, 2018.

[3] Ery Arias-Castro and Meng Wang. Distribution-free tests for sparse heterogeneous mixtures. *TEST*, 26(1):71–94, 2016.

[4] Ery Arias-Castro and Meng Wang. Distribution-free tests for sparse heterogeneous mixtures. *arXiv preprint arXiv:1308.0346*, 2018.

[5] Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen–Stein method. *Statistical Science*, pages 403–424, 1990.

[6] Peter C. Austin and Elizabeth A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34:3661–3679, 2015.

[7] Jan Beyersmann, Aurelien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28(6):956–971, 2009.

[8] Anders E Bilgrau, Poul S Eriksen, Jakob G Rasmussen, Hans E Johnsen, Karen Dybkær, and Martin Bøgsted. GMCM: Unsupervised clustering and meta-analysis using Gaussian mixture copula models. *Journal of Statistical Software*, 70(2):1–23, 2016.

[9] Marina Bogomolov and Ruth Heller. Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3):505–516, 2018.

[10] M Bohm, R Xu, Y Zhang, S Varma, M Fischer, S Kadire, G Tran, M Rahal, S Aniwan, J Meserve, A Weiss, G Kochhar, J L Koliani-Pace, J P Campbell, B Boland, S Singh, D Faleck, A Winters, S Chablaney, R Hirten, R Ungaro, E Shmidt, K Lasch, V Jairaith, D Hudesman, S Chang, D Lukin, A Swaminath, B E Sands, J Colombel, S Kane, E V Loftus, B Shen, C A Siegel, W J Sandborn, and P S Dulai. Comparative effectiveness of vedolizumab and tumor necrosis factor-antagonist therapy in Crohn’s disease: A multicenter

consortium propensity score-matched analysis. *Gastroenterology*, 154(6):S369 – S370, 2019.

[11] T Tony Cai, X Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011.

[12] Tony T Cai and Yihong Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232, 2014.

[13] Paul L Canner. A simulation study of one-and two-sample kolmogorov-smirnov statistics with a particular weight function. *Journal of the American Statistical Association*, 70(349):209–211, 1975.

[14] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

[15] Sourav Chatterjee. Stein's method for concentration inequalities. *Probability Theory and Related Fields*, 138(1):305–321, 2007.

[16] William J Conover and David S Salsburg. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment. *Biometrics*, pages 189–196, 1988.

[17] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[18] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, 2008.

[19] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[20] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[21] David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

[22] David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30(1):1–25, 2015.

[23] Parambir S Dulai, Siddharth Singh, Xiaoqian Jiang, Farhad Peerani, Neeraj Narula, Khadija Chaudrey, Diana Whitehead, David Hudesman, Dana Lukin, Arun Swaminath, Eugenia Shmidt, Shuang Wang, Brigid S Boland, John T Chang, Sunanda Kane, Corey A Siegel,

Edward V Loftus, William J Sandborn, Bruce E Sands, and Jean-Frederic Colombel. The real-world effectiveness and safety of vedolizumab for moderate–severe crohn's disease: results from the us victory consortium. *The American journal of gastroenterology*, 111(8):1147, 2016.

[24] Helmut Finner and Veronika Gontscharuk. Two-sample Kolmogorov-Smirnov-type tests revisited: old and new tests in terms of local levels. *The Annals of Statistics*, 46(6A):3014–3037, 2018.

[25] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*, volume 168. Marcel Dekker, Inc., 4th edition, 2003.

[26] Peter Hall and Jiashun Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010.

[27] Wassily Hoeffding. A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, 22(4):558–566, 1951.

[28] Yuri I Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[29] D Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics*, 7(1):108–115, 1979.

[30] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data (2nd ed.)*. John Wiley & Sons, 2011.

[31] Jeffry A Katz, Gil Melmed, and Bruce E Sands. The facts about inflammatory bowel diseases. *Crohn's & Colitis Foundation of America, New York*, 2011.

[32] Eric L Lehmann. Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pages 165–179, 1951.

[33] Erich L Lehmann. The power of rank tests. *The Annals of Mathematical Statistics*, 24(1):23–43, 1953.

[34] Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

[35] Lingling Li, Changyu Shen, Ann C Wu, and Xiaochun Li. Propensity score-based sensitivity analysis method for uncontrolled confounding. *American journal of epidemiology*, 174(3):345–353, 2011.

[36] Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[37] Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963, 1998.

[38] Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14(6):570–580, 2013.

[39] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, 1982.

[40] D Lukin, D Faleck, R Xu, Y Zhang, A Weiss, S Aniwan, S Kadire, G Tran, M Rahal, A Winters, S Chablaney, J L Koliani-Pace, J Meserve, J P Campbell, G Kochhar, M Bohm, S Varma, M Fischer, B Boland, S Singh, R Hirten, R Ungaro, K Lasch, E Shmidt, V Jairaith, D Hudesman, S Chang, A Swaminath, B Shen, S Kane, E V Loftus, B E Sands, J Colombel, C A Siegel, W J Sandborn, and P S Dulai. Comparative safety profile of vedolizumab and tumor necrosis factor-antagonist therapy for inflammatory bowel disease: A multicenter consortium propensity score-matched analysis. *Gastroenterology*, 154(6):S68, 2019.

[41] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[42] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[43] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. On the exact Berk–Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[44] Neeraj Narula, Farhad Peerani, Joseph Meserve, Gursimran Kochhar, Khadija Chaudrey, Justin Hartke, Prianka Chilukuri, Jenna Koliani-Pace, Adam Winters, Leah Katta, et al. Vedolizumab for ulcerative colitis: treatment outcomes from the victory consortium. *The American journal of gastroenterology*, 113(9):1345, 2018.

[45] Søren Feodor Nielsen. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

[46] Ao No Pettitt. A two-sample Anderson-Darling rank statistic. *Biometrika*, 63(1):161–168, 1976.

[47] W Ken Redekop and Deirdre Mladsi. The faces of personalized medicine: A framework for understanding its meaning and scope. *Value in Health*, 16(6):S4–S9, 2013.

[48] Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.

[49] Paul R. Rosenbaum. *Observational Studies*. Springer-Verlag New York, 2nd edition, 2002.

[50] Changyu Shen, Xiaochun Li, Lingling Li, and Martin C Were. Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical Journal*, 53(5):822–837, 2011.

[51] Nikolai Vasilyevich Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Mathematics University Moscow*, 2:3–16, 1939.

[52] Florin Vaida and Ronghui Xu. Proportional hazards model with random effects. *Statistics in medicine*, 19(24):3309–3324, 2000.

[53] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[54] Ronghui Xu, Gordon Honerkamp-Smith, and Christina D Chambers. Statistical sensitivity analysis for the estimation of fetal alcohol spectrum disorders prevalence. *Reproductive Toxicology*, 86:62–67, 2019.

[55] Weichao Xu, Yunhe Hou, Y S Hung, and Yuexian Zou. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276, 2013.

[56] Sihai Dave Zhao. False discovery rate control for identifying simultaneous signals. *arXiv preprint arXiv:1512.04499*, 2015.

[57] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Optimal detection of weak positive latent dependence between two sequences of multiple tests. *Journal of Multivariate Analysis*, 160:169–184, 2017.