Subsampling and Integrating Modeling Uncertainty in Statistical Analysis

by

Xinlian Zhang

(Under the Direction of Wenxuan Zhong and Gauri Sankar Datta)

Abstract

The advent of the age of big data poses challenges for statistical data analysis. On the one hand, the ultra-large size of datasets renders the application of many classical statistical methods computationally demanding. On the other hand, with the system studied getting more complicated, the model set-up for some popular off-the-shelf methods may not be applicable anymore; modeling of model uncertainty, in particular, becomes important since it is challenging to put one single (set of) assumption(s) on complicated systems. Developing new theoretically justifiable and computationally efficient methods for tackling big data problems from computational and modeling perspectives is the primary motivation for my research. The first focus of my work is on studying the theoretical properties of subsampling methods for dealing with the sheer size of big data. In the framework of the linear model, I show the asymptotic normality of subsampling estimators for both estimating the parameter (unconditional inference) and approximating full sample estimate (conditional inference) with certain regularity conditions satisfied. Based on these asymptotic results, I propose optimal subsampling estimators under different scenarios. The second focus is to propose a Bayesian hierarchical model for integrating the model uncertainty in statistical inference. Under the smoothing spline model, I incorporate the uncertainty in model assumption, i.e., choice of penalty, as a mixture prior for the function to be estimated, and carefully choose innovative (partially) noninformative priors for the parameters in the model. The propriety

of the resulting posterior distribution is established to provide theoretical underpinnings. Advantages of the proposed methods are shown using both simulated and real-world examples. In the end, I also discuss the application part of my research in the generation of small RNAs and their function in gene silencing in *C. elegans*.

INDEX WORDS: subsampling estimator; asymptotic distribution; mean squared error; Bayesian smoothing spline; model uncertainty.

Subsampling and Integrating Modeling Uncertainty in Statistical Analysis

by

Xinlian Zhang

B.S., Beijing Forestry University, China, 2011

M.S., Central University of Finance and Economics, China, 2014

A Dissertation Submitted to the Graduate Faculty of

The University of Georgia

in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2019

Subsampling and Integrating Modeling Uncertainty in Statistical Analysis

by

Xinlian Zhang

Major Professor: Wenxuan Zhong

Co-major Professor: Gauri Sankar Datta

Committee: Ping Ma

Ying Xu

Abhyuday Mandal

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

May 2019

Acknowledgments

This work was made possible through the support of many people. I want to thank my advisors Professor Wenxuan Zhong and Professor Gauri Sankar Datta for all the continuous and enormous support they have given me throughout my time as a Ph.D. student; I have been extremely fortunate knowing and working with them. I am grateful to Professor Zhong and Professor Datta for their immense knowledge and always motivating guidance in overcoming numerous obstacles I faced in my study and my life, especially during my job search. Many thanks to Professor Ping Ma for his tremendous patience, encouragement, and guidance throughout my program. My sincere thanks also go to Professor Ying Xu and Professor Abhyuday Mandal, who provided insightful comments to my works and stimulating discussions on many broad research topics. I would like to thank Professor Bin Yu for many helpful discussions and Dr. Shusen Wang for reading and providing constructive comments on an earlier version of the works in Chapter 2 and 3. Many thanks to everyone in the big data analytics lab. It was fantastic to have the opportunity to work with these wonderful young researchers during the last five years. I would also like to express my genuine gratitude to the faculty members in the Department of Statistics at the University of Georgia; they have truly influenced me through their extraordinary dedication to teaching and molding of young minds. The works throughout my Ph.D. study are supported by the National Institute of Health R01 GM113242 to PI Wenxuan Zhong, National Institute of Health R01GM122080 to PI Ping Ma, National Science Foundation DMS-1222718 to PI Ping Ma, National Science Foundation DMS-1438957 to PI Ping Ma, National Science Foundation DMS-1228288 to PI Wenxuan Zhong, National Science Foundation DMS-1440037 to PI Ping Ma, and National Science Foundation DMS-1228288 to leading PI Wenxuan Zhong.

Contents

Acknowledgments **iv**

1 Introduction **1**

1.1 Subsampling for large-scale regression problems 2

1.2 Model uncertainty in smoothing spline 6

2 Optimal Subsampling Estimators in Unconditional Inference **12**

2.1 Asymptotic properties of subsampling estimators in unconditional inference . 13

2.2 Empirical studies 17

2.3 Summary and discussion 25

3 Optimal Subsampling Estimators in Conditional Inference **27**

3.1 Asymptotic properties of subsampling estimators in conditional inference . . 28

3.2 Relationship between the subsampling estimators 31

3.3 Empirical studies 35

3.4 Summary and discussion 43

4 Bayesian Spline Smoothing with Ambiguous Penalties **45**

4.1 Introduction 46

4.2 Accounting ambiguity of J using mixture distribution as a prior for η 47

4.3 Algorithm for computing posterior probabilities 51

4.4 Empirical studies 53

4.5 Summary and discussion 61

5 Small RNAs-Dependent Gene Silencing in *C. elegans* 62

5.1 Background 63

5.2 Preliminary data analysis 65

6 Concluding Remarks and Future Works 68

A Proofs of Asymptotic Properties of Optimal Subsampling Estimators 70

A.1 Proofs of Theorem 1 and Theorem 2 70

A.2 Unweighted subsampling estimators 81

B Proofs of Bayesian Spline Smoothing with Ambiguous Penalties 87

B.1 Proof of Theorem 3 87

Bibliography 97

Chapter 1

Introduction

Chapter summary: With the rapid development of science and technology, large and complex data have been generated in many areas, such as genomics, social media, economics, and neuroscience. The extraordinary amount of big and complicated datasets provide unprecedented opportunities for data-driven knowledge discovery and decision making. However, the task of analyzing these data itself becomes a significant challenge. On the one hand, the ultra-large size of datasets renders the application of many statistical methods computationally impossible. On the other hand, with the system being studied getting more complicated, the model set-up for some popular off-the-shelf methods may not be applicable anymore. In this chapter, I briefly summarize the large-scale computation through subsampling and integrating model uncertainty problems in big and complex data.

1.1 Subsampling for large-scale regression problems

The rapid advance in science and technology within the past few decades has brought an extraordinary amount of previously inaccessible data. To deal with the sheer large size of the data, one powerful family of methods is the subsampling methods. In subsampling methods, instead of reading in and analyzing the full sample of huge size, one first takes a random subsample from the original full sample, then uses this subsample as a surrogate in the subsequent computation and estimation. *The key to the success of subsampling methods is to construct data-adaptive subsampling probability, which gives preference to those data points that are influential to model fitting and statistical inference of interest.* In this section, we study the linear regression problem, one of the most fundamental problems in statistics and machine learning.

Observe the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ and the predictor matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$, we consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the coefficient vector to be estimated, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$, and ε_is are independently and identically distributed with mean 0 and variance $\sigma^2 (< \infty)$ for $i = 1, \ldots, n$. We assume that sample size n is large and that the predictor matrix $\mathbf{X}$ has full column rank, i.e., $\mathbf{X}^T\mathbf{X}$ is invertible. A popular method to fit model (1.1) to data is the least squares (LS) method. The resulting ordinary least squares (OLS) estimator, $\hat{\boldsymbol{\beta}}_{OLS}$, of $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{1.2}$$

where $\|\cdot\|$ represents the Euclidean norm.

The computation cost for calculating $\hat{\boldsymbol{\beta}}_{OLS}$ is $O(np^2)$, which is daunting when n and/or p are large. Algorithm 1 summarizes the steps of general subsampling methods for regression

problems.

Algorithm 1 Random Subsampling Estimation Algorithm

- **Step 1 (Subsampling) Subsample from the full data.**
  Draw a random subsample of size $r \ll n$ with replacement from the full data using $\{\pi_i\}_{i=1}^n$ as the subsampling probabilities. The resulting subsample is denoted as $(\mathbf{X}^*, \mathbf{Y}^*)$. The corresponding subsampling probability is recorded as $\{\pi_i^*\}_{i=1}^r$.

- **Step 2. (Estimation) Calculate the following weighted least squares using the subsample to get the estimator $\tilde{\boldsymbol{\beta}}$.**

$$\tilde{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\Phi}^*\mathbf{Y}^* - \boldsymbol{\Phi}^*\mathbf{X}^*\boldsymbol{\beta}\|^2 = (\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{Y}^*, \tag{1.3}$$

  where $\boldsymbol{\Phi}^* = \text{diag}(1/\sqrt{r\pi_i^*})$.

Remark 1. It is worth mentioning that Algorithm 1 is also referred to as the *weighted* subsampling algorithm. In contrast, Ma et al. (2014, 2015) developed *unweighted* subsampling methods, in which one takes a subsample with certain subsampling probabilities and the estimates in Step 2 are not weighted by the subsampling probability. Ma et al. (2014, 2015) observed that unweighted subsampling estimators displayed some attractive properties, such as smaller variances than those of the weighted leverage-based subsampling estimators. In this thesis, we focus on studying the weighted subsampling estimators. Some discussion of unweighted subsampling methods and the comparison between weighted and unweighted estimators are provided in Appendix A.2.

Subsampling methods have received a significant amount of attention, both traditionally within statistics and more recently within the area of Randomized Linear Algebra (RLA), since they can achieve high-quality solutions at lower computation cost compared to traditional computing packages. With different subsampling probabilities, Algorithm 1 defines a family of subsampling estimators. The uniform subsampling estimator (UNIF) uses uniform subsampling probability, i.e., $\pi_i^{UNIF} = 1/n$ for $i = 1, \ldots, n$. Drineas et al. (2006, 2011) proposed the basic leverage-based subsampling estimator (BLEV), and the sampling probability

is proportional to the leverage score h_{ii} of the predictor matrix $\mathbf{X}$, i.e.,

$$\pi_i^{BLEV} = \frac{h_{ii}}{\sum_{i=1}^{n} h_{ii}},$$

where $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ for $i = 1, \ldots, n$. Within RLA, various subsampling methods with leverage-score-based subsampling probability distributions have recently been proved, both theoretically and empirically, to lead to estimators for approximating the full sample least squares estimates in linear models (Drineas et al., 2006, 2011, 2012; Mahoney, 2011; Drineas and Mahoney, 2016).

Most early efforts for the study of subsampling estimators focused on algorithmic aspects, e.g., running time bounds, independent of inferential bounds on the problem solution (Drineas et al., 2006; Mahoney, 2011; Drineas et al., 2012; Woodruff et al., 2014; Yang et al., 2015). Recently, a significant amount of efforts have been devoted to the study of statistical properties of subsampling estimators. In the literature, there are two classes of approaches to this study. The first class is based on finite sample concentration inequalities. Using concentration inequalities on a given data set, Drineas et al. (2006, 2011, 2012), Clarkson and Woodruff (2013), and Meng and Mahoney (2013) showed that the approximation error of subsampling estimators could be bounded. Raskutti and Mahoney (2015) designed criteria of prediction efficiency (PE) and residual efficiency (RE) for the sketched LS estimator and provided upper bounds for several types of random projection and random sampling algorithms. The second class is based on the mean squared error (MSE). Ma et al. (2014, 2015), in particular, employed a Taylor expansion to study the estimator's MSE, based on which they proposed improved subsampling estimators, i.e., shrinkage leverage estimator (SLEV) with $\pi_i = \lambda h_{ii}/\sum_{i=1}^{n} h_{ii} + (1-\lambda)/n$, where λ is a constant weight between 0 and 1. In the course to minimize the variance of the subsampling estimator, Wang et al. (2016) proposed relaxed optimal criteria from the perspective of experiment design. They suggested a simplified approximation to the traditional A-optimality criterion. However, it is worth pointing out that

the approximated A-optimality criterion in Wang et al. (2016) is based on considering the *conditional variance of subsampling estimator given a subsample*; thus, the randomness from subsampling is not considered in this criterion. Chen et al. (2016) employed the generalized inverse of random matrices and provided new subsampling estimator for linear regression. Their approach did incorporate the randomness of both subsampling and error in the model. Unfortunately, their proposed estimator relies on the unknown true parameters $\boldsymbol{\beta}_0$ and σ^2, rendering their estimator impractical. Along with this line of thinking, it remains elusive whether there exists a practically applicable optimal subsampling estimator in the sense of minimizing MSE or its variants.

To answer this question, we need to have a deeper understanding of the statistical properties of the random subsampling estimators. The main challenge is that there are two sources of randomness contributing to the statistical performance of subsampling estimators. The first source of randomness is the random errors in the model, i.e., ε_is, which are typically attributed to measurement error or random noise inherited in random observations $\mathbf{Y}$. The second source of randomness is the random subsampling procedure. Furthermore, these two sources of randomness coupled together in the estimator, that is, the subsampling estimator is expressed as $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^{*})^{-1}\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{Y}^{*}$, as in Algorithm 1. See the text above Eqn (A.2) in Appendix A.1.1 for another representation of $\tilde{\boldsymbol{\beta}}$. The term $(\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^{*})$ involves the random subsampling procedure, and it is engaged in the subsampling estimator in a nonlinear fashion (inversion). Due to the presence of randomness introduced by subsampling and its nonlinear involvement in the subsampling estimator, it is nontrivial to directly analyze the exact distribution of subsampling estimator.

When dealing with estimators with complicated forms such as the case with subsampling estimators, one common practice in statistics is to rely on asymptotic theory to simplify the complicated analytical expressions of the quantity of interest based on the behavior of estimators in large samples, i.e., the asymptotic theory. In our quest of optimal subsampling estimators, we exploit the same strategy. In Chapter 2, we derive the asymptotic distribution

of subsampling estimators for a large sample n with certain regularity conditions satisfied. Aside from considering the subsampling estimator for estimating the true parameter, which we refer to as the unconditional inference, we also consider the case of using subsampling estimator for approximating the full sample estimates, which we refer to as the conditional inference, in Chapter 3. In conditional inference, we consider the full sample as given, and the goal of taking the subsample is to approximate any calculations based on the full sample, e.g., the OLS estimate in (1.2). With the asymptotic results derived, we propose the *optimal* subsampling estimators that minimize various versions of the asymptotic mean squared error (AMSE).

1.2 Model uncertainty in smoothing spline

Model uncertainty is one of the challenging areas of statistical analysis, especially for big data which are usually complicated and challenging to put a unified model assumption on. In practice, when faced with multiple possible candidate models, the popular approach nowadays is to perform model selection using criteria such as AIC, BIC and MSE, and make statistical inference solely based on the selected model as if it were the true model. In this case, the uncertainty in the previous model selection step is not reflected in statistical inference. Conditioning on a single selected model ignores model uncertainty and thus leads to a poor estimation of uncertainty in statistical inference. In this section, we study and illustrate this problem of model uncertainty in the framework of smoothing spline, one of the most popular nonparametric methods for flexible function estimation.

Consider a regression problem in which we observe x_i $(i = 1, \ldots, n)$ on the domain $\mathscr{X}$ and the response Y_i that depends on x_i through an unknown function $\eta(\cdot)$ defined on $\mathscr{X}$, i.e.,

$$Y_i = \eta(x_i) + \varepsilon_i, \quad x_i \in \mathscr{X}, \quad i = 1, \ldots, n, \tag{1.4}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. One popular approach to estimating η is the smoothing spline method, in which one minimizes the following penalized least squares

$$\min_{\eta} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \lambda J(\eta), \tag{1.5}$$

where the first term measures the goodness-of-fit of the model to the data, the second term $J(\eta)$ is a penalty functional on η, and λ is a finite nonnegative tuning parameter. See Wahba (1990), Ramsay and Silverman (2005), Wang (2011), Gu (2013) and references therein for comprehensive overviews of the vast field of nonparametric estimation through spline smoothing.

In the classical application of the smoothing spline method, the functional $J(\eta)$ in (1.5) is pre-specified to reflect users' assumptions on η. The form of $J(\eta)$ defines the space in which we carry out the minimization, i.e., $\mathcal{H}_J = \{\eta : J(\eta) < \infty\}$ or a subspace of $\mathcal{H}_J$. The choice of λ determines an estimate by balancing the constraint specified by J and the goodness-of-fit of the model to the observed data. Given a penalty functional $J(\eta)$, a significant amount of efforts have been devoted to the estimation of the λ (Wahba and Craven, 1978; Wahba, 1985; Gu and Wahba, 1991). As important as a good choice of λ is, it is worth noting that the effectiveness of tuning λ in estimating η is subject to an appropriate choice of $J(\eta)$ in the first place. Thus the choice of the functional $J(\eta)$ serves as a higher level of tuning for the estimation of η and is of similar importance to the tuning of λ subsequently. In the following, we discuss the choice of $J(\eta)$ and the associated ambiguity penalty challenges from two practical perspectives.

In the context of data smoothing using smoothing spline, the form of $J(\eta)$ specifies a constraint on the roughness of η. When $\mathscr{X}$ is an interval of the real line, e.g., $[0, 1]$, the penalty $J(\eta) = \int_0^1 (\eta'')^2 \mathrm{d}x$ and $J(\eta) = \int_0^1 (\eta')^2 \mathrm{d}x$ are both popular choices. These yield the cubic spline and linear spline estimates for η, respectively. More generally, we may use $J(\eta) = \int_0^1 (\frac{\mathrm{d}^m \eta}{\mathrm{d}x^m} + a_{m-1}\frac{\mathrm{d}^{m-1}\eta}{\mathrm{d}x^{m-1}} + ... + a_1 \frac{\mathrm{d}\eta}{\mathrm{d}x} + a_0)^2 \mathrm{d}x$, where a_is are constant coefficients. Using higher or-

der splines penalty, i.e., larger m, implies that one tends to favor smoother functions. When $\mathscr{X}$ is a product domain, e.g., $[0,1]^d$ with $d > 1$, the tensor product cubic spline and thin-plate spline are two popular approaches. In the tensor product cubic spline with $d = 2$, we take $J(\eta) = \theta_1 \int_0^1 (\int_0^1 \frac{\partial^2 \eta}{\partial x_{\langle 1 \rangle}^2} \mathrm{d}x_{\langle 2 \rangle})^2 \mathrm{d}x_{\langle 1 \rangle} + \theta_2 \int_0^1 (\int_0^1 \frac{\partial^2 \eta}{\partial x_{\langle 2 \rangle}^2} \mathrm{d}x_{\langle 1 \rangle})^2 \mathrm{d}x_{\langle 2 \rangle} + \theta_3 \int_0^1 (\int_0^1 \frac{\partial^3 \eta}{\partial x_{\langle 1 \rangle} \partial x_{\langle 2 \rangle}^2} \mathrm{d}x_{\langle 1 \rangle})^2 \mathrm{d}x_{\langle 2 \rangle} + \theta_4 \int_0^1 (\int_0^1 \frac{\partial^3 \eta}{\partial x_{\langle 1 \rangle}^2 \partial x_{\langle 2 \rangle}} \mathrm{d}x_{\langle 2 \rangle})^2 \mathrm{d}x_{\langle 1 \rangle} + \theta_5 \int_0^1 \int_0^1 (\frac{\partial^4 \eta}{\partial x_{\langle 1 \rangle}^2 \partial x_{\langle 2 \rangle}^2})^2 \mathrm{d}x_{\langle 1 \rangle} \mathrm{d}x_{\langle 2 \rangle}$, where $\theta_1, \ldots, \theta_5$ are tuning parameters like λ in (1.5). For m-th order thin-plate spline, the penalty functional $J(\eta)$ is of the form $\sum_{\alpha_1 + \ldots + \alpha_d = m} \frac{m!}{\alpha_1! \ldots \alpha_d!} \int \ldots \int \left(\frac{\partial^m \eta}{\partial x_{\langle 1 \rangle}^{\alpha_1} \ldots \partial x_{\langle d \rangle}^{\alpha_d}} \right)^2 \mathrm{d}x_{\langle 1 \rangle} \ldots \mathrm{d}x_{\langle d \rangle}$. In this case, using the thin plate spline explicitly implies that we assume η has rotational invariance, whereas using the cubic spline does not. Taking different penalties implies different assumptions on the smoothness of the function η. In practice, using different penalties may lead to different conclusions.

Below we illustrate the influence of the choice of J using a toy example. A random sample of size 100 is generated according to the model (1.4) and the scatter plot of raw data is shown in the left panel of Figure 1.1. The scatter plot displays a general parabolic pattern with a few repeated peaks and troughs. Based on the observation, we model the data through penalized least squares using two reasonable choices of splines: periodic cubic spline with $J(\eta) = \int_0^1 (\eta'')^2 \mathrm{d}x$ and linear spline with $J(\eta) = \int_0^1 (\eta')^2 \mathrm{d}x$. The tuning parameters for both splines are selected by generalized cross-validation (GCV) (Wahba and Craven, 1978). The fitted curves are plotted in the middle and right panels of Figure 1.1, respectively. From the periodic cubic spline fit, we conclude that the relationship is smooth and the data generating system is noisy; from the linear spline fit, we conclude that the relationship is wiggly and the data generating system is not too noisy. Based on the different choice of penalties, the conclusions we make are totally different. For cases like this, there exists ambiguity in choosing between the two estimates and there requires a unifying model for resolving such conflicting results. Indeed, with limited information at hand, it is more reasonable to take both functionals into consideration, rather than choosing one over the other. Another motivation for considering multiple penalties in data smoothing problem is that using a

single penalty functional specifies a constraint on which family of basis function to be used. It is expected that the estimation procedure using multiple penalties should benefit from a variety of basis functions from multiple families and be able to adapt to the complex features of the function to be estimated.

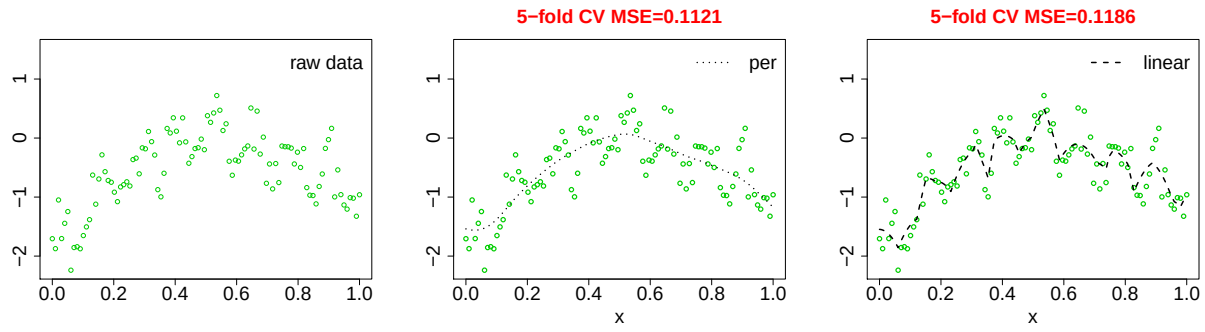


Figure 1.1: A toy example. From left to right: scatter plot of the raw data; curve fitted using periodic cubic spline (per); curve fitted using linear spline (linear). For both of the spline fit, the tuning parameter λ is selected by the generalized cross-validation (GCV) (Wahba and Craven, 1978). We also calculated MSE based on five-fold cross-validation. The MSEs for periodic cubic and linear spline are 0.1121 and 0.1186, respectively.

In the context of modeling a general dynamic system using smoothing spline, the choice of $J(\eta)$ is based on one's belief in the mechanism of the data generating system, e.g., a system of ordinary differential equations (ODEs). In some applications, there is definitive guiding information. For example, the distribution of heat (or variation in temperature) in a given region over time is governed by the heat equation. Using this information, technical tools are available to define the penalty (Gu, 2013, Chapter 4). However, a suitable choice of differential equation systems is often not unique. This is particularly true in biological, ecological, and social and economic sciences, where the principles used to deduce the differential equation systems tend to be less exact than those used in physics and chemistry. There might be multiple competitive ODE systems that can be used to describe the same underlying data generating system. In this case, our belief in the dynamic system is ambiguous. Take the popular dynamic system of interacting predator and prey species as an example. We use H and P to denote the population sizes of the prey and predator, respectively.

There are two popular and well-studied models: the Lotka-Volterra model, which assumes $\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP$, and $\frac{\mathrm{d}P}{\mathrm{d}t} = eHP - vP$, and the predator-dependent functional response model, which assumes $\frac{\mathrm{d}H}{\mathrm{d}t} = rH - aHP/(1+zP)$, and $\frac{\mathrm{d}P}{\mathrm{d}t} = eHP/(1+zP) - vP$, where r, a, e, v, and z are pre-specified constants (Murdoch et al., 2003). Based on different ODEs, the interpretation of the underlying dynamic system and estimation of the quantity of interest will be different. Another example is the modeling of gene expression data. In the literature, there are efforts in modeling the time course gene expression data using ODEs (Chen et al., 1999; Zhang et al., 2018). In a genome-wide analysis, a large number of genes will be analyzed under the same assumption. There naturally exist genes associated with essential periodic biological processes, such as cell cycle or circadian rhythm regulation known to be rhythmic (Whitfield et al., 2002; Luan and Li, 2004); in the same time frame there are other genes that can be non-periodically activated with irregular intervals in a living eukaryotic cell, like pulses turning on and off rapidly and discontinuously (Chubb et al., 2006). In other words, the genes do not necessarily follow the same set of ODEs. In both of the examples described above, unified methods that reconcile different assumptions or beliefs for the same dynamic system are in need.

To the best of our knowledge, the problem of dealing with two or more ambiguous penalties is not well studied for nonparametric models. One possible solution to the problem of ambiguous penalties, as mentioned in the beginning of this section, is through model selection. Direct application of classical criteria, such as AIC (Akaike, 1974) and BIC (Schwarz, 1978), is nontrivial since the dimensionality for smoothing spline fit is hard to define, mostly because that the concept of degree of freedom for nonparametric models still remain illusive. A practical approach is to evaluate the prediction performance of estimators based on different penalties using cross-validation. However, this strategy can easily fail, especially when the performance of estimators corresponding to different penalties are very similar. In the toy example shown in Figure 1.1, we also evaluated the prediction performance of two methods using five-fold cross-validation. The MSEs of the periodic cubic spline and linear

spline are 0.1121 and 0.1186, respectively. The prediction accuracy of different estimators are close to each other; it is still unclear how to select among two candidate penalties. If we proceed to choose a penalty with the lowest MSE in spite of the ambiguity, then great care needs to be taken on the statistical inference made thereon. The topic of post-selection inference is still an ongoing topic largely focusing on the problem of variable selection in linear regression problems (Berk et al., 2013; Tibshirani et al., 2016; Lee et al., 2016). Currently available results are not readily applicable to the nonparametric setting.

Besides model selection, another loosely related topic is the analysis of spatially inhomogeneous data. Luo and Wahba (1997) and Sklar et al. (2013) developed computation algorithms using multiple libraries of basis to adapt to certain local structures of data. Overall, a self-contained model for combining multiple penalties is still lacking.

In this thesis, instead of performing model selection, we aim to incorporate the ambiguity in choosing penalty functional in model inference for smoothing spline. In Chapter 4, we take a Bayesian perspective on the problem of ambiguous penalties, make use of the connection between penalized least squares and Bayesian estimation, and build hierarchical Bayesian model for properly integrating the model uncertainty in the context of a smoothing spline.

Chapter 2

Optimal Subsampling Estimators in Unconditional Inference

Chapter summary: One popular method for dealing with the computational burden brought by the sheer size of large-scale data sets is through random subsampling. While subsampling methods have attracted a significant amount of attention in the literature in the past few years, statistical works that study the large sample/asymptotic properties of these subsampling methods are still lacking. In Section 2.1 we outline the regularity conditions and establish the asymptotic normality of the subsampling estimator for estimating parameters in linear regression, and we propose several criteria and optimal subsampling probability distributions; and in Sections 2.2, we present some empirical results. All of the technical proofs are listed in Appendix A.

In this chapter, we study the asymptotic properties of the subsampling estimator $\tilde{\boldsymbol{\beta}}$ in the scenario of unconditional inference, i.e., estimating the true model parameter $\boldsymbol{\beta}_0$. Then, we evaluate the performance of subsampling estimators using AMSE under different scenarios and propose the corresponding optimal estimators.

2.1 Asymptotic properties of subsampling estimators in unconditional inference

In unconditional inference, we are interested in the setting where the subsample size r is a fraction of full sample size n, thus we write the subsample size $r = O(n^{1-\alpha})$, where $0 \leq \alpha < 1$. A larger α gives rise to a smaller r. Moreover, to satisfy $\sum_{i=1}^{n} \pi_i = 1$, the minimum subsampling probability should be less than or equal to $1/n$, i.e., $\pi_{min} = \min\{\pi_i\}_{i=1}^{n} \leq 1/n$, and we write $\pi_{min} = O(n^{-\gamma_0})$, where $\gamma_0 \geq 1$. A larger γ_0 gives rise to a smaller π_{min}. Using these notations, we show the asymptotic normality of $\tilde{\boldsymbol{\beta}}$ in unconditional inference in Theorem 1.

Theorem 1. *We assume that the following regularity conditions hold.*

(A1). There exist positive constants b and B such that

$$b \leq \lambda_{min}(\mathbf{X}^T\mathbf{X}/n) \leq \lambda_{max}(\mathbf{X}^T\mathbf{X}/n) \leq B,$$

where $\lambda_{max}(\mathbf{X}^T\mathbf{X}/n)$ and $\lambda_{min}(\mathbf{X}^T\mathbf{X}/n)$ denote the maximum and minimum eigenvalue of matrix $\mathbf{X}^T\mathbf{X}/n$.

(A2). $\gamma_0 + \alpha < 2$.

As $n \to \infty$, we have

$$(\sigma^2\boldsymbol{\Sigma}_0)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \xrightarrow{d} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p), \tag{2.1}$$

where $\mathbf{\Sigma}_0 = (\mathbf{X}^T\mathbf{X})^{-1}[\mathbf{X}^T(\mathbf{I}+\mathbf{\Omega})\mathbf{X}](\mathbf{X}^T\mathbf{X})^{-1}$, $\mathbf{\Omega} = diag\{1/r\pi_i\}_{i=1}^n$, *and* $\mathbf{I}_p$ *denotes a* $p \times p$ *identity matrix. Thus, in unconditional inference,* $\tilde{\boldsymbol{\beta}}$ *is an asymptotically unbiased estimator of* $\boldsymbol{\beta}_0$*, i.e.,*

$$AE(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0, \tag{2.2}$$

and the asymptotic variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ *is*

$$AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{\Sigma}_0. \tag{2.3}$$

The proof of Theorem 1 is provided in Appendix A.1.2.

The assumption in *(A1)* indicates that $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ converges to a positive definite constant matrix. This requires that predictor matrix $\mathbf{X}$ is of full column rank and the values of elements in $\mathbf{X}$ are not overdispersed. This condition ensures the consistency of full sample OLS estimator, and it is commonly used to help ease the technical handling of proof in regression model (Lai et al., 1978; Zou, 2006).

The assumption in *(A2)* puts a constraint on the smallest subsampling probability and subsample size. When $\gamma_0 + \alpha < 2$ holds, we have $n^{-\gamma_0} > n^{-(2-\alpha)}$, i.e., the smallest subsampling probability cannot be too small and in particular it should be $O(n^{\alpha})$ away from $O(n^{-2})$. Condition *(A2)* can also be rewritten as $n^{1-\alpha}n^{-\gamma_0} > n^{-1}$, which states that when the smallest subsampling probability is small, the subsample size needs to be large so that data points with the smallest subsampling probability can get a fair chance to be selected into the subsample. We also note that the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ in Theorem 1 has a "sandwich-type" expression. The center term depends on the reciprocal subsampling probabilities. Thus the direct constraint on the minimum subsampling probability is to avoid blowing up the variance.

Using the asymptotic results in Theorem 1, we propose the following three estimators for different purposes.

Estimating $\boldsymbol{\beta}_0$. We plug in the asymptotic variance and asymptotic squared bias in

Theorem 1 and get the $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$,

$$AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \;=\; \sigma^2 \mathrm{tr}\{(\mathbf{X}^T\mathbf{X})^{-1}\} + \frac{1}{r}\sum_{i=1}^{n} \frac{\sigma^2}{\pi_i} ||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{2.4}$$

Thus, $AMSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ can be considered as a function of $\{\pi_i\}_{i=1}^n$. It is trivial to use the Lagrange multipliers to find the minimizer of the right-hand side of (2.4) with the constraint $\sum_{i=1}^n \pi_i = 1$. The following proposition gives the minimum $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$ subsampling estimator.

Proposition 1. The subsampling estimator with the subsampling probabilities

$$\pi_i = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, i = 1, \ldots, n, \tag{2.5}$$

has the smallest $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$. The estimator with the subsampling probabilities in (2.5) is referred to as inverse-covariance (IC) subsampling estimator.

Remark 2. Using the main Algorithm 1 in Drineas et al. (2012), the calculation of subsampling probabilities in IC can be reduced to $O(nplog(n)/\epsilon)$, where ϵ is the desired approximation error parameter.

Other than making inference on $\boldsymbol{\beta}_0$, we are also interested in linear functions of $\boldsymbol{\beta}_0$ of the form $\mathbf{L}\boldsymbol{\beta}_0$, where $\mathbf{L}$ is any constant matrix of suitable dimension. In particular, we present the inference on $\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$.

Predicting Y (Estimating $\mathbf{X}\boldsymbol{\beta}_0$). In regression analysis, prediction, i.e., inference on the true regression line $\mathbf{X}\boldsymbol{\beta}_0$, is crucially important. In this case, simple algebra yields that

$$AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0) \;=\; p\sigma^2 + \frac{1}{r}\sum_{i=1}^{n} \frac{\sigma^2}{\pi_i} ||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{2.6}$$

Thus, $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$ can be considered as a function of $\{\pi_i\}_{i=1}^n$ and used to develop the optimal subsample estimator. The following proposition gives the minimum $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0)$ subsampling estimator using the Lagrange multiplier.

Proposition 2. The subsampling estimator with the subsampling probabilities

$$\pi_i \;=\; \frac{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} = \frac{\sqrt{h_{ii}}}{\sum_{i=1}^n \sqrt{h_{ii}}}, i = 1, \ldots, n, \tag{2.7}$$

has the smallest $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0)$ and is referred to as root leverage (RLEV) subsampling estimator.

The second equality in (2.7) is by

$$\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|^2 = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i)^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = h_{ii}.$$

Remark 3. Chen et al. (2016) proposed optimal sampling estimators for estimating $\boldsymbol{\beta}_0$ and predicting $\mathbf{Y}$. Their optimal sampling probability depends on the unknown parameters, and they proposed to use (2.7) as a rough approximation of their subsampling probability without any rigorous justification.

Estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$. In this case, we examine

$$AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0) \;=\; \sigma^2\mathrm{tr}(\mathbf{X}^T\mathbf{X}) + \frac{\sigma^2}{r}\sum_{i=1}^n \frac{1}{\pi_i}||\mathbf{x}_i||^2. \tag{2.8}$$

Proposition 3. The subsampling estimator with the subsampling probabilities

$$\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{x}_i\|}, i = 1, \ldots, n, \tag{2.9}$$

has the smallest $AMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0)$ and is referred to as predictor-length (PL) subsampling estimator.

The computation costs of BLEV and RLEV subsampling estimators using exact leverage scores are in the order of $O(np^2)$. In contrast, the computation cost of PL subsample estimate is only $O(np)$, which is linear to both sample size and predictor size. This is the lowest computation cost among all non-uniform subsample estimators to our knowledge.

2.2 Empirical studies

To assess the performance of the proposed subsampling methods, we present the main results of our empirical analysis in this section.

2.2.1 Simulation studies

Simulation setting

We generated synthetic data from $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i$, where $\mathbf{x}_i$s and $\boldsymbol{\beta}_0$ are $p \times 1$ predictor vectors and $p \times 1$ coefficient vector, and random error $\varepsilon_i \overset{iid}{\sim} N(0, 1)$, $i = 1, \ldots, n$. Here we let $p = 10$, and sample size $n = 5000$. We set the first and last two entries of the coefficient vector $\boldsymbol{\beta}_0$ to be 1, and rest to be 0.1. The predictor $\mathbf{x}_i$s were independently generated from one of the following multivariate distributions.

- Multivariate normal distribution $\mathbf{N}(\mathbf{1}, \mathbf{D})$, where $\mathbf{1}$ is a $p \times 1$ column vector of 1, and the (i, j)th element of $\mathbf{D}$ is set to $1 \times 0.7^{|i-j|}$ for $i, j = 1, \ldots, p$. We refer to it as MN data.

- Multivariate noncentral t-distribution with 3 degrees of freedom, noncentrality parameter $\mathbf{1}$, and scale matrix $\mathbf{D}$, i.e., $t_3(\mathbf{1}, \mathbf{D})$. We refer to it as T3 data.

- Log-normal distribution $\mathbf{LN}(\mathbf{1}, \mathbf{D})$. We refer to it as LN data.

- Multivariate noncentral t-distribution with 1 degree of freedom, noncentrality parameter $\mathbf{1}$, and scale matrix $\mathbf{D}$, i.e., $t_1(\mathbf{1}, \mathbf{D})$. We refer to it as T1 data.

MN and T3 are two commonly seen examples of symmetric distributions used for generating the data (predictors). We use LN as an example of nonsymmetric data generating distributions. We use T1 as an example of distribution for which the expectation and variance do not exist, i.e., we do not expect the condition that $\mathbf{X}^T\mathbf{X}/n$ converges to a constant

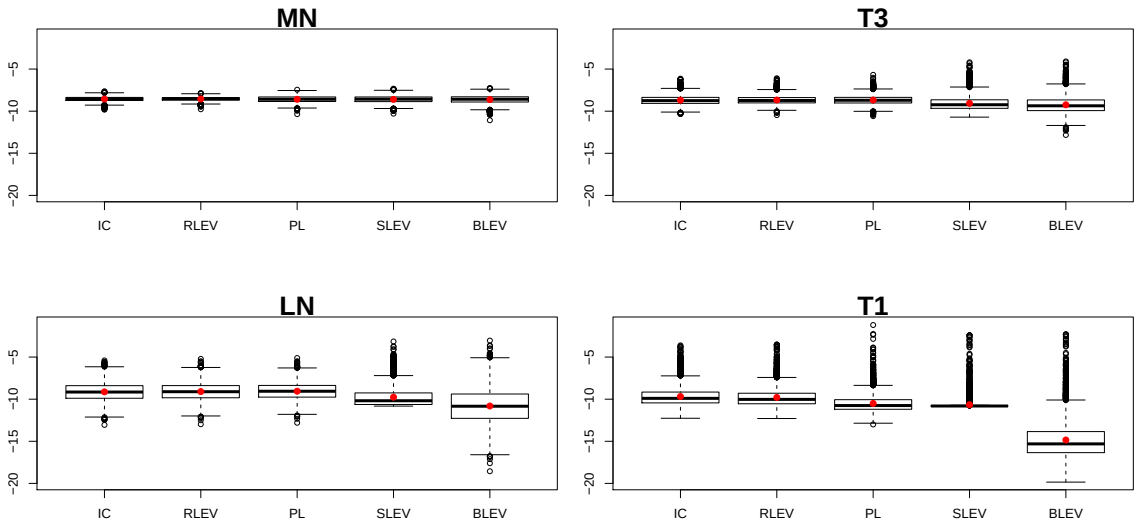


Figure 2.1: Box plots of the subsampling probabilities (in log scale) of all data points in IC, RLEV, PL, SLEV, and BLEV (from left to right in each panel) in MN, T3, LN, and T1 data for p=10 and n =5000. In each box plot, the red dot inside the box indicates the mean of corresponding subsampling probabilities (in log scale).

matrix to be satisfied. This is beyond the scope of Theorem 1; thus the good asymptotic performance of $\tilde{\boldsymbol{\beta}}$ might not be guaranteed. In particular, the bias and variances might not always be guaranteed to converge fast to zero as subsample sizes increase. Further, the variances of $\mathbf{X}\tilde{\boldsymbol{\beta}}$ and $\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}$ might not be properly defined as subsample sizes increase.

In Figure 2.1, we present box plots of the subsampling probabilities (in log scale) of all the data points in IC, RLEV, PL, SLEV, and BLEV (from left to right) in MN, T3, LN, and T1. The subsampling probability distributions of BLEV are generally more dispersive than those of other estimators. There exist a significant number of extremely small sampling probabilities in BLEV, especially when the data distribution has an heavier tail, such as the cases of LN and T1. These extremely small sampling probabilities in BLEV are effectively augmented in SLEV. However, the medians of sampling probabilities in SLEV are still smaller than the first quantiles of sampling probabilities in IC, RLEV, and PL in T3, LN, and T1. The large number of relatively small subsampling probabilities in BLEV and SLEV will

inflate the variances of subsampling estimators considering the expression of asymptotic variances in Theorem 1. Thus it is expected that BLEV and SLEV give rise to estimates with relatively large variances, especially when data was generated from distributions with heavy tails, such as LN and T1. As the tails of generating distributions of $\mathbf{x}_i$ get heavier (from MN to T3 to LN to T1), the sampling probability distributions in all methods get more dispersed. It is expected that the differences among different estimators, if any, will be more significant in T3, LN, and T1 than in MN.

Subsampling estimators for estimating true parameter

In this section, we evaluate the performance of proposed subsampling estimators in estimating $\boldsymbol{\beta}_0$, $\boldsymbol{X}\boldsymbol{\beta}_0$, and $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0$. Following the simulation setting stated above, we generated 100 replicates of MN, T3, LN, and T1 data. We apply IC, RLEV, PL, SLEV, and BLEV to each replicated data set to obtain $\tilde{\boldsymbol{\beta}}_b$ at subsample sizes $r = 100, 200, 500, 700, 1000$. Then, we calculated squared bias and variance for estimating $\boldsymbol{\beta}_0$ for each subsampling method.

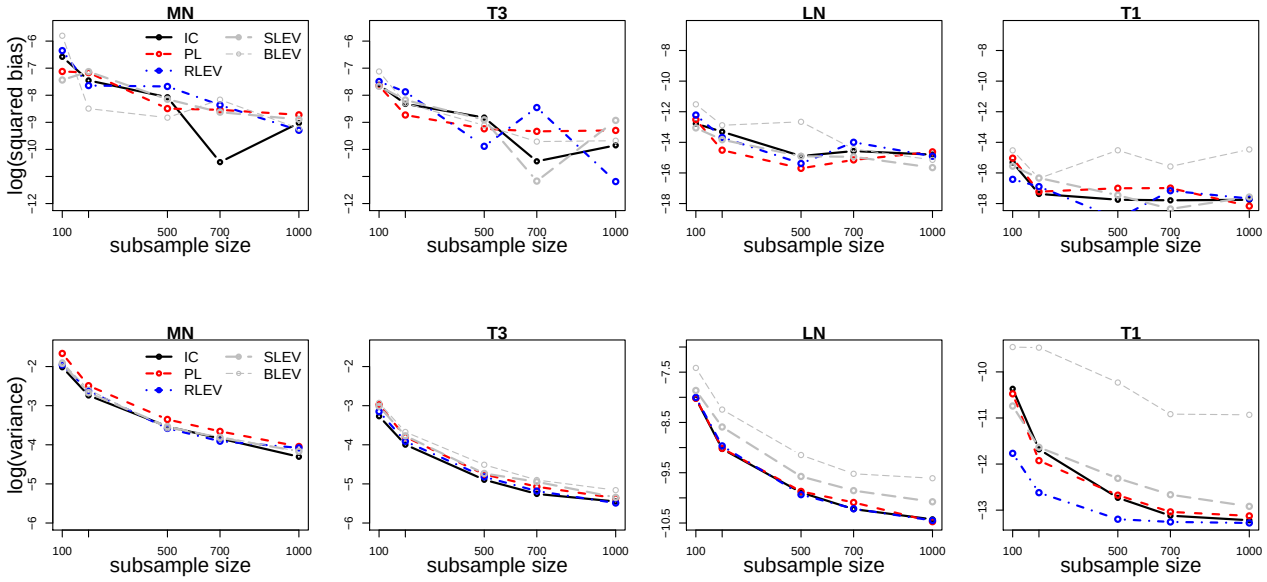


Figure 2.2: The squared biases (the first row) and the variances (the second row) of IC, RLEV, PL, SLEV, and BLEV estimates in estimating $\boldsymbol{\beta}_0$ (in log scale) at different subsample sizes.

In Figure 2.2, we plot the squared biases (the first row) and the variances (the second row) of IC, RLEV, PL, SLEV, and BLEV estimates in estimating $\boldsymbol{\beta}_0$ in MN, T3, LN and T1 in log scale. First, both the squared biases and the variances show decreasing patterns as subsample size increases, and the squared biases are much smaller than the corresponding variances. These observations are within expectation since Theorem 1 states that the weighted subsample estimators are asymptotic unbiased and consistent estimators of $\boldsymbol{\beta}_0$ provided that regularity conditions are satisfied. Second, the variances of estimates using IC, whose subsampling probabilities minimize $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$, are slightly smaller than the variances of estimates using other methods in MN and T3 at most subsample sizes. It should be noted that the variances of estimates using IC, RLEV, and PL are all smaller than those of BLEV and SLEV estimates in T3 and LN. As mentioned below Figure 2.1, the larger variances of BLEV estimates are caused by the existence of extremely small subsampling probabilities in BLEV. Taking a weighted average of subsampling probability distributions in BLEV and UNIF show a beneficial effect on the variances for SLEV estimators, but the variances of SLEV estimators are still larger than those of IC in T3, LN, and T1. It is worth mentioning that in T1, despite the violation of regularity conditions, the proposed estimators IC, RL, and PL still outperform BLEV and SLEV in terms of variances when the subsample size is greater than 200. Third, the squared biases and the variances of all estimates gets smaller from MN to T3 to LN and T1. Besides, we note that the proposed estimators IC, RL and PL perform extremely similar to each other. The close relationship among these estimators and those proposed in the next chapter will be discussed in Section 3.2 altogether.

In Figure 2.3, we plot the variances of IC, RLEV, PL, SLEV, and BLEV in predicting $\mathbf{Y}$ at different subsample sizes in MN, T3, LN, and T1. The variances of estimates using RLEV, whose subsampling probabilities minimize $AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0)$, are smaller than the variances of estimates using other methods in LN and T1 consistently when subsample sizes are larger than 200.

In Figure 2.4, we plot the variances of IC, RLEV, PL, SLEV, and BLEV estimates

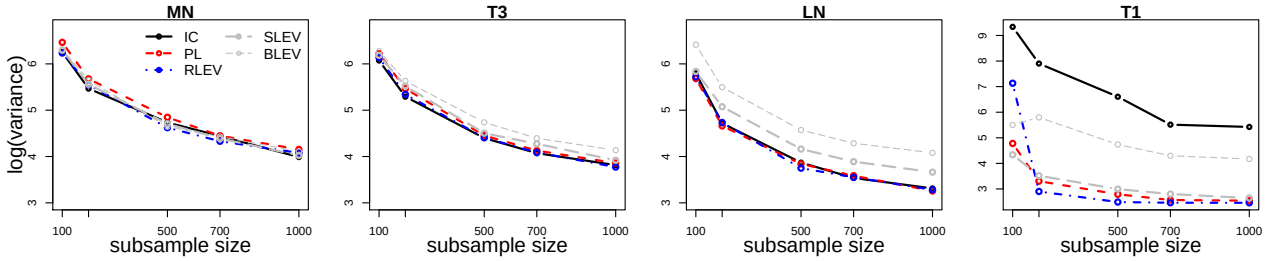


Figure 2.3: The variances of IC, RLEV, PL, SLEV, and BLEV estimates in predicting $\boldsymbol{Y}$ (in log scale) at different subsample sizes.

in estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ in MN, T3, LN, and T1. For estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, the variances of estimates using PL, IC, and RLEV are smaller than the variances of estimates using SLEV and BLEV in T3 and LN at most subsample sizes.

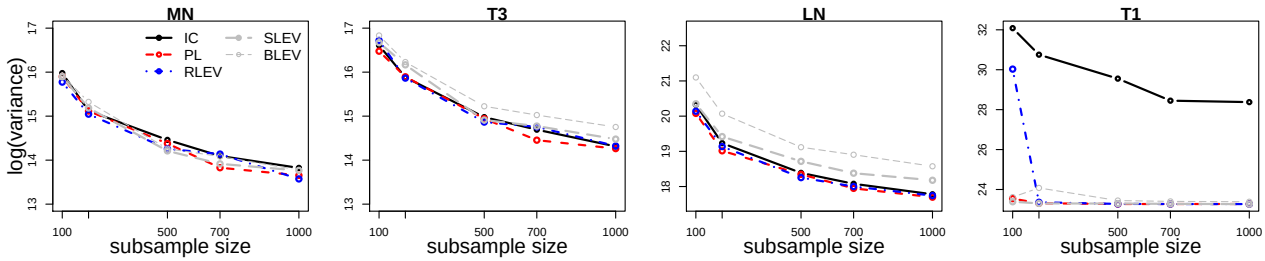


Figure 2.4: The variances of IC, RLEV, PL, SLEV, and BLEV estimates in estimating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ (in log scale) at different subsample sizes.

Next, we illustrate the performance of the proposed subsampling methods using two real-world examples.

2.2.2 Real data analysis

Airline delay

We compiled a flight delay data set from the website of the US Department of Transportation (USDOT, 2017). The data set contains records of $3,274,894$ US domestic flights during

weekdays from Mondays to Thursdays in 2017. There are five variables for each flight record: arrival delay (difference in minutes between scheduled and actual arrival time, and early arrivals show negative numbers), arrival taxi in time (in minutes), departure taxi out time (in minutes), departure delays (difference in minutes between scheduled and actual departure time, and early departures show negative numbers), and computer reservation system based elapsed time of the flight (in minutes; a measure for the distance of the flight). We are interested in predicting the arrival delay of each flight using the rest variables. We fit the linear regression model (1.1), with the response being flight arrival delay. In addition to the four variables (other than arrival delay) in our data set as linear predictors, we also included their quadratic and all pairwise interaction terms. We thus have 14 predictors in total. Considering the large number of flights, we used the subsampling methods for the estimation. The response was centered, and predictors were standardized prior to the analysis.

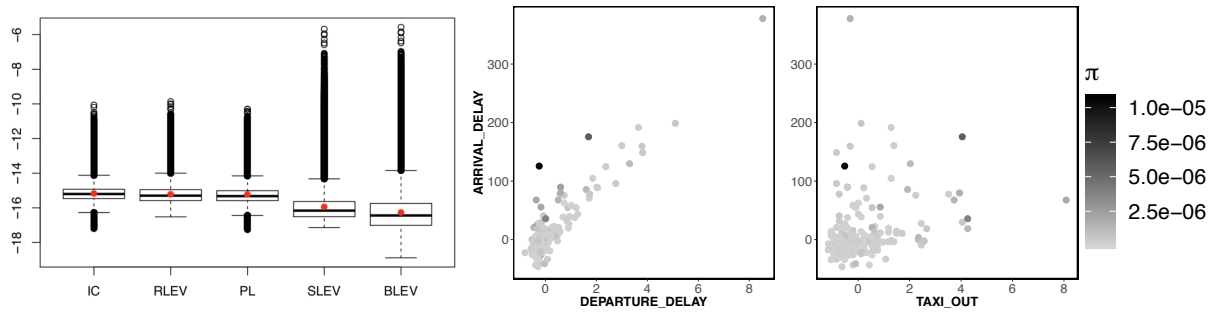


Figure 2.5: Exploratory analysis for Airline delay data. The left panel is the box plots of subsampling probabilities (in log scale) of all data points in IC, RLEV, PL, SLEV, and BLEV. The middle panel and right panel are the scatter plots of the 200 sampled response and two predictors using the subsampling probability distribution in IC.

In the left panel of Figure 2.5, we present the box plots of subsampling probabilities (in log scale) of all data points in IC, RLEV, PL, SLEV, and BLEV. We note that the subsampling probability distributions are right-skewed, similar to those in Figure 2.1 in the simulation study. Using the subsampling probability distribution in IC, we took a subsample of size 200 from the full data. The middle and right panels in Figure 2.5 are the scatter plots of the sampled response and the two predictors, respectively. These scatter plots provide a

visual sketch of the full sample data. We observe a somewhat linear relationship between the response and these two predictors, consolidating our linear model assumption.

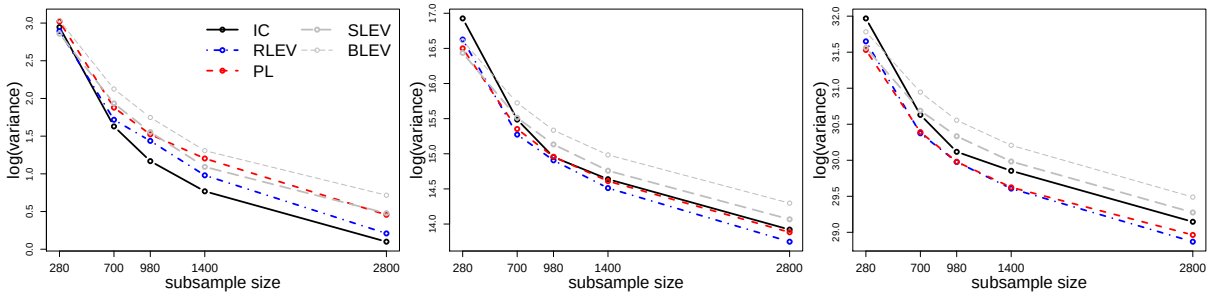


Figure 2.6: The variances of IC, RLEV, PL, SLEV, and BLEV estimates for estimating $\boldsymbol{\beta}_0$ (the left panel), $\mathbf{X}\boldsymbol{\beta}_0$ (the middle panel) and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ (the right panel) (in log scale) at different subsample sizes for airline delay data.

Next, we repeatedly apply IC, RLEV, PL, SLEV, and BLEV to this data set for 100 times at subsample size $r = 20p, 50p, 70p, 100p, 200p$ with $p = 14$. Since we do not have information on $\boldsymbol{\beta}_0$, we only present the variance of the resulting estimates for estimating $\boldsymbol{\beta}_0$, $\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ for each method. The results are summarized in Figure 2.6. For estimating $\boldsymbol{\beta}_0$, the estimates using IC, whose sampling probability minimizes the $AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$, show smaller variances and consistently outperform all other methods as subsample size increases. For estimating $\boldsymbol{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, the estimates using RLEV and PL show competitive performance as subsample sizes increases. IC, PL, and RLEV, all consistently outperform BLEV and SLEV in terms of variances at subsample sizes greater than 700.

"YearPredictionMSD"

The "YearPredictionMSD" data set (Bertin-Mahieux et al. (2011)) was downloaded from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). It consists of records of 515,345 songs released between the year 1922 and 2011. For each song, multiple segments are taken, and each segment is characterized by 12 timbre features. These timbre features capture timbral characteristics, such as brightness and flatness, of each segment. The mean and

variance of each timbre feature, as well as the covariances between every two timbre features, are calculated. Our primary interest for analysis is to use all timbre feature information to predict the year of release. We fitted the linear regression model (1.1) to this data set, where y_i is the year of releasing of the i-th song in log scale, and $\mathbf{x}_i$ is a 90×1 vector containing all timbre feature predictors, $i = 1, \ldots, 515,345$. The response was centered, and the predictors were standardized before the analysis.

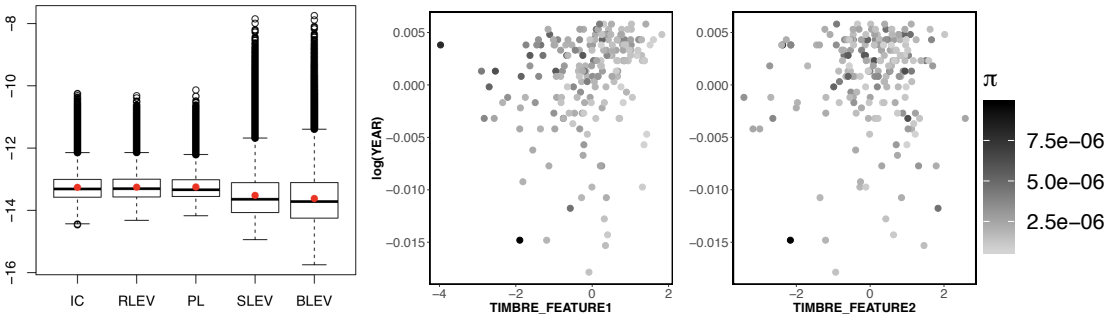


Figure 2.7: Exploratory analysis for "YearPredictionMSD" data. The left panel is for the box plots of subsampling probabilities of all data points (in log scale) for IC, RLEV, PL, SLEV, and BLEV. A subsample of size 200 is taken from the full data using the subsampling probabilities of IC. The middle panel and right panel are the scatter plots of sampled response and two timbre feature predictors.

In the left panel of Figure 2.7, we present the box plots of subsampling probability distributions in IC, RLEV, PL, SLEV, and BLEV in "YearPredictionMSD" data set. Inspecting the box plots reveals that all sampling distributions are right-skewed. Using the subsampling probability distribution in IC, we took a subsample of size 200 from the full data. The middle and right panels of Figure 2.7 are the scatter plots of the sampled response and two timbre features respectively. These scatter plots provide a visual sketch of the full sample data.

We repeatedly apply IC, RLEV, PL, SLEV, and BLEV methods to the data set for 100 times at subsample sizes $r = 10p, 20p, 50p, 70p, 100p$, with $p = 90$. In Figure 2.8, we plot the variances of the estimates for all subsample methods for estimating $\boldsymbol{\beta}_0$, $\boldsymbol{Y}\boldsymbol{\beta}_0$, and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ in log scale. For estimating $\boldsymbol{\beta}_0$, the variances of IC, RLEV, and PL estimates are

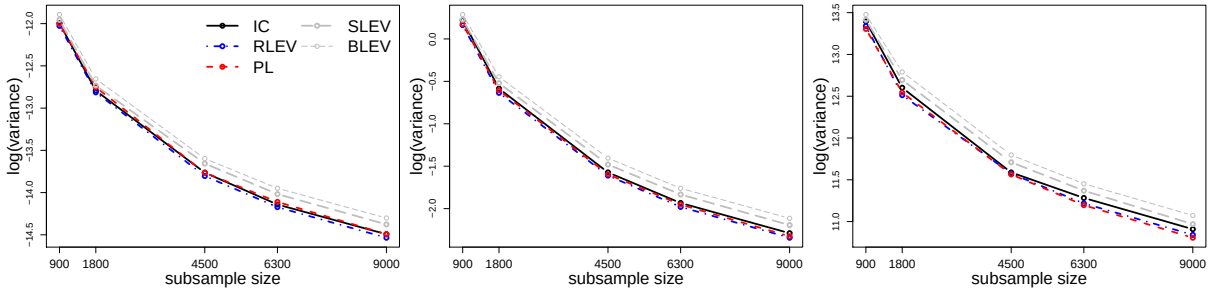


Figure 2.8: The variances of IC, RLEV, PL, SLEV, and BLEV estimates for estimating $\boldsymbol{\beta}_0$ (the left panel), $\boldsymbol{X}\boldsymbol{\beta}_0$ (the middle panel), and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$(the right panel) (in log scale) at different subsample sizes for "YearPredictionMSD" data.

comparable to each other and consistently smaller than those of SLEV and BLEV estimates at all subsample sizes. For estimating $\boldsymbol{Y}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, the variances of RLEV and PL estimates are consistently smaller the variances of estimates using other methods.

2.3 Summary and discussion

In this chapter, we study the properties of subsampling estimator for estimating parameters in the linear model. We establish the asymptotic normality of subsampling estimators under general regularity conditions. We show that subsampling estimators are asymptotically unbiased estimators of the true coefficient, and obtain an explicit form of the asymptotic variance in this case (Theorem 1). Since the normal distribution is uniquely determined by mean and variance, we then seek to construct subsampling estimators with subsampling probability distributions minimizing the AMSE. This leads us to propose two interesting subsampling estimator, the root leverage (RLEV) estimator (Proposition 2), whose subsampling probability is constructed using the square root of leverage scores, and the predictor-length (PL) estimator (Proposition 3), whose subsampling probability is constructed using the row norms of the predictor matrix. The RLEV and PL estimator minimize the AMSE for estimating

$\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$, respectively. In particular, the computation cost of PL estimator is only $O(np)$, which is the best possible one. Our results are fundamentally different from previous work on the statistical properties of the subsampling estimators since we provide asymptotic distribution for the estimators, rather than finite-sample concentration inequality on how the subsample estimators deviate from the full sample estimates based on the first two moments of the estimators. Although not the focus of this paper, it is worth pointing out that the asymptotic distribution can be used to perform statistical inference tasks such as hypothesis testing and constructing confidence intervals, whereas finite sample concentration inequality results may not.

We conduct a comprehensive empirical evaluation of the performance of subsampling estimators for estimating the parameters on synthetic data sets. We use predictor matrices generated from various distributions, including a heavy tail distribution and an asymmetric distribution. For all settings under consideration, we calculate the squared bias and variance of subsampling estimators. We demonstrate that the squared bias of subsampling estimators are much smaller than the correspoinding variances, and the squared bias decreases as subsample size increases. It is consistent with our theory's conclusion that the subsampling estimators are asymptotically unbiased. The variance of subsampling estimators decreases as subsample size increases, indicating the consistent property of subsampling estimators. In addition, we demonstrate that the novel estimators have better performance, i.e., smaller variances, than existing ones, confirming the optimality results we proposed.

Chapter 3

Optimal Subsampling Estimators in Conditional Inference

Chapter summary: In this chapter, we focus on the conditional inference, in which the goal of taking the subsample is to approximate the full sample estimates, e.g., the OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$. In Section 3.1, we outline the regularity conditions and establish the asymptotic normality of the subsampling estimator for approximating the full sample estimate, and we propose several criteria and optimal subsampling probability distributions. As a summary, in Section 3.2, we discuss the relationship between all the proposed estimators in this and the previous chapter, and several other subsampling estimators in the literature. In Sections 3.3, we present empirical results on simulated and two real-world examples. All of the technical proofs are listed in Appendix A.

In this chapter, we are interested in the conditional inference using the data set, i.e., the goal of the analysis is to approximate the full sample estimates, say $\hat{\boldsymbol{\beta}}_{OLS}$.

3.1 Asymptotic properties of subsampling estimators in conditional inference

In Theorem 2, we show the asymptotic properties of $\tilde{\boldsymbol{\beta}}$ when the data set is given.

Theorem 2. *Given the full sample data* $\{\mathbf{X}, \mathbf{Y}\}$*, we assume that* $\mathbf{X}$ *is of full column rank and* $\|\mathbf{x}_i\| < \infty$ *for* $i = 1, \ldots, n$*, where* $\mathbf{x}_i$ *is the* i*th row of* $\mathbf{X}$*. If the subsampling probabilities* $\{\pi_i\}_{i=1}^n$ *are nonzero, i.e.,* $\pi_i > 0$ *for* $i = 1, \ldots, n$*, then as subsample size* $r \to \infty$*, we have*

$$(\sigma^2 \boldsymbol{\Sigma}_c)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \quad \xrightarrow{d} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p)\,, \tag{3.1}$$

where $\boldsymbol{\Sigma}_c = \frac{1}{r}(\mathbf{X}^T\mathbf{X})^{-1}[\sum_{i=1}^n \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T](\mathbf{X}^T\mathbf{X})^{-1}$*. Thus, in conditional inference,* $\tilde{\boldsymbol{\beta}}$ *is an asymptotically unbiased estimator for* $\boldsymbol{\beta}_{OLS}$*, i.e.,*

$$AE(\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}_{OLS}, \tag{3.2}$$

and the asymptotic variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ *is*

$$AVar(\tilde{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{\Sigma}_c. \tag{3.3}$$

The proof is presented in Appendix A.1.3.

Theorem 2 shows that as subsample size r gets larger, the distribution of $\tilde{\boldsymbol{\beta}}$ can be well approximated by a normal distribution. Subsampling estimator $\tilde{\boldsymbol{\beta}}$ is asymptotically unbiased to $\hat{\boldsymbol{\beta}}_{OLS}$. Similar to the case of unconditional inference, the asymptotic variance $AVar(\tilde{\boldsymbol{\beta}})$ also has "sandwich-type" expression, where both of side terms $(\mathbf{X}^T\mathbf{X})^{-1}$ are free of the subsampling probability, and the center term depends on the reciprocal subsampling probabilities.

Thus, much like the case in Theorem 1 in unconditional inference, we also expect that the extremely small probabilities may result in large variances of the corresponding estimators, e.g., BLEV, and further the potential advantage of SLEV over BLEV for approximating full sample OLS estimate.

Note that AVar($\tilde{\boldsymbol{\beta}}$) in conditional inference depends on the full sample least square residuals e_i^2s, which are not readily available from the subsample. We tackle this problem by taking the expectation of e_i^2s. The metric we thus use in this case is Expected MASE (EAMSE),

$$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = E_{\mathbf{Y}}(AMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})). \tag{3.4}$$

Based on the asymptotic properties, we establish the EAMSE as a function of subsampling probabilities $\{\pi_i\}_{i=1}^n$ and propose the associated optimal subsampling estimators.

Approximating $\hat{\boldsymbol{\beta}}_{OLS}$. To propose optimal subsampling estimator for approximating $\hat{\boldsymbol{\beta}}_{OLS}$, we use asymptotic results in Theorem 2 to get $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$. Note that $\mathrm{E}_{\mathbf{Y}}(e_i^2) = (1 - h_{ii})\sigma^2$, then

$$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS}) = \mathrm{E}_{\boldsymbol{Y}}(\mathrm{tr}(AVar(\tilde{\boldsymbol{\beta}}))) = \frac{1}{r}\sum_{i=1}^{n} \frac{(1 - h_{ii})\sigma^2}{\pi_i} ||(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{3.5}$$

Using the Lagrange multipliers method, the following proposition gives subsampling estimator that minimizes $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$.

Proposition 4. The subsample estimator with the subsampling probabilities

$$\pi_i = \frac{\sqrt{1 - h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1 - h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}, i = 1, \ldots, n \tag{3.6}$$

has the smallest $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$ and is referred to as the inverse-covariance negative-leverage (ICNLEV) estimator.

Parallel to the inference on $\mathbf{X}\boldsymbol{\beta}_0$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0$ in unconditional inference, we present the analysis for approximating $\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

Approximating $\hat{\mathbf{Y}}$ ($\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$). In this case, we propose optimal subsampling estimator for approximating $\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, i.e., we want to approximate the full sample prediction. Using asymptotic results in Theorem 2, we get

$$EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}};\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^{n}\frac{(1-h_{ii})\sigma^2}{\pi_i}||\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i||^2. \tag{3.7}$$

Proposition 5. The subsample estimator with the subsampling probabilities

$$\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1-h_{ii}}\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|} = \frac{\sqrt{(1-h_{ii})h_{ii}}}{\sum_{i=1}^{n}\sqrt{(1-h_{ii})h_{ii}}}, i = 1,\ldots,n \tag{3.8}$$

has the smallest $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}};\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}))$, and is referred to the estimator with sampling probabilities in (3.8) as multiplication negative leveraging (MNLEV) estimator.

Approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. In this case, we focus on approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and note

$$EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}};\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) = \frac{1}{r}\sum_{i=1}^{n}\frac{(1-h_{ii})\sigma^2}{\pi_i}||\mathbf{x}_i||^2. \tag{3.9}$$

Proposition 6. The subsampling estimator with the subsampling probabilities

$$\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}, i = 1,\ldots,n \tag{3.10}$$

has the smallest $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}};\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$ and is referred to as predictor-length negative-leverage estimator (PLNLEV).

As a summary, the six proposed estimators (IC, RLEV, PL, ICNLEV, MNLEV, and PLNLEV), along with three others (UNIF, BLEV, and SLEV) that we include for completeness, are outlined in Table 3.1.

Estimator	Sampling Probabilities	Criterion	Results
UNIF	$\pi_i = \frac{1}{n}$	$--$	$--$
BLEV	$\pi_i = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}}$	$--$	–
SLEV	$\pi_i = \lambda \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} + (1-\lambda)\frac{1}{n}$	$--$	–
IC	$\pi_i = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$	$AMSE(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}_0)$	Eqn(2.5)
RLEV	$\pi_i = \frac{\sqrt{h_{ii}}}{\sum_{i=1}^n \sqrt{h_{ii}}}$	$AMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\boldsymbol{\beta}_0)$	Eqn(2.7)
PL	$\pi_i = \frac{\|\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{x}_i\|}$	$AMSE(\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}; \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0)$	Eqn(2.9)
ICNLEV	$\pi_i = \frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$	$EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$	Eqn(3.6)
MNLEV	$\pi_i = \frac{\sqrt{(1-h_{ii})h_{ii}}}{\sum_{i=1}^n \sqrt{(1-h_{ii})h_{ii}}}$	$EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$	Eqn(3.8)
PLNLEV	$\pi_i = \frac{\sqrt{1-h_{ii}}\|\mathbf{x}_i\|}{\sum_{i=1}^n \sqrt{1-h_{ii}}\|\mathbf{x}_i\|}$	$EAMSE(\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}; \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS})$	Eqn(3.10)

Table 3.1: A summary on three existing subsampling estimators (UNIF, BLEV, SLEV) and six subsampling estimators (IC, RLEV, PL, ICNLEV, MNLEV, PLNLEV) proposed in this work.

3.2 Relationship between the subsampling estimators

In this section, we elaborate on the relationships among IC, RLEV, PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV.

"Shrinking" properties of proposed estimators. In this part, we illustrate the "shrinking" property of proposed optimal subsampling estimators compared to the BLEV estimator. For the convenience of description, in this section we refer the numerators of the subsampling probabilities in subsampling estimators as scores, e.g., the RLEV score is $\sqrt{h_{ii}}$ and the MNLEV score is $\sqrt{(1-h_{ii})h_{ii}}$. In Figure 3.1, we show the relationships between leverage score h_{ii} (also named BLEV score in Figure 3.1) and RLEV score, MNLEV score,

and SLEV score ($0.9h_{ii} + 0.1p/n$ with $p/n = 0.2$). We observe that the MNLEV score $\sqrt{(1-h_{ii})h_{ii}}$ amplifies small h_{ii}s but shrinks the h_{ii}s that are close to one back to zero. But it is worth noting that since $\sum_{i=1}^{n} h_{ii} = p$, we expect h_{ii}s to be very small and bounded away from 1. The MNLEV score slightly shrinks the large leverage and greatly amplifies the moderate leverage scores. Both MNLEV and RLEV scores are nonlinear shrinkage of the leverage scores. The SLEV score ($0.9h_{ii} + 0.1p/n$) also shrinks the small h_{ii}s and amplifies the large h_{ii}s, but in a linear fashion.

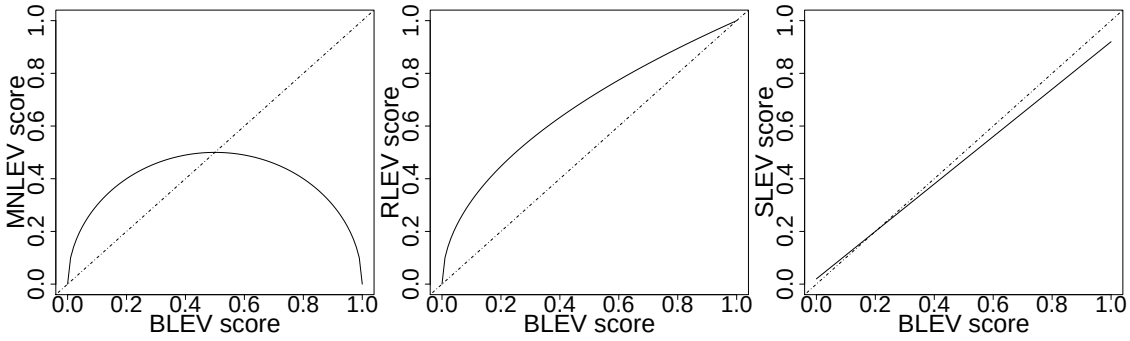
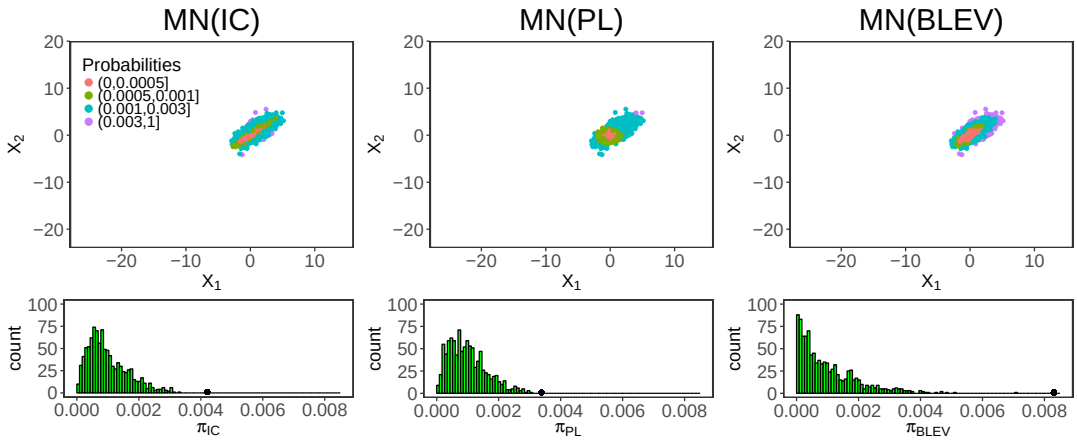


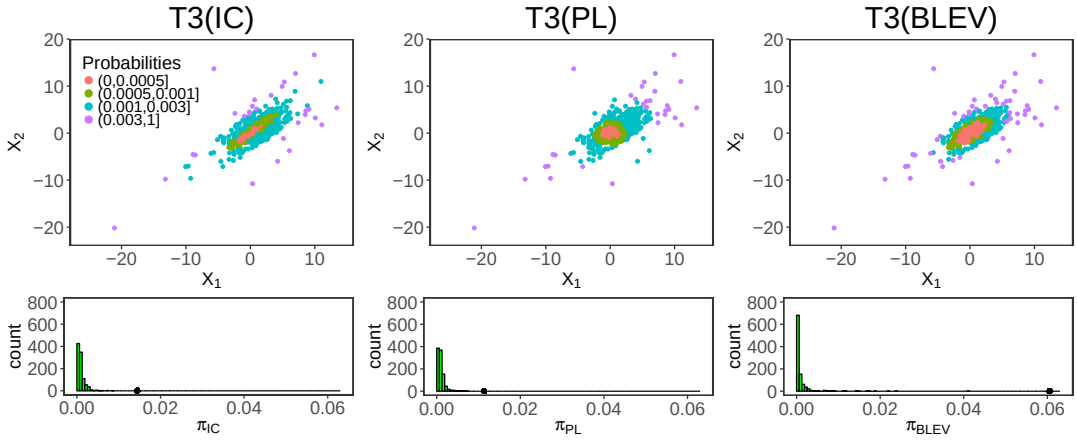
Figure 3.1: The relationship between the scores of subsampling methods. Left panel: MNLEV score $\sqrt{(1-h_{ii})h_{ii}}$ vs BLEV score h_{ii}. Middle panel: RLEV score $\sqrt{h_{ii}}$ vs BLEV score h_{ii}. Right panel: SLEV score $0.9h_{ii} + 0.1p/n$, where $p/n = 0.2$, vs BLEV score h_{ii}.

The advantage of such "shrinking" is two-fold. On the one hand, the data with high leverage scores could be "outliers". Shrinking the subsampling probabilities of high leverage data points helps reduce the risk of selecting outliers into subsamples. On the other hand, amplifying the subsampling probabilities of low leverage data points reduces the variance of the resulting subsampling estimator.

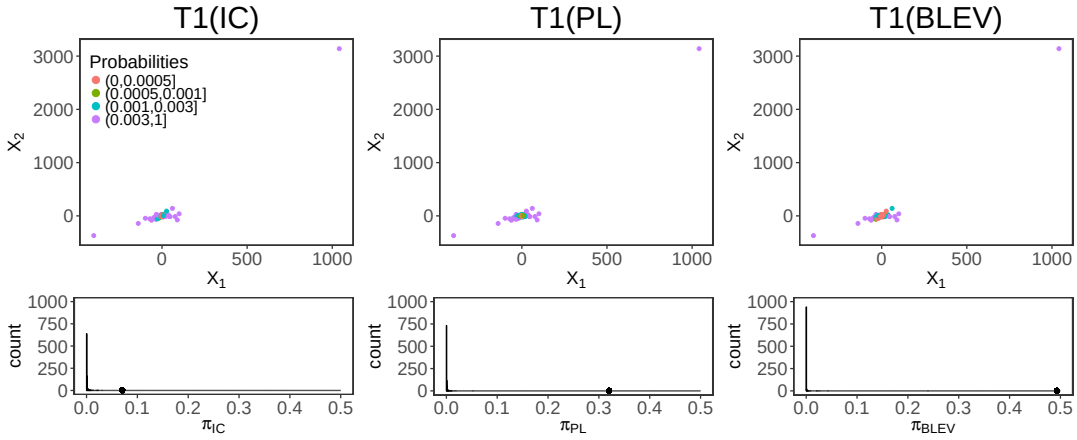
The role of h_{ii}s. If h_{ii}s are homogeneous, as in the case when $\mathbf{x}_i$s are Gaussian distributed, the subsampling probabilities of ICNLEV estimator ($\frac{\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\sqrt{1-h_{ii}}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) and those of IC estimator ($\frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^{n}\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$) will be similar to each other.



(a) The scatterplots of data points generated from a bivariate normal distribution with colors coding the subsampling probabilities in the IC (the left panel), PL (the middle panel), and BLEV (the right panel). Below each scatterplot is the histogram of corresponding subsampling probabilities with the dot representing the maximum value.



(b) Same as in (a), except that the data points are generated from a two-dimensional noncentral t distribution with three degrees of freedom.



(c) Same as in (a), except that the data points are generated from a two-dimensional noncentral t distribution with one degree of freedom.

Figure 3.2: The scatter plots of 1000 data points generated from three distributions in Example 2 in Section 3.2.

Moreover, since $\sum_{i=1}^{n} h_{ii} = p$, given a fixed value of p, one expects that h_{ii}s will become smaller as sample size n increases and that the subsampling probabilities of ICNLEV estimator and those of IC estimator will also be similar to each other. Similarly, when $h_{ii} = o(1)$ for all $i = 1, \ldots, n$, i.e., when h_{ii}s are extremely small compared to 1, the subsampling probabilities of ICNLEV estimator and those of IC estimator will again also be similar to each other. Analogous arguments also apply to the subsampling probabilities of PLNLEV and PL, and RLEV and MNLEV.

Two Examples.

In the following, we use two examples to further illustrate the relationship between various subsampling probabilities.

Example 1. Consider linear regression model with orthogonal predictor matrix, i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. In this case, we have $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \|\mathbf{x}_i\|^2$. Further, ICNLEV score, MNLEV score and PLNLEV score all equal to $\sqrt{(1-h_{ii})h_{ii}}$. Analogously, IC score coincides with RLEV score and PL score, and they all equal to $\|\mathbf{x}_i\|$.

Example 2. In practice, orthogonal predictor matrix in Example 1 is rare. Thus we further consider a toy example of linear regression model with two correlated predictors. We generate 1000 data points from a multivariate normal distribution, a multivariate noncentral t distribution with three degrees of freedom, and a multivariate noncentral t distribution with one degree of freedom, respectively.

We plot these data points in the top row of Figure 3.2. In each scatterplot, the color of the point indicates the magnitude of subsampling probabilities in IC, PL and BLEV methods from left to right. Below each scatterplot, we also show the histogram of the corresponding subsampling probabilities. Examination of Figure 3.2 reveals a pattern shared by all sampling distributions, i.e., the subsampling probabilities of data points in the center are smaller than those of data points in the boundary. Also, it is worth noting that compared to $\pi_i^{PL} \propto \|\mathbf{x}_i\|$, both $\pi_i^{IC} \propto \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|$ and $\pi_i^{BLEV} \propto \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ depends on $(\mathbf{X}^T\mathbf{X})^{-1}$,

which normalize the scale of predictors. Thus we notice that high probability points, i.e., the purple points, in PL scatter around the upper right and lower left corner while the high probability points in IC and BLEV form a contour on the outer. The difference is caused by the normalization using $(\mathbf{X}^T\mathbf{X})^{-1}$. The histograms in each row also show the critical difference between the subsampling probabilities of BLEV and those of IC and PL, i.e., the subsampling probability distribution of BLEV is more dispersed than others'. In other words, there are a significant number of data points with a large number of extremely small probabilities and a ver small number of extremely large probabilities in BLEV.

3.3 Empirical studies

3.3.1 Simulation studies

In this section, we following the setup in section 2.2.1 and assess the performance of the proposed subsampling methods.

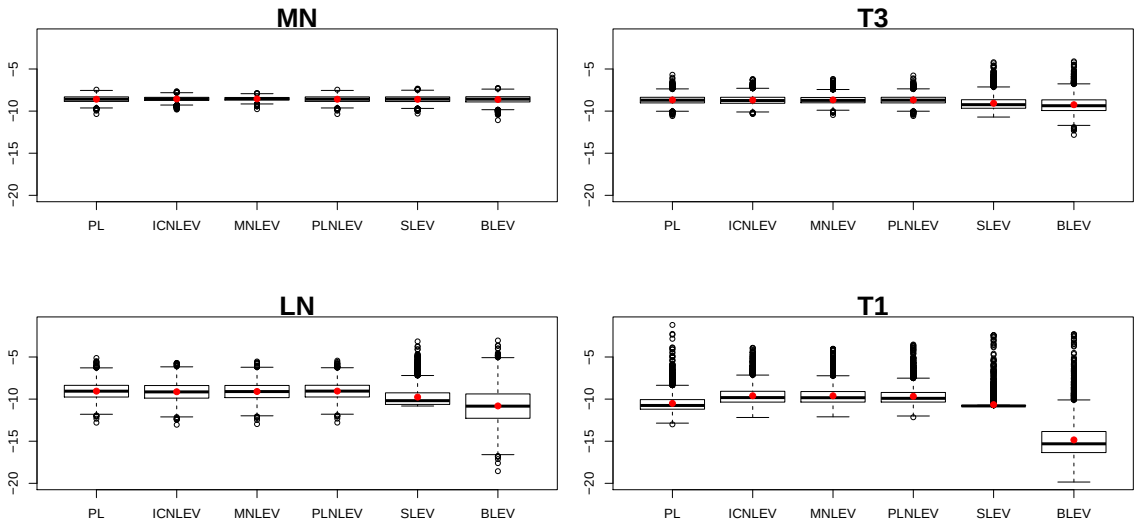


Figure 3.3: Box plots of the subsampling probabilities (in log scale) of all data points in PL, ICNLEV, PLNLEV, MNLEV, SLEV, and BLEV (from left to right in each panel) in MN, T3, LN, and T1 data for $p = 10$ and $n = 5000$. In each box plot, the red dot represents the mean of the corresponding subsampling probabilities (in log scale).

In Figure 3.3, we present box plots of the subsampling probabilities (in log scale) of all the data points in methods proposed in this chapter in MN, T3, LN, and T1 (from left to right). In particular, we incorporate the subsampling probabilities in PL, SLEV, and BLEV for comparison. We still observe that the distribution of subsampling probabilities (in log scale) in all estimators are somewhat right-skewed, i.e., there exist some large probabilities in the subsampling probability distributions. The subsampling probability distributions of PL, ICNLEV, PLNLEV, and MNLEV are similar to each other. Analogous to the case shown in Figure 2.1, the subsampling probability distributions of BLEV and SLEV are more dispersive than those of other estimators in T3, LN and T1, and there exist a significant amount of extremely small sampling probabilities. The medians of sampling probabilities in SLEV are still lower than the first quantiles of sampling probabilities in PL, ICNLEV, PLNLEV and MNLEV in T3, LN and T1.

Next, we evaluate the performance of the subsampling estimators for approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$. Following the simulation setting in Section 2.2.1, we generate one MN data set, one T3 data set, one LN data set and one T1 data set. For each data set, the full sample OLS estimate is calculated. We set subsample sizes $r = 100, 200, 500, 700, 1000$ and repeatedly apply ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV methods 100 times at each subsample size to get subsampling estimates $\tilde{\boldsymbol{\beta}}_b$, where $b = 1, \ldots, 100$. Using these estimates, we calculate the squared bias and the variance of each method for approximating $\hat{\boldsymbol{\beta}}_{OLS}$.

In Figure 3.4, we plot squared biases and variances (in log scale) of ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ at different subsample sizes in all data sets. Several observations are worth noting in Figure 3.4. First, the squared biases are negligible compared to the corresponding variances. For all subsampling methods, both the squared biases and the variances generally decrease as subsample size increases. These observations are in agreement with Theorem 2, which states that subsampling estimators are asymptotical unbiased estimators of $\hat{\boldsymbol{\beta}}_{OLS}$ provided that regularity conditions are satisfied.

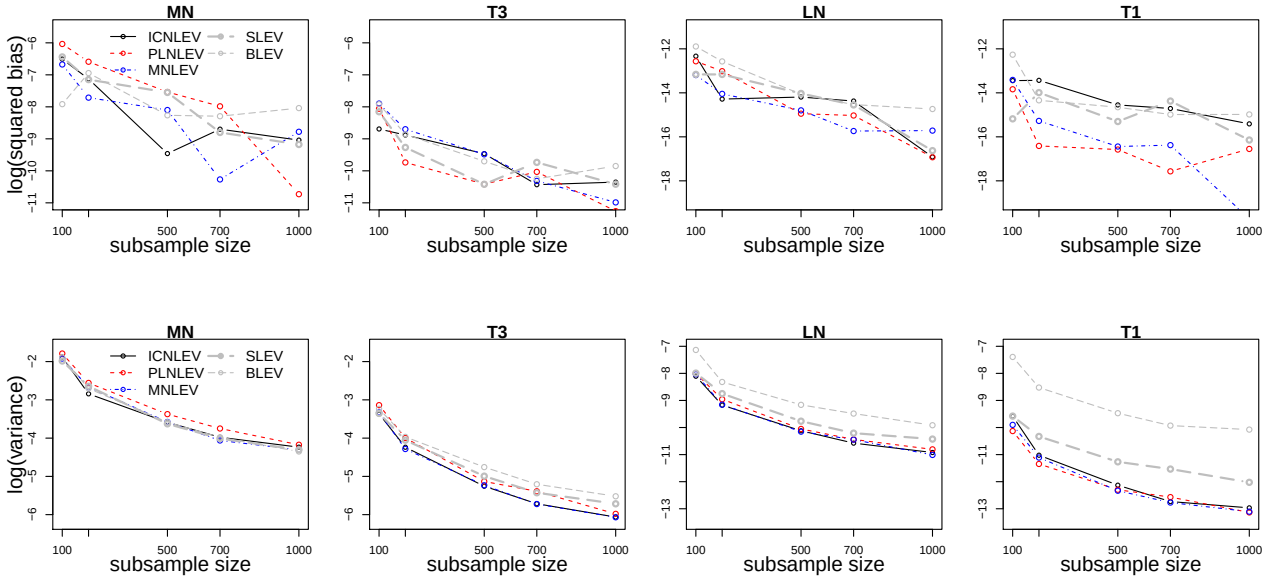


Figure 3.4: The squared biases (the first row) and the variances (the second row) of ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (in log scale) at different subsample sizes.

Second, the variances of estimates using ICNLEV, whose subsampling probabilities minimize $EAMSE(\tilde{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}}_{OLS})$, are slightly smaller than the variances of estimates using other methods in T3 and LN at most subsample sizes. The variances of estimates using ICNLEV, MNLEV and PLNLEV are consistently smaller than those of SLEV and BLEV in LN and T1. Third, all subsampling estimators perform better in LN than in T3 and MN, i.e., the squared biases and the variances of all estimates in LN and T1 are smaller compared to those in T3 and MN. Fourth, in T1, technically the definition of EAMSE is not proper, since the expectation of predictors $\mathbf{x}_i$s does not exist, but all the proposed estimators still outperform the BLEV and SLEV.

To examine the performance of proposed subsampling estimators for approximating $\hat{\mathbf{Y}}_{OLS}(= \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, we plot the variances (in log scale) of $\mathbf{X}\tilde{\boldsymbol{\beta}}_b$ at different subsample sizes for all subsampling estimators in Figure 3.5. The variances of estimates using MNLEV, whose subsampling probabilities minimize $EAMSE(\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$ are slightly smaller than those

of estimates using other methods in T3 and LN at most subsample sizes. But in T1, the abnormal behavior of predictors interferes with the performance of MNLEV and renders it not optimal.

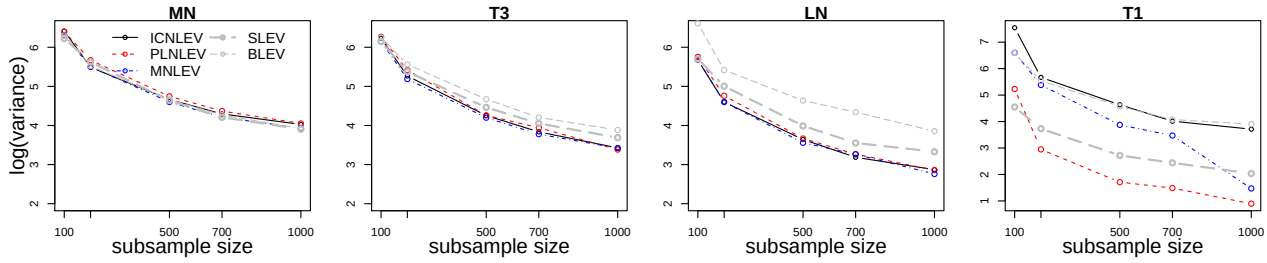


Figure 3.5: The variances of ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\hat{\mathbf{Y}}_{OLS}(= \mathbf{X}\tilde{\boldsymbol{\beta}}_{OLS})$ (in log scale) at different subsample sizes.

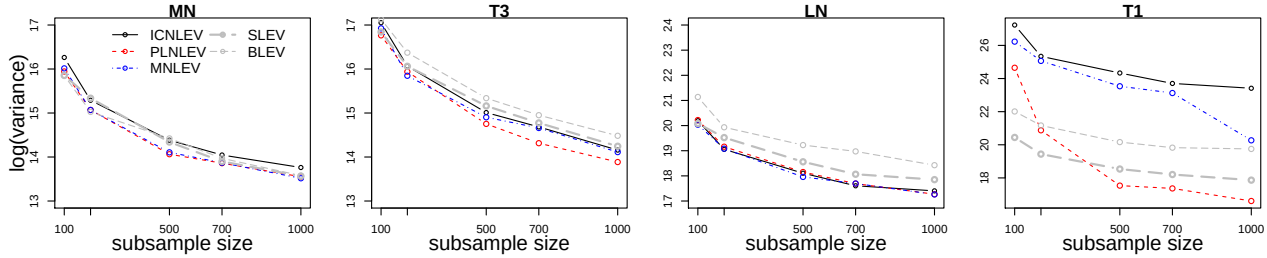


Figure 3.6: The variances of ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates in approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (in log scale) at different subsample sizes.

To assess the performance of various subsampling estimators for approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. We plot the variances of $\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}_b$ for all subsampling estimators in Figure 3.6. For all estimators, the variances decrease as subsample size increases in all data sets. In T3, the variances of estimates using PLNLEV, whose subsampling probabilities minimize $EAMSE(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}; \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$, are smaller than the variances of estimates using other methods at most subsample sizes. In this case, despite the violation of conditions for the proper definition of EMASE in T1, the variance of PLNLEV estimates are still the smallest when the subsample size is greater than 200.

At the end of this section, we report the empirical computational costs of the subsampling

	$p = 100$			$p = 200$		
subsample size	$20p$	$50p$	$100p$	$20p$	$50p$	$100p$
IC	41.6	41.7	41.8	93.9	93.9	95.1
RLEV	50.4	50.4	50.5	72.8	72.8	74.4
PL	9.2	9.3	9.4	12.3	12.3	13.9
ICNLEV	62.9	63.0	63.0	121.7	121.7	123.3
MNLEV	50.4	50.4	50.5	72.8	72.8	74.4
PLNLEV	52.0	52.1	52.1	82.7	82.7	84.3
SLEV	50.4	50.4	50.5	72.8	72.8	74.4
BLEV	50.4	50.4	50.5	72.8	72.8	74.4
Full Sample OLS	111.1			400.1		

Table 3.2: The computing time of ICNLEV, IC, PLNLEV, PL, and BLEV in CPU seconds in T3 using $n = 1,000,000$ and $p = 100$, 200.

estimators. We recorded the computing time in CPU seconds using T3 data with $n = 1,000,000$, $p = 100$ and 200. All computations were performed by R 3.2.2 on a Linux operating system on a computer server, which has 256GB of RAM and 8TB of storage space and two (each with ten cores) Intel Xeon E5-2670 v2 2.50GHz. The recorded CPU seconds are summarized in Table 3.2. The computing time of full sample OLS is included as a benchmark. We note that the CPU seconds for calculating all subsampling estimators are less than that of full sample OLS. The advantage of subsampling estimators gets more obvious as the number of predictors p increases. As subsample size increases, the change in computing time is minimal. We note that ICNLEV shows significantly longer computing time than IC, PLNLEV, SLEV, and BLEV since it requires calculating both $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and h_{ii}, however, others require only one of these two. The table also shows the significant computational advantage of PL over all other subsampling estimators. PL has the shortest computing time in each case and scales better than other estimators as p increases.

3.3.2 Real data analysis

In this part, we illustrate the performance of the proposed subsampling methods using two real-world examples in the previous chapter.

Airline delay dataset revisited

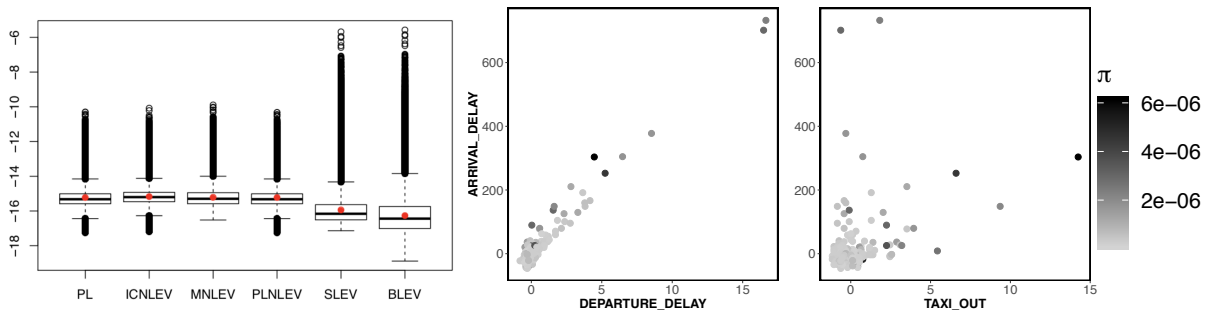


Figure 3.7: Exploratory analysis for the Airline delay data. The left panel is the box plots of subsampling probabilities (in log scale) of all data points in PL, ICNLEV, MNLEV, PLNLEV, and BLEV. The middle panel and right panel are the scatter plots of the 200 sampled response values and two predictors using the subsampling probability distribution in ICNLEV.

In the left panel of Figure 3.7, we plot the box plots of subsampling probabilities (in log scale) of all data points in PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV. We note that the subsampling probability distributions are right-skewed, similar to those in Figure 3.3 in simulation study. Using the subsampling probability distribution in ICNLEV, we took a subsample of size 200 from the full data. The middle and right panels in Figure 3.7 are the scatter plots of the sampled response and the first two predictors respectively. These scatter plots provide a visual sketch of the full sample data.

Next, we repeatedly apply PL, ICNLEV, IC, PLNLEV, SLEV, and BLEV to this data set for 100 times at subsample size $r = 20p, 50p, 70p, 100p, 200p$, where $p = 14$. Then we calculated the squared bias and the variance of resulting estimates in approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\boldsymbol{Y}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ for each method. The results are summarized in Figure 3.8. We notice

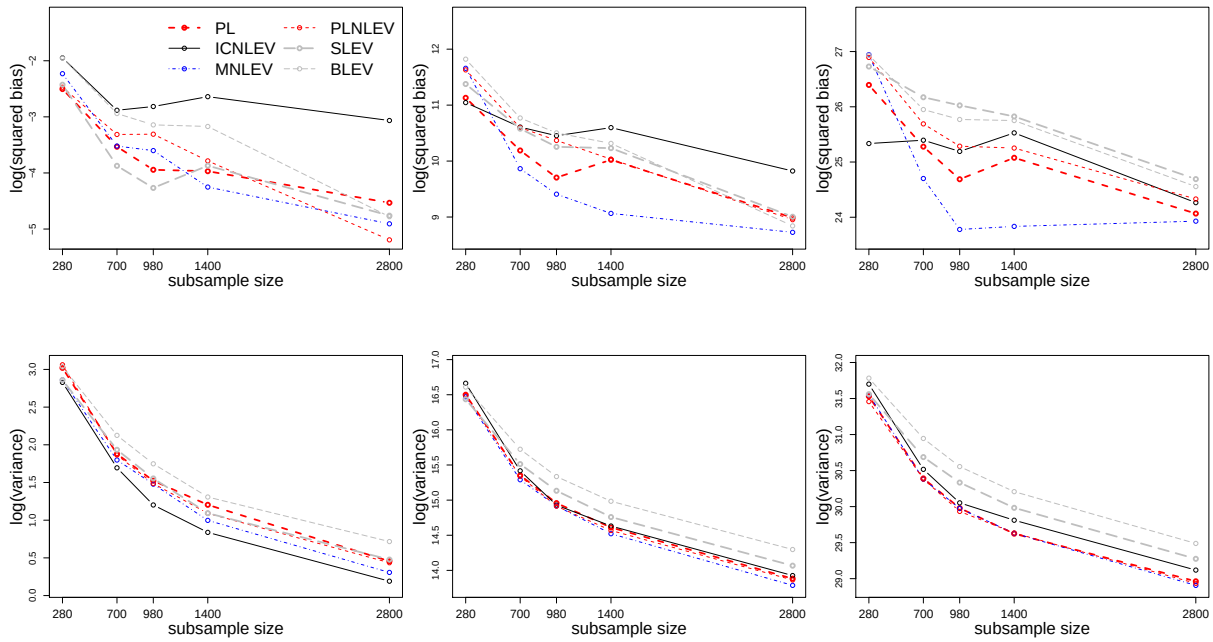


Figure 3.8: The squared biases (the first row) and variances (the second row) of PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (the first column), $\hat{\boldsymbol{Y}}_{OLS}$ (the second column) and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ (the third column) (in log scale) at different subsample sizes for Airline Delay data.

that the squared biases of all methods are all much smaller compared to the corresponding variances for all methods at all subsample sizes. For approximating $\hat{\boldsymbol{\beta}}_{OLS}$, the estimates using ICNLEV and MNLEV show consistently smaller variances than estimates using other methods when subsample size is greater than 500. For approximating $\hat{\boldsymbol{Y}}_{OLS}$, the variances of estimates using PLNLEV are among the smallest. PL, ICNLEV, MNLEV, and PLNLEV all consistently outperform SLEV and BLEV in terms of variance at subsample sizes greater than 700. For approximating $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, the estimates using PLNLEV, and its approximation PL, show competitive performance as subsample sizes gradually increases.

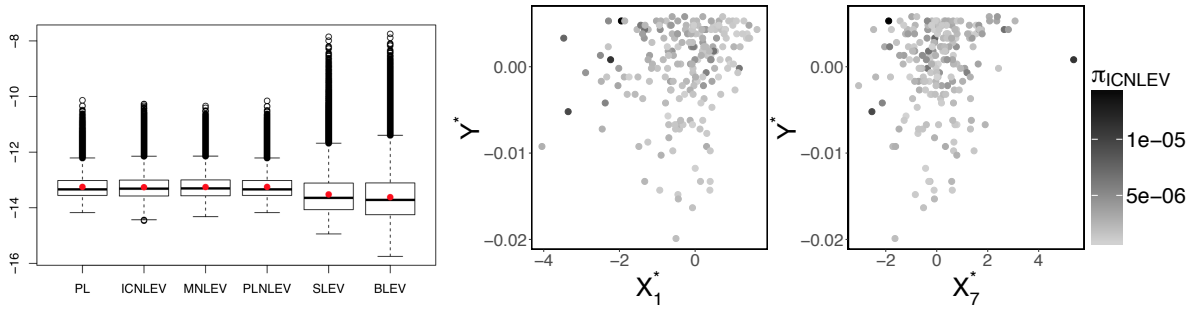


Figure 3.9: Exploratory analysis for "YearPredictionMSD" data. The left panel is the box plots of subsampling probabilities of all data points (in log scale) for PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV. A subsample of size 200 is taken from the full data using the subsampling probabilities of ICNLEV. The middle panel and right panel are the scatter plots of sampled response versus two timbre feature predictors.

"YearPredictionMSD" dataset revisited

In the left panel of Figure 3.9, we present the box plots of subsampling probability distributions in PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV in "YearPredictionMSD" data set. Inspecting the box plots reveals that all sampling distributions are right-skewed. Using the subsampling probability distribution in ICNLEV, we took a subsample of size 200 from the full data. The middle and right panels of Figure 3.9 are the scatter plots of the sampled response and two timbre features respectively. These scatter plots provide a visual sketch of the full sample data.

We repeatedly apply ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV methods to the data set for 100 times at subsample sizes $r = 10p, 20p, 50p, 70p, 100p$, where $p = 90$. In Figure 3.10, we plot the squared biases and the variances of the estimates for all weighted subsample methods in approximating $\hat{\boldsymbol{\beta}}_{OLS}$, $\hat{\boldsymbol{Y}}_{OLS}$, and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ in log scale. For all three scenarios, the squared biases are much smaller than the corresponding variances for all methods at all subsample sizes. For approximating $\hat{\boldsymbol{\beta}}_{OLS}$, the variances of ICNLEV, MNLEV, and PLNLEV estimates are comparable to each other and consistently smaller than those of SLEV and BLEV estimates at all subsample sizes. For approximating $\hat{\boldsymbol{Y}}_{OLS}$

and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$, the variances of PLNLEV and PL estimates are consistently smaller than the variances of estimates using other methods when subsample size is greater than 1800.

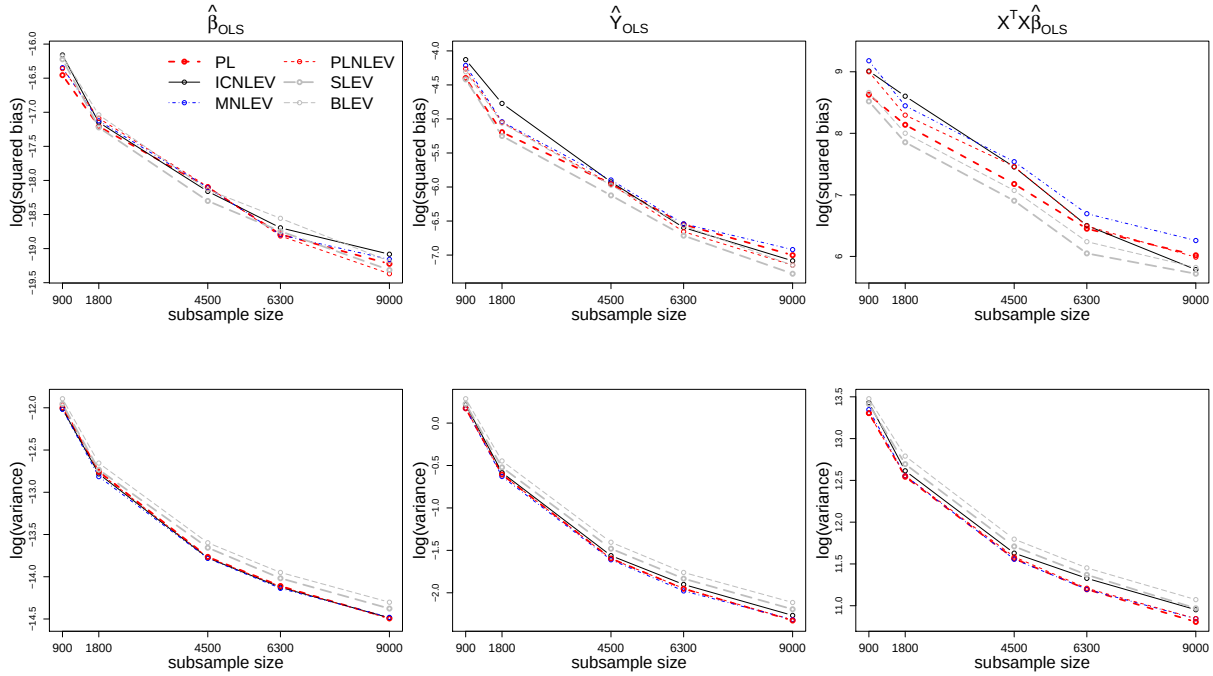


Figure 3.10: The squared biases (the first row) and variances (the second row) of PL, ICNLEV, MNLEV, PLNLEV, SLEV, and BLEV estimates for approximating $\hat{\boldsymbol{\beta}}_{OLS}$ (the first column), $\hat{\boldsymbol{Y}}_{OLS}$ (the second column), and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$(the third column) (in log scale) at different subsample sizes for "YearPredictionMSD" data.

3.4 Summary and discussion

In this chapter, we study the performance of subsampling estimators in conditional inference, in which the data is considered given, and the goal is to approximate the full sample estimate. We establish the asymptotic normality of subsampling estimators for the linear model under general regularity conditions. It is shown that subsampling estimators are asymptotically unbiased and consistent estimators of the ordinary least squares (OLS) estimate when the data is given and fixed. We obtain an explicit form of asymptotic variance and EAMSE of subsampling estimators in this case (Theorem 2). We then seek to construct subsampling

estimators with subsampling probability distributions aimed at minimizing the EAMSE. This yields several interesting estimators. For example, we notice that the estimator for minimizing the EAMSE for approximating $\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ and $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ both involve square root of negative leverage scores ($\sqrt{1-h_{ii}}$), which are in significant contrast to basic leveraging estimator whose sampling probability is proportional to leverage scores. In addition, we also conduct a comprehensive empirical evaluation of the performance of subsampling estimators for conditional inference, confirming the asymptotic unbiasedness, consistency, and optimality.

Acknowledgement

We would like to thank Bin Yu for many helpful discussions. We also thank Shusen Wang for reading and providing constructive comments on an earlier version of this work.

Chapter 4

Bayesian Spline Smoothing with Ambiguous Penalties

Chapter summary: A popular method for flexible function estimation in nonparametric models is the smoothing spline method. Much like any other statistical methods, when applying this method, one needs to specify a set of model assumptions, penalty functional in the case of smoothing spline method, which puts a soft constraint on the function to be estimated, e.g., the $J(\eta) = \int_0^1 (\eta'')^2 \, \mathrm{d}x$ in the cubic spline, where η is the function to be estimated. A reasonable choice of penalty functional, i.e., an appropriate set of assumptions on the function to be estimated, is the key to the success of the smoothing spline method discussed in this chapter. In practice, for many dynamic systems, there naturally exist multiple sets of widely accepted assumptions. Thus, we have multiple applicable penalties. We refer to this problem as the problem of ambiguous penalties. Neglecting the model uncertainty in penalties and proceeding to model with any one of the candidates may produce misleading results of statistical inference. In this chapter, we take a Bayesian perspective and propose a fully Bayesian approach that takes into consideration all the penalties as well as the ambiguity in choosing them in our inference. The outline of this chapter is as follows. We describe the proposed Bayesian model and discuss the priors used in Section 4.2. In Section 4.3, we provide the details of the algorithm for model estimation. Simulation and real data analysis follow in Section 4.4. A few remarks in Section 4.5 conclude the chapter.

4.1 Introduction

In smoothing spline models, as outlined in Section 1.2, the unknown function η in a regression problem $Y_i = \eta(x_i) + \varepsilon_i, \quad x_i \in \mathscr{X}, \quad i = 1, \ldots, n$, where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, is estimated through the following penalized least squares

$$\min_{\eta} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \lambda J(\eta),$$

where the first term measures the goodness-of-fit of the model to the data, the second term $J(\eta)$ is a penalty functional on η, and λ is a finite nonnegative tuning parameter.

The penalized least squares method for function estimation has a convenient Bayesian interpretation (Wahba, 1978, 1990; Gu, 1992; Berry et al., 2002). The term $J(\eta)$ in penalized least squares can be viewed as prior information about η in a Bayesian setup. It can be shown that choosing the commonly used quadratic penalty, e.g., $J(\eta) = \int_0^1 (\eta'')^2 \mathrm{d}x$ in cubic spline, is equivalent to setting a partially improper Gaussian prior for the η (Wahba, 1978, 1990). As a consequence, the smoothing spline estimator through penalized least squares with a quadratic roughness penalty is equivalent to the posterior mean of the Gaussian posterior of η. This results will be formally stated in Section 4.2.

Following this line of thinking, the ambiguity in choosing $J(\eta)$ can be incorporated as ambiguity in choosing the prior for η. In particular, we propose a mixture distribution as prior for η. Although straightforward in principle, setup and practical implementations of the full hierarchical model require attention to details. Prior information for the variance parameters must be interpretable and computationally manageable. We discuss the details in the next section.

4.2 Accounting ambiguity of J using mixture distribution as a prior for η

We first present necessary notations for the Bayesian interpretation of smoothing spline estimates through penalized least squares for one given penalty J.

For the smoothing spline model, the penalty functional J defines a seminorm in $\mathcal{H}_J = \{\eta : J(\eta) < \infty\}$ and the seminorm induces a semi-inner-product $J(\eta) = J(\eta, \eta)$ on $\mathcal{H}_J$. The null space of J, defined as $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$, is a finite dimensional linear subspace of $\mathcal{H}_J$. We denote the dimension of $\mathcal{N}_J$ as m and let $\{\phi_i\}_{i=1}^m$ be a basis of $\mathcal{N}_J$. Given the nondegenerate $\mathcal{N}_J$, the semi-inner-product and seminorm can be made a full inner product and full norm in $\mathcal{H}_J$ via augmentation by adding extra terms that define appropriate inner product and norm for $\mathcal{N}_J$. Provided that the evaluation functional in $\mathcal{H}_J$ is continuous, then the space $\mathcal{H}_J$ is a reproducing kernel Hilbert space (RKHS). We denote the reproducing kernel associated with $\mathcal{H}_J$ as $R(\cdot, \cdot)$. The form of $R(\cdot, \cdot)$ depends on the form of J. For the penalty that results in cubic spline estimator, $J(\eta) = \int_0^1 (\eta'')^2 \mathrm{d}x$, its null space $\mathcal{N}_J$ is composed of polynomial functions up to order 1. If we use the $(h, g) = \sum_{j=0}^1 (\int_0^1 f^{(j)} \mathrm{d}x)(\int_0^1 g^{(j)} \mathrm{d}x)$ as the inner product in $\mathcal{N}_J$, then the full inner product on $\mathcal{H}_J$ is $(h, g) = \sum_{j=0}^1 (\int_0^1 f^{(j)} \mathrm{d}x)(\int_0^1 g^{(j)} \mathrm{d}x) + \int_0^1 f'' g'' \mathrm{d}x$. The reproducing kernel $R(\cdot, \cdot)$ corresponding to this inner product can be analytically written out using the scaled Bernoulli polynomials (Gu, 2013, Section 2.3.3). Using the definitions above, we present the following lemma, which is adapted from Theorem 1.5.3 in Wahba (1990).

Lemma 1. *Consider model* $Y_i = \eta(x_i) + \varepsilon_i$, *where* $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, *let*

$$\begin{aligned} \eta(x) &= \sum_{j=1}^m d_j \phi_j(x) + b^{1/2} Z(x), \\ d_j &\sim N(0, \tau^2), j = 1, \ldots, m, \end{aligned} \tag{4.1}$$

where b and τ^2 are constants, $\phi_j(\cdot)s$ are the basis functions of $\mathcal{N}_J$, and $Z(x)$ is a zero-mean Gaussian stochastic process with covariance $R(\cdot,\cdot)$, the reproducing kernel of $\mathcal{H}_J$. Denote $\hat{\eta}_{\tau^2}(x) = E(\eta(x)|Y_i = y_i, i = 1,\ldots,n)$. For each fixed x, we have

$$\lim_{\tau^2\to\infty} \hat{\eta}_{\tau^2}(x) \;=\; \eta_\lambda(x), \tag{4.2}$$

where $\eta_\lambda(x)$ is the smoothing spline estimates in (1.5) through penalized least squares with $\lambda = \sigma^2/b$.

Lemma 1 establishes the equivalence between Bayesian estimates using Gaussian process as prior and smoothing splines estimates through penalized least squares. It is worth mentioning that letting $\tau^2 \to \infty$ in (4.2) is equivalent to setting an improper prior for d_js in (4.1), i.e., $f(d_j) \propto 1_{d_j\in\mathbb{R}}$.

Next, we adapt the prior settings in Lemma 1 to the case with two candidate penalties, say J_1 and J_0. The generalization to multiple candidate penalties is immediate. We suppose that with probability $0 < \nu < 1$, the prior of η has the same distribution as a Gaussian process $W_{(1)}(x)$,

$$W_{(1)}(x) = \sum_{j=1}^{p} d_j\phi_j(x) + \sum_{j=1}^{p_1} d_{1j}\phi_{1j}(x) + b_1^{1/2}Z_{(1)}(x); \tag{4.3}$$

and with probability $1-\nu$, the prior of η has the same distribution as a Gaussian process $W_{(0)}(x)$,

$$W_{(0)}(x) = \sum_{j=1}^{p} d_j\phi_j(x) + \sum_{j=1}^{p_0} d_{0j}\phi_{0j}(x) + b_0^{1/2}Z_{(0)}(x). \tag{4.4}$$

In (4.3) and (4.4), $\{\{\phi_j(x)\}_{j=1}^{p}, \{\phi_{1j}(x)\}_{j=1}^{p_1}\}$ is the basis for the null space of J_1, and $\{\{\phi_j(x)\}_{j=1}^{p}, \{\phi_{0j}(x)\}_{j=1}^{p_0}\}$ is the basis for the null space of J_0. In (4.3) and (4.4), $Z_{(1)}(x)$ and $Z_{(0)}(x)$ are zero mean Gaussian stochastic processes with covariance functions $R_{(1)}(s,t)$ and $R_{(0)}(s,t)$, which are the reproducing kernels of RKHSs defined using J_1 and J_0, respectively. The probability ν represents our belief in choosing between the two stochastic processes as

the prior for η. We now complete the setup by specifying the prior for $\boldsymbol{d}_1 = (d_{11}, \ldots, d_{1p_1})^T$, $\boldsymbol{d}_0 = (d_{01}, \ldots, d_{0p_0})^T$, $\boldsymbol{d} = (d_1, \ldots, d_p)^T$, σ^2, b_1, b_0, and ν as

$$\boldsymbol{d}_u | g_u, \sigma^2 \sim \boldsymbol{N}(\boldsymbol{0}, g_u \sigma^2 (\boldsymbol{S}_u^T \boldsymbol{S}_u)^{-1}), \quad u = 0, 1, \tag{4.5}$$

$$f(g_u) \propto g_u^{-\alpha_u - 1} \exp\left(-\frac{\beta_u}{g_u}\right), \quad u = 0, 1, \tag{4.6}$$

$$f(\boldsymbol{d}, \sigma^2, b_1, b_0, \nu) \propto \frac{1_{\boldsymbol{d} \in \mathscr{R}^p} 1_{0<\nu<1}}{(\sigma^2)^s (b_0 + b_1)^a}, \tag{4.7}$$

where $\alpha_1 > 0$, $\alpha_0 > 0$, $\beta_1 > 0$, $\beta_0 > 0$, s, and a are given hyperparameters, $\boldsymbol{S}_0$ is the $n \times p_0$ matrix with (i, j)-th entry being $\phi_{0j}(t_i)$, and $\boldsymbol{S}_1$ is the $n \times p_1$ matrix with (i, j)-th entry being $\phi_{1j}(t_i)$. Generally, p_1, p_0 and p are small integers. For example, if we use the cubic spline penalty as J_1(or J_0), then $p_1 + p = 2$ (or $p_0 + p = 2$). Considering the ambiguity in choosing among penalties, there might be significant overlap between the null space of J_1 and J_0. Thus it is often possible that $p_1 = 0$, or $p_0 = 0$. In that case, the parameters $\boldsymbol{d}_1$, or $\boldsymbol{d}_0$ is not defined, thus we do not need to further introduce priors as we did in (4.5) and (4.6).

The parameters in $\boldsymbol{d}$ are common parameters associated with both penalties. We choose a flat prior in (4.7), and this is equivalent to the choice of d_js in Lemma 1 after letting $\tau^2 \to \infty$ in (4.2). Both $\boldsymbol{d}_1$ and $\boldsymbol{d}_0$ are associated with only one of the candidate penalties. Independently giving $\boldsymbol{d}_1$ and $\boldsymbol{d}_0$ flat priors will render the posterior density function not integrable. To circumvent this problem, in (4.5) we set the g priors, which are commonly used priors for regression coefficients in Bayesian model selection (Tiao and Zellner, 1964a,b; Zellner, 1986; Liang et al., 2008; Maruyama and George, 2011). In (4.6), we specify the priors for g_0 and g_1, the scale parameters of the variance-covariance matrix of $\boldsymbol{d}_0$ and $\boldsymbol{d}_1$, using the inverse gamma distribution. It is recommended to set the hyperparameters α_0, β_0, α_1, and β_1 such that the mean for priors of g_1 and g_0, if exist, are greater than 1. This implies that the scale of variances for priors of $\boldsymbol{d}_1$ and $\boldsymbol{d}_0$ is larger compared to σ^2. In particular, as default, we recommend setting $\alpha_1 = 1/2$, $\beta_1 = 1/2$. In this case, the

mean for the inverse gamma prior for g_1 does not exist and the mode is 1/3. Furthermore, using priors in (4.5) and (4.6) is equivalent to specifying a Cauchy prior for $\boldsymbol{d}_1$. Similar argument applies to $\boldsymbol{d}_0$. Through introducing g_1 and g_0, we mimic flat priors for $\boldsymbol{d}_1$ and $\boldsymbol{d}_0$ using proper but reasonably flat distributions, and this choice can be seen as a compromise between noninformative distribution and a stringent multivariate normal distribution.

Similar to the case of $\boldsymbol{d}_1$ and $\boldsymbol{d}_0$, the prior distribution for b_1 and b_0 is also of key importance to the integrability of posterior density function. In this case, we specify a partially proper distribution for b_1 and b_0. By a partially proper distribution function, we mean that the function is only integrable with respect to b_1 given b_0, and vice versa. This prior distribution allows either b_1 or b_0 to take values close to 0.

The prior of σ^2 is based on the kernel of an inverse gamma distribution, which includes the Jeffery's prior for model variance parameter as a special case. For the choice of prior for the parameter ν, we opt to use Uniform(0,1) giving no preference on the two penalties. This can be modified if strong subjective preference information is available. For example, we can modify so that ν follows a Beta distribution with known parameters, the sufficient conditions provided in the following Theorem 3 will remain intact.

For hyper-parameters a and s involved in prior distributions of the variance parameters b_1, b_0 and σ^2 in (4.7), we discuss proper choices in Theorem 3 that ensure the resulting posterior distribution from the proposed model being proper. Before that, we define a few necessary matrix notations. Denote $\boldsymbol{S}$ as the $n \times p$ matrix with (i,j)-th entry being $\phi_j(x_i)$, $\boldsymbol{Q}_1$ as the $n \times n$ matrix with (i,j)-th entry being $R_{(1)}(x_i, x_j)$, and $\boldsymbol{Q}_0$ the as $n \times n$ matrix with (i,j)-th entry being $R_{(0)}(x_i, x_j)$. Further let $\boldsymbol{S}_{(1)} = (\boldsymbol{s}_{11}, \ldots, \boldsymbol{s}_{1n})^T$ and $\boldsymbol{S}_{(0)} = (\boldsymbol{s}_{01}, \ldots, \boldsymbol{s}_{0n})^T$, where $\boldsymbol{s}_{1i} = (\phi_1(x_i), \ldots, \phi_p(x_i), \phi_{11}(x_i), \ldots, \phi_{1p_1}(x_i))^T$ and $\boldsymbol{s}_{0i} = (\phi_1(x_i), \ldots, \phi_p(x_i), \phi_{01}(x_i), \ldots, \phi_{0p_0}(x_i))^T$, $i = 1, \ldots, n$. In addition, we use $Colsp(\boldsymbol{M})$ to denote the column space of matrix $\boldsymbol{M}$.

Theorem 3. *The posterior distribution from model (1.4) and prior distributions in (4.3) to (4.7) is proper on the sample points if the following conditions are satisfied.*

Condition (I): $1 < a < 2$*;*

Condition (II): $N - p + 2a + 2s > 6$*;*

Condition (III): for $u = 0, 1$*, either*

(a) if $Colsp(\boldsymbol{S}_{(u)}) \bigoplus Colsp(\boldsymbol{Q}_u) \subset \mathscr{R}^n$*,* $2(k_u - p) - p_u + 4a > 8$*, or*

(b) if $Colsp(\boldsymbol{S}_{(u)}) \bigoplus Colsp(\boldsymbol{Q}_u) = \mathscr{R}^n$*,* $2(n - k_u) + p_u + 4s < 4$*.*

In Theorem 3, we consider the case with no repeated measurements on each sample point x_i. In addition, we have incorporated the case with replicates on each distinct x_i in Appendix B. In that case, $Y_{ij} = \eta(x_i) + \varepsilon_{ij}, \quad x_i \in \mathscr{X}, \quad i = 1, \ldots, n, j = 1, \ldots, n_i$. Denote $SSE = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$. If $SSE > 0$, then Condition (III) can be relaxed as simply $\min\{2k_1 - p_1, 2k_0 - p_0\} - 2p + 4a > 8$.

4.3 Algorithm for computing posterior probabilities

Due to the complex prior settings, the analytical form of posterior estimates is not readily available. We resort to the Markov chain Monte Carlo (MCMC) method for posterior inference. Denote $\boldsymbol{V}_1$ and $\boldsymbol{V}_0$ as $n \times k_1$ and $n \times k_0$ full column rank matrices such that $\boldsymbol{Q}_1 = \boldsymbol{V}_1 \boldsymbol{V}_1^T$ and $\boldsymbol{Q}_0 = \boldsymbol{V}_0 \boldsymbol{V}_0^T$. Note that given ν, $\boldsymbol{d}$, $\boldsymbol{d}_0$, $\boldsymbol{d}_1$, b_0 and b_1, the prior distribution of $\eta(\cdot)$ evaluated at n sample points is a mixture multivariate Gaussian distribution, i.e., $\nu \boldsymbol{N}(\boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_0\boldsymbol{d}_0, b_0\boldsymbol{Q}_0) + (1-\nu)\boldsymbol{N}(\boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_1\boldsymbol{d}_1, b_1\boldsymbol{Q}_1)$. We reparameterize and write the prior distribution of $\eta(\cdot)$ evaluated at n sample points as the distribution of $(\boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_1\boldsymbol{d}_1 + \boldsymbol{V}_1\boldsymbol{z}_1)$ with probability ν, and as the distribution of $(\boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_0\boldsymbol{d}_0 + \boldsymbol{V}_0\boldsymbol{z}_0)$ with probability $1 - \nu$, where $\boldsymbol{z}_1 \sim \boldsymbol{N}(\boldsymbol{0}, b_1\boldsymbol{I})$ and $\boldsymbol{z}_0 \sim \boldsymbol{N}(\boldsymbol{0}, b_0\boldsymbol{I})$. Next, we write the full posterior after integrating out ν as

$$f(\boldsymbol{d}, \boldsymbol{d}_0, \boldsymbol{d}_1, \sigma^2, g_0, g_1, b_0, b_1, \boldsymbol{z}_1, \boldsymbol{z}_2 | \boldsymbol{Y}) \propto (f_0 + f_1)\pi \tag{4.8}$$

where

$$
\begin{aligned}
f_0 &= \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_0\boldsymbol{d}_0-\boldsymbol{V}_0\boldsymbol{z}_0)^T(\boldsymbol{Y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_0\boldsymbol{d}_0-\boldsymbol{V}_0\boldsymbol{z}_0)}{2\sigma^2}\right) \\
f_1 &= \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_1\boldsymbol{d}_1-\boldsymbol{V}_1\boldsymbol{z}_1)^T(\boldsymbol{Y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_1\boldsymbol{d}_1-\boldsymbol{V}_1\boldsymbol{z}_1)}{2\sigma^2}\right) \\
\pi &= \frac{1}{(\sigma^2)^{\frac{n}{2}+s}(b_1+b_0)^a} g_1^{-\alpha_1-1}\exp\left(-\frac{\beta_1}{g_1}\right) g_0^{-\alpha_0-1}\exp\left(-\frac{\beta_0}{g_0}\right) \\
&\quad \frac{1}{\sqrt{|g_1\sigma^2(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1}|}}\exp\left(-\frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)\boldsymbol{d}_1}{2g_1\sigma^2}\right)\frac{1}{\sqrt{|g_0\sigma^2(\boldsymbol{S}_0^T\boldsymbol{S}_0)^{-1}|}}\exp\left(-\frac{\boldsymbol{d}_0^T(\boldsymbol{S}_0^T\boldsymbol{S}_0)\boldsymbol{d}_0}{2g_0\sigma^2}\right) \\
&\quad \frac{1}{b_1^{k_1/2}}\exp\left(-\frac{\boldsymbol{z}_1^T\boldsymbol{z}_1}{2b_1}\right)\frac{1}{b_0^{k_0/2}}\exp\left(-\frac{\boldsymbol{z}_0^T\boldsymbol{z}_0}{2b_0}\right).
\end{aligned}
$$

Our choices of priors are mostly based on the kernel of conjugate priors; thus the implementation of Gibbs sampler is convenient. However, because we introduce the mixture prior for the η, the full conditional distributions of some parameters will also be mixture distributions. This situation is problematic if vanilla Gibbs sampler is applied. Take the parameters $\boldsymbol{d}$ and $\boldsymbol{d}_1$ as an example. The full conditional distribution of $\boldsymbol{d}$ is a mixture of $\boldsymbol{N}((\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{Y}-\boldsymbol{V}_1\boldsymbol{z}_1-\boldsymbol{S}_1\boldsymbol{d}_1),\sigma^2(\boldsymbol{S}^T\boldsymbol{S})^{-1})$ and $\boldsymbol{N}((\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{Y}-\boldsymbol{V}_0\boldsymbol{z}_0-\boldsymbol{S}_0\boldsymbol{d}_0),\sigma^2(\boldsymbol{S}^T\boldsymbol{S})^{-1})$. The full conditional distribution of $\boldsymbol{d}_1$ is a mixture of $\boldsymbol{N}((1/\sigma^2+1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^T(\boldsymbol{y}-\boldsymbol{V}_1\boldsymbol{z}_1-\boldsymbol{S}\boldsymbol{d})/\sigma^2,(1/\sigma^2+1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1})$ and $\boldsymbol{N}(\boldsymbol{0},g_1\sigma^2(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1})$. We observe that one of the mixture components of the full conditional distribution of $\boldsymbol{d}_1$ depends solely on the prior $\boldsymbol{N}(0,g_1\sigma^2(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1})$, which contains no data information and taking samples directly from this component lowers the efficiency of the algorithm. Another problem is that it is possible that in m-th run of Gibbs sampler, the value of $\boldsymbol{d}_1$, denoted as $\boldsymbol{d}_1^{(m)}$, is sampled according to $\boldsymbol{N}((1/\sigma^2+1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^T(\boldsymbol{Y}-\boldsymbol{V}_1\boldsymbol{z}_1-\boldsymbol{S}\boldsymbol{d})/\sigma^2,(1/\sigma^2+1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1})$ while the value of $\boldsymbol{d}$, denoted as $\boldsymbol{d}^{(m)}$ is sampled according to the $\boldsymbol{N}((\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{Y}-\boldsymbol{V}_0\boldsymbol{z}_0-\boldsymbol{S}_0\boldsymbol{d}_0),\sigma^2(\boldsymbol{S}^T\boldsymbol{S})^{-1})$. The estimate of posterior mean using this sample contains $\boldsymbol{S}\boldsymbol{d}^{(m)}+\boldsymbol{S}_1\boldsymbol{d}_1^{(m)}$, and this is not reasonable.

To circumvent the problems above, we modify the Gibbs sampler so that we update the

parameters by groups and avoid direct sampling from the component of mixture distributions that consist of only prior distributions. We introduce a new indicator variable U, which takes 0 or 1. The parameters in our model are grouped into two overlapping parts $\boldsymbol{\theta}_U = (\boldsymbol{d}, \boldsymbol{d}_U, b_U, g_U, \boldsymbol{z}_U, \sigma^2)$, $U = 0, 1$. The parameters in $\boldsymbol{\theta}_U$ are concerned in the model with only penalty J_U. The algorithm outlined below is used for effectively taking samples from the posterior distribution $p(\boldsymbol{d}, \boldsymbol{d}_1, \boldsymbol{d}_0, b_1, b_0, g_1, g_0, \boldsymbol{z}_1, \boldsymbol{z}_0, \sigma^2|\boldsymbol{Y})$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

1. Set initial values for the parameters $(\boldsymbol{d}^{(0)}, \boldsymbol{d}_1^{(0)}, \boldsymbol{d}_0^{(0)}, b_1^{(0)}, b_0^{(0)}, g_1^{(0)}, g_0^{(0)}, \boldsymbol{z}_1^{(0)}, \boldsymbol{z}_0^{(0)}, (\sigma^2)^{(0)}, U^{(0)})$.

2. For $m = 1, 2, \ldots$, iterate the following two steps.

   (a) Given $U = u$, we use Gibbs sampler to update the parameters in $\boldsymbol{\theta}_u$ according to $f_u\pi$. The full conditional distributions for $\boldsymbol{\theta}_u$ are deferred to the Appendix.

   (b) Denote the indicator variable in current state as $U = u$. We move to the state with indicator variable $U = u'$ with probability $\delta = \min\left\{1, \frac{\int f_{u'}\pi \mathrm{d}\boldsymbol{d}\mathrm{d}\boldsymbol{d}_0\mathrm{d}\boldsymbol{d}_1\mathrm{d}\boldsymbol{z}_0\mathrm{d}\boldsymbol{z}_1}{\int f_u\pi \mathrm{d}\boldsymbol{d}\mathrm{d}\boldsymbol{d}_0\mathrm{d}\boldsymbol{d}_1\mathrm{d}\boldsymbol{z}_0\mathrm{d}\boldsymbol{z}_1}\right\}$.

4.4 Empirical studies

4.4.1 Simulation studies

In this section, we present simulation studies to assess the performance of our proposed method. We generate data according to model (1.4), in which the predictors are generated uniformly from the domain of interest. In the following, we consider one-dimensional, two-dimensional and three-dimensional functions. In each case, the function $\eta(\cdot)$ is randomly set to be one of two different expressions with equal probability.

Case I: One-dimensional Data. The domain of interest is $[0, 1]$ and sample size $n = 100$. We randomly set η as one of the following two functions with equal probability:

(i) $\eta(x) = (\sin(2\pi(2x + x^2)) + 1)/2$.

(ii) $\eta(x) = (\sin(16\pi x) - 8(x - 0.5)^2 + 8(x - 0.5)^3 1_{(0.5,1]}(x) + 297/128)/(13567/4096)$.

Case II: Two-dimensional Data. The domain of interest is $[0, 6]^2$, and sample size $n = 100$. We randomly set η as one of the following two functions with equal probability:

(i) $\eta(x_1, x_2) = (\exp(-(\boldsymbol{x} - \mathbf{3}_2)^T(\boldsymbol{x} - \mathbf{3}_2)/2) - \exp(-9/2))/(1 - \exp(-9/2))$, where $\boldsymbol{x} = (x_1, x_2)^T$, and $\mathbf{3}_2 = (3, 3)^T$.

(ii) $\eta(x_1, x_2) = (\sin(2\pi x_1) + 1)/2$.

Case III: Three-dimensional Data. The domain of interest is $[0, 6]^3$, and sample size $n = 1000$. We randomly set η as one of the following two functions with equal probability:

(i) $\eta(x_1, x_2, x_3) = (\exp(-(\boldsymbol{x} - \mathbf{3}_3)^T(\boldsymbol{x} - \mathbf{3}_3)/2) - \exp(-27/2))/(1 - \exp(-27/2))$, where $\boldsymbol{x} = (x_1, x_2, x_3)^T$, and $\mathbf{3}_3 = (3, 3, 3)^T$.

(ii) $\eta(x_1, x_2, x_3) = (\sin(2\pi x_1) + 1)/2$.

The functions in Case I are generated combining polynomial and trigonometric functions. The function (i) in Case II and III are selected so that it is symmetric with respect to the arguments, and the function (ii) in Case II and III are selected so that it is not symmetric with respect to the arguments. For all functions considered, they are normalized so that the range is 1. The true functions used in the first two cases are plotted in Figure 4.1. We set the signal-to-noise ratio (SNR), defined as $\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2/(n\sigma^2)$, to three levels: 3, 5, 7.

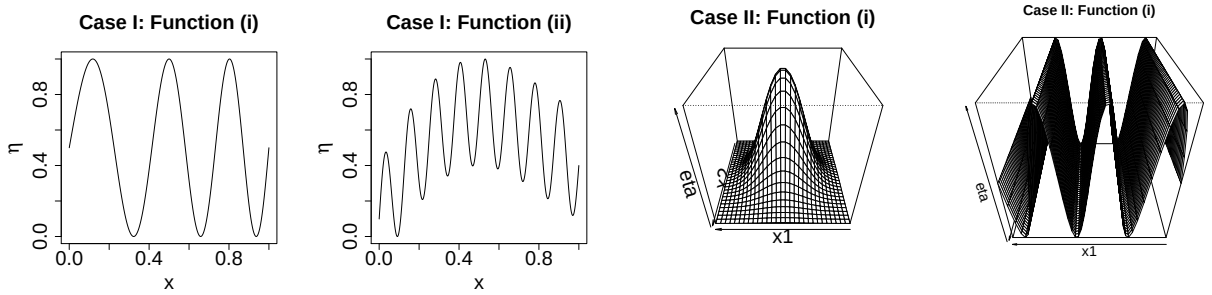


Figure 4.1: True functions used in simulation Case I and Case II. The true functions in Case III are the three-dimensional counterparts of the true functions in Case II.

In Case I, the two candidate penalties are the penalties for linear spline and periodic cubic spline. In Case II and III, the two candidate penalties are the penalties for thin-plate spline and tensor product cubic spline. For the three cases studied, we compare the

performance of two smoothing spline estimators based on one of two chosen penalties and our mixture prior estimator. In all three cases, $p_1 = p_0 = 0$. We opt to use $a = 3/2$, $s = 1/4$, and $\alpha_1 = \alpha_0 = \beta_1 = \beta_0 = 1/2$, for the hyperparameters, which satisfy all the conditions in Theorem 3. To assess the estimation accuracy of each method, we calculate the MSE, which is defined as $\sum_{j=1}^{n}(\eta(x_j) - \hat{\eta}(x_j))^2/n$, where x_js are newly generated grid points on the domain of interest. Figure 4.2 presents box plots of the MSEs based on 50 runs for each case under three SNR levels. The estimates of all the smoothing spline methods using one penalty term are carried out using the algorithms in Gu (2014). In all cases, for different SNR levels, the median of MSEs of the proposed method is the lowest. The interquartile range of MSEs of the proposed method is generally smaller than those of other methods, suggesting that our method is more stable. This is expected since the data generating process implies that there might not be one penalty choice that dominates the other. Therefore, the two estimators that are based on solely one penalty tend to display larger variances. However, the proposed method takes modeling uncertainty into consideration and performs closer to the estimator of the penalty with better performance in each replicated data.

4.4.2 Tweet trend data

In the past decade, social media have experienced an explosive rise in our society. Upon its wide popularity, the massive streaming social media service also provide an excellent alternative data collection procedure for collecting information concerning peoples' opinions on many societal matters. Assisted by the wide usage of location aware mobile devices, social media bear the advantage of generating data with exact location information and updating data with a minimum time lag. Recent studies have demonstrated the power of data generated by social media in many areas such as discovering influenza trend (Helwig et al., 2015), predicting election results (Broersma and Graham, 2012) and studying financial market (Bing et al., 2014). In this section, we use the Twitter trend data reported in Helwig

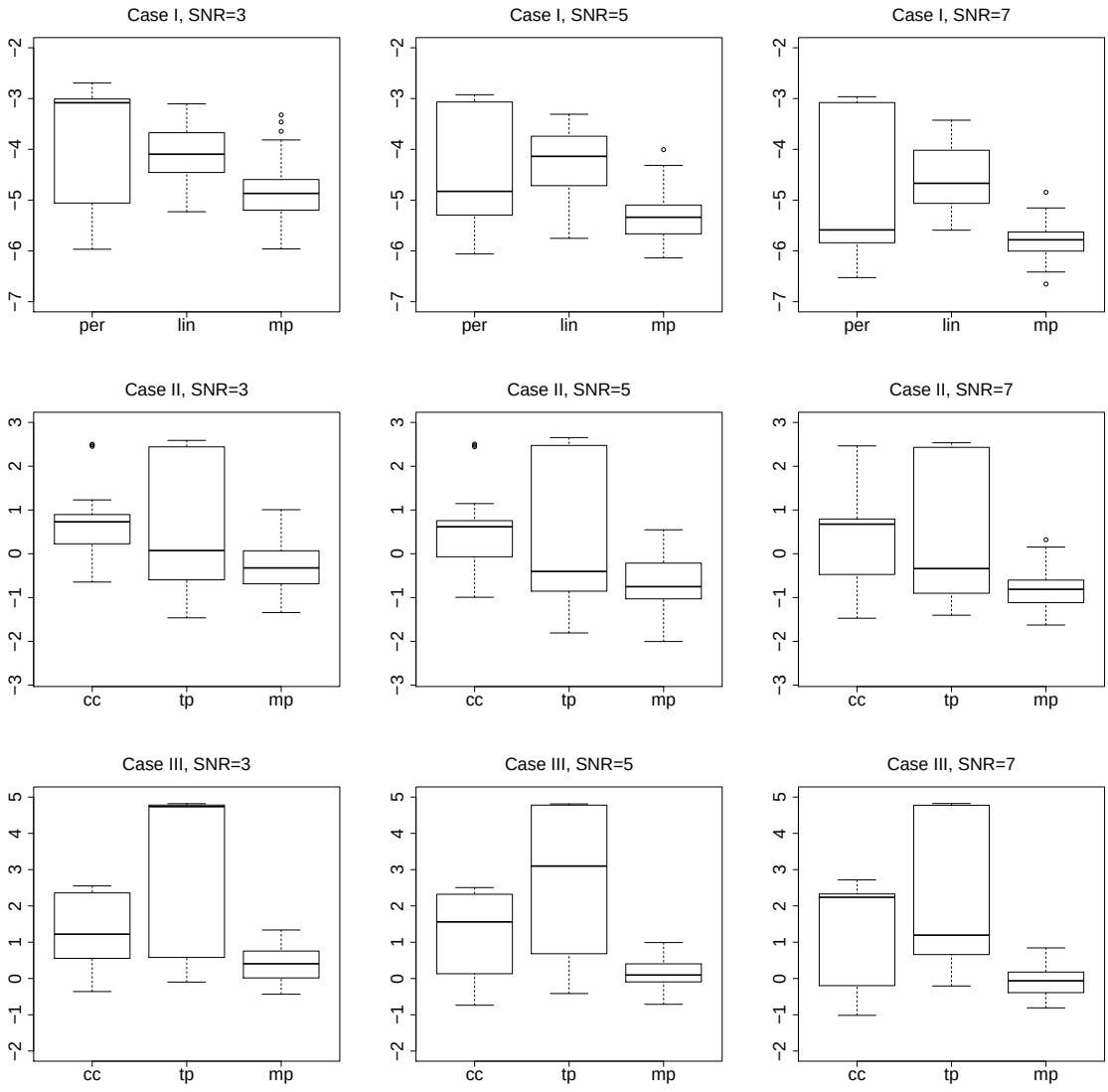


Figure 4.2: Box plots of the MSEs after logarithm transformation for three cases based on 50 simulation runs under three SNR levels, $3, 5, 7$. "per" stands for periodic cubic spline; "lin" stands for linear spline; "cc" stands for cubic spline; "tp" stands for thin-plate spline; "mp" stands for our method using mixture prior.

et al. (2015). Our primary interest for analyzing this data is to model the intensity of tweet counts within a region as a function of its geographic information.

Our raw data contain a total of over 8 million of tweet records (with GPS information) from the mainland of the United States of America, over a typical work week (Monday to

Thursday) in January 2013. We first map all of the tweet timestamps recorded in central standard time to local standard time. We bin the data according to both space and time. In particular, we use 50 bins for longitude values, 25 bins for latitude values. We follow the time binning in Helwig et al. (2015) and defined 13 time bins, denoted as 0:00, 2:00, 4:00, $\cdots$, 24:00, throughout the day.[1] Using these spatial and temporal bin sizes, we end up with n =8563 bins with tweet intensity, spatial and temporal information recorded.

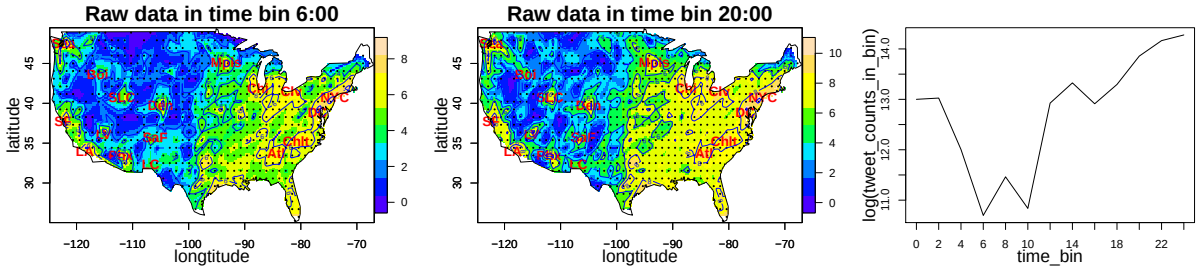


Figure 4.3: Tweet trend data. From left to right: Tweet intensity map at time bin 6:00; Tweet intensity map at time bin 20:00; Tweet intensity for all time bins. The maps are smoothed by a linear interpolation of the observed points. In the left and middle panels, "Sea" is Seattle; "SF" is San Francisco; "LA" is Los Angeles; "Boi" is Boise; "LV" is Las Vegas; "SLC" is Salt Lake City; "Phx" is Phenix City; "Den" is Denver; "SaF" is Santa Fe; "LC" is Las Cruces; "Chi" is Chicago; "Clv" is Cleveland; "Atl" is Atlanta; "Chlt" is Charlotte; "NYC" is New York City.

In Figure 4.3, we plot spatial patterns of tweet intensity using the raw data in time bin 6:00 and 20:00, as examples of the less busy and busy time bins of day, respectively. It is expected that spatial patterns highlight big cities, the positions of which are indicated by red letters on the map. In time bin 6:00, spatial patterns of tweet intensity are smoother compared to that in time bin 20:00, especially in the north-western area. In Figure 4.3, we also plot the temporal trend throughout the 13 time bins on the right panel. The total number of tweets decrease from early morning and reached the minimum at time bin 6:00. Then it starts to climb up in the afternoon and reach the maximum at midnight.

Given a time bin, we model the tweet trend intensity, defined as tweet counts after log

[1]The 13 time bins are 0:00-1:00, 1:00-3:00, 3:00-5:00, 5:00-7:00, 7:00-9:00, 9:00-11:00, 11:00-13:00, 13:00-15:00, 15:00-17:00, 17:00-19:00, 19:00-21:00, 21:00-23:00, 23:00-24:00, denoted as 0:00, 2:00, 4:00, 6:00, 8:00, 10:00, 12:00, 14:00, 16:00, 18:00, 20:00, 22:00, 24:00 respectively.

transformation, using model (1.4), in which the predictors are the longitude and latitude of the center of each bin. Similar to Case II in the simulation study, both thin-plate spline and cubic spline penalties are popular choices for the estimation of two-dimensional functions in the Tweet trend data. Thus, we use the setting in Case II in the simulation study and apply thin-plate spline, cubic spline, and proposed mixture prior method to all 13 time bins. The MSEs of three methods are plotted in Figure 4.4. The MSEs of thin-plate spline and cubic spline methods are mostly close to each other along all time bins. The mixture penalty method outperformed both cubic spline and thin-plate spline in all 13 time bins.

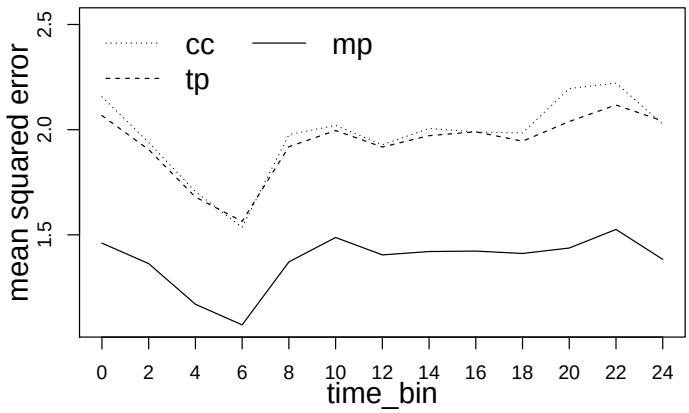


Figure 4.4: Twitter trend data: MSE of cubic spline (cc), thin-plate spline (tp), and mixture prior method (mp) in 13 time bins of day.

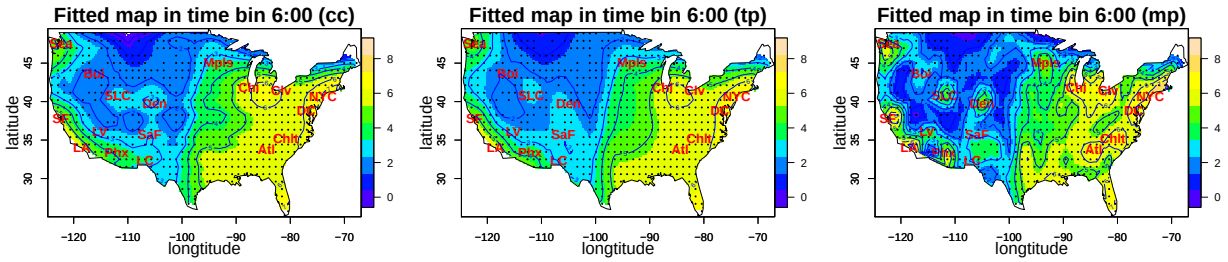


Figure 4.5: Tweet trend data in time bin 6:00. From left to right: fitted map using the cubic spline (cc), thin-plate spline (tp), mixture prior method (mp).

In Figure 4.5 and Figure 4.6, we plot the fitted maps of cubic spline, thin-plate spline,

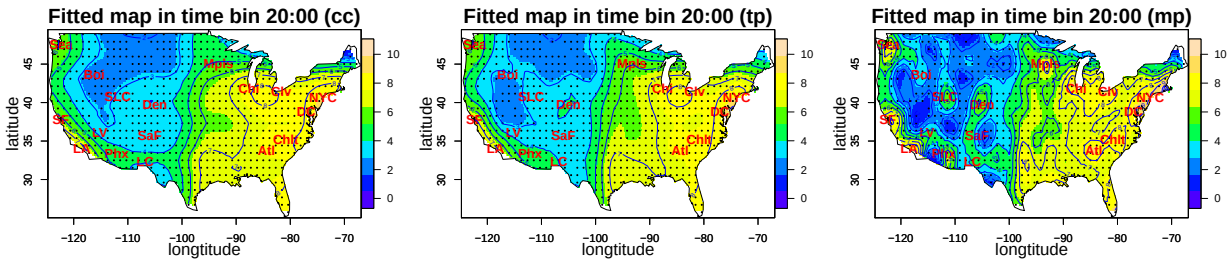


Figure 4.6: Tweet trend data in time bin 20:00. From left to right: fitted maps using the cubic spline (cc), thin-plate spline (tp), and mixture prior method (mp).

and mixture penalty methods in time bin 6:00 and 20:00. For both time bins, the fitted maps of cubic spline and thin-plate spline differ slightly in the middle part around Santa Fe and Denver area. Both thin-plate and cubic spline methods are over-smoothed in this case and have failed to identify big cities such as Boise, Las Vegas, Santa Fe, Las Cruces, Phoenix City, and Atlanta areas. All these big cities are identified by our method.

4.4.3 Methylation profiling in tomato

Tomato (*Solanum lycopersicum*) is an important vegetable crop. It is also one of the important model plants for studying the development of fleshy fruits, which are unique because they typically go through a ripening process after seed maturation. Regulation of this process ensures accurate and tissue-specific control of key developmental transitions (Zhong et al., 2013). DNA methylation is one of several epigenetic mechanisms that cells use to control gene expression without changing the DNA sequence. It is a process in which additional methyl groups are added to the DNA molecules. For DNA's four bases, cytosine(C), adenine(A), guanine(G), and thymine(T), cytosine and adenine can be methylated. The methylation of cytosine is widespread in both eukaryotes and prokaryotes. In the literature, Manning et al. (2006) and Seymour et al. (2008) hypothesized that DNA methylation contributes highly to the regulation of fruit ripening, e.g., the hypermethylation of the gene locus *Cnr* (encoding the transcription factor colorless nonripening) arrests fruit development. It is of

great interest to perform a precise whole genome methylation profiling to study the role of methylation throughout the development of tomatos. In this project, we have acquired the whole genome bisulfite sequencing data of a sample from colorless nonrippening (Cnr) fruits at 60 dpa (days postanthesis). The goal of the analysis here is to estimate the methylation level throughout the whole genome. In this study, we look at the methylation of cytosine in the sequence context CHG, where H corresponds to A, T or C. The CHG methylation is one type of the methylation that is commonly observed in plants but not in mammalians.

The genome of tomato has 12 chromosomes. The raw data, i.e., the methylation level, is plotted in the band on the outmost of left panel in Figure 4.7. In the plot, we have also included the density of genes. We notice that the methylation level is largely related to the density of genes, i.e., in the areas where the genes are annotated, the methylation level is relatively low. The patterns of methylation level throughout different chromosomes are quite different. For example, the methylation level in chromosome 6 and 7 are much more fluctuative than other chromosomes. Considering the different smoothness structures in 12 chromosomes, we consider using our mixture penalty method. In particular, we use periodic cubic spline and linear spline as in the one-dimensional case in the simulation study. The estimated whole genome methylation level using periodic cubic spline, linear spline, and mixture penalty method are plotted in Figure 4.7. We notice that the two spline methods are all over-smoothed and have overlooked some of the important hypermethylated/demetholated regions. For example, on Chromosome 7, the mixture penalty method has identified the demethylation area missed by both periodic spline and linear spline. In Figure 4.7, we have plotted the MSEs on 12 chromosomes. The mixture penalty method has smaller MSE than the other two methods in all but Chromosome 1. The precise profiling of whole genome methylation level of Cnr fruits will be integrated into comparison with earlier stage of fruits or sample of plants with other treatment, such as virus invaded plants, to identify differentially methylated regions (DMRs) that contribute to fruit development.

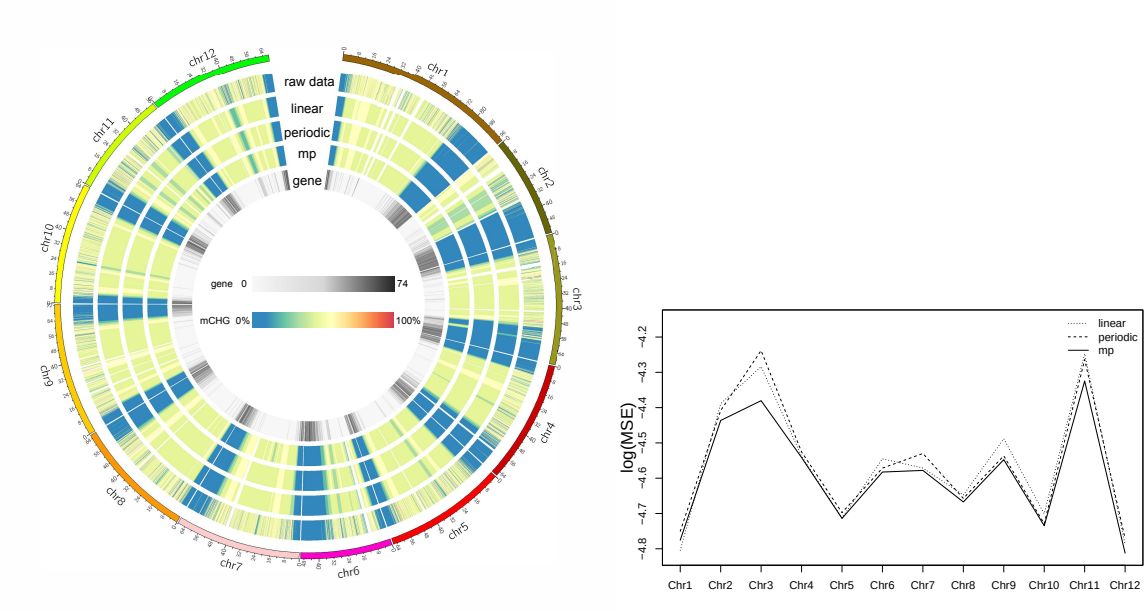


Figure 4.7: Genome-wide methylation profiling of tomato. Left: From inner to outer circle, the heatmap plots the number of genes, the methylation level fitted using mixture penalty method, the methylation level fitted using linear spline, the methylation level fitted using periodic cubic spline, and raw data. Right: MSEs of linear spline, periodic cubic spline, and mixture penalty method on 12 chromosomes.

4.5 Summary and discussion

In this chapter, we propose a Bayesian model for incorporating the ambiguity in choosing penalties in smoothing spline models. In particular, we use a mixture distribution based on the available choices of penalties as a prior for the function to be estimated. We propose partial noninformative priors on the parameters involved and provided sufficient conditions for the propriety of the resulting posterior distributions. The proposed estimator takes both penalties into consideration instead of performing model selection. Our method outperforms the methods that use only one penalty in both simulated and real-world data analysis.

Chapter 5

Small RNAs-Dependent Gene Silencing in *C. elegans*

Chapter summary: A continuing theme of my research is applying classical statistical tools to solve scientific problems arising from the bioinformatics disciplines. I have been actively engaged in collaborative and interdisciplinary work with a focus on the studies related to small RNAs. In this chapter, I present one topic in *C. elegans*. Small RNAs play an important role in guiding Argonaute proteins to nascent RNA transcripts to induce co-transcriptional gene silencing in *C. elegans*. The protein Aquarius helicase was known to be required to initiate this small RNA-induced heritable gene silencing process, and in this chapter, we demonstrate that this reliance is related to the existence of introns in the gene.

5.1 Background

The RNA interference (RNAi) is a biological process in which RNA molecules participate in the repressing of gene expression or translation. Small RNA pathways in eukaryotes are related to RNAi mechanisms(Fire et al., 1998). In this process, small RNAs (sRNAs) of 21-32 nucleotide (nt) long are bound by Argonaute superfamily proteins, interact with target RNAs through Watson-Crick base-pairing and initiate silencing of these targets. Such sRNA-mediated gene silencing can be post-transcriptional on mRNAs in the cytoplasm (PTGS) or co-transcriptional on nascent transcripts in the nucleus (coTGS). In particular, the latter provides the potential to couple sRNA-mediated silencing to DNA and chromatin-based gene regulatory pathways.

It has been shown that *C. elegans* has coTGS mechanisms in the soma and the germline (Weick and Miska, 2014). *C. elegans* piRNAs are 21 nt RNAs with a 5′ uracil (21U-RNAs) that are bound by the PRG-1 Piwi protein in the germline cytoplasm (Batista et al., 2008; Das et al., 2008; Wang and Reinke, 2008). In the germline, sRNAs and piRNAs can initiate coTGS through a two-step mechanism. First, a PRG-1 protein forms a complex with the piRNA, and the complex recognizes its target RNAs through base pairing. Then, the complex recruits an RNA-dependent RNA polymerase (RdRp)-containing complex to generate the secondary 22 nt antisense sRNAs with a 5′ guanine (22G-RNAs) (Pak and Fire, 2007; Bagijn et al., 2012). These secondary 22G siRNAs align antisense to the targets in a site close to the direct target site for piRNAs, and they are then bound by the Argonaute HRDE-1 and imported back into the nucleus (Shirayama et al., 2012; Ashe et al., 2012; Buckley et al., 2012). An HRDE-1/22G-RNA complex is then thought to directly interact with nascent transcripts. During this process, additional 22G-RNAs (tertiary siRNAs) are generated. These tertiary siRNAs align antisense to the targets in a site away from the direct target site for piRNAs. This stage is also called 22G-RNAs spreading, as shown in left panel of Figure 5.1. As tertiary sRNAs can promote the production of further 22G-RNAs, they could

lead to potentially unlimited stages and even multi-generation of silencing (Sapetschnig et al., 2015). However, the detailed molecular basis of this process of how HRDE-1 links sRNA-mediated silencing to coTGS and chromatin modifications remains unclear. In this chapter, we are interested in exploring the role of a conserved RNA helicase Aquarius/EMB-4, a specific protein that interacts with HRDE-1, in small RNA pathways in *C. elegans*.

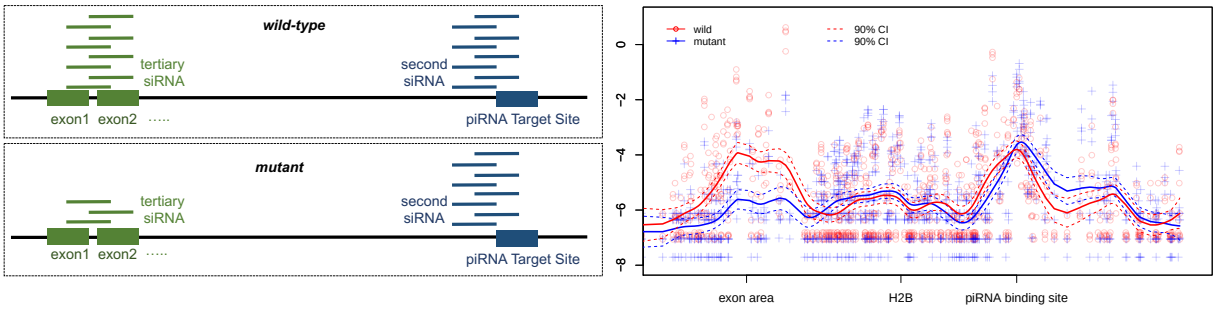


Figure 5.1: Left panel: An illustration of 22G RNA spreading: in wild-type samples, both second and tertiary siRNA are produced at a normal level for piRNA to perform gene silencing; in mutant samples, certain proteins necessary for the production of tertiary siRNA is eliminated so the tertiary siRNAs have low abundance. Right panel: The expression level of EMB-4 mutant and wild-type samples on a sensor gene. Circles and crosses are observed expression level after normalization. The dataset was downloaded from Akay et al. (2017). The solid lines are curves fitted by a generalized smoothing spline ANOVA model (SSANOVA). The dashed lines are the 90% Bayesian confidence intervals. We note that the fitted curves in the right panel faithfully recover the model suggested in the left panel.

Eukaryotic mRNA transcription is an elaborate and multi-step process in which the genetic information in DNA is translated into complementary RNA. The first step is initiation. The RNA polymerase II along with some general transcription factors assembles and binds to transcription starting site. The second step is elongation, i.e., nascent transcripts are produced by the elongating RNA polymerase II. Third, nascent RNA transcripts are processed to mature mRNAs through the assembly of multiple large ribonucleic acid protein (RNP) complexes to perform tasks including $5'$ capping, splicing and $3'$ poly(A) tailing. All of these steps are required to protect transcripts from degradation and to ensure that mRNAs are successfully exported into the cytoplasm for downstream translation. Of these steps, RNA splicing is probably the most complex of all. In this process, the newly made precursor mes-

senger RNA (pre-mRNA) transcript is transformed into a mature messenger RNA (mRNA), and it requires more than hundreds of proteins and many different non-coding RNAs functioning in a delicate way (Nilsen, 2003; Wahl and Lührmann, 2015). Pre-mRNA splicing by the spliceosome is immediately followed by the assembly of a set of proteins known as the exon-junction complex (EJC). EJC functions in export, localization, and translation of mRNAs. Assembly of different EJC components can determine the fate of mRNAs, and EJC can be considered a regulatory hub between transcription and translation (Le Hir et al., 2016).

In coTGS, one of the important HRDE-1 interacting factors is the conserved RNA helicase Aquarius/EMB-4, which binds introns and recruits the EJC to newly spliced transcripts (Tyc et al., 2017). As Aquarius is known to bind to introns and is required for RNP re-modeling during spliceosome and EJC assembly, it is reasonable to hypothesize that nuclear RNAi machinery needs to overcome the intronic barriers for efficient and complete silencing of target sequences, and that introns can influence the function of Aquarius/EMB-4 during gene silencing in a negative direction.

5.2 Preliminary data analysis

In order to examine the influence of introns, we used the data reported in Akay et al. (2017). We first group the transcripts according to the number of exons and plot the 22G RNA level in different groups in Figure 5.3. As shown, it is generally harder for transcripts with a higher number of exons, i.e., more introns, to accumulate the 22G-RNA populations. This indicates that introns can potentially form a barrier to nuclear RNAi by limiting 22G RNA levels.

For studying the role of Aquarius/EMB-4, we focus on the transcripts that do somehow show a decrease in 22G RNA levels in EMB-4 mutants. For each transcript, we divide the total length into 20 equal bins, calculate a fold change for each bin, and identify the

transcripts with at least one bin that has fold change less than 0.5. We plot the fold change for corresponding 22G-RNA and RNA in Figure 5.3. Indeed, we observe that gene de-silencing in EMB-4 mutants is correlated with 22G-RNA depletion in transcripts with more introns. Furthermore, the negative correlation between 22G-RNA abundance and mRNA levels is stronger as intron number increases in EMB-4 mutant animals.

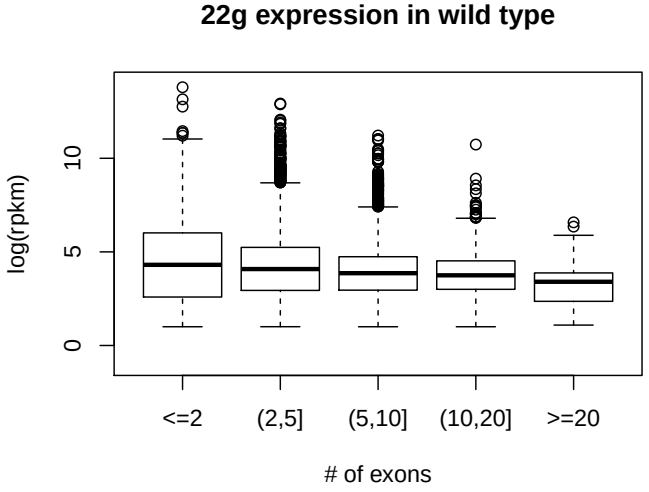


Figure 5.2: The level of 22G RNA, measured in $\log_2$ rpkm (read counts per million per kilobase pair), in groups of genes with different numbers of exons in wild type *C. elegans*.

In fact, in a separate experiment, it is shown that the removal of introns from a coTGS target removes the requirement for Aquarius/EMB-4 in sRNA-mediated silencing using the piRNA sensor gene (Akay et al., 2017). These evidence corroborate a model where the RNA helicase Aquarius/EMB-4 is essential for providing the co-transcriptional silencing complex access to nascent transcripts, especially those undergoing splicing (Figure 5.3). In plants, similar to our results, it has been shown that intron-containing transgenes are protected from nuclear RNAi pathway in comparison to the strongly silenced intronless transgenes (Christie et al., 2011). In summary, our results provide supporting evidence to the growing body of literature showing that pre-mRNA processing is a powerful barrier to co-transcriptional gene silencing in eukaryotes.

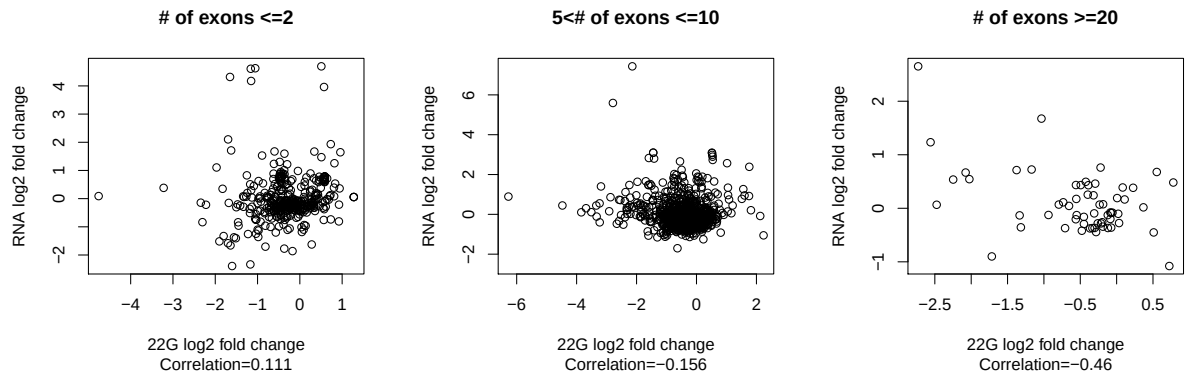


Figure 5.3: mRNA expression levels correlate negatively with 22G-RNA abundance and correlation increases when the transcripts have more exons. X-axis shows $\log_2$ fold change of 22G-RNAs in mutants/wild type, Y-axis shows $\log_2$ fold change of mRNA in mutants/wild type (correlation coefficient and p-values are shown on graphs). The correlation coefficients refer to the Pearson correlation correlation, and all the correlation coefficients are significant at the significant level of 0.05.

Chapter 6

Concluding Remarks and Future Works

The emergence of massive and complex data in different fields of science has posted numerous challenges and opportunities for statisticians. To deal with the computation issues that come with large scale data, we studied the subsampling estimators and their asymptotic properties. We provided the asymptotic distribution of the subsampling estimators. We showed that the subsampling estimators are unbiased and consistent estimators for approximating full sample OLS estimate and for estimating true coefficients provided certain regularity conditions are satisfied. Using these asymptotic results, we proposed optimal criteria and several optimal subsampling estimators under various settings for minimizing the criteria. Compared to simple linear regression, the computation cost of smoothing spline is much higher, at the order of $O(n^3)$, and the cost increases exponentially as the number of predictors increases. Compared to the simple linear regression, the computation burden of nonparametric regression model is even more pressing. Thus, some future topics along this line of research would be studying the asymptotic properties of subsampling estimators in smoothing spline and other nonparametric regression models.

Dealing with ambiguity/uncertainty in making model assumptions is a challenging problem for statistical analysis. We studied the case of choosing ambiguous penalties in smoothing spline models. We used a mixture distribution based on the available choices of penalties as a prior for the function to be estimated and partial noninformative priors on the parameters involved. The proposed estimator takes both penalties into consideration, and thus achieve the goal of incorporating model uncertainty. It outperforms all the other methods that use only one penalty in both simulated and real-world data analysis. One future extension is to apply our paradigm to the generalized nonparametric models. Chen and Ibrahim

(2003), Bové and Held (2011) and many others have provided conjugate and extended g-prior for regression coefficients in generalized linear models, which can be incorporated into the Bayesian model for generalized nonparametric models. Further efficient MCMC algorithms will be studied alongside.

As a second step for studying the small RNA related gene silencing, we are in the process of developing a constrained multisample deconvolution method for simultaneously isoform assembly and expression estimation using the tools of matrix decomposition. In the end, we intend to build novel statistical models and estimation methods to integrate spatial patterns of small RNA levels with isoform expressions.

Appendix A

Proofs of Asymptotic Properties of Optimal Subsampling Estimators

A.1 Proofs of Theorem 1 and Theorem 2

In this part, we collect the proofs of the Theorem 1 and Theorem 2.

A.1.1 Notation and technical preliminaries

Define $(K_1, \ldots, K_n)$ as a random vector which follows a multinomial distribution, $\text{Mult}(r, \{\pi_i\}_{i=1}^n)$, with the subsample size r being total number of trials, and $\{\pi_i\}_{i=1}^n$ being the probability of events. We use the random vector $(K_1, \ldots, K_n)$ to denote the outcome of subsampling, i.e., K_i represents the number of times the i-th observation being subsampled.

Define $\mathbf{K} = \text{diag}\{K_i\}_{i=1}^n$, $\mathbf{\Omega} = \text{diag}\{1/r\pi_i\}_{i=1}^n$, and $\mathbf{W} = \mathbf{\Omega K}$. For the ith diagonal element of matrix $\mathbf{W}$, denoted as W_i, we have

$$\text{E}(W_i) = 1, \quad \text{Var}(W_i) = \frac{(1-\pi_i)}{r\pi_i}, \quad \text{Cov}(W_i, W_j) = -\frac{1}{r}, \quad i \neq j, \quad i, j = 1, \ldots, n. \quad \text{(A.1)}$$

Simple algebra yields that the subsampling estimator of Algorithm 1 can be written as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\mathbf{\Phi}^{*2}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{\Phi}^{*2}\mathbf{Y}^* = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}. \quad \text{(A.2)}$$

Throughout the rest of the proof, for matrix $\mathbf{A}$ we write $\mathbf{A} = O_p(n^\delta)$ to denote that all elements of $\mathbf{A}$ are in the order of $O_p(n^\delta)$. The Cramer-Wold Device and Lemma 2 below

govern the proof for Theorem 1 and Theorem 2.

Cramer-Wold Device. *For random vectors* $\mathbf{Z}_n = (Z_{n1}, \ldots, Z_{np})^T$ *and* $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$*, a necessary and sufficient condition for* $\mathbf{Z}_n \xrightarrow{d} \mathbf{Z}$ *is that* $\boldsymbol{b}^T\mathbf{Z}_n \xrightarrow{d} \boldsymbol{b}^T\mathbf{Z}$ *as* $n \to \infty$*, for each* $\boldsymbol{b} \in \mathbb{R}^p$*.*

Remark 4. If the goal is to derive the asymptotic distribution of a random vector, then by Cramer-Wold Device it suffices to derive the asymptotic distribution on any linear combinations of the random vector, which is a scalar, i.e., one-dimensional case. Our estimator $\tilde{\boldsymbol{\beta}}$ in (A.2) is a vector of random variables, and we intend to derive the asymptotic distribution of $\tilde{\boldsymbol{\beta}}$. The Cramer-Wold device reduces derivation of asymptotic distribution for *vectors* to the usual *scalar* case. See more details about the Cramer-Wold device in (Billingsley, 1995, sec. 29).

Convergence of Geometric Series of Matrices. *Let* $\mathbf{A}$ *be a* $n \times n$ *square matrix. We use* $\rho(\mathbf{A})$ *to denote the spectral radius of matrix* $\mathbf{A}$*, i.e.,* $\rho(\mathbf{A}) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$*, where* $\lambda_1, \ldots, \lambda_n$ *are the eigenvalues of matrix* $\mathbf{A}$*. If* $\rho(\mathbf{A}) < 1$*, then* $(\mathbf{I} - \mathbf{A})$ *is invertible, and the series*

$$\mathbf{S} \;=\; \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots$$

converges to $(\mathbf{I} - \mathbf{A})^{-1}$*.*

Remark 5. The convergence of geometric series of matrices will be used in the proof of Lemma 2 below. See more details and proof of this conclusion in (Hubbard and Hubbard, 1999, sec. 1.5).

Lemma 2. *Assume that* $0 < \pi_i < 1$ *for* $i = 1, \ldots, n$*. If*

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X} \;=\; O_p\left(\frac{1}{r^{\frac{\delta}{2}}}\right), \tag{A.3}$$

where δ *is a positive constant, then the weighted subsample estimator in (A.2) can be written*

as

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r^{\delta}), \tag{A.4}$$

where $\boldsymbol{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

Proof. By (A.3), we have

$$((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X})^2 \;=\; O_p(1/r^{\delta}). \tag{A.5}$$

Therefore,

$$[\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}]^{-1} = \mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p(1/r^{\delta}). \tag{A.6}$$

We expand (A.2) as follows,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Y}) \\
&= [\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Y}) \\
&= [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p(1/r^{\delta})](\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}) \quad \text{(A.7)} \\
&= [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} + O_p(1/r^{\delta})](\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{Y}) \\
&= \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\boldsymbol{e} + O_p(1/r^{\delta}) \\
&= \hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r^{\delta}), \quad \text{(A.8)}
\end{aligned}$$

where the expansion in (A.7) is by the convergence of geometric series of matrices and the assumption $\delta > 0$, and the equality in (A.8) holds since $\mathbf{X}^T\boldsymbol{e} = 0$. □

Remark 6. Lemma 2 relates the subsampling estimator $\tilde{\boldsymbol{\beta}}$ to the $\hat{\boldsymbol{\beta}}_{OLS}$ with an order constraint on the residual term, i.e., $O_p(1/r^{\delta})$. In application of Lemma 2 to the proof of Theorem 1 (asymptotic normality of $\tilde{\boldsymbol{\beta}}$ for estimating $\boldsymbol{\beta}_0$), we subtract $\boldsymbol{\beta}_0$ from both sides of (A.8) to relate $\tilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$. Lemma 2 is directly applied in the proof of Theorem 2 (asymptotic normality of $\tilde{\boldsymbol{\beta}}$ for approximating $\hat{\boldsymbol{\beta}}_{OLS}$).

Remark 7. The assumption of $\delta > 0$ implies that $\rho((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}) \to 0$ as $r \to \infty$. By the convergence of geometric series of matrices, the inverse of $[\mathbf{I}+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}]$, i.e., $\mathbf{X}^T\mathbf{W}\mathbf{X}$, exists and the expansion in $(A.7)$ is valid asymptotically. In the proof of Theorems 1 and 2, we will verify the condition in Lemma 2, i.e., $\delta > 0$. The exact magnitude of δ depends on $(\mathbf{W}-\mathbf{I})$, and it differs in Theorems 1 and 2.

Hajek-Sidak Central Limit Theorem. *Let $X_1, \ldots, X_n$ be identical and independently distributed (i.i.d.) random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ are both finite. Define $T_n = d_1X_1 + \ldots + d_nX_n$, then*

$$\frac{T_n - \mu\sum_{i=1}^n d_i}{\sigma\sqrt{\sum_{i=1}^n d_i^2}} \xrightarrow{d} N(0,1)\,, \tag{A.9}$$

whenever the Noether condition

$$\frac{\max_{1\le i\le n} d_i^2}{\sum_{i=1}^n d_i^2} \to 0, \quad \text{as } n \to \infty, \tag{A.10}$$

is satisfied.

Remark 8. The Hajek-Sidak Central Limit Theorem deals with the asymptotic normality of a weighted average of identically and independently distributed random variables (Saleh, 2006). It is used in the proof of Lemma 3 below and thus facilitate the final proof of Theorem 1.

For statement and proof of Lemma 3, Lemma 4, and Theorem 1 below, recall from main text that we denote $r = O(n^{1-\alpha})$, where $0 \le \alpha < 1$ and $\pi_{min} = O(n^{-\gamma_0})$, $\gamma_0 \ge 1$. Also recall that $\mathbf{\Omega} = \text{diag}\{1/r\pi_i\}_{i=1}^n$.

Lemma 3. *Define $\mathbf{U} = diag(U_1, \ldots, U_n)$ such that $U_i \overset{iid}{\sim} Poisson(r\pi_i), i = 1, \ldots, n$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, and ε_is follow independent and identical distribution with mean 0 and variance σ^2. If (A1) there exist b and B such that $b \le \lambda_{min}(\mathbf{X}^T\mathbf{X}/n) \le \lambda_{max}(\mathbf{X}^T\mathbf{X}/n) \le B$, and*

(A2) $\gamma_0 + \alpha < 2$ *holds, as* $n \to \infty$,

$$\boldsymbol{\Sigma}_0^{-\frac{1}{2}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \xrightarrow{d} \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p), \tag{A.11}$$

where $\boldsymbol{\Sigma}_0 = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}_p + \boldsymbol{\Omega})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$.

Proof. We establish the asymptotic normality of random vector $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$ using the Cramer-Wold device. That is, we first convert the random vector to a scalar random variable as follows and show its asymptotic normality. For any constant vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{b} \neq \mathbf{0}$, we evaluate

$$\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} = \sum_{i=1}^{n} d_i\zeta_i, \tag{A.12}$$

where $d_i = \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{\sqrt{r\pi_i + r^2\pi_i^2}}{r\pi_i}$ and $\zeta_i = U_i\varepsilon_i/\sqrt{r\pi_i + r^2\pi_i^2}$, $\mathrm{E}(\zeta_i) = 0$ and $\mathrm{Var}(\zeta_i) = \sigma^2$.

Thus Equation (A.12) is a weighted average of independent random variables ζ_i, it suffices to verify the Noether condition (A.10) of Hajek-Sidak central limit theorem to show the asymptotic normality of $\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$.

For d_i^2, we have

$$d_i^2 \leq (1 + \frac{1}{r\pi_{min}})(\boldsymbol{a}^T\mathbf{x}_i)^2 \leq (1 + \frac{1}{r\pi_{min}})\boldsymbol{a}^T\boldsymbol{a}M_x, \tag{A.13}$$

where $\boldsymbol{a} = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{b}$, $M_x = \max\{\mathbf{x}_i^T\mathbf{x}_i\}_{i=1}^n$ and the last inequality is derived using Cauchy-Schwarz inequality. Thus $\max_{1\leq i\leq n} d_i^2 \leq (1 + \frac{1}{r\pi_{min}})\boldsymbol{a}^T\boldsymbol{a}M_x$.

For $\sum_{i=1}^n d_i^2$, we have

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(1 + \frac{1}{r\pi_i})\boldsymbol{a}^T\mathbf{x}_i\boldsymbol{a}^T\mathbf{x}_i \geq (1 + \frac{1}{r\pi_{max}})\boldsymbol{a}^T\mathbf{X}^T\mathbf{X}\boldsymbol{a} \geq (n + \frac{n}{r\pi_{max}})\lambda_0\boldsymbol{a}^T\boldsymbol{a}, \tag{A.14}$$

where we use λ_0 to denote the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}/n$ for ease of notation.

Combining (A.13) and $(A.14)$, we have

$$\lim_{n\to\infty} \frac{\max_{1\le i\le n} d_i^2}{\sum_{i=1}^n d_i^2} \le \lim_{n\to\infty} \frac{(1+\frac{1}{r\pi_{min}})M_x}{(n+\frac{n}{r\pi_{max}})\lambda_0} \le \frac{M_x}{\lambda_0} \lim_{n\to\infty} \frac{1+r\pi_{min}}{(nr\pi_{min}+\frac{n\pi_{min}}{\pi_{max}})} = 0, \qquad (A.15)$$

where we used the condition $\alpha+\gamma_0<2$, i.e., $nr\pi_{min}\to\infty$ as $n\to\infty$. Since

$$\sum_{i=1}^n \mathrm{Var}(d_i\zeta_i) = \sigma^2 \sum_{i=1}^n (\boldsymbol{a}^T\mathbf{x}_i)^2(1+\frac{1}{r\pi_i}) = \sigma^2 \boldsymbol{a}^T\mathbf{X}^T(\mathbf{I}_p+\boldsymbol{\Omega})\mathbf{X}\boldsymbol{a},$$

by Cramer-Wold device, the proof is thus complete.

□

In the following statement and proof of Lemma 4 and Theorem 1, we use $A|B$ to denote the conditional distribution of random variable A given random variable B.

Lemma 4. *Given fixed $\boldsymbol{b}\in\mathbb{R}^p$ such that $\boldsymbol{b}\neq\mathbf{0}$, where $\boldsymbol{\Omega}$, $\mathbf{U}$, and $\boldsymbol{\Sigma}_0$ are defined as in Lemma 3, as $n\to\infty$ we have*

$$(\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^n U_i = r \quad \xrightarrow{d} \quad N(0,1). \qquad (A.16)$$

Proof. For $i=1,\ldots,n$,

$$\mathrm{Cov}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}U_i\varepsilon_i, \sum_{i=1}^n U_i) = \sum_{i=1}^n \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathrm{Cov}(U_i\varepsilon_i, U_i) = 0, \qquad (A.17)$$

we have

$$\mathrm{Cov}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}, \sum_{i=1}^n U_i) \quad = \quad 0.$$

Using the results in Lemma 3, we have

$$\begin{pmatrix} (\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon} \\ \frac{1}{\sqrt{r}}(\sum_{i=1}^n U_i - r) \end{pmatrix} \quad \xrightarrow{d} \quad \mathbf{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right). \qquad (A.18)$$

Further we have

$$(\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^{n}U_i = r \quad \xrightarrow{d} \quad N(0,1). \tag{A.19}$$

□

Remark 9. The proof of above Lemma 4 is analogous to that of Lemma 2.2 and Theorem 2.1 in Morris (1975). The case in Morris (1975) is more complicated than our case here. The simplification is that in (A.18) due to existence of $\boldsymbol{\varepsilon}$ in $(\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$, the covariance of $\frac{1}{\sqrt{r}}(\sum_{i=1}^{n}U_i - r)$ and $(\boldsymbol{b}^T\boldsymbol{\Sigma}_0\boldsymbol{b})^{-\frac{1}{2}}\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}$ is naturally zero. Morris (1975) has considered more complicated cases for construction of similar set of independent random variables. In addition, we refer to Morris (1975) for the technical details for the validity of getting asymptotic conditional distribution in (A.19) from the asymptotic joint distribution in (A.18).

Remark 10. The key difference between the conclusion in Lemma 3 and Lemma 4 is that in Lemma 4 we consider conditional distribution whereas in Lemma 3 we consider unconditional distribution.

Lemma 5. *If random variables* $U_i \sim Poisson(\lambda_i)$, $i = 1, \ldots, n$, *then*

$$(U_1, \ldots, U_n)|\sum_{i=1}^{n}U_i = r \sim Mult\left(r, \left\{\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}\right\}_{i=1}^{n}\right).$$

Proof. The key is to note that $\sum_{i=1}^{n}U_i \sim \text{Poisson}(\sum_{i=1}^{n}\lambda_i)$. Thus

$$\begin{aligned} P(U_1 = u_1, \ldots, U_n = u_n|\sum_{i=1}^{n}U_i = r) &= \frac{P(U_1 = u_1, \ldots, U_n = u_n, \sum_{i=1}^{n}U_i = r)}{P(\sum_{i=1}^{n}U_i = r)} \\ &= \frac{N!}{u_1!\ldots u_n!}\Pi_{i=1}^{n}\left(\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}\right)^{u_i}, \quad \text{if.} \sum_{i=1}^{n}u_i = r. \end{aligned}$$

□

Remark 11. The random variables for denoting subsampling process in $\tilde{\boldsymbol{\beta}}$ are $(K_1, \ldots, K_n)$, which form a random vector following a multinomial distribution. Lemma 3 and Lemma 4 are results of $U_1, \ldots, U_n$, which are random variables following Poisson distribution. Lemma 5 makes the connection between a multinomial distribution and a Poisson distribution.

In the following section A.1.2 and A.1.3 we provide the proof of Theorem 1 and Theorem 2, respectively. The proof of Theorem 1 is substantially more complicated than that of Theorem 2. In conditional inference, i.e., the case of Theorem 2, the data observed are considered as given and the only randomness comes from subsampling. However, in unconditional inference, i.e., the case of Theorem 1, the analysis concerns the unobserved hypothetical data as well as their subsampling probabilities, thus there exist more randomness to be quantified.

A.1.2 Proof of Theorem 1

In this section, we present the **proof of Theorem 1** using results of aforementioned lemmas.

Proof. We first verify the condition of $\delta > 0$ in Lemma 2. To do that, we give the magnitude of δ in (A.3) in the condition of Lemma 2. Note

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X} = (\mathbf{X}^T\mathbf{X}/n)^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}/n,$$

where $(\mathbf{X}^T\mathbf{X}/n)^{-1}$ converges to a constant matrix by assumption (A1). Thus, the order of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}$ depends on the order of $\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}/n$. We then derive the order of the (s,t)th element of $\mathbf{X}^T(\mathbf{W}-\mathbf{I})\mathbf{X}/n$, i.e., $\frac{\sum_{i=1}^n x_{si}x_{it}(W_i-1)}{n}$. We have

$$
\begin{aligned}
\mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_{si}x_{it}(W_i-1)}{n}\right) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} x_{si}x_{it}(W_i-1)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}(x_{si}x_{it})^2\frac{1-\pi_i}{r\pi_i} - 2\sum_{i<j} x_{si}x_{it}x_{sj}x_{tj}\frac{1}{r}\right) \\
&= \frac{1}{rn^2}\left[\sum_{i=1}^{n}(x_{si}x_{it})^2\frac{1-\pi_i}{\pi_i} - \left(\left(\sum_{i=1}^{n} x_{si}x_{it}\right)^2 - \sum_{i=1}^{n}(x_{si}x_{it})^2\right)\right] \\
&= \frac{1}{r}\left[\sum_{i=1}^{n}\frac{(x_{si}x_{it})^2}{n^2\pi_i} - \left(\sum_{i=1}^{n}\frac{x_{si}x_{it}}{n}\right)^2\right] \\
&= O_p\left(\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\right). \qquad \text{(A.20)}
\end{aligned}
$$

For $\sum_{i=1}^{n}\frac{1}{\pi_i}$, one has $n^2 \le \sum_{i=1}^{n}\frac{1}{\pi_i} \le n/\pi_{min}$ and plug these inequalities in (A.20) to note $0 < (2-\gamma_0-\alpha) \le \delta$ in (A.3). We have validated the assumption in Lemma 2.

We subtract $\boldsymbol{\beta}_0$ from both sides of equation (A.4) in Lemma 2 and get

$$
\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 + O_p\left(\frac{1}{r^\delta}\right), \qquad \text{(A.21)}
$$

where we recall that $\boldsymbol{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$. For the right-hand side of (A.21), we note that $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) = O\left(\frac{1}{n}\right)$, which is of higher order (i.e., smaller in magnitude) than the residual term $O_p\left(\frac{1}{r^\delta}\right)$, i.e., $O_p\left(\frac{1}{n^{2-\alpha-\delta}}\right)$. Thus both $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0$ and the residual term can be omitted and the asymptotic distribution of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is equivalent to that of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}$. For the rest of the proof, we establish asymptotic normality of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}$.

Note that

$$
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon}), \qquad \text{(A.22)}
$$

where recall that $\boldsymbol{\varepsilon}$ is the random noise in the model (1.1). The elements in vector $\boldsymbol{\varepsilon}$ are uncorrelated and the elements in vector $\boldsymbol{e}$ are correlated. We will show that the order of

$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon})$ by calculating the variances of the sth element of $\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\varepsilon})/n$. We have

$$\begin{aligned}&\mathrm{Var}(\frac{\sum_{i=1}^n x_{si}W_i(e_i-\varepsilon_i)}{n})\\&\quad=\frac{1}{n^2}\left(\sum_{i=1}^n x_{si}^2\mathrm{Var}(W_i(e_i-\varepsilon_i))+2\sum_{i<j}x_{si}x_{sj}\mathrm{Cov}\left[W_i(e_i-\varepsilon_i),W_j(e_j-\varepsilon_j)\right]\right). \quad \text{(A.23)}\end{aligned}$$

Now we shall analyze the two components on the right-hand side of (A.23).

For the first component, we have

$$\begin{aligned}\sum_{i=1}^n\mathrm{Var}(W_i(e_i-\varepsilon_i)) &= \sum_{i=1}^n\mathrm{E}(W_i^2(e_i-\varepsilon_i)^2)=\sum_{i=1}^n\mathrm{Var}(W_i)\mathrm{Var}(e_i-\varepsilon_i)+(\mathrm{E}W_i)^2\mathrm{Var}(e_i-\varepsilon_i)\\&= \sum_{i=1}^n\frac{1-\pi_i}{r\pi_i}h_{ii}\sigma^2+h_{ii}\sigma^2=O_p\left(\frac{1}{r\pi_{min}}\right), \quad \text{(A.24)}\end{aligned}$$

where the last equality holds since $\sum_{i=1}^n h_{ii}=p$.

For the second component, we have

$$\begin{aligned}\sum_{i<j}\mathrm{Cov}\left(W_i(e_i-\varepsilon_i),W_j(e_j-\varepsilon_j)\right) &= \sum_{i<j}\mathrm{E}(W_iW_j(e_i-\varepsilon_i)(e_j-\varepsilon_j))\\&= \sum_{i<j}\mathrm{E}(W_iW_j)\mathrm{E}((e_i-\varepsilon_i)(e_j-\varepsilon_j))=O_p\left(\frac{n}{r}\right). \quad \text{(A.25)}\end{aligned}$$

Substituting (A.24) and (A.25) into (A.23),

$$\mathrm{Var}(\frac{\sum_{i=1}^n x_{si}W_i(e_i-\varepsilon_i)}{n})=O_p\left(\frac{1}{n^2r\pi_{min}}\right). \quad \text{(A.26)}$$

Combining (A.22) and (A.26) we want to argue that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{e}-\boldsymbol{\epsilon})$ is of higher order than $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}$. Thus, if we establish the asymptotic normality of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$, then the asymptotical normality of $\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ in (A.21) will follow directly.

Note that $\mathbf{W}$ can be written as $\mathbf{W}=\boldsymbol{\Omega}\mathbf{K}$. By Lemma 5, $(K_1,\ldots,K_n)$ and $[(U_1,\ldots,U_n)|\sum_{i=1}^n U_i=r]$ are identically distributed. So $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$ and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{U}\boldsymbol{\varepsilon}|\sum_{i=1}^n U_i$ are identically distrubted. Thus, the conclusion from Lemma 4 can be applied.

We then establish the asymptotic normality using the Cramer-Wold device. That is, for any constant vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{b} \neq \mathbf{0}$, we evaluate $\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varepsilon}$.

Finally, combining (A.21), Lemma 3, and Lemma 4, we have

$$\boldsymbol{\Sigma}_0^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \xrightarrow{d} \quad \mathbf{N}(\mathbf{0}, \mathbf{I}_p), \quad \text{as } n \to \infty, \tag{A.27}$$

where $\boldsymbol{\Sigma}_0 = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}_p + \boldsymbol{\Omega})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$.

□

A.1.3 Proof of Theorem 2

In this section, we provide the proof of Theorem 2. It is much easier than the proof of Theorem 1 in the unconditional inference.

Proof. Given data $\{\mathbf{X}, \mathbf{Y}\}$, we first calculate δ in (A.3) so that we could apply Lemma 2. In this case, since $\|\mathbf{x}_i\| < \infty$, where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$, each element of $\mathbf{X}^T\mathbf{X}$ is considered as finite by assumption. The (s,t)-th element of $\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{X}$ is $\sum_{i=1}^n x_{is}x_{it}(W_i - 1)$.

$$\text{Var}(\sum_{i=1}^n x_{is}x_{it}(W_i - 1)) \;=\; \frac{1}{r}\left(\sum_{i=1}^n (x_{is}x_{it})^2 \frac{1-\pi_i}{\pi_i} - 2\sum_{i<j} x_{is}x_{it}x_{js}x_{tj}\right) = O_p\left(\frac{1}{r}\right), \tag{A.28}$$

i.e., $\delta = 1$ in (A.3).

Note that $\mathbf{K}$ can be write as $\mathbf{K} = \sum_{j=1}^r \mathbf{K}^{(j)}$, where $\mathbf{K}^{(j)} = \text{Diag}\{K_i^{(j)}\}_{i=1}^n$, $(K_1^{(j)}, \ldots, K_n^{(j)}) \overset{iid}{\sim} \text{Mult}(1, \{\pi_i\})_{i=1}^n$ for $j = 1, \ldots, n$. By (A.4) in Lemma 2 and (A.28), we can show

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e} + O_p(1/r) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^r \mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e} + O_p(1/r).
\end{aligned}$$

Then, Cramer-Wold device is used to establish asymptotic normality of $(\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^r \mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e}$,

i.e., for any constant vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{b} \neq \mathbf{0}$, we consider $\sum_{j=1}^{r} \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e}$, which is a summation of r independent random variables. Since the elements in $\boldsymbol{X}$ and $\boldsymbol{e}$ are considered as finite numbers and $\pi_i > 0$, the Noether condition in Hajek-Sidek CLT is satisfied. Without loss of generality, verify

$$
\begin{aligned}
& \mathrm{Var}(\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(1)}\boldsymbol{e}) \\
= \; & \mathrm{Var}(\sum_{i=1}^{n} \boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \frac{1}{r\pi_i} K_i^{(1)} e_i) \\
= \; & \sum_{i=1}^{n} (\boldsymbol{a}^T\mathbf{x}_i e_i \frac{1-\pi_i}{r\pi_i} e_i \mathbf{x}_i^T\boldsymbol{a}) - 2\sum_{i\leq j} \boldsymbol{a}^T\mathbf{x}_i e_i \frac{1}{r} e_j \mathbf{x}_j^T\boldsymbol{a} \\
= \; & \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n} \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a} - \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n} \mathbf{x}_i e_i^2 \mathbf{x}_i^T + 2\sum_{i<j} \mathbf{x}_i e_i e_j \mathbf{x}_j^T\right)\boldsymbol{a} \\
= \; & \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n} \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a} - \frac{1}{r}\boldsymbol{a}^T\mathbf{X}^T\boldsymbol{e}\boldsymbol{e}^T\mathbf{X}\boldsymbol{a} \\
= \; & \frac{1}{r}\boldsymbol{a}^T\left(\sum_{i=1}^{n} \frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T\right)\boldsymbol{a}, \qquad \text{(A.29)}
\end{aligned}
$$

where $\boldsymbol{a} = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{b}$, (A.29) is derived by the fact that $\mathbf{X}^T\boldsymbol{e} = \mathbf{0}$. By CLT, we have $\boldsymbol{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^{r}\mathbf{X}^T\boldsymbol{\Omega}\mathbf{K}^{(j)}\boldsymbol{e} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{b}^T\boldsymbol{\Sigma}_c\boldsymbol{b})$, where $\boldsymbol{\Sigma}_c = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\Sigma}_e(\mathbf{X}^T\mathbf{X})^{-1}$, $\boldsymbol{\Sigma}_e = \frac{1}{r}\sum_{i=1}^{n}\frac{e_i^2}{\pi_i}\mathbf{x}_i\mathbf{x}_i^T$. Thus by the Cramer-Wold device, Theorem 2 follows. □

A.2 Unweighted subsampling estimators

In this section, we focus on the unweighted subsampling method, which is presented in Algorithm 2. Note that Equations (1.3) and (A.30) require that $\mathbf{X}^{*T}\boldsymbol{\Phi}^{*2}\mathbf{X}^*$ and $\mathbf{X}^{*T}\mathbf{X}^*$ are invertible. If they are not invertible, generalized inverse may be used.

In the following, we study the properties of the unweighted subsampling estimators in Algorithm 2. Theorem 4 establishes the asymptotic normality of the unweighted sampling estimators $\tilde{\boldsymbol{\beta}}_u$ for approximating the full sample OLS when the data $\{\mathbf{X}, \mathbf{Y}\}$ are considered

Algorithm 2 Unweighted Subsampling Estimation Algorithm

- **Step 1. (Subsampling) Subsample with replacement from the full data.** This step is the same as Step 1 in weighted subsampling estimation Algorithm 1
- **Step 2. (Estimation) Calculate unweighted least squares using the subsample.** Solve unweighted least squares on the subsample to get the *Unweighted Subsample Estimator* $\tilde{\boldsymbol{\beta}}^u$, i.e.,

$$\tilde{\boldsymbol{\beta}}^u \quad = \quad \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}\|^2 = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{Y}^*. \tag{A.30}$$

as given. Theorem 5 establishes the asymptotic normality of $\tilde{\boldsymbol{\beta}}_u$ for estimating $\boldsymbol{\beta}_0$.

Theorem 4. *Given data* $\{\mathbf{X}, \mathbf{Y}\}$*, we assume that* $\mathbf{X}$ *is of full column rank, and* $\{\|\mathbf{x}_i\|\}_{i=1}^n < \infty$*. Moreover, we assume that all sampling probabilities are nonzero. Then we have*

$$(\mathbf{V}_c^u)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_u - \hat{\boldsymbol{\beta}}_{WLS}) \quad \xrightarrow{d} \quad \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p)\,, \quad \text{as } r \to \infty, \tag{A.31}$$

where the full sample weighted least squares (WLS) estimator $\hat{\boldsymbol{\beta}}_{WLS}$ *is defined as*

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{Y}),$$

$\mathbf{V}_c^u = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{V}_e^u(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$, $\mathbf{V}_e^u = r\sum_{i=1}^n \pi_i(e_i^u)^2\boldsymbol{x}_i\boldsymbol{x}_i^T$, $e_i^u = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{WLS}$.

Theorem 4 states that $\tilde{\boldsymbol{\beta}}_u$ is an asymptotic unbiased estimator of the full sample weighted least squares (WLS) estimator $\hat{\boldsymbol{\beta}}_{WLS}$. Unless $\hat{\boldsymbol{\beta}}_{WLS} = \hat{\boldsymbol{\beta}}_{OLS}$, the unweighted subsampling estimators are asymptotic biased for $\hat{\boldsymbol{\beta}}_{OLS}$.

Theorem 5. *Recall* $r = O(n^{1-\alpha})$ *and* $\pi_{min} = \min\{\pi_i\}_{i=1}^n = O(n^{-\gamma_0})$*, where* $0 \le \alpha < 1$*, and* $\gamma_0 \ge 1$*. Denote* $\pi_{max} = \max\{\pi_i\}_{i=1}^n = O(n^{-\gamma_1})$*,* $\lambda_{min}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}) = O(n^{\beta})$*, where* $\gamma_1 \le 1$ *and* $\beta > 0$*. If*

Condition UNW1: $\alpha + 2\beta > 1$; $\alpha + \gamma_1 > 1$; $\alpha + 2\gamma_0 - \gamma_1 < 2$

hold, then we have

$$(\mathbf{V}_0^u)^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}}^u - \boldsymbol{\beta}_0) \xrightarrow{d} \boldsymbol{N}(\mathbf{0}, \mathbf{I}_p), \quad \textit{as } n \to \infty,$$

where $\mathbf{V}_0^u = \sigma^2(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{V}_p^u(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$, $\mathbf{V}_p^u = \mathbf{X}^T(\boldsymbol{\Omega}^{-1} + \boldsymbol{\Omega}^{-2})\mathbf{X}$.

Theorem 5 states that $\tilde{\boldsymbol{\beta}}_u$ is an asymptotic unbiased estimator of the true model parameter $\boldsymbol{\beta}_0$. Since both weighted and unweighted subsampling estimators are asymptotically unbiased to the true model parameter, then the natural question is how to choose among them. To answer this question, we compare their asymptotic variances.

Lemma 6. *Define* $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}$, *then*

$$\mathbf{V}_0 - \mathbf{V}_0^u = \mathbf{A}(\boldsymbol{\Omega} - \mathbf{I}_n)\mathbf{A}^T.$$

If $r\pi_i < 1$ for all $i = 1, \ldots, n$, then $\mathbf{V}_0 - \mathbf{V}_0^u$ is positive definite. Consequently, the unweighted subsampling estimators have smaller asymptotic variances than the weighted subsampling estimators. We thus recommend using the unweighted subsampling estimators for estimating true model parameters.

A.2.1 Proof of Theorem 4

Proof. The proof of Theorem 4 is analogous to that of Theorem 2. In particular, analogous to Equation (A.4), we have the following expansion for $\tilde{\boldsymbol{\beta}}^u$,

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^u &= (\mathbf{X}^T\mathbf{K}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{K}\mathbf{Y}) \\ &= \hat{\boldsymbol{\beta}}_{WLS} + (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{K}\boldsymbol{e}^u + O(1/r^{2\delta^u}), \end{aligned} \tag{A.32}$$

where $\boldsymbol{e}^u = \mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}_{WLS}$ and δ^u is similar to δ. The rest of the proof is straightforward by following the steps in the poof of Theorem 2. □

Lemma 7. *If the Condition UNW1 is satisfied, then $\hat{\boldsymbol{\beta}}_{WLS}$ is a consistent estimator of $\boldsymbol{\beta}_0$.*

Proof. Let $\boldsymbol{b} \in \mathbb{R}^p$, we show that $\mathrm{Var}(\boldsymbol{b}^T\hat{\boldsymbol{\beta}}_{WLS}) \to 0$ if Condition UNW1 is satisfied. This is true since

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{b}^T\hat{\boldsymbol{\beta}}_{WLS}) &= \boldsymbol{b}^T\mathrm{Var}(\hat{\boldsymbol{\beta}}_{WLS})\boldsymbol{b} \\
&= \boldsymbol{b}^T((\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1})\boldsymbol{b} \\
&= O(\lambda_{max}((\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1})). \qquad \text{(A.33)}
\end{aligned}$$

For $\lambda_{max}((\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1})$,

$$\lambda_{max}((\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}) \leq \frac{\lambda_{max}(\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X})}{\lambda^2_{min}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})}. \qquad \text{(A.34)}$$

For the middle part of the sandwich type variance-covariance matrix, $\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X} = r^2\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\pi_i^2 \leq M_x\sum_{i=1}^n r^2\pi_i^2 < M_x\sum_{i=1}^n r\pi_i = M_x r$. In Condition UNW1, $1-\alpha-2\beta<0$ implies that we have $\lambda_{max}((\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-2}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}) \to 0$, as $n \to \infty$. □

A.2.2 Proof of Theorem 5

Proof. The proof here is similar to that of Theorem 1.

First, we verify the magnitude of δ^u by examining order of $(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{K}-\boldsymbol{\Omega}^{-1})\mathbf{X}$.

$$\begin{aligned}
\mathrm{Var}(\sum_{i=1}^n \tilde{X}_{si}(\mathbf{K}-\boldsymbol{\Omega}^{-1})X_{it}) &= \sum_{i=1}^n r\pi_i(1-\pi_i)\tilde{X}_{si}^2X_{it}^2 - 2\sum_{i<j} r\pi_i\pi_j\tilde{X}_{si}X_{it}\tilde{X}_{sj}X_{jt} \\
&= \sum_{i=1}^n r\pi_i\tilde{X}_{si}^2X_{it}^2 - (\sum_{i=1}^n \pi_i\tilde{X}_{si}X_{it})^2 \\
&= O(r\sum_{i=1}^n \pi_i\tilde{X}_{si}^2),
\end{aligned}$$

where $\sum_{i=1}^{n} r\pi_i \tilde{X}_{si}^2 \leq \sum_{i=1}^{n} r\pi_i / \lambda_{min}^2(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}) = r/\lambda_{min}^2(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}) \to 0$ as $n \to \infty, r \to \infty$ by assumption. Thus we have implicitly verified that $\delta^u > 0$.

We subtract $\boldsymbol{\beta}_0$ from both side of (A.32) and get

$$\tilde{\boldsymbol{\beta}}^u - \boldsymbol{\beta}_0 \;=\; \hat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{K}\boldsymbol{e}^u + O(1/r^{\delta^u}).$$

We have established the consistency of $\hat{\boldsymbol{\beta}}_{WLS}$ in Lemma 7, the variance of $\boldsymbol{e}^u$ is bounded and $\mathrm{Var}((\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{K}-\mathbf{\Omega}^{-1})\boldsymbol{e}^u)$ is of the same order of $(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{K}-\mathbf{\Omega}^{-1})\mathbf{X}$, which is $O(M_x \sum_{i=1}^{n} r\pi_i/\lambda_{min}^2(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}))$. For $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{WLS}-\boldsymbol{\beta}_0)$, we have shown that $O(\mathrm{Var}(\hat{\boldsymbol{\beta}}_{WLS}-\boldsymbol{\beta}_0)) = O(M_x \sum_{i=1}^{n} r^2\pi_i^2/\lambda_{min}^2(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X}))$. By $\alpha+\gamma_1 > 1$ in Condition UNW1, $r\pi_{max} \to 0$, thus $\mathrm{Var}((\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{K}-\mathbf{\Omega}^{-1})\boldsymbol{e}^u)$ is of lower order of $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{WLS}-\boldsymbol{\beta}_0)$.

Then we adopt similar approach in the proof of Theorem 1 and substitute $\boldsymbol{e}^u$ and $\boldsymbol{K}$ with $\boldsymbol{\epsilon}$ and $\boldsymbol{U}$, and prove the conclusion through establishing the asymptotic normality by Cramer-Wold Device. That is, for any $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{b} \neq \mathbf{0}$, we examine

$$\boldsymbol{b}^T(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{U}\boldsymbol{\varepsilon}. \tag{A.35}$$

We also define independent random variables $\zeta_i = U_i\varepsilon_i/\sqrt{r\pi_i + r^2\pi_i^2}$ such that $\mathrm{E}(\zeta_i) = 0$ and $\mathrm{Var}(\zeta_i) = \sigma^2$. We then have

$$\boldsymbol{b}^T(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{U}\boldsymbol{\varepsilon} \;=\; \sum_{i=1}^{n} d_i^u\zeta_i,$$

where $d_i^u = \boldsymbol{b}^T(\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\boldsymbol{x}_i\sqrt{r\pi_i + r^2\pi_i^2}$. To verify the Noether condition, we examine

$$\max_{1\leq i\leq n}(d_i^u)^2 \;\leq\; (r\pi_{max} + r^2\pi_{max}^2)\boldsymbol{a}_u^T\boldsymbol{a}_u M_x, \tag{A.36}$$

where $\boldsymbol{a}_u = (\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\boldsymbol{b}$. We also have

$$\begin{aligned}\sum_{i=1}^{n}(d_i^u)^2 &= \sum_{i=1}^{n}\boldsymbol{a}_u{}^T\boldsymbol{x}_i(r\pi_i + r^2\pi_i^2)\boldsymbol{x}_i^T\boldsymbol{a}_u \\ &\geq (r\pi_{min} + r^2\pi_{min}^2)\boldsymbol{a}_u^T\mathbf{X}^T\mathbf{X}\boldsymbol{a}_u \\ &\geq (nr\pi_{min} + nr^2\pi_{min}^2)\lambda_0\boldsymbol{a}_u^T\boldsymbol{a}_u, \end{aligned} \tag{A.37}$$

where λ_0 is the the smallest eigenvalue of $\mathbf{X}\mathbf{X}^T/n$ as defined in the proof of Theorem 1. Combining (A.36), (A.37), and Condition UW1 in Theorem 5, we get

$$\lim_{n\to\infty}\frac{(d_{max}^u)^2}{\sum_{i=1}^{n}(d_i^u)^2} \leq \lim_{n\to\infty}\frac{M_x}{\lambda_0}\left(\frac{r\pi_{max} + r^2\pi_{max}^2}{nr\pi_{min} + nr^2\pi_{min}^2}\right) < \frac{M_x}{\lambda_0}\lim_{n\to\infty}\left(\frac{\pi_{max}}{nr\pi_{min}^2}\right) = 0, \tag{A.38}$$

since $r\pi_{min} \to 0$ as $n \to \infty, r \to \infty$. We thus verified the Noether condition for Equation (A.36).

The rest of the proof is analogous to that of Theorem 1. □

Appendix B

Proofs of Bayesian Spline Smoothing with Ambiguous Penalties

B.1 Proof of Theorem 3

To facilitate the final proof of Theorem 3, we first state two useful lemmas.

Lemma 8. *For a matrix $\boldsymbol{X}$ of dimension $n \times m$ and of full column rank, we define a matrix $\boldsymbol{R}$ of dimension $n \times (n-m)$ such that $\boldsymbol{X}^T\boldsymbol{R} = \boldsymbol{0}$ and $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$. We let $\boldsymbol{\Sigma}$ be a positive definite matrix, then*

$$\begin{aligned} & \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1} \\ = \;& \boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R})^{-1}\boldsymbol{R}^T = \boldsymbol{R}(\lambda\boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{R} + \boldsymbol{I})^{-1}\boldsymbol{R}^T. \end{aligned} \tag{B.1}$$

Proof. To see (B.1), we let $\boldsymbol{K}_1 = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{K}_2 = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. Note that both $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are idempotent, and that $\boldsymbol{K}_1\boldsymbol{K}_2 = \boldsymbol{0}$, thus $\boldsymbol{K}_1 + \boldsymbol{K}_2$ is idempotent. Since $rank(\boldsymbol{K}_1 + \boldsymbol{K}_2) = trace(\boldsymbol{K}_1 + \boldsymbol{K}_2) = rank(\boldsymbol{K}_1) + rank(\boldsymbol{K}_2) = n - m + m = n$, we know that $\boldsymbol{K}_1 + \boldsymbol{K}_2 = \boldsymbol{I}$, i.e.,

$$\boldsymbol{I} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{\Sigma}^{\frac{1}{2}} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}. \tag{B.2}$$

Then we pre and post multiply both sides of (B.2) by $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and get (B.1). □

Lemma 9. *Given a $n \times p$ matrix $\boldsymbol{S}$ and a positive definite $n \times n$ matrix $\tilde{\boldsymbol{M}}_1$, we let $\tilde{\boldsymbol{H}}_1 = \tilde{\boldsymbol{M}}_1 - \tilde{\boldsymbol{M}}_1\boldsymbol{S}(\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1$. Suppose we have another $n \times p_1$ matrix $\boldsymbol{S}_1$ of full column*

rank, if the columns in $\boldsymbol{S}$ *and* $\boldsymbol{S}_1$ *are linearly independent, then matrix* $\boldsymbol{S}_1^T\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1$ *is positive definite.*

Proof. Note that $[\ \boldsymbol{S}\ \ \boldsymbol{S}_1\ ]^T\tilde{\boldsymbol{M}}_1[\ \boldsymbol{S}\ \ \boldsymbol{S}_1\ ] = \begin{bmatrix} \boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S} & \boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}_1 \\ \boldsymbol{S}_1^T\tilde{\boldsymbol{M}}_1\boldsymbol{S} & \boldsymbol{S}_1^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}_1 \end{bmatrix}$ is positive definite. Let $\boldsymbol{A}_{11} = \boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}$, $\boldsymbol{A}_{12} = \boldsymbol{A}_{21}^T = \boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}_1$, $\boldsymbol{A}_{22} = \boldsymbol{S}_1^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}_1$. Then, $\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ is positive definite. However, $\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = \boldsymbol{S}_1^T\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1$.

□

In the following, we restate the model and prior settings in the main text and provide the proof to Theorem 1. In particular, we incorporate the case of repeated observations at distinct sample points, i.e.,

$$Y_{ij} = \eta(x_i) + \epsilon_{ij}, \tag{B.3}$$

where $i = 1\ldots, n$, $j = 1, \ldots, n_i$, and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. We denote $\boldsymbol{y} = (\bar{Y}_1, \ldots, \bar{Y}_n)^T$, where $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$. Further, let $SSE = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ and $N = \sum_{i=1}^{n} n_i$.

On n distinct sample points, we evaluate $n\times p$ matrix $\boldsymbol{S}$ of full column rank, $n\times p_0$ matrix $\boldsymbol{S}_0$ of full column rank, $n \times p_1$ matrix $\boldsymbol{S}_1$ of full column rank, $n \times n$ nonnegative definite matrix $\boldsymbol{Q}_0$ with rank $k_0 \leq n$, $n \times n$ nonnegative definite matrix $\boldsymbol{Q}_1$ with rank $k_1 \leq n$, as described in Section 2 in the main text. By construction, the columns of $\boldsymbol{S}$, $\boldsymbol{S}_0$, and $\boldsymbol{S}_1$ are linearly independent. To accommodate the possibility of reduced rank of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_0$, we let $\boldsymbol{V}_1$ and $\boldsymbol{V}_0$ be full column rank $n \times k_1$ and $n \times k_0$ matrices such that $\boldsymbol{Q}_1 = \boldsymbol{V}_1\boldsymbol{V}_1^T$ and $\boldsymbol{Q}_0 = \boldsymbol{V}_0\boldsymbol{V}_0^T$. We use $\boldsymbol{W}$ to denote the n-dimensional diagonal matrix with the i-th diagonal element being $1/n_i$, and rewrite the model on n sample points as

$$\boldsymbol{y}|\boldsymbol{\eta}, \sigma^2 \sim \boldsymbol{N}(\boldsymbol{\eta}, \sigma^2 \boldsymbol{W}), \tag{B.4}$$

$$\text{Given } \boldsymbol{d}_1, \boldsymbol{d}_0, \boldsymbol{z}_1, \boldsymbol{z}_0, \nu, \text{ we let } \boldsymbol{\eta} = \begin{cases} \boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_1\boldsymbol{d}_1 + \boldsymbol{V}_1\boldsymbol{z}_1, \text{ with probability } \nu, \\ \boldsymbol{S}\boldsymbol{d} + \boldsymbol{S}_0\boldsymbol{d}_0 + \boldsymbol{V}_0\boldsymbol{z}_0, \text{ with probability } 1-\nu, \end{cases} \tag{B.5}$$

$$\boldsymbol{z}_u | b_u \sim \boldsymbol{N}(\boldsymbol{0}, b_u \boldsymbol{I}), u = 0, 1, \tag{B.6}$$

$$\boldsymbol{d}_u | g_u, \sigma^2 \sim \boldsymbol{N}(\boldsymbol{0}, g_u \sigma^2 (\boldsymbol{S}_u^T \boldsymbol{S}_u)^{-1}), u = 0, 1, \tag{B.7}$$

$$g_u \sim \text{Inv-Gamma}(\alpha_u, \beta_u), u = 0, 1, \tag{B.8}$$

$$f(\boldsymbol{d}, \sigma^2, \tau_0^2, \tau_1^2, b_0, b_1, \nu) \propto \frac{1_{\nu \in [0,1]} 1_{\sigma^2 \in (0,\infty)} 1_{b_0 \in (0,\infty)} 1_{b_1 \in (0,\infty)} 1_{\nu \in [0,1]} 1_{\boldsymbol{d} \in \mathscr{R}^p}}{(\sigma^2)^s (b_1 + b_0)^a}, \tag{B.9}$$

where we use Inv-Gamma(α_u, β_u) to denote inverse gamma distribution with shape parameter α_u and scale parameter β_u, α_0, α_1, β_0, β_1, and a, s are all given hyperparameters.

We outline the sufficient conditions for the resulting posterior distribution from model in (B.3) and prior settings in (B.5) to (B.9) to be proper on the sample points. In particular, we consider two disjoint cases.

Case 1: $SSE > 0$. The conditions for propriety are

(1.1) $1 < a < 2$,

(1.2) $N - p + 2a + 2s > 6$,

(1.3) $\min\{2k_1 - p_1, 2k_0 - p_0\} - 2p + 4a > 8$.

Case 2: $SSE = 0$. The conditions for propriety are

(2.1) Same as (1.1),

(2.2) Same as (1.2),

(2.3) For u=1,0, we further need to consider two cases:

(a) if $Colsp(\boldsymbol{S}_{(u)}) \bigoplus Colsp(\boldsymbol{Q}_u) \subset \mathscr{R}^n$, $2(k_u - p) - p_u + 4a > 8$;

(b) if $Colsp(\boldsymbol{S}_{(u)}) \bigoplus Colsp(\boldsymbol{Q}_u) = \mathscr{R}^n$, $2(n - k_u) + p_u + 4s < 4$.

Proof. To begin with, we note

$$\int f(\boldsymbol{d}, \boldsymbol{d}_0, \boldsymbol{d}_1, \sigma^2, g_0, g_1, b_0, b_1, \boldsymbol{\eta}, \nu | \boldsymbol{Y}) \mathrm{d}\boldsymbol{z}_1 \mathrm{d}\boldsymbol{z}_0 \mathrm{d}\boldsymbol{\eta} \mathrm{d}\nu = C_0(f_0 + f_1),$$

where C_0 is a positive constant that does not concern parameters,

$$\begin{aligned} f_0 &= \frac{1}{\sqrt{|b_0\boldsymbol{Q}_0 + \sigma^2\boldsymbol{W}|}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_0\boldsymbol{d}_0)^T(b_0\boldsymbol{Q}_0 + \sigma^2\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_0\boldsymbol{d}_0)}{2} - \frac{SSE}{2\sigma^2}\right) \\ &\quad \frac{1}{(g_1\sigma^2)^{p_1/2}} \exp\left(-\frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)\boldsymbol{d}_1}{2g_1\sigma^2}\right) \frac{1}{(g_0\sigma^2)^{p_0/2}} \exp\left(-\frac{\boldsymbol{d}_0^T(\boldsymbol{S}_0^T\boldsymbol{S}_0)\boldsymbol{d}_0}{2g_0\sigma^2}\right) \\ &\quad \frac{1}{(\sigma^2)^{\frac{\sum_{i=1}^n(n_i-1)}{2}}(\sigma^2)^s(b_1+b_0)^a} g_1^{-\alpha_1-1} \exp\left(-\frac{\beta_1}{g_1}\right) g_0^{-\alpha_0-1} \exp\left(-\frac{\beta_0}{g_0}\right), \qquad \text{(B.10)} \\ f_1 &= \frac{1}{\sqrt{|b_1\boldsymbol{Q}_1 + \sigma^2\boldsymbol{W}|}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_1\boldsymbol{d}_1)^T(b_1\boldsymbol{Q}_1 + \sigma^2\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_1\boldsymbol{d}_1)}{2} - \frac{SSE}{2\sigma^2}\right) \\ &\quad \frac{1}{(g_1\sigma^2)^{p_1/2}} \exp\left(-\frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)\boldsymbol{d}_1}{2g_1\sigma^2}\right) \frac{1}{(g_0\sigma^2)^{p_0/2}} \exp\left(-\frac{\boldsymbol{d}_0^T(\boldsymbol{S}_0^T\boldsymbol{S}_0)\boldsymbol{d}_0}{2g_0\sigma^2}\right) \\ &\quad \frac{1}{(\sigma^2)^{\frac{\sum_{i=1}^n(n_i-1)}{2}}(\sigma^2)^s(b_1+b_0)^a} g_1^{-\alpha_1-1} \exp\left(-\frac{\beta_1}{g_1}\right) g_0^{-\alpha_0-1} \exp\left(-\frac{\beta_0}{g_0}\right). \qquad \text{(B.11)} \end{aligned}$$

We only show $\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}g_1 \mathrm{d}\boldsymbol{d}_1 \mathrm{d}\sigma^2 \mathrm{d}b_1 < \infty$ below. A similar argument can be made for f_0.

First, we integrate out $\boldsymbol{d}_0$, g_0, b_0 from right-hand side of (B.11). Let us denote $\tilde{\boldsymbol{M}}_1 = (b_1\boldsymbol{Q}_1 + \sigma^2\boldsymbol{W})^{-1}$. Given $a > 1$ in Condition (1.1) and Condition (2.1) we have

$$\begin{aligned} &\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \\ &= \frac{C}{\sqrt{|\tilde{\boldsymbol{M}}_1^{-1}|}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_1\boldsymbol{d}_1)^T\tilde{\boldsymbol{M}}_1(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{d} - \boldsymbol{S}_1\boldsymbol{d}_1)}{2} - \frac{SSE}{2\sigma^2}\right) \\ &\quad \frac{1}{(g_1\sigma^2)^{p_1/2}} \exp\left(-\frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)\boldsymbol{d}_1}{2g_1\sigma^2}\right) \frac{1}{(\sigma^2)^{s+\frac{N-n}{2}}(b_1)^{a-1}} g_1^{-\alpha_1-1} \exp\left(-\frac{\beta_1}{g_1}\right), \qquad \text{(B.12)} \end{aligned}$$

where C is a positive constant that does not concern parameters.

Second, we integrate out $\boldsymbol{d}$ from the right-hand side of (B.12).

$$\begin{aligned}
&\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}\boldsymbol{d} \\
= \quad & \frac{C}{\sqrt{|\tilde{\boldsymbol{M}}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}|}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)^T(\tilde{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1\boldsymbol{S}(\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1)(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)}{2} - \frac{SSE}{2\sigma^2}\right) \\
& \frac{1}{(g_1\sigma^2)^{p_1/2}} \exp\left(-\frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)\boldsymbol{d}_1}{2g_1\sigma^2}\right) \frac{1}{(\sigma^2)^{s+\frac{N-n}{2}}(b_1)^{a-1}} g_1^{-\alpha_1-1} \exp\left(-\frac{\beta_1}{g_1}\right). \qquad \text{(B.13)}
\end{aligned}$$

Third, we integrate out g_1 from the right-hand side of (B.13).

$$\begin{aligned}
&\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}\boldsymbol{d} \mathrm{d}g_1 \\
= \quad & \frac{C}{\sqrt{|\tilde{\boldsymbol{M}}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}|}} \frac{1}{(\frac{\boldsymbol{d}_1^T\boldsymbol{S}_1^T\boldsymbol{S}_1\boldsymbol{d}_1}{2\sigma^2}+\beta_1)^{\frac{p_1+2\alpha_1}{2}}(\sigma^2)^{s+\frac{N-n+p_1}{2}}(b_1)^{a-1}} \\
& \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)^T(\tilde{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1\boldsymbol{S}(\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1)(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)}{2} - \frac{SSE}{2\sigma^2}\right) \qquad \text{(B.14)}
\end{aligned}$$

Fourth, we integrate out $\boldsymbol{d}_1$ from the right-hand side of (B.14).

$$\begin{aligned}
&\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}g_1 \mathrm{d}\boldsymbol{d}_1 \\
\leq \quad & \frac{C}{\sqrt{|\tilde{\boldsymbol{M}}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}|}(\sigma^2)^{s+\frac{N-n+p_1}{2}}(b_1)^{a-1}} \sqrt{\int \frac{1}{(\frac{\boldsymbol{d}_1^T\boldsymbol{S}_1^T\boldsymbol{S}_1\boldsymbol{d}_1}{2\sigma^2}+\beta_1)^{(p_1+2\alpha_1)}} \mathrm{d}\boldsymbol{d}_1} \\
& \sqrt{\int \exp\left(-\frac{2(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)^T(\tilde{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1\boldsymbol{S}(\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1)(\boldsymbol{y}-\boldsymbol{S}_1\boldsymbol{d}_1)}{2} - \frac{SSE}{\sigma^2}\right) \mathrm{d}\boldsymbol{d}_1} \qquad \text{(B.15)} \\
= \quad & \frac{C}{\sqrt{|\tilde{\boldsymbol{M}}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S}|}\sqrt[4]{|\boldsymbol{S}_1^T\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1|}(\sigma^2)^{s+\frac{N-n}{2}+\frac{p_1}{4}}(b_1)^{a-1}} \\
& \exp\left(-\frac{\boldsymbol{y}^T(\tilde{\boldsymbol{H}}_1-\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1(\boldsymbol{S}_1^T\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1\tilde{\boldsymbol{H}}_1)\boldsymbol{y}}{2} - \frac{SSE}{2\sigma^2}\right), \qquad \text{(B.16)}
\end{aligned}$$

where we use the Cauchy-Schwartz Inequality in (B.15), use the kernel of Multivariate t-

distribution in (B.16), and denote $\tilde{\boldsymbol{H}}_1 = \tilde{\boldsymbol{M}}_1 - \tilde{\boldsymbol{M}}_1\boldsymbol{S}(\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\tilde{\boldsymbol{M}}_1$. From Lemma 9, we know that the matrix $\boldsymbol{S}_1^T\tilde{\boldsymbol{H}}_1\boldsymbol{S}_1$ is positive definite in (B.16).

Fifth, we intend to integrate out σ^2 from the right-hand side of (B.16). For this purpose, we reparameterize (σ^2, b_1) to (σ^2, λ_1), where $b_1 = \lambda_1\sigma^2$. Denote $\boldsymbol{M}_1 = (\lambda_1\boldsymbol{Q}_1 + \boldsymbol{W})^{-1}$ and $\boldsymbol{H}_1 = \boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{M}_1$. Recall that $\boldsymbol{S}_{(1)} = (\boldsymbol{S}, \boldsymbol{S}_1)$, and it is trivial to verify that $\boldsymbol{H}_1 - \boldsymbol{H}_1\boldsymbol{S}_1(\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^T\boldsymbol{H}_1 = \boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1$. Using the (σ^2, λ_1) parameterization, we have

$$\begin{aligned}
&\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}g_1 \mathrm{d}\boldsymbol{d}_1 \\
\leq\ & \frac{C}{\sqrt{|\boldsymbol{M}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}|}\sqrt[4]{|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1|}(\sigma^2)^{s+a+\frac{N-p}{2}-2}(\lambda_1)^{a-1}} \\
& \exp\left(-\frac{\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1)\boldsymbol{y}}{2\sigma^2} - \frac{SSE}{2\sigma^2}\right). \qquad \text{(B.17)}
\end{aligned}$$

Assuming $s + a + \frac{N-p}{2} - 2 > 1$, i.e., $N - p + 2a + 2s > 6$ in Condition (2.1) and (2.2), we now proceed to integrate out σ^2.

$$\begin{aligned}
&\int f_1 \mathrm{d}\boldsymbol{d}_0 \mathrm{d}g_0 \mathrm{d}b_0 \mathrm{d}g_1 \mathrm{d}\boldsymbol{d}_1 \mathrm{d}\sigma^2 \\
\leq\ & \frac{C}{\sqrt{|\boldsymbol{M}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}|}\sqrt[4]{|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1|}(\lambda_1)^{a-1}} \\
& \left(\frac{\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1)\boldsymbol{y} + SSE}{2}\right)^{-(s+a+\frac{N-p}{2}-3)}. \qquad \text{(B.18)}
\end{aligned}$$

Finally, we integrate out λ_1 from the right-hand side of (B.18).

1. $SSE > 0$.

   In this case, the term $\left(\frac{\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1)\boldsymbol{y} + SSE}{2}\right)^{-(s+a+\frac{N-p}{2}-3)}$ is a bounded function of λ_1.

   For $|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}|$, we notice that there exist some positive constants C_1 and ξ_1 such that $|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}| \geq C_1(1 + \xi_1\lambda_1)^p$. Next we consider $|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1|$. We define $\boldsymbol{R}$ as $n \times (n - p)$

matrix such that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, $\boldsymbol{R}^T\boldsymbol{S} = \boldsymbol{0}$. Note $\boldsymbol{H}_1 = \boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{M}_1 = \boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{M}_1^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T$ by (B.1) in Lemma 8.

Then, $|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1| = |\boldsymbol{S}_1^T\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{M}_1^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{S}_1| = |\boldsymbol{S}_1^T\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{W}\boldsymbol{R}+\lambda_1\boldsymbol{R}^T\boldsymbol{Q}_1\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{S}_1|$. It is obvious that $\boldsymbol{R}^T\boldsymbol{W}\boldsymbol{R}$ is positive definite, there exists $(n-p)\times(n-p)$ positive definite matrix $\boldsymbol{U}$ such that $\boldsymbol{R}^T\boldsymbol{W}\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}^T$. Considering that the columns of $\boldsymbol{S}$ and $\boldsymbol{S}_1$ are linearly independent, the matrix $\boldsymbol{S}_1^T\boldsymbol{R}$ is of full row rank. (If not, then there exists $\boldsymbol{c}$ such that $\boldsymbol{R}^T\boldsymbol{S}_1\boldsymbol{c} = \boldsymbol{0}$. Moreover, by construction, $\boldsymbol{R}$ is such that $\boldsymbol{R}^T\boldsymbol{S} = \boldsymbol{0}$. Thus $\boldsymbol{S}_1\boldsymbol{c}$ is in the column space of $\boldsymbol{S}$, which is a contradiction.) Thus there exist some constants C_2, and ξ_2 such that

$$\begin{aligned} |\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1| &= |\boldsymbol{S}_1^T\boldsymbol{R}(\boldsymbol{U}^T)^{-1}(\boldsymbol{I}+\lambda_1\boldsymbol{U}^{-1}\boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{R}(\boldsymbol{U}^T)^{-1})^{-1}\boldsymbol{U}^{-1}\boldsymbol{R}^T\boldsymbol{S}_1| \\ &\geq C_2(1+\xi_2\lambda_1)^{-p_1}. \end{aligned}$$

Furthermore, there exist some positive constants C and ξ such that

$$\int f_1 \mathrm{d}\boldsymbol{d}_0\mathrm{d}g_0\mathrm{d}b_0\mathrm{d}g_1\mathrm{d}\boldsymbol{d}_1\mathrm{d}\sigma^2\mathrm{d}\lambda_1 \tag{B.19}$$

$$\leq \int \frac{C}{\sqrt{|\boldsymbol{M}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}|}\sqrt[4]{|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1|}(\lambda_1)^{a-1}}\mathrm{d}\lambda_1 \tag{B.20}$$

$$\leq \int \frac{C}{(1+\xi\lambda_1)^{\frac{k_1}{2}-\frac{p}{2}-\frac{p_1}{4}}(\lambda_1)^{a-1}}\mathrm{d}\lambda_1. \tag{B.21}$$

For the integral on the right-hand side of (B.21) to be proper, we need $a-1<1$, i.e., $a<2$ and $\frac{k_1}{2}-\frac{p}{2}-\frac{p_1}{4}+a-1>1$, i.e., $2(k_1-p)-p_1+4a>8$.

2. SSE=0, and $Colsp(\boldsymbol{S}_{(1)})\bigoplus Colsp(\boldsymbol{V}_1)\subset\mathscr{R}^n$.

Recall that we let $\boldsymbol{V}_1$ be a full column rank matrix of dimension $n\times k_1$ such that $\boldsymbol{Q}_1 = \boldsymbol{V}_1\boldsymbol{V}_1^T$. When SSE=0, it is obvious that $N=n$ and $\boldsymbol{W}=\boldsymbol{I}$. In this case, the quadratic term $\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1)\boldsymbol{y}$ requires more investigation. It is noted that in this case $k_1 < n-p-p_1$. We denote a matrix $\boldsymbol{R}_1$ of

dimension $n \times (n-p-p_1)$ such that $\boldsymbol{R}_1^T\boldsymbol{R}_1 = \boldsymbol{I}$ and $\boldsymbol{R}_1^T\boldsymbol{S}_{(1)} = \boldsymbol{0}$. Then there exist $(p+p_1) \times k_1$ and $(n-p-p_1) \times k_1$ dimensional matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ such that $\boldsymbol{V}_1 = \boldsymbol{S}_{(1)}\boldsymbol{T}_1 + \boldsymbol{R}_1\boldsymbol{T}_2$. Then $\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{R}_1^T\boldsymbol{V}_1\boldsymbol{V}_1^T\boldsymbol{R}_1 = \boldsymbol{T}_2\boldsymbol{T}_2^T$ is rank deficient, i.e., there exist constants k_j and orthonomal vectors $\boldsymbol{\alpha}_j$, $j = 1, \ldots, l$, $l < (n-p-p_1)$, such that $\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 = \sum_{j=1}^{l} k_j\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^T$. We further define orthonormal vectors $\boldsymbol{\alpha}_{j'}$, $j' = l+1, \ldots, n-p-p_1$, such that $\boldsymbol{\alpha}_{j'}^T\boldsymbol{\alpha}_j = 0, \forall j' = l+1, \ldots, n-p-p_1$ and $j = 1, \ldots, l$. Thus,

$$
\begin{aligned}
& \boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}\boldsymbol{M}_1)\boldsymbol{y} \\
= \; & \boldsymbol{y}^T\boldsymbol{R}_1(\lambda_1\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 + \boldsymbol{I})^{-1}\boldsymbol{R}_1^T\boldsymbol{y}^T \\
= \; & \boldsymbol{y}^T\boldsymbol{R}_1\left(\sum_{j=1}^{l}(\lambda_j k_j + 1)\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^T + \sum_{j'=l+1}^{n-p-p_1}\boldsymbol{\alpha}_{j'}\boldsymbol{\alpha}_{j'}^T\right)^{-1}\boldsymbol{R}_1^T\boldsymbol{y}^T \\
> \; & \boldsymbol{y}^T\boldsymbol{R}_1\left(\sum_{j'=l+1}^{n-p-p_1}\boldsymbol{\alpha}_{j'}\boldsymbol{\alpha}_{j'}^T\right)\boldsymbol{R}_1^T\boldsymbol{y}.
\end{aligned}
$$

Thus $\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}\boldsymbol{M}_1)\boldsymbol{y} > 0$ with probability 1. It reduces to the same case as in $SSE > 0$ and the discussion from (B.19) to (B.21) follows. The integral on the right-hand side of (B.18) is proper if $a < 2$ and $2(k_1 - p) - p_1 + 4a > 8$.

3. $SSE = 0$, and $Colsp(\boldsymbol{S}_{(1)}) \bigoplus Colsp(\boldsymbol{V}_1) = \mathscr{R}^n$.

   It is noted that in this case $k_1 \geq n-p-p_1$. Recall that we defined $\boldsymbol{R}_1$ as an $n \times (n-p-p_1)$ matrix such that $\boldsymbol{R}_1^T\boldsymbol{R}_1 = \boldsymbol{I}$ and $\boldsymbol{R}_1^T\boldsymbol{S}_{(1)} = \boldsymbol{0}$. By (B.1),

$$
\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1 = \boldsymbol{R}_1(\lambda_1\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 + \boldsymbol{I})^{-1}\boldsymbol{R}_1^T. \quad \text{(B.22)}
$$

   Next, we show that matrix $\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1$ is of full rank. Since $Colsp(\boldsymbol{S}_{(1)}) \bigoplus Colsp(\boldsymbol{V}_1) = \mathscr{R}^n$, then there exist $(p+p_1) \times k_1$ and $(n-p-p_1) \times k_1$ matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ such that $\boldsymbol{R}_1 = \boldsymbol{S}_{(1)}\boldsymbol{A}_1 + \boldsymbol{V}_1\boldsymbol{A}_2$. Further, $\boldsymbol{R}_1^T\boldsymbol{R}_1 = \boldsymbol{R}_1^T\boldsymbol{V}_1\boldsymbol{A}_2$, thus we know that $\text{rank}(\boldsymbol{R}_1^T\boldsymbol{R}_1) = (n-p-p_1) \leq \text{rank}(\boldsymbol{R}_1^T\boldsymbol{V}_1) \leq (n-p-p_1)$, i.e., rank $(\boldsymbol{R}_1^T\boldsymbol{V}_1) = n-p-p_1$. Then,

$\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1 = (\boldsymbol{R}_1^T\boldsymbol{V}_1)(\boldsymbol{R}_1^T\boldsymbol{V}_1)^T$ is positive definite.

Using (B.22) and the fact that $\boldsymbol{R}_1^T\boldsymbol{Q}_1\boldsymbol{R}_1$ is positive definite, we know that there exist some constants ξ'' and δ'' such that

$$\boldsymbol{y}^T(\boldsymbol{M}_1 - \boldsymbol{M}_1\boldsymbol{S}_{(1)}(\boldsymbol{S}_{(1)}^T\boldsymbol{M}_1\boldsymbol{S}_{(1)})^{-1}\boldsymbol{S}_{(1)}\boldsymbol{M}_1)\boldsymbol{y} \;\geq\; \frac{\delta''}{\xi'' + \lambda_1} \tag{B.23}$$

and further

$$\int f_1 \mathrm{d}\boldsymbol{d}_0\mathrm{d}g_0\mathrm{d}b_0\mathrm{d}g_1\mathrm{d}\boldsymbol{d}_1\mathrm{d}\sigma^2\mathrm{d}\lambda_1 \leq \int_0^\infty \frac{C(\xi'' + \lambda_1)^{(s+a+\frac{N-p}{2}-3)}}{\sqrt{|\boldsymbol{M}_1^{-1}|}\sqrt{|\boldsymbol{S}^T\boldsymbol{M}_1\boldsymbol{S}|}\sqrt[4]{|\boldsymbol{S}_1^T\boldsymbol{H}_1\boldsymbol{S}_1|}(\lambda_1)^{a-1}}\mathrm{d}\lambda_1 \tag{B.24}$$

For the integral on the right-hand side of (B.24) to be proper, we need $a - 1 < 1$, i.e., $a < 2$, and $\frac{k_1}{2} - \frac{p}{2} - \frac{p_1}{4} - (s + a + \frac{n-p}{2} - 3) + a - 1 > 1$, i.e., $2(n - k_1) + p_1 + 4s < 4$.

Combining a similar argument for f_0 in (B.10), the conclusion in Theorem 1 can be obtained.

□

Full conditional distributions in Gibbs sampller

Suppose our current state variable $U = 1$. Gibbs sampler is used to generate sample of $\boldsymbol{\theta}_u$ according to f_u. To do that, we derive the set of full conditional distributions from the posterior joint density of $(\boldsymbol{d}, \boldsymbol{d}_1, \boldsymbol{z}_1, b_1, \sigma^2)$ in the case of no repeated measurements as follows.

- $\boldsymbol{d}|. \sim \boldsymbol{N}((\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T(\boldsymbol{Y} - \boldsymbol{V}_1\boldsymbol{z}_1 - \boldsymbol{S}_1\boldsymbol{d}_1), \sigma^2(\boldsymbol{S}^T\boldsymbol{S})^{-1})$.
- $\boldsymbol{d}_1|. \sim \boldsymbol{N}((1/\sigma^2 + 1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^T(\boldsymbol{y} - \boldsymbol{V}_1\boldsymbol{z}_1 - \boldsymbol{S}\boldsymbol{d})/\sigma^2, (1/\sigma^2 + 1/(g_1\sigma^2))^{-1}(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1})$.
- $\boldsymbol{z}_1|. \sim \boldsymbol{N}((\boldsymbol{I}/b_1 + (\boldsymbol{V}_1^T\boldsymbol{V}_1)/\sigma^2)^{-1}\boldsymbol{V}_1^T(\boldsymbol{y} - \boldsymbol{S}_1\boldsymbol{d}_1 - \boldsymbol{S}\boldsymbol{d})/\sigma^2, (\boldsymbol{I}/b_1 + (\boldsymbol{V}_1^T\boldsymbol{V}_1)/\sigma^2)^{-1})$.
- $g_1|. \sim$ Inv-Gamma$(\alpha_{g_1}, \beta_{g_1})$, where $\alpha_{g_1} = \alpha_u + \frac{p_1}{2}$, $\beta_{g_1} = \beta_u + \frac{\boldsymbol{d}_1^T(\boldsymbol{S}_1^T\boldsymbol{S}_1)^{-1}\boldsymbol{d}_1}{2\sigma^2}$.

- $\sigma^2|. \sim$ Inv-Gamma$(\alpha_{\sigma^2}, \beta_{\sigma^2})$, with $\alpha_{\sigma^2} = \frac{n+p_1}{2}+s-1$, $\beta_{\sigma^2} = \frac{(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_1\boldsymbol{d}_1-\boldsymbol{V}\boldsymbol{z}_1)^T\boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{S}\boldsymbol{d}-\boldsymbol{S}_1\boldsymbol{d}_1-\boldsymbol{V}\boldsymbol{z}_1)}{2}$.
- $f(b_1|.) \propto \frac{1}{(b_1+b_2)^a b_1^{k_1/2}} \exp\left(-\frac{\boldsymbol{v}_1^T\boldsymbol{v}_1}{2b_1}\right)$. In order to take samples from the above distribution, we use the rejection sampling method. We use C_0 to denote a predefined large constant, e.g., 10000,
  - If $b_1/(\frac{\boldsymbol{v}_1^T\boldsymbol{v}_1}{2}/(\frac{k_1}{2}-2)) >= C_0$, we use the proposal distribution Inv-Gamma$(\frac{k_1}{2}-1, \frac{\boldsymbol{v}_1^T\boldsymbol{v}_1}{2})$.
  - If $b_1/(\frac{\boldsymbol{v}_1^T\boldsymbol{v}_1}{2}/(\frac{k_1}{2}-2)) <= C_0$, we use the proposal distribution Inv-Gamma$(\frac{k_1}{2}+a-1, \frac{\boldsymbol{v}_1^T\boldsymbol{v}_1}{2})$.

The case with $U = 0$ can be similarly derived. The initial values of $\boldsymbol{d}$, $\boldsymbol{d}_1$, $\boldsymbol{z}_1$, $\boldsymbol{d}_0$, and $\boldsymbol{z}_0$ can be set as the respective ordinary least squares estimates. The initial values of b_1, b_0, and σ^2 can be provided using a random sample from $Unif(0, 1)$. The initial values of U can be provided using a random sample from $Bernoulli(0.5)$.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Akay, A., T. Di Domenico, K. M. Suen, A. Nabih, G. E. Parada, M. Larance, R. Medhi, A. C. Berkyurek, X. Zhang, C. J. Wedeles, et al. (2017). The helicase aquarius/emb-4 is required to overcome intronic barriers to allow nuclear rnai pathways to heritably silence transcription. *Developmental Cell 42*(3), 241–255.

Ashe, A., A. Sapetschnig, E. Weick, J. Mitchell, M. Bagijn, A. Cording, A. Doebley, L. Goldstein, N. Lehrbach, J. Le Pen, and G. Pintacuda (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell 150*(1), 88–89.

Bagijn, M. P., L. D. Goldstein, A. Sapetschnig, E.-M. Weick, S. Bouasker, N. J. Lehrbach, M. J. Simard, and E. A. Miska (2012). Function, targets, and evolution of caenorhabditis elegans pirnas. *Science 337*(6094), 574–578.

Batista, P. J., J. G. Ruby, J. M. Claycomb, R. Chiang, N. Fahlgren, K. D. Kasschau, D. A. Chaves, W. Gu, J. J. Vasale, S. Duan, et al. (2008). Prg-1 and 21u-rnas interact to form the pirna complex required for fertility in c. elegans. *Molecular Cell 31*(1), 67–78.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*(2), 802–837.

Berry, S. M., R. J. Carroll, and D. Ruppert (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97–457,160–169.

Bertin-Mahieux, T., D. P. Ellis, B. Whitman, and P. Lamere (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley.

Bing, L., K. C. Chan, and C. Ou (2014). Public sentiment analysis in twitter data for prediction of a company's stock price movements. In *e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on*, pp. 232–239. IEEE.

Bové, D. S. and L. Held (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis 6*(3), 387–410.

Broersma, M. and T. Graham (2012). Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice 6*(3), 403–419.

Buckley, B., K. Burkhart, S. Gu, G. Spracklin, A. Kershner, K. Fritz, H., A. J., Fire, and S. Kennedy (2012). A nuclear argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature 489*(7416), 447.

Chen, M.-H. and J. G. Ibrahim (2003). Conjugate priors for generalized linear models. *Statistica Sinica*, 461–476.

Chen, S., R. Varma, A. Singh, and J. Kovačcević (2016). A statistical perspective of sampling scores for linear regression. In *Information Theory (ISIT), 2016 IEEE International Symposium*, pp. 1556–1560. IEEE.

Chen, T., H. L. He, and G. M. Church (1999). Modeling gene expression with differential equations. In *Biocomputing'99*, pp. 29–40. World Scientific.

Christie, M., L. J. Croft, and B. J. Carroll (2011). Intron splicing suppresses rna silencing in arabidopsis. *The Plant Journal 68*(1), 159–167.

Chubb, J. R., T. Trcek, S. M. Shenoy, and R. H. Singer (2006). Transcriptional pulsing of a developmental gene. *Current Biology 16*(10), 1018–1025.

Clarkson, K. L. and D. P. Woodruff (2013). Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 81–90.

Das, P. P., M. P. Bagijn, L. D. Goldstein, J. R. Woolford, N. J. Lehrbach, A. Sapetschnig, H. R. Buhecha, M. J. Gilchrist, K. L. Howe, R. Stark, et al. (2008). Piwi and pirnas act upstream of an endogenous sirna pathway to suppress tc3 transposon mobility in the caenorhabditis elegans germline. *Molecular Cell 31*(1), 79–90.

Drineas, P., M. Magdon-Ismail, M. Mahoney, and D. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research 13*, 3475–3506.

Drineas, P. and M. Mahoney (2016). RandNLA: Randomized Numerical Linear Algebra. *Commun. ACM 59*(6), 80–90.

Drineas, P., M. Mahoney, and S. Muthukrishnan (2006). Sampling algorithms for l_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136.

Drineas, P., M. Mahoney, S. Muthukrishnan, and T. Sarlos (2011). Faster least squares approximation. *Numerische Mathematik 117*, 219–249.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature 391*(6669), 806.

Gu, C. (1992). Penalized likelihood regression: a Bayesian analysis. *Statistica Sinica*, 255–264.

Gu, C. (2013). *Smoothing spline ANOVA models*, Volume 297. Springer Science & Business Media.

Gu, C. (2014). Smoothing spline *anova* models: R package gss. *Journal of Statistical Software 58*(5), 1–25.

Gu, C. and G. Wahba (1991). Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing 12*(2), 383–398.

Helwig, N. E., Y. Gao, S. Wang, and P. Ma (2015). Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance. *Spatial Statistics 14*, 491–504.

Hubbard, J. and B. Hubbard (1999). *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Prentice Hall.

Lai, T. L., H. Robbins, and C. Z. Wei (1978). Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences 75*(7), 3034–3036.

Le Hir, H., J. Saulière, and Z. Wang (2016). The exon junction complex as a node of post-transcriptional networks. *Nature Reviews Molecular Cell Biology 17*(1), 41.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics 44*(3), 907–927.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423.

Luan, Y. and H. Li (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics 20*(3), 332–339.

Luo, Z. and G. Wahba (1997). Hybrid adaptive splines. *Journal of the American Statistical Association 92*(437), 107–116.

Ma, P., M. Mahoney, and B. Yu (2014). A statistical perspective on algorithmic leveraging. In *Proceedings of the 31th ICML Conference*, pp. 91–99.

Ma, P., M. Mahoney, and B. Yu (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research 16*, 861–911.

Mahoney, M. (2011). *Randomized Algorithms for Matrices and Data.* Foundations and Trends in Machine Learning. Boston: NOW Publishers. Also available at: arXiv:1104.5557.

Manning, K., M. Tör, M. Poole, Y. Hong, A. J. Thompson, G. J. King, J. J. Giovannoni, and G. B. Seymour (2006). A naturally occurring epigenetic mutation in a gene encoding an sbp-box transcription factor inhibits tomato fruit ripening. *Nature Genetics 38*(8), 948.

Maruyama, Y. and E. I. George (2011). Fully Bayes factors with a generalized g-prior. *Annals of Statistics 39*(5), 2740–2765.

Meng, X. and M. W. Mahoney (2013). Low-distortion subspace embeddings in input-sparsity

time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 91–100. ACM.

Morris, C. (1975). Central limit theorems for multinomial sums. *Annals of Statistics 3*(1), 165–188.

Murdoch, W. W., C. J. Briggs, and R. M. Nisbet (2003). *Consumer-resource dynamics*. New York: Princeton University Press.

Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays 25*(12), 1147–1149.

Pak, J. and A. Fire (2007). Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science 315*(5809), 241–244.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics.

Raskutti, G. and M. Mahoney (2015). A statistical perspective on randomized sketching for ordinary least-squares. In *Proceedings of the 32th ICML Conference*.

Saleh, A. M. E. (2006). *Theory of Preliminary Test and Stein-type Estimation with Applications*. John Wiley & Sons.

Sapetschnig, A., P. Sarkies, N. J. Lehrbach, and E. A. Miska (2015). Tertiary sirnas mediate paramutation in c. elegans. *PLoS Genetics 11*(3), e1005078.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Seymour, G., M. Poole, K. Manning, and G. J. King (2008). Genetics and epigenetics of fruit development and ripening. *Current Opinion in Plant Biology 11*(1), 58–63.

Shirayama, M., M. Seth, H. Lee, W. Gu, T. Ishidate, D. Conte Jr, and C. Mello (2012). pirnas initiate an epigenetic memory of nonself rna in the c. elegans germline. *Cell 150*(1), 65–77.

Sklar, J. C., J. Wu, W. Meiring, and Y. Wang (2013). Nonparametric regression with basis selection from multiple libraries. *Technometrics 55*(2), 189–201.

Tiao, G. C. and A. Zellner (1964a). Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika 51*(1/2), 219–230.

Tiao, G. C. and A. Zellner (1964b). On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society. Series B*, 277–285.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association 111*(514), 600–620.

Tyc, K. M., A. Nabih, M. Z. Wu, C. J. Wedeles, J. A. Sobotka, and J. M. Claycomb (2017). The conserved intron binding protein emb-4 plays differential roles in germline small rna pathways of c. elegans. *Developmental Cell 42*(3), 256–270.

USDOT, B. (2017). Rita airline delay data download. `https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236`.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B*, 364–372.

Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 1378–1402.

Wahba, G. (1990). *Spline models for observational data*, Volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Wahba, G. and P. Craven (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–404.

Wahl, M. C. and R. Lührmann (2015). Snapshot: spliceosome dynamics i. *Cell 161*(6), 1474–1474.

Wang, G. and V. Reinke (2008). A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Current Biology 18*(12), 861–867.

Wang, Y. (2011). *Smoothing splines: methods and applications*. Boca Raton: Chapman and Hall.

Wang, Y., A. W. Yu, and A. Singh (2016). On computationally tractable selection of experiments in regression models. *ArXiv e-prints, 1601.02068*.

Weick, E.-M. and E. A. Miska (2014). pirnas: from biogenesis to function. *Development 141*(18), 3458–3471.

Whitfield, M. L., G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell 13*(6), 1977–2000.

Woodruff, D. P. et al. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science 10*(1–2), 1–157.

Yang, T., L. Zhang, R. Jin, and S. Zhu (2015). An explicit sampling dependent spectral error bound for column subset selection. In *International Conference on Machine Learning*, pp. 135–143.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, Volume 6, pp. 233–243.

Zhang, Q., Y. Yu, J. Zhang, and H. Liang (2018). Using single-index odes to study dynamic gene regulatory network. *PloS ONE 13*(2), e0192833.

Zhong, S., Z. Fei, Y.-R. Chen, Y. Zheng, M. Huang, J. Vrebalov, R. McQuinn, N. Gapper, B. Liu, J. Xiang, et al. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nature Biotechnology 31*(2), 154.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.