Pretest and Stein-type Shrinkage Estimators in Linear and Generalized Partial Linear Models

by

Le An Lac

A thesis submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment for the degree of

MASTER OF SCIENCE

Department of Statistics

The University of Manitoba

Winnipeg

Abstract

In this thesis, we consider two estimation problems of the regression parameters in generalized partial linear regression model and multiple linear regression model with many covariates. We consider the situation where some of the regression parameters may be suspected to satisfy some restrictions and the nonparametric part is considered as nuisance.

We first propose some novel and improved methods to estimate the regression coefficients of generalized partial linear models (GPLM). This model extends the generalized linear model by adding a nonparametric component. Like parametric models, variable selection is important in the GPLM to single out the inactive covariates. Instead of deleting inactive covariates, we use them as an auxiliary information. We define two models, the unrestricted model includes all the covariates whereas the restricted one includes the active covariates only. We then combine these two model estimators optimally to form the pretest and shrinkage estimators. We study the asymptotic properties to derive the asymptotic biases and risks of the estimators. We show that the asymptotic risks of the shrinkage estimators are strictly less than that of the full model estimators. A simulation study is conducted to assess the performance of the proposed estimators. We then apply our proposed methods to analyze a real credit scoring data. Both simulation study and real data example corroborate with the theoretical result.

Optimal design plays an important role in achieving good estimation of the parameters. Motivated by this fact, we propose another novel method to further improve the pretest and shrinkage estimators. The results are very promising. Apart from the modeling and post-modeling procedures, pre-modeling stage plays a key role in achieving efficient estimators of the parameters. The optimal combinations of values of inputs which are normally numeric must be chosen before running an experiment. We consider the most popular D-optimality criterion and construct the optimal design using a class of algorithms. We then generate the data according to the optimal design and finally obtain our pretest and shrinkage estimators in multiple linear regression models. Our studies evidently show that our proposed estimators using optimal design theory outperform the estimators without using optimal design.

Acknowledgements

First and foremost, I would like to sincerely express my deepest gratitude to my supervisors, Professor Saumen Mandal and Professor Shakhawat Hossain, who have the substance and attitude of a genius. I am thankful for their continuing encouragement, persistent support, and great advice throughout my studies. The door to their assistance are always open whenever I have concerns regarding advice. Without their supportive and persistent guidance, this thesis definitely would not have been completed.

I am also thankful to my committee members, Professor Aerambamoorthy Thavaneswaran and Professor Steven Zheng, for their careful review, insightful questions, and valuable suggestions to better my thesis. I also would extend my appreciation to Margaret, Rosa, and all the supporting staff from Department of Statistics and Faculty of Graduate Studies who have assisted me with all administrative work.

Last but not the least, I gratefully acknowledge the generous suport from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Faculty of Graduate Studies for awarding me Canada Graduate Scholarship (CGS-M) and University of Manitoba Graduate Fellowship (UMGF). I am also thankful for the financial support from Faculty of Science, Department of Statistics, and Professor Mandal and Professor Hossain's NSERC research grants.

Dedication

I would like to take this opportunity to extend my profound appreciation from my deep heart to my beloved family and friends for their unceasing encouragement and support.

To my beloved family, thank you for your endless love, trust, patience, understanding, and constant supports. Your thoughtful encouragement have made grow better and stronger.

To my childhood, highschool and college friends, thank you for the lifelong friendships, being my source of inspiration and motivation.

To my friends and classmates in Canada, thank you for making my educational jouney become more meaningful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The linear regression model is a commonly used statistical tool for finding the relationship between the response and covariates. It has been studied rigorously and the simplicity of linearity makes it easily applicable to various fields, including medical science, finance, environmental science, econometrics, social science and computer science. Regression analysis may include variety of methods to model the relationship between variables and response. For the parametric model, several estimation techniques have been considered, specifically least squares estimation (LSE) and maximum likelihood estimation (MLE) methods to estimate the parameters. LSE has received a significant attention in both theory and applications and no distributional assumptions are required for deriving the parameter estimates. On the other hand, maximum likelihood estimation deliver a similar estimation as LSE, however, assumption of distribution is required. Both LSE and MLE lead to the best estimate among linear unbiased estimates which is called unrestricted estimator.

One common problem in regression analysis frequently occurs in selecting active covariates for the response, particularly when a large number of active covariates is under investigation. This tells the researchers for two choices. An unrestricted model with all covariates and a restricted model that contains only active covariates. The pretest estimation technique defined by Bancroft (1944) tests whether the coefficients of covariates or the linear restrictions of coefficients are zero or not and include only those covariates that are rejected by the test. Another way to select the active covariates or restricted model is to use the existing variable selection techniques, such as AIC or BIC, among others, when regression models are assumed to be sparse. The goal here is to minimize the prediction error while reducing the number of covariates in the model. The James-Stein shrinkage estimation method is the improved estimation method that allows the researcher to achieve this goal since it uses information from the inactive covariates for estimating the coefficients of the active covariates.

To fix the idea of pretest and shrinkage estimators, let us consider the estimation problem in a multiple linear regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots n. \tag{1.1}$$

where y_i is the response for the ith individual, $\boldsymbol{\beta} = (\beta_1,\, \beta_2,\, \ldots,\, \beta_p)^\top$, $\boldsymbol{x}_i^\top =$ $(x_{i1},\, x_{i2},\, \ldots, x_{ip})$, and ε_i is normally distributed with mean 0 and variance σ^2. Based on the sample information, the unrestricted maximum likelihood estimator (MLE) of the regression coefficient is given by

$$\hat{\boldsymbol{\beta}}^{ML} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{1.2}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\top$ is an $n \times p$ design matrix.

The unrestricted model includes all the p covariates and we estimate the parameters of the model based on available sample data. The restricted model includes p_2 covariates when the parameters satisfies a set of p_2 linear restrictions $\boldsymbol{R\beta} = \boldsymbol{r}$ (auxiliary information), where $\boldsymbol{R}$ is a $p_2 \times p$ matrix of rank $p_2 \leq p$, and r is a given $p_2 \times 1$ vector of known constants. Accordingly, for a particular $\boldsymbol{R}$ matrix, we may partition the regression parameter vector $\boldsymbol{\beta}$ into two components as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to have dimensions $p_1 \times 1$ and $p_2 \times 1$, respectively, such that $p = p_1 + p_2$. We consider $\boldsymbol{\beta}$ in this way as we are interested in estimating $\boldsymbol{\beta_1}$ by incorporating the auxiliary information of $\boldsymbol{\beta_2}$ into the estimation procedure. The restricted maximum likelihood estimator (RMLE) under the restriction $\boldsymbol{R\beta} = \boldsymbol{r}$ can be written as

$$\hat{\boldsymbol{\beta}}^{RML} = \hat{\boldsymbol{\beta}}^{ML} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top(\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}}^{ML}).$$

Similary, we can construct restricted estimator under the restriction $\boldsymbol{\beta_2} = \mathbf{0}$.

1.0.1 Pretest and Shrinkage Estimator

Once we get the restricted model estimator along with the unrestricted model estimator, we can test the validity of the restriction using a suitable test statistic, say Λ_L. In the pretest estimation method, we test the restriction in the form of null hypothesis $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$. The pretest estimator of $\boldsymbol{\beta}$ denoted as $\hat{\boldsymbol{\beta}}_P$ based on $\hat{\boldsymbol{\beta}}^{ML}$ and $\hat{\boldsymbol{\beta}}^{RML}$ is defined as

$$\hat{\boldsymbol{\beta}}_P = \hat{\boldsymbol{\beta}}^{ML} - I(\Delta_L \leq \chi^2_{p_1,\alpha})(\hat{\boldsymbol{\beta}}^{ML} - \hat{\boldsymbol{\beta}}^{RML}),$$

where $I(\cdot)$ is an indicator function that selects the unrestricted or restricted model based on H_0 is false or true and $\chi^2_{p_1,\alpha}$ is the α-level critical value of the distribution of test statistic Λ_L under H_0. In a two-step procedure, one would test the hypothesis $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ first, based on the outcome of the test result one should adapt the estimator. The pretest estimator combines the resulted estimators from these two steps to a single estimator. That is, testing and estimation are done simultaneously. For details, see Hossain and

Lac (2021), Hossain et al. (2009), Ahmed et al. (2007), and others.

Unfortunately, the pretest estimator is not a continuous function that changes based on the α-level. To avoid this discontinuity, we use the shrinkage estimator which is a continuous function and expresses the MLE and the RMLE in the form of a linear combination given as

$$\hat{\boldsymbol{\beta}}_S = (1 - (q-2)\Lambda_L^{-1})\hat{\boldsymbol{\beta}}^{ML} + (1-\lambda)\hat{\boldsymbol{\beta}}^{RML}, \quad q > 2.$$

When the value of Λ_L is small, λ can be negative. Thus, the shrinkage estimator will be less attractive and become over-shrinkage. To overcome this issue, positive shrinkage estimator is considered. It is defined by

$$\hat{\boldsymbol{\beta}}_{S+} = \lambda_+\hat{\boldsymbol{\beta}}^{ML} + (1-\lambda_+)\hat{\boldsymbol{\beta}}^{RML},$$

where λ_+ takes positve value of λ.

More discussions about pretest and shrinkage estimation can be found in Ahmed and Fallahpour (2012), Ahmed (2014), Hossain et al. (2015), Hossain et al. (2016), Hossain and Lac (2021), Fourdrinier et al. (2018), Battauz and Bellio (2021), and Mandal et al. (2019).

1.0.2 Partially Linear Models

A partially linear regression model is defined as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + m(\boldsymbol{t}_i) + \varepsilon_i, \quad i = 1, 2, \ldots n. \tag{1.3}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$ are $p \times 1$ and $q \times 1$ covariate vectors, respectively, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, $m(\cdot)$ is an unknown real-valued function defined on $[0, 1]$, and the ε_i's are unobservable random errors.

Model (1.3) has many applications in medicine, economics, finance, social and biological sciences. For example, in a clinical trial to compare two treatments, a patient's response may be related to the treatment received and on some covariates (e.g. cholesterol level). In this situation, the researcher may not know the effect of cholesterol level on the response, but may want to estimate the treatment differences which are believed to be constant and independent of cholesterol level. More details and estimation methods of partially linear models can be found in Boente et al. (2006), Ni et al. (2009), Hossain et al. (2009), and Härdle et al. (2012a).

1.0.3 Generalized Partial Linear Models

The generalized linear model is a tool which covers many possible nonlinear relations between covariates $\boldsymbol{x}_i$ and the response variable $\boldsymbol{y}_i$. The partially linear model in (1.3) can be extended to generalized partial linear models (GPLM) which has the form as

$$\mathrm{E}(y_i|\boldsymbol{x}_i, \boldsymbol{t}_i) = \mathrm{G}\left\{\boldsymbol{x}_i^\top \boldsymbol{\beta} + m(\boldsymbol{t}_i)\right\}, \quad i = 1, 2, \ldots n, \tag{1.4}$$

where $\mathrm{G}(\cdot)$ is a link function. These models allow a nonparametric inclusion of a part of the covariates. We assume a decomposition of the covariates into two vectors $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$, where $\boldsymbol{x}_i$ cover discrete and continuous covariables and $\boldsymbol{t}_i$ cover continuous covariables only. The details of these models and estimation methods will be discussed in Chapters 2 and 3. More details of these model can also be found in Boente et al. (2006), Boente et al. (2016), Carroll et al. (1997), and Liang (2008).

1.1 Motivation and a Brief Summary

The motivation of this thesis is two-fold:

(i) to develop the pretest and shrinkage estimators in generalized partial linear models,

(ii) to further improve the pretest and shrinkage estimators using the theory of optimal design.

Our first motivation came from a real credit scoring data. Credit scoring data are quite important in the risk assessment process. We came across a real dataset called German credit scoring data set, available at https://archive.ics.uci.edu/ml/datasets/ statlog + (german+credit+data) which contains observations on 20 socioeconomic variables for 1000 individuals. All individuals belong to the same bank. The individuals have been classified as good or bad credit risks. The response variable, `Creditability`, in the dataset corresponds to the risk label, 0 has been classified as bad credit risk (300 cases) and 1 has been classified as good credit risk (700 cases). It was needed to determine if a new applicant for the bank is in good or bad credit risk situation based on a set of socioeconomic variables. There were many covariates. Because of insufficient number of observations in each category and we only used ten covariates, in which nine were categorial and one was continuous covariate. We tried with a backward elimination procedure based on AIC and residual deviance criteria. This motivated us to model the above credit scoring data based on a reduced GPLM that contains only the significant covariates from the unrestricted GPLM.

GPLM is a flexible model in the sense that it extends the generalized linear model by adding a nonparametric component. In many situations, the relationship between the response variable and its associated covariates may not be addressed by either linear model or non-linear model. In addition, the form of the function $m(\cdot)$ in the model may not always be defined in advance. Semiparametric models, such as the partially linear model (PLM) and GPLM become meaningful in finding this type of relationship because of their robustness to model misspecification. As a tool for estimating the parameters, James-Stein shrinkage estimation method shrink the unrestricted estimator towards the restricted estimator. From the best of our knowledge, we have not found any research in the reviewed literature that develops the pretest and shrinkage estimators for GPLM. Therefore, in this study, we develop pretest and shrinkage estimation methods for GPLM and compare their performance based on the asymptotic bias and asymptotic risk functions. A Monte Carlo simulation study is conducted to compare the performance of pretest and shrinkage estimators with the unrestricted generalized Speckman estimator. The German credit scoring data is analyzed to illustrate the usefulness of our proposed methods.

Our second motivation came from the fact that optimal design theory is a powerful tool to obtain best estimation of the parameters of a statistical model. Motivated by this, we attempt to further improve the pretest and shrinkage estimators using the optimal design theory. Applying the pretest and shrinkage estimation methods in the optimal design technique in multiple linear regression is a new concept in the literature. The idea is that the optimal combinations of values of inputs are chosen before running an experiment according to a chosen criterion. There are a variety of criteria in the literature, the most popular one is the D-optimality criterion. We generate the data according to the optimal design and finally obtain our pretest and shrinkage estimators. We conduct extensive simulation studies to compare the performance of the pretest and shrinkage estimators under the setup with and without optimal design. In overall comparison, our studies show that the pretest and shrinkage estimators applying optimal design perform better as compare with these estimators under regular design.

1.2 Organization of the Thesis

The rest of this thesis is organized as follows.

- Chapter 2 provides the introduction of GPLM, generalized Speckman method, and backfitting procedures for estimating parametric component.

- Chapter 3 derives the restricted, pretest, and shrinkage estimators for the parametric part of GPLM. The asymptotic distribution, asymptotic biases and risks of the proposed estimators will also be derived. The performance of the proposed estimators is investigated by extensive simulation studies. We further illustrate the proposed methodology through an analysis of credit scoring data.

- Chapter 4 is devoted to combine the shrinkage estimators with optimal design theory and then obtain the optimal estimators under multiple linear regression. The performance of the proposed estimators is investigated by extensive simulation studies as well.

- Chapter 5 presents our conclusions of this thesis and a brief discussion of future works.

Chapter 2

Generalized Partial Linear Models (GPLM)

2.1 Introduction

The linear regression model is widely studied to find the relationship between the response and the covariates. However, in some situations the linear model is not sufficient to explain the relationship between the response variable and the covariates. The stringent requirement of linearity can increase the risk of model misspecification that leads to invalid parameter estimates. There are many statistical problems where the continuous response depends on the covariates in a nonlinear way (Engle et al., 1986). The nonparametric methods assume no predetermined functional form of covariates and this form is estimated entirely using the information from the data. These methods suffer from the curse of dimensionality which requires the sample size to increase exponentially with the number of covariates. As a tradeoff, a semiparametric model that includes parametric and nonparametric parts can avoid the curse of dimensionality to a large extent.

Semiparametric models have been widely applied in medicine, economics, finance, social and biological sciences because of their excellent scientific utility, novelty and flexibility. An excellent discussion and assessment of semiparametrics is given in Wellner et al. (2006). A landmark book on semiparametrics is Bickel et al. (1993). Fitting parametric models instead of semiparametric models generally leads to inconsistent estimators and faulty inference due to a high probability of model misspecification. Researchers consider the partially linear framework that allows most predictors to be modelled linearly while one or a small number of covariates enter in the model nonparametrically. Partially linear model (PLM) has been extensively studied (Härdle et al., 2012a,b; Ahmed et al., 2007) and several approaches have been developed to construct the estimators (Ni et al., 2009; Chen and Shiau, 1991; Engle et al., 1986). A nice summary about PLM can be found in the monograph of Härdle et al. (2012a). The other form of PLMs include partially nonlinear regression models (Li and Nie, 2008), partially linear single-index regression models (Yu and Ruppert, 2002; Carroll et al., 1997) and partially linear varying coefficient model (Ahmad et al., 2005).

Generalized linear models (GLM) are the extension of the linear regression model that allow the response variable to be count and categorical and it follows different distributions (Dunn and Smyth, 2018). For the GLM, different link functions can be used that would denote a different relationship between covariates and the response variable (e.g. log, logit, etc). We are interested in a particular semiparametric model, so-called generalized partial linear models (GPLM) which extend the GLM by adding one or a few continuous covariates that may behave nonparametrically (Boente and Rodriguez, 2010). The PLM and GLM are the special cases of GPLM. For non-normal responses, including binary, Poisson, and gamma (non-negative and positive-skewed) responses, the classical PLM is not appropriate, and thus the GPLM is adopted and extended by incorporating a link function. By introducing the non-parametric function along with the parametric component, GPLM have been used to explore the complicated relationship between the response and the covariates of interest.

Many authors have tried to introduce algorithms to estimate the parameters of the GPLM. Müller (2001) reviewed different estimation procedures based on kernel methods using profile likelihood and backfitting algorithm for estimating parametric and nonparametric components simultaneously. Boente et al. (2006) introduced a family of robust estimates in GPLM. Severini and Wong (1992) and Severini and Staniswalis (1994) considered the estimation of GPLM using generalized profile likelihoods, and they also provided a review of the literature on generalized profile likelihoods. Liang (2008) studied GPLM with missing covariates and discussed a method combining local linear regression, quasi-likelihood and weighted estimating equation to estimate the parametric and non-parametric components. Leng et al. (2011) proposed the use of GPLM and developed a least squares approximation-based variable selection procedure to study the significant predictors of condom use in HIV-infected adults. Rahman et al. (2020) considered the semiparametric efficient inferences in the GPLM and developed a bias-corrected estimating procedure and a bias-corrected empirical log-likelihood ratio for point estimation and confidence regions for the parameters of interest.

2.2 Generalized Partial Linear Models (GPLM)

For the response variable $\boldsymbol{Y}$ and covariates $(\boldsymbol{X}, \boldsymbol{T})$, the model structure of the PLM, which is widely used for continuous responses $\boldsymbol{Y}$, is described as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + m(\boldsymbol{T}) + \boldsymbol{\epsilon}. \tag{2.1}$$

In this form,

- $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\top$, $\boldsymbol{x}_i \in \mathbb{R}^p$ for $i = 1, 2, \ldots, n$ and $\boldsymbol{T} \in \mathbb{R}^q$,

- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ is a vector of unknown parameters,
- $m(\cdot)$ is an unknown smooth function, and
- $\boldsymbol{\epsilon}$ are the independent errors, with a conditional mean zero given the covariates and finite variance.

The generalization of PLM is defined by a known monotone link function $G(\cdot)$. This link function explains more complicated relationship between covariates and response in the model. The extension of model (2.1) and its variance are as follows.

$$E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{T}) = G\{\boldsymbol{X}\boldsymbol{\beta} + m(\boldsymbol{T})\}, \quad Var(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{T}) = \sigma^2 V(\mu), \tag{2.2}$$

where

- $\mu = G(\eta) = E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{T}) = G\{\boldsymbol{X}\boldsymbol{\beta} + m(\boldsymbol{T})\}$,
- $\boldsymbol{X}$ denotes a p-dimensional covariate vector,
- $\boldsymbol{T}$ is a continuous q-dimensional covariate vector which is modeled nonparametrically,
- $\boldsymbol{\beta}$ is the parameter of interest,
- $m(\cdot)$ is considered as a nuisance parameter,
- $V(\cdot)$ is a known function and σ^2 is a dispersion parameter.

The possible estimate of σ^2 can be obtained from

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{2.3}$$

where $\hat{\mu}_i = G(\hat{\eta}_i)$ and $\hat{\eta}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{m}(t_i)$.

The estimation methods for the GPLM are based on the idea that an estimate $\hat{\boldsymbol{\beta}}$ can be obtained by holding $m(\cdot)$ fixed in the log-likelihood and the resulting estimate $\hat{\boldsymbol{\beta}}$ can be then used to estimate the nonparametric part $m(\cdot)$. The estimation methods are based on kernel smoothing for estimating the nonparametric part of the model.

There are three methods to estimate the parametric and nonparametric parts for GPLM: profile likelihood, generalized Speckman, and backfitting algorithm. In the literature, it was found that the profile likelihood method and generalized Speckman method for identity link are equivalent as they produce similar results when the bandwidth is small or when the nonparametric part, $m(\boldsymbol{T})$, is small in magnitude compared to the parametric part, $\boldsymbol{X\beta}$ (Härdle et al., 2012b).

Using Nadaraya-Watson type kernel smoothing, Müller (2001) showed in the simulation study that the generalized Speckman method performs best among three methods for small sample sizes and gives similar results for parametric and nonparametric part estimates to the profile likelihood when the sample size increases. In this thesis we consider the generalized Speckman and backfitting algorithm for large sample sizes based on Nadaraya-Watson kernel smoothing (see Müller, 2001; Härdle et al., 2012b).

2.3 Estimation

2.3.1 Generalized Speckman Estimator

Generalized Speckman estimation (GSE) method refers to Speckman (1988). For the identity link function G and normally distributed responses $\boldsymbol{Y}$, the GSE and profile likelihood methods coincide each other. No iterations are required for simultaneous estimation of both $\boldsymbol{\beta}$ and $m(\bullet)$.

Following we will summarize the GSE method for identity link and extend this method to other link functions based on log-likelihood function.

a. Partially Linear Model (PLM)

For the identity link, the GPLM becomes the PLM (2.1). Taking the conditional expectation of (2.1) with respect to $\boldsymbol{T}$ and differencing the two equations leads to $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\varepsilon}$, where $\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X}$ and $\widetilde{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y}$ (Härdle et al., 2012b, Section 7.1). The elements of smoother matrix $\boldsymbol{S}$ are given by

$$S_{ij} = \frac{\mathcal{K}_{\boldsymbol{h}}(\boldsymbol{t}_i - \boldsymbol{t}_j)}{\sum_{k=1}^{n} \mathcal{K}_{\boldsymbol{h}}(\boldsymbol{t}_k - \boldsymbol{t}_j)},$$

where $\mathcal{K}$ denotes a multidimensional kernel function and $\boldsymbol{h}$ is the respective bandwidth vector and $\mathcal{K}_{\boldsymbol{h}}(\boldsymbol{u}) = \frac{1}{\boldsymbol{h}}\mathcal{K}(\boldsymbol{u}/\boldsymbol{h})$ for q-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{h}$. The Speckman estimators for $\boldsymbol{\beta}$ and m are, respectively

$$\hat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}^{\top}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^{\top}\widetilde{\boldsymbol{Y}}, \quad \text{and} \quad \hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{2.4}$$

These are the estimators for the partially linear model proposed by Speckman (1988).

b. Extend to Generalized Partial Linear Model (GPLM)

To extend Speckman's approach to GPLM, we have to take inverse link function of G and the distribution of $\boldsymbol{Y}$ takes into account. The estimation method thus combines with the iteratively reweighted least squares (IRLS) that used in the estimation of GLM (McCullagh and Nelder, 1989) and the Speckman approach for the PLM. We will use the same maximum likelihood method for the GPLM as we have used to estimate parameters in GLM.

To define the likelihood, we consider observation values $(y_i, \boldsymbol{x}_i, \boldsymbol{t}_i), \quad i = 1, 2, \cdots, n$. Let $\mu_{i,\boldsymbol{\beta}} = G(\boldsymbol{x}_i^{\top}\boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t}_i))$ from (2.2). We now denote the log-likelihood in a GPLM by

$$\ell(\boldsymbol{\mu}, \boldsymbol{y}) = \log \prod_{i=1}^{n} L(\mu_{i,\boldsymbol{\beta}}, y_i) = \sum_{i=1}^{n} \ell\left(G(\boldsymbol{x}_i^\top \boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t}_i)), y_i\right). \tag{2.5}$$

We maximize the above log-likelihood to estimate $\boldsymbol{\mu}$ which includes parametric and nonparametric parts of GPLM. The log-likelihood (2.5) differs from the log-likelihood for GLM (McCullagh and Nelder, 1989) in terms of nonparametric part $m_{\boldsymbol{\beta}}(\boldsymbol{t}_i)$.

In GLM we use an IRLS algorithm to implement the Newton-Raphson method with Fisher scoring for an iterative solution to the likelihood equations $\partial\ell/\partial\boldsymbol{\beta} = \mathbf{0}$ without nonparametric part of (2.5). The estimated parameter vector from these equations through an iterative updating is $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}^{new} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z},$$

where $\boldsymbol{z}$ denotes the adjusted dependent variable and $z = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{W}^{-1}\boldsymbol{v}$. Here $\boldsymbol{v}$ is an $n \times 1$ vector and $\boldsymbol{W}$ is a $n \times n$ diagonal matrix containing the first and the second derivative of $\ell(\boldsymbol{\mu}, \boldsymbol{y})$, respectively. Both derivatives are evaluated at $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

For the GPLM, the Speckman estimator is combined with the IRLS method used in the estimation of GLM. In the IRLS algorithm we solve the weighted least square problem at each iteration step of a GLM on the adjusted response. The same IRLS algorithm will be used for the GPLM when we replace the weighted least squares fit with a partially linear fit on the adjusted response given by

$$\boldsymbol{z} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{m}(\boldsymbol{t}) - \boldsymbol{W}^{-1}\boldsymbol{v}, \tag{2.6}$$

where $\boldsymbol{v}$ and $\boldsymbol{W}$ are defined as before but now evaluated at $\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} + \hat{m}(\boldsymbol{t}_i)$ and S is a smoothing matrix with elements

$$S_{ij} = \frac{\ell_i''(\boldsymbol{x}_i^\top\boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t}_i))\mathcal{K}_H(\boldsymbol{t}_i - \boldsymbol{t}_j)}{\sum_{i=1}^{n}\ell_i''(\boldsymbol{x}_i^\top\boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t}_i))\mathcal{K}_H(\boldsymbol{t}_i - \boldsymbol{t}_j)}. \tag{2.7}$$

The generalized Speckman algorithm for the GPLM is summarized in Table (2.1).

Table 2.1: Summary of generalized Speckman algorithm for GPLM.

```
1.  Choose an initial value for $\boldsymbol{\beta}$.
    This value can be obtained from fitting a parametric GLM.
2.  In each step, calculate:
    $\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X}$, and
    $\widetilde{\boldsymbol{z}} = \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \boldsymbol{W}^{-1}\boldsymbol{v}$.
3.  Update $\hat{\boldsymbol{\beta}}$:
    $\hat{\boldsymbol{\beta}}^{new} = (\widetilde{\boldsymbol{X}}^\top\boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\top\boldsymbol{W}\widetilde{\boldsymbol{z}}$.
4.  Update $\hat{\boldsymbol{m}}$:
    $\hat{\boldsymbol{m}}^{new} = \boldsymbol{S}(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.
```

2.3.2 Backfitting Algorithm

The backfitting algorithm was proposed proposed by Buja et al. (1989) and Hastie and Tibshirani (1990, Section 5.3). This algorithm regresses the additive components separately on the partial residuals.

a. Partially Linear Model (PLM)

The PLM consists of additive parametric and nonparametric components. The backfitting algorithm for estimating PLM can be thought of having two smoothers:

1. a projection matrix $\mathcal{P} = \boldsymbol{X}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}$ from least squares fit $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ on one or more covariates, and
2. a smoother $\boldsymbol{S}$ producing an estimate $\hat{\boldsymbol{m}}$.

The backfitting steps are $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathcal{P}(\boldsymbol{Y} - \boldsymbol{m})$ and $\hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. The iteration is unnecessary since we can solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{m}}$ explicitly:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y}, \quad \hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

b. Extend to Generalized Partial Linear Model (GPLM)

The above idea for the PLM estimation method can be extended to the GPLM. We can apply the PLM approach using a weighted smoother matrix on the adjusted response (Hastie and Tibshirani, 1990, Section 6.7). Computational steps of this method are summarized in Table (2.2).

Table 2.2: Summary of generalized backfitting algorithm for GPLM.

1. Compute starting value:
   $\boldsymbol{\beta} = \boldsymbol{0}$ and $m_{\boldsymbol{\beta}}(\boldsymbol{t}) = G^{-1}(\overline{Y})$.
2. In each step, calculate:
   $\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{X}$, and
   $\widetilde{\boldsymbol{z}} = \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \boldsymbol{W}^{-1}\boldsymbol{v}$.
3. Update $\hat{\boldsymbol{\beta}}$:
   $\hat{\boldsymbol{\beta}}^{new} = (\widetilde{\boldsymbol{X}}^{\top}\boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^{\top}\boldsymbol{W}\widetilde{\boldsymbol{z}}$.
4. Update $\hat{\boldsymbol{m}}$:
   $\hat{\boldsymbol{m}}^{new} = \boldsymbol{S}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})$.

The final estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{m}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{m}$ are called the unrestricted GSE (UG) estimators based on the GSE and backfitting methods. Under some regularity conditions, Härdle et al. (2012b) mentioned in Section 7 that the estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, see details in Section 3.4. The nonparametric function $m_{\boldsymbol{\beta}}(\boldsymbol{t})$ can be estimated with the usual univariate rate of convergence.

Chapter 3

Pretest and Shrinkage Estimations for Generalized Partial Linear Models

3.1 Introduction

In this chapter we consider the problem of pretest and shrinkage estimation methods for GPLM when there exist many covariates in the parametric part and some of them may be inactive for the response. Researchers should therefore exclude the inactive covariates from the parametric part of the model to make the predictor space sparse so as to achieve the goal of good prediction accuracy (Hastie et al., 2015).

Initially we are interested in estimating both parametric and nonparametric parts but our primary interest is on the parametric part while the nonparametric part serves as a nuisance function. To implement this interest researchers fit two models:

1. full or unrestricted model which includes all the covariates in the parametric part, and
2. submodel or restricted model which includes reduced number of covariates.

The restricted model can be formed when the prior/auxilliary information about the covariates in the model is available. We may get this prior information from similar previous studies or expert knowledge. We can express this information in terms of linear restrictions on the parameters of parametric part and test whether any of the parameters or linear combinations of parameters are insignificant through a pretesting strategy. However, if the information is not available, researchers can apply backward elimination methods with AIC, BIC or other criteria to select active covariates. The constraint on the full parameter vector is then placed by using the remaining inactive predictors.

Specifically, let $\boldsymbol{\beta}$ be a $p \times 1$ parameter vector for the parametric part and p_2 be the number of inactive predictors. We would like to find the restricted model that satisfies a set of p_2 linear restrictions

$$\boldsymbol{R\beta} = \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $p_2 \times p$ matrix of rank $p_2 \leq p$ and $\boldsymbol{r}$ is a $p_2 \times 1$ vector of constants.

We then combine the unrestricted and restricted model estimators optimally to obtain shrinkage estimators. This approach is only effective for moderate values of p_2. For large p_2, researchers may consider penalty methods to obtain sparse models and apply the shrinkage estimation strategy to obtain efficient estimators.

Several authors studied the pretest and shrinkage estimators for semiparametric linear models. Ahmed et al. (2007) used a profile least squares approach based on kernel smoothing estimation of the nonparametric component to develop the pretest, shrinkage, and penalty estimators of the parametric part when nonparametric part is nuisance. Later on Hossain et al. (2016) extended this work for longitudinal data. Raheem et al. (2012) extended this study to estimate the nonparametric part using the B-splines basis function and proposed estimators based on shrinkage strategies. Xu and Yang (2012) introduced the preliminary test backfitting estimator and preliminary test Speckman estimator when the validity of the linear restrictions on the parameters is suspected. Hossain et al. (2015) introduced shrinkage and penalty estimators in a GLM when there are many active predictors and some of them may not have influence on the response. Hossain and Lac (2021) developed optimal estimation strategies such as, pretest and shrinkage methods, for the analysis of binary longitudinal data under the partially linear single-index model where some regression parameters are subject to restrictions.

3.2 Restricted Estimator

Suppose we have a linearly independent restriction $\boldsymbol{R\beta} = \boldsymbol{r}$. Since $\boldsymbol{R}$ has rank p_2, this motivates us to determine the restricted parameter space $\Omega = \{\boldsymbol{\beta}, m(\boldsymbol{t}) | \boldsymbol{R\beta} = \boldsymbol{r}\}$. In order to maximize log-likelihood function (2.5) under Ω, we define modified likelihood

$$F(\boldsymbol{\beta}, m_{\boldsymbol{\beta}}(t), \lambda) = \sum_{i=1}^{n} \ell\left(G(\boldsymbol{x}_i^\top \boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t}_i)), y_i\right) + \sum_{j=1}^{p_2} (\boldsymbol{R}_j^\top \boldsymbol{\beta} - r_j)\lambda_i,$$

where $\boldsymbol{R}_j$ is the j^{th} column of the $\boldsymbol{R}$ matrix, and r_j is the j^{th} component of $\boldsymbol{r}$, $j = 1, 2, \cdots, p_2$.

Let l_i' and l_i'' denote the first and second derivatives of the log-likelihood

l with respect to the first argument, we have

$$\begin{aligned}\frac{\partial F}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} l_i'(\boldsymbol{x}_i + m_i') + \sum_{l=1}^{p_2} R_{lj}\lambda_l(\boldsymbol{R}_j^\top \boldsymbol{\beta} - r_j) = \widetilde{\boldsymbol{X}}\boldsymbol{v} - \boldsymbol{R}^\top \boldsymbol{\Lambda} \boldsymbol{r} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\beta}, \\ \frac{\partial^2 F}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} &= \sum_{i=1}^{n} l_i''(\boldsymbol{x}_i + m_i')(\boldsymbol{x}_i + m_i')^\top + \sum_{i=1}^{p_2} \lambda_i R_{ij} R_{il} = \widetilde{\boldsymbol{X}}^\top \boldsymbol{W} \widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R},\end{aligned}$$

where

- $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i + m_i'$,
- $\boldsymbol{v} = (l_1', l_2', \cdots, l_n')$, $\boldsymbol{W} = (l_1'', l_2'', \cdots, l_n'')$, and
- $\boldsymbol{\Lambda}$ is a $p_2 \times p_2$ diagonal matrix with λ_j, $j = 1, 2, \cdots, p_2$, as diagonal elements.

Finding the restricted estimator

Same IRLS algorithm in Section (2.3.1) will be applied to find the restricted estimator at $(d+1)$th iteration. We have

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) &= \hat{\boldsymbol{\beta}}^{(d)} - \left(\frac{\partial^2 F}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right)^{-1} \times \left(\frac{\partial F}{\partial \boldsymbol{\beta}}\right) \\ &= \hat{\boldsymbol{\beta}}^{(d)} - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}\boldsymbol{v} - \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R}\hat{\boldsymbol{\beta}}^{(d)}).\end{aligned}$$

In Section (2.3) we have $\widetilde{\boldsymbol{z}} = \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(d)} - \boldsymbol{W}^{-1}\boldsymbol{v}$ and $\hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(d)})$ that implies $\boldsymbol{v} = \boldsymbol{W}\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(d)} - \boldsymbol{W}\widetilde{\boldsymbol{z}}$. Therefore, $\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda)$ can be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) &= \hat{\boldsymbol{\beta}}^{(d)} - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}\boldsymbol{v} - \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R}\hat{\boldsymbol{\beta}}^{(d)}) \\
&= \hat{\boldsymbol{\beta}}^{(d)} - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}\boldsymbol{W}\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(d)} - \widetilde{\boldsymbol{X}}\boldsymbol{W}\widetilde{\boldsymbol{z}} - \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R}\hat{\boldsymbol{\beta}}^{(d)}) \\
&= \hat{\boldsymbol{\beta}}^{(d)} - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}\boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})\hat{\boldsymbol{\beta}}^{(d)} \\
&\quad + (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}\boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R}).
\end{aligned}$$

Thus, the $\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda)$ can be simplified as

$$\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) = (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r}). \tag{3.1}$$

We first compute the 1st argument, $(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}$. For convenience in computation, let $\boldsymbol{M} = \widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}}$, then we have

$$\begin{aligned}
(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1} &= (\boldsymbol{M} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1} \\
&= \boldsymbol{M}^{-1} - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1} \\
&= \boldsymbol{M}^{-1} - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I} + \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda})^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}.
\end{aligned}$$

Now, we can compute $\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda)$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) &= (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r}) \\
&= (\boldsymbol{M} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r}) \\
&= \left(\boldsymbol{M}^{-1} - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I} + \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda})^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}\right)(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{R}^\top \boldsymbol{\Lambda}\boldsymbol{r})
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} \\
&\quad + \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})^{-1}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})\boldsymbol{r} \\
&\quad - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} \\
&\quad - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda}\boldsymbol{r} \\
&= \boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})^{-1}\boldsymbol{r} \\
&\quad - \boldsymbol{M}^{-1}\boldsymbol{R}^\top \boldsymbol{\Lambda}(\boldsymbol{I}+\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top\boldsymbol{\Lambda})^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} \\
&= \boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} - \boldsymbol{M}^{-1}\boldsymbol{R}^\top(\boldsymbol{\Lambda}^{-1}+\boldsymbol{R}\boldsymbol{M}^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\boldsymbol{M}^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} - \boldsymbol{r}).
\end{aligned}
$$

Substituting $\boldsymbol{M}$ as $\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}}$, the $\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda)$ in (3.1) can be written as

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) &= (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}} + \boldsymbol{R}^\top\boldsymbol{\Lambda}\boldsymbol{R})^{-1}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} + \boldsymbol{R}^\top\boldsymbol{\Lambda}\boldsymbol{r}) \\
&= (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} \\
&\quad - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{R}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top\right)^{-1} \\
&\quad \times \left(\boldsymbol{R}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} - \boldsymbol{r}\right).
\end{aligned}
$$

According to Nyquist (1991), if the components of Λ are large, the restricted generalized Speckman estimates of parametric and nonparametric parts are

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \lim_{\Lambda\to\infty} \hat{\boldsymbol{\beta}}^{(d+1)}(\lambda) \\
&= (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} \\
&\quad - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top \left(\boldsymbol{R}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top\right)^{-1} \left(\boldsymbol{R}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{z}} - \boldsymbol{r}\right). \\
&= \hat{\boldsymbol{\beta}} - (\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top \left(\boldsymbol{R}(\widetilde{\boldsymbol{X}}^\top \boldsymbol{W}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \quad \text{and}
\end{aligned}
$$

$\widetilde{\boldsymbol{m}}(\boldsymbol{t}) = \widetilde{S}(\boldsymbol{z} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$.

The estimators $\tilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{m}}$ are called restricted GSE (RG) for $\boldsymbol{\beta}$ and $\boldsymbol{m}$, respectively. It is clear that the estimators $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ are updated by a parametric method with non-parametrically modified design matrix. Analogously, we can derive $\tilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{m}}$ using the backfitting algorithm.

3.3 Pretest and Shrinkage Estimators

When the null hypothesis $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ is true, the restricted estimator has smaller mean squared error than the unrestricted estimator. However, for $\boldsymbol{R\beta} \neq \boldsymbol{r}$ the restricted estimator $\tilde{\boldsymbol{\beta}}$ may be biased and inconsistent in many cases. For this reason, Ahmed (2014) suggested to consider pretest estimator by taking $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$ based on whether H_0 is rejected or accepted.

a. Pretest estimator

The pretest estimator (PT) denoted as $\hat{\boldsymbol{\beta}}^{PT}$ for the parameters $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}}^{PT} = \hat{\boldsymbol{\beta}} - I(\hat{\Lambda} \leq \chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}),$$

where

- $I(\cdot)$ is the indicator function that chooses whether to use the unrestricted or restricted estimators, and
- $\hat{\Lambda}$ is the test statistic which follows an approximate $\chi^2_{p_2}$ distribution with p_2 degrees of freedom under the null hypothesis. This test statistic is given in Section (3.4.1).

Based on the indicator function $I(\cdot)$, PT is a discontinuous function and chooses between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. This estimator depends on the acceptance or rejection of H_0 as well as the choice of α. To avoid this limitation of discontinuity, we define the shrinkage estimator (SE).

b. Shrinkage estimator

The shrinkage estimator is a continuous function which shrinks the $\hat{\boldsymbol{\beta}}$ towards $\tilde{\boldsymbol{\beta}}$. Shrinkage estimator is expressed as a linear combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}}^{SE} = \tilde{\boldsymbol{\beta}} + (1 - (p_2 - 2)\hat{\Lambda}^{-1})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}). \tag{3.2}$$

Let $\phi = 1 - (p_2 - 2)\hat{\Lambda}^{-1}$, the equation (3.2) can be rewritten as

$$\hat{\boldsymbol{\beta}}^{SE} = \phi\hat{\boldsymbol{\beta}} + (1 - \phi)\tilde{\boldsymbol{\beta}}.$$

- If $\phi = 1$, no shrinkage occurs and the estimates are the same as the unrestricted estimator UG.
- If $\phi = 0$, the restricted estimator RG is selected.

The scaler quantity (p_2-2) controls the degree of shrinkage. If H_0 is true, the value of the test statistic is small and more weight is placed on $\hat{\boldsymbol{\beta}}$, otherwise more weight is placed on $\tilde{\boldsymbol{\beta}}$.

However, the drawback of the shrinkage estimator is that the factor $1-(p_2-2)\hat{\Lambda}^{-1}$ can be negative (a phenomenon known as over-shrinkage) and it is not a convex combination of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, although it performs well over the entire parameter space relative to $\hat{\boldsymbol{\beta}}$. This issue can be improved by taking its positive part of the estimator.

c. Positive shrinkage estimator

The positive part shrinkage estimator (PSE) is defined as

$$\hat{\boldsymbol{\beta}}^{PSE} = \tilde{\boldsymbol{\beta}} + \left(1 - \frac{(p_2-2)}{\hat{\Lambda}}\right)_+ (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}),\ p_2 \geq 3,$$

where $(\cdot)_+ = 1$ if $(\cdot) > 0$, else $(\cdot)_+ = 0$.

3.4 Asymptotic Properties of the Estimators

In this section we will investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}$ and $\hat{m}_{\boldsymbol{\beta}}(\cdot)$ under some regularity conditions. We will show in Theorems 1 and 2 that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{m}_{\boldsymbol{\beta}}(\cdot)$ are consistent. The following regularity conditions are needed to discuss the asymptotic properties of the estimators.

1. $\boldsymbol{T}$, $\boldsymbol{x}_i$, and σ take values in compact sets $\mathbb{R}^q$, $\mathbb{R}^p$, $(0, \infty)$, respectively.

2. The parameter $\boldsymbol{\beta}$ takes values in a compact set $\mathbb{R}^p$ and the parameter m takes values in the set $M = \{\boldsymbol{g} \in \mathbb{R}^q : \ \boldsymbol{g} \text{ is bounded}\}$.

3. The functions $G(\cdot)$ is twice differentiable with bounded derivatives.

4. The functions $\hat{m}_{\boldsymbol{\beta}}(\boldsymbol{t}_i)$ and $m_{\boldsymbol{\beta}}(\boldsymbol{t}_i)$ are continuously differentiable with respect to $(\boldsymbol{\beta}, \boldsymbol{t})$ and twice continuously differentiable with respect to $\boldsymbol{\beta}$.

5. The kernel function $\mathcal{K}(\cdot)$ is a symmetric, continuously differentiable and bounded probability density function on $[-1, 1]$. That is, $\int_{-1}^{1} \mathcal{K}(u)du = 1, \mathcal{K} \geq 0, \mathcal{K}(u) = \mathcal{K}(-u)$.

3.4.1 Asymptotic Distribution of Unrestricted GPLM Estimator $\hat{\boldsymbol{\beta}}$

Let $\boldsymbol{\beta}_0$, m_0 and σ_0^2 are the true parameter values of $\boldsymbol{\beta}$, m and σ^2, respectively, we have

$$E_0(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{T}) = G\{\boldsymbol{X}\boldsymbol{\beta}_0 + m_0(\boldsymbol{T})\},$$

$$\text{Var}_0(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{T}) = \sigma_0^2 V(\mu_0)$$

where $\mu_0 = G(\boldsymbol{X}^\top \boldsymbol{\beta}_0 + m_0(\boldsymbol{T}))$; E_0 and Var_0 denote the expectation and variance under the true model, respectively.

The estimator $\hat{m}_{\boldsymbol{\beta}}(\boldsymbol{t})$ plays an important role in the large sample properties of $\hat{\boldsymbol{\beta}}$. Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix such that $\boldsymbol{\Sigma}^{-1}$ has (i, j)th value as

$$E_0 \left(\frac{\partial^2}{\partial \beta_i \partial \beta_j} \ell(G(\boldsymbol{X\beta} + \hat{m}_{\boldsymbol{\beta}}(\boldsymbol{T})), \boldsymbol{y})) \right). \tag{3.3}$$

- If the regularity conditions $1 - 5$ hold and there exists a compact set $\mathcal{A} \subseteq \mathbb{R}^p$ such that $\lim_{n\to\infty} \mathrm{P}(\hat{\boldsymbol{\beta}} \in \mathcal{A}) = 1$, then $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$.

- If there exists $t_0 \subseteq \mathcal{B} \subseteq \mathbb{R}^q$, then $||\hat{m}_{\boldsymbol{\beta}}(\boldsymbol{t}) - m_0(\boldsymbol{t})||_{t_0,\infty} \xrightarrow{P} 0$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma_0^2$ (Severini and Staniswalis, 1994).

Based on the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, the estimation of covariance matrix $\boldsymbol{\Sigma}$ is required for testing

$$H_0 : \boldsymbol{R\beta} = \boldsymbol{r}.$$

The next theorem will show the asymptotic normality of unrestricted GPLM.

Theorem 1: (Asymptotic normality distribution of unrestricted GPLM estimator)

Assume that the regularity conditions $1 - 5$, $nh^8 \to 0$, and $\log(1/h)/nh \to 0$ are satisfied. As $n \to \infty$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0),$$

where $\mathbf{\Sigma}_0 = \sigma_0^2 \mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ is the $p \times p$ asymptotic covariance matrix. The inverse of $\mathbf{\Sigma}$ is given by

$$E_0 \left[(\boldsymbol{I} - \boldsymbol{S}) \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^\top (\boldsymbol{I} - \boldsymbol{S}) \rho_2 (\boldsymbol{X} \boldsymbol{\beta}_0 + m_0(\boldsymbol{T})) \right] \tag{3.4}$$

where $\rho_2(g) = \left\{ \frac{d\mu(g)}{dg} \right\}^2 V^{-1}\{\mu(g)\}$ with $\mu = G\left(\boldsymbol{X}\boldsymbol{\beta} + m_{\boldsymbol{\beta}}(\boldsymbol{t})\right)$.

Let $\hat{\mathbf{\Sigma}}_0$ be the estimate of $\mathbf{\Sigma}_0$. According to Severini and Staniswalis (1994), the estimate of $\mathbf{\Sigma}_0$ can be obtained by replacing $\boldsymbol{\beta}_0$ and $m_0(\boldsymbol{T})$ in (3.4) by $\hat{\boldsymbol{\beta}}$ and $\hat{m}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{T}))$, respectively.

Above results can be used to construct a Wald-type statistic to make inferences involving a subset of the regression parameter, that is, when the parameters are in linear restriction, $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$.

Suppose that the regularity conditions $1 - 5$ and the conditions of Theorem 1 hold, then

$$\hat{\Lambda} = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top (\boldsymbol{R}(\hat{\mathbf{\Sigma}}_0)\boldsymbol{R}^\top)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) + o_p(1) \tag{3.5}$$

Under the null hypothesis, the test statistic $\hat{\Lambda}$ follows an approximate $\chi^2_{p_2}$ distribution with p_2 degrees of freedom.

Note: Under the local alternative (3.6) and the regularity conditions $1 - 5$, as $n \to \infty$, the test statistic $\hat{\Lambda}$ converges to a non-central $\chi^2_q(\Delta)$ distribution with non-centrality parameter $\Delta = \boldsymbol{\delta}^{\top}(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^{\top})^{\top}\boldsymbol{\delta}$ (Ahmed and Fallahpour, 2012), where δ is defined in Section (3.4.2).

3.4.2 Asymptotic Joint Distribution of Unrestricted and Restricted Estimators

To study the asymptotic bias (AB) and asymptotic risk (AR) of the proposed estimators (Sections 3.4.3 and 3.4.4) using the local asymptotic normality approach of Vaart (1998), we will start with asymptotic joint normality of UG and RG.

Under the fixed alternative $H_a : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{r}+\boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_{p_2})$ is a real fixed vector; the estimators $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$ are asymptotically equivalent in distribution to $\hat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}$ has unbounded risk. As the result, we cannot differentiate the bias and the risk properties of first three estimators $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$. In order to differentiate, we consider the following sequence of local alternative,

$$H_{(n)} : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r} + \frac{\boldsymbol{\delta}}{\sqrt{n}}, \; n > 0. \tag{3.6}$$

The magnitude of the distance $||\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r}||$ is determined by localizing the fixed vector $\boldsymbol{\delta}$ and the sample size n. For any fixed $\boldsymbol{\delta}$, the distance $\boldsymbol{\delta}/\sqrt{n}$ shrinks as the sample size increases. Under the local alternative (3.6) and the regularity conditions $1-5$, we have

$$\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}) \xrightarrow{D} N(\boldsymbol{\delta}, \boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top).$$

The following joint asymptotic normality of UG and RG under (3.6) allows us to facilitate the theoretical and numerical comparison of AB and AR of the estimators $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$ in next sections.

Theorem 2: (Aymptotic joint normality of UG and RG)

Under the local alternative (3.6) and the regularity conditions $1-5$, we have the joint distribution, as $n \to \infty$

$$\begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{pmatrix} \xrightarrow[n\to\infty]{\mathcal{L}} N_{3p}\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{\gamma} \\ -\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0 - \boldsymbol{J}_0 & \boldsymbol{J}_0 \\ (\boldsymbol{\Sigma}_0 - \boldsymbol{J}_0)^\top & \boldsymbol{\Sigma}_0 - \boldsymbol{J}_0 & \mathbf{0} \\ \boldsymbol{J}_0^\top & \mathbf{0} & \boldsymbol{J}_0 \end{bmatrix} \right)$$

where

- $\boldsymbol{\eta}_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}), \boldsymbol{\eta}_2 = \sqrt{n}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}), \boldsymbol{\eta}_3 = \sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})$,
- $\boldsymbol{J}_0 = \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0$,
- $\boldsymbol{\gamma} = -\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta}$.

Proof of Theorem 2:

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\eta}_1) &= \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) = \mathbf{0}. \\
\mathrm{E}(\boldsymbol{\eta}_2) &= \mathrm{E}\left(\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) \\
&= \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}-\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}))\right) \\
&= -\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\right) \\
&= -\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta} = \boldsymbol{\gamma}. \\
\mathrm{E}(\boldsymbol{\eta}_3) &= \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right) = \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})-\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) = -\boldsymbol{\gamma}. \\
\mathrm{Var}(\boldsymbol{\eta}_1) &= \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) = \boldsymbol{\Sigma}_0. \\
\mathrm{Var}(\boldsymbol{\eta}_2) &= \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}-\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}))\right) \\
&= \mathrm{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) + \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\mathrm{Var}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\right)(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
&\quad - 2\mathrm{Cov}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right), \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}))\right) \\
&= \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 - 2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
&= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
&= \boldsymbol{\Sigma}_0 - \boldsymbol{J}_0. \\
\mathrm{Var}(\boldsymbol{\eta}_3) &= \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right) = \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}+\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-r))\right) \\
&= \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\mathrm{Var}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\right)(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
&= \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 = \boldsymbol{J}_0. \\
\mathrm{Cov}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) &= \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}), \sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) \\
&= \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}-\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}))\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) - \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}\right)\\
&= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{J}_0.\\
\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_3) &= \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right)\\
&= \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\\
&= \mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) - \mathrm{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\\
&= \boldsymbol{\Sigma}_0 - \mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \boldsymbol{J}_0.\\
\mathrm{Cov}(\boldsymbol{\eta}_2, \boldsymbol{\eta}_3) &= \mathrm{Cov}\left(\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right)\\
&= \mathrm{Cov}\left(\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\\
&= \mathrm{Cov}\left(\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) - \mathrm{Var}\left(\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\\
&= \boldsymbol{\Sigma}_0 - \boldsymbol{J}_0 - (\boldsymbol{\Sigma}_0 - \boldsymbol{J}_0) = \boldsymbol{0}.
\end{aligned}
$$

3.4.3 Asymptotic Bias of Proposed Estimators

We now discuss the asymptotic bias of the proposed estimators. Let $\hat{\boldsymbol{\beta}}^*$ be any of the proposed estimators. Then the asymptotic bias of $\hat{\boldsymbol{\beta}}^*$ is defined by

$$\mathrm{AB}(\hat{\boldsymbol{\beta}}^*) = \int \boldsymbol{z} d\tilde{F}_{\hat{\boldsymbol{\beta}}^*}(\boldsymbol{z}),$$

where $\tilde{F}_{\hat{\boldsymbol{\beta}}^*}(\boldsymbol{z})$ is the asymptotic distribution function of $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})$.

Note: Since $\mathrm{AB}(\hat{\boldsymbol{\beta}}^*)$ is not a scaler form, in order to do comparison, we will use its asymptotic quadratic bias (QB) where $\mathrm{QB}(\hat{\boldsymbol{\beta}}^*)$ is defined as

$$\mathrm{QB}(\hat{\boldsymbol{\beta}}^*) = \mathrm{AB}(\hat{\boldsymbol{\beta}}^*)^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)\mathrm{AB}(\hat{\boldsymbol{\beta}}^*).$$

We can find QBs and derive ABs using the following theorem.

Theorem 3: (Asymptotic bias of proposed estimators)

Let Z_1 and Z_2 be $\chi^2_{p_2+2}(\Delta)$ and $\chi^2_{p_2+4}(\Delta)$ random variables, respectively and their distributions are generally denoted by $\Psi_g(x,\Delta) = P\left(\chi^2_g(\Delta) \leq x\right)$. Let $\chi^2_{p_2,\alpha}$ be the α-level critical value of central χ^2-distribution. Under the local alternative (3.6) and the regularity conditions $1-5$, we have

- $$\mathrm{AB}(\hat{\boldsymbol{\beta}}\quad) = \mathbf{0}$$
- $$\mathrm{AB}(\tilde{\boldsymbol{\beta}}\quad) = \boldsymbol{\gamma}$$
- $$\mathrm{AB}(\hat{\boldsymbol{\beta}}^{PT}) = \boldsymbol{\gamma}\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)$$
- $$\mathrm{AB}(\hat{\boldsymbol{\beta}}^{SE}) = \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}\right)$$
- $$\begin{aligned}\mathrm{AB}(\hat{\boldsymbol{\beta}}^{PSE}) &= \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}\right) + \boldsymbol{\gamma}\Psi_{p_2+2}(p_2-2,\Delta) \\ &\quad - \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}I(Z_1 < p_2-2)\right).\end{aligned}$$

Proof of Theorem 3:

We will use the following Lemma 1 to derive the proof of Theorem 3.

Lemma 1

Let $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\Sigma}_p$ is a nonnegative definite matrix with rank $p_2 \leq$

p. Let $\boldsymbol{Q}$ be a $p \times p$ symmetric and positive definite matrix with rank p such that $\boldsymbol{\Sigma}_p\boldsymbol{Q}$ is an idempotent matrix and $\boldsymbol{\Sigma}_p\boldsymbol{Q}\boldsymbol{\delta} = \boldsymbol{\delta}$. Let $\boldsymbol{W} = \boldsymbol{Q}^{1/2}\boldsymbol{W}^*\boldsymbol{Q}^{1/2}$, where $\boldsymbol{W}^*$ is a non-negative definite matrix.

For all h, Borel measurable and real-valued integrable function, we have

1. $$\mathrm{E}\left(h\left(\boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X}\right)\boldsymbol{W}\boldsymbol{X}\right) = \mathrm{E}\left(\varphi\left(\chi^2_{p+2}(\boldsymbol{\delta}^\top\boldsymbol{Q}\boldsymbol{\delta})\right)\right)\boldsymbol{W}\boldsymbol{\delta}$$

2. $$\begin{aligned}\mathrm{E}\left(\varphi\left(\boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X}\right)\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right) &= \mathrm{E}\left(h\left(\chi^2_{p+2}(\boldsymbol{\delta}^\top\boldsymbol{Q}\boldsymbol{\delta})\right)\right)\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}_{\boldsymbol{p}})\\ &\quad+\mathrm{E}\left(h\left(\chi^2_{p+4}(\boldsymbol{\delta}^\top\boldsymbol{Q}\boldsymbol{\delta})\right)\right)\boldsymbol{\delta}^\top\boldsymbol{W}\boldsymbol{\delta}.\end{aligned}$$

For the details and the proof of this Lemma, see Nkurunziza and Chen (2013).

Using the above Lemma 1, we now can derive ABs of the proposed estimators.

$$\begin{aligned}
\mathrm{AB}(\hat{\boldsymbol{\beta}}) &= \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) = \mathbf{0}.\\
\mathrm{AB}(\tilde{\boldsymbol{\beta}}) &= \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right)\\
&= \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}+\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)\right)\\
&= \mathbf{0}-\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)\right)\\
&= -\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta} = \boldsymbol{\gamma}.\\
\mathrm{AB}(\hat{\boldsymbol{\beta}}^{PT}) &= \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{PT}-\boldsymbol{\beta}\right)\right)\\
&= \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})-\boldsymbol{\beta}\right)\right)\\
&= \mathbf{0}-\lim_{n\to\infty}\mathrm{E}\left(I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right)\\
&= \boldsymbol{\gamma}\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta).
\end{aligned}$$

$$
\begin{aligned}
\mathrm{AB}(\hat{\boldsymbol{\beta}}^{SE}) &= \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{SE} - \boldsymbol{\beta}\right)\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}\left(\tilde{\boldsymbol{\beta}} + (1-(p_2-2)\hat{\Lambda}^{-1})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right)\right) \\
&= \mathbf{0} - \lim_{n\to\infty} \mathrm{E}\left((p_2-2)\hat{\Lambda}^{-1}\sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right) \\
&= \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}\right). \\
\mathrm{AB}(\hat{\boldsymbol{\beta}}^{PSE}) &= \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{PSE} - \boldsymbol{\beta}\right)\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{SE} - \boldsymbol{\beta}\right)\right) - \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}(1-(p_2-2)\hat{\Lambda}^{-1})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})I(\hat{\Lambda} \le p_2-2)\right) \\
&= \mathrm{AB}(\hat{\boldsymbol{\beta}}^{SE}) - \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(1-(p_2-2)\hat{\Lambda}^{-1})I(\hat{\Lambda} \le p_2-2)\right) \\
&= \mathrm{AB}(\hat{\boldsymbol{\beta}}^{SE}) + \boldsymbol{\gamma}\mathrm{E}\left(I(\hat{\Lambda} \le p_2-2)\right) - \boldsymbol{\gamma}(p_2-2)\mathrm{E}\left(\hat{\Lambda}^{-1}I(\hat{\Lambda} \le p_2-2)\right) \\
&= \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}\right) + \boldsymbol{\gamma}\Psi_{p_2+2}(p_2-2,\Delta) - \boldsymbol{\gamma}(p_2-2)E\left(Z_1^{-1}I(Z_1 < p_2-2)\right).
\end{aligned}
$$

Remarks:

The comparisons of ABs of the esitmators are summarized as follows.

- If $\boldsymbol{\delta} = \mathbf{0}$, the AB of any estimator is a $\mathbf{0}$ vector.
- If $\boldsymbol{\delta} > \mathbf{0}$ and let $\Delta\boldsymbol{\omega} = \boldsymbol{\gamma}$, where $\Delta = \boldsymbol{\delta}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^\top\boldsymbol{\delta}$, all the ABs are scaler multiple of Δ along with $\boldsymbol{\omega}$ except the AB of $\hat{\boldsymbol{\beta}}$.
- While the AB of $\tilde{\boldsymbol{\beta}}$ is an unbounded function of Δ, the ABs of $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$ are bounded in Δ as $\mathrm{E}(Z_1^{-1})$ is a non-increasing log-convex function of Δ.
- Although the AB of $\hat{\boldsymbol{\beta}}^{SE}$ is close to $\hat{\boldsymbol{\beta}}^{PSE}$, the bias curve of $\hat{\boldsymbol{\beta}}^{PSE}$ stays

below the curve of $\hat{\boldsymbol{\beta}}^{SE}$.

- When $\Delta > 0$, the AB of $\hat{\boldsymbol{\beta}}^{PT}$ will increases to a point then decrease towards zero.

3.4.4 Asymptotic Risk of Proposed Estimators

We now discuss the *asymptotic risk* of the proposed estimators. To derive expressions for the ARs of the estimators, we define a quadratic loss function

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_*; \boldsymbol{Q}) = \left(\sqrt{n}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})\right)^\top \boldsymbol{Q} \left(\sqrt{n}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})\right),$$

where $\boldsymbol{Q}$ is a nonnegative definite weight matrix.

Typically when considering $\boldsymbol{Q}$ as a $p \times p$ identity matrix, $\mathcal{L}(\hat{\boldsymbol{\beta}}_*)$ is the usual quadratic loss function. However, a loss function with general $\boldsymbol{Q}$ will give different weights to different $\boldsymbol{\beta}$'s. It leads to shrinkage estimators may not outperform in the entire parameter space with respect to unrestricted GPLM estimator. Therefore, for simplicity, we consider $\boldsymbol{Q} = \boldsymbol{I}_{p \times p}$ in the simulation studies.

The mean squared error (MSE) matrix for any estimator $\hat{\boldsymbol{\beta}}^*$ under the quadratic loss function is

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = \lim_{n\to\infty} \mathrm{E}\big\{n(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^\top\big\} = \int \boldsymbol{z}\boldsymbol{z}^\top d\tilde{F}_{\hat{\boldsymbol{\beta}}^*}(\boldsymbol{z}), \tag{3.7}$$

and the AR is defined as

$$\mathrm{AR}(\hat{\boldsymbol{\beta}}^*;\boldsymbol{Q}) = \int \boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z} d\tilde{F}_{\hat{\boldsymbol{\beta}}^*}(\boldsymbol{z}) = \mathrm{trace}\big(\boldsymbol{Q}\ \mathrm{MSE}(\hat{\boldsymbol{\beta}})\big). \tag{3.8}$$

We will use the formula (3.8) to calculate the numerical risks of all proposed estimators in the simulation studies.

Theorem 4: (Asymptotic risk of proposed estimators)

If the condition of Theorem 3 holds, then the asymptotic risks of proposed estimators are as follows.

- $\mathrm{AR}(\hat{\boldsymbol{\beta}}\quad;\boldsymbol{Q}) = \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{\Sigma}_0\right),$
- $\mathrm{AR}(\tilde{\boldsymbol{\beta}}\quad;\boldsymbol{Q}) = \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) - \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{J}_0\right) + \boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma},$
- $$\begin{aligned}\mathrm{AR}(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{Q}) &= \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) - \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{J}_0\right)\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta),\\ &- \boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}\left(\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta) - 2\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)\right),\end{aligned}$$
- $$\begin{aligned}\mathrm{AR}(\hat{\boldsymbol{\beta}}^{SE};\boldsymbol{Q}) &= \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) + (p_2-2)\mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{J}_0\right)\left((p_2-2)\mathrm{E}(Z_1^{-2}) - 2\mathrm{E}(Z_1^{-1})\right),\\ &+ (p_2-2)\left((p_2-2)\mathrm{E}(Z_2^{-2}) - 2\mathrm{E}(Z_2^{-1} - Z_1^{-1})\right)\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma},\end{aligned}$$
- $$\begin{aligned}\mathrm{AR}(\hat{\boldsymbol{\beta}}^{PSE};\boldsymbol{Q}) &= \mathrm{AR}(\hat{\boldsymbol{\beta}}^{SE};\boldsymbol{Q}) - \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{J}_0\right)\mathrm{E}\left((1-(p_2-2)Z_1^{-1})^2 I(Z_1 < p_2-2)\right)\\ &- \mathrm{E}\left((1-(p_2-2)Z_2^{-1})^2 I(Z_2 < p_2-2)\right)\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}\end{aligned}$$

$$+ 2\mathrm{E}\left((1-(p_2-2)Z_1^{-1})I(Z_1 < p_2-2)\right)\boldsymbol{\gamma}^\top \boldsymbol{Q}\boldsymbol{\gamma}.$$

Proof of Theorem 4:

Based on the definition of AR function, to derive the ARs of the proposed estimators, we will first derive their **covariance matrices**. The covariance matrix of any estimator $\hat{\boldsymbol{\beta}}^*$ is defined as

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) = \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^\top\right).$$

a. AR of unrestricted estimator $\hat{\boldsymbol{\beta}}$

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right) \\
&= \mathrm{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top) = \boldsymbol{\Sigma}_0, \quad \text{and} \\
\mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) &= \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{\Sigma}_0\right).
\end{aligned}$$

b. AR of restricted estimator $\tilde{\boldsymbol{\beta}}$

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{\beta}}) &= \lim_{n\to\infty} \mathrm{E}\left(n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left(n\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\right)\right. \\
&\times \left.\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\right)^\top\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right) \\
&+ \lim_{n\to\infty} \mathrm{E}\left(n\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\right)
\end{aligned}$$

$$
\begin{aligned}
& - && 2\lim_{n\to\infty} \mathrm{E}\left(n\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right), \quad \text{where} \\
\text{1st term} & = && \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) = \mathrm{Var}(\hat{\boldsymbol{\beta}}), \\
\text{2nd term} & = && \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\right) \\
& = && \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
& = && \boldsymbol{J}_0 + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top, \text{and} \\
\text{3rd term} & = && -2\lim_{n\to\infty} \mathrm{E}\left(n\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) \\
& = && -2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 = -2\boldsymbol{J}_0, \quad \text{Therefore, we have} \\
\mathrm{Var}(\tilde{\boldsymbol{\beta}}) & = && \mathrm{Var}(\hat{\boldsymbol{\beta}}) - \boldsymbol{J}_0 + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top, \quad \text{and} \\
\mathrm{AR}(\tilde{\boldsymbol{\beta}};\boldsymbol{Q}) & = && \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) - \mathrm{trace}\,(\boldsymbol{Q}\boldsymbol{J}_0) + \boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}.
\end{aligned}
$$

c. AR of pretest estimator $\hat{\boldsymbol{\beta}}^{PT}$

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{PT}) & = && \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}}^{PT}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{PT}-\boldsymbol{\beta})^\top\right) \\
& = && \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) \\
& + && \lim_{n\to\infty} \mathrm{E}\left(I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\right) \\
& - && 2\lim_{n\to\infty} \mathrm{E}\left(nI(\hat{\Lambda}\le\chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right), \quad \text{where} \\
\text{1st term} & = && \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) = \mathrm{Var}(\hat{\boldsymbol{\beta}}) \\
\text{2nd term} & = && \lim_{n\to\infty} \mathrm{E}\left(nI(\hat{\Lambda}\le\chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\right) \\
& = && \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\lim_{n\to\infty} \mathrm{E}\left(I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})n(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})^\top\right) \\
& \times && (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0 \\
& = && \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)
\end{aligned}
$$

$$
\begin{aligned}
& + && \boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta) \\
& = && \boldsymbol{J}_0\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta) + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta). \\
\text{3rd term} & = && -2\lim_{n\to\infty}\mathrm{E}\left(nI(\hat{\Lambda}\le\chi^2_{p_2,\alpha})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) \\
& = && -2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \\
& + && \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\left(\sqrt{n}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right)^\top I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})\right) \\
& - && \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\left(\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\sqrt{n}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r})\right)^\top I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})\right) \\
& = && -2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \\
& + && \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\left(\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\right)^\top I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})\right) \\
& - && \lim_{n\to\infty}\mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\left(\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\sqrt{n}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r})\right)^\top I(\hat{\Lambda}\le\chi^2_{p_2,\alpha})\right) \\
& = && -2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta) \\
& - && 2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_0\left(\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta)-\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)\right) \\
& = && -2\boldsymbol{J}_0\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta) - 2\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\left(\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta)-\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)\right),
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{PT}) & = && \mathrm{Var}(\hat{\boldsymbol{\beta}}) - \boldsymbol{J}_0\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta) - \boldsymbol{\gamma}\boldsymbol{\gamma}^\top\left(\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta)-2\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)\right) \\
\mathrm{AR}(\boldsymbol{\beta}^{\hat{PT}};\boldsymbol{Q}) & = && \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) - \mathrm{trace}\left(\boldsymbol{Q}\boldsymbol{J}_0\right)\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta) \\
& - && \boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}\left(\Psi_{p_2+4}(\chi^2_{p_2,\alpha}\Delta)-2\Psi_{p_2+2}(\chi^2_{p_2,\alpha},\Delta)\right).
\end{aligned}
$$

d. AR of shrinkage estimator $\hat{\boldsymbol{\beta}}^{SE}$

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) & = && \lim_{n\to\infty}\mathrm{E}\left(n(\hat{\boldsymbol{\beta}}^{SE}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{SE}-\boldsymbol{\beta})^\top\right) \\
& = && \lim_{n\to\infty}\mathrm{E}\left(n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right)
\end{aligned}
$$

$$
\begin{aligned}
& + && \lim_{n\to\infty} \mathrm{E}\left(n(p_2-2)^2 \hat{\Lambda}^{-2} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \right) \\
& - && 2 \lim_{n\to\infty} \mathrm{E}\left(n(p_2-2) \hat{\Lambda}^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \right), \text{where} \\
\text{1st term} & = && \lim_{n\to\infty} \mathrm{E}\left(n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \right) = \mathrm{Var}(\hat{\boldsymbol{\beta}}). \\
\text{2nd term} & = && \lim_{n\to\infty} \mathrm{E}\left(n(p_2-2)^2 \hat{\Lambda}^{-2} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \right) \\
& = && (p_2-2)^2 \lim_{n\to\infty} \left(\boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \right. \\
& \times && \left. \mathrm{E}\left(\hat{\Lambda}^{-2} \sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \right) (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{R}\boldsymbol{\Sigma}_0 \right) \\
& = && (p_2-2)^2 \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{R}\boldsymbol{\Sigma}_0 \mathrm{E}(Z_1^{-2}) \\
& + && (p_2-2)^2 \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{\delta}\boldsymbol{\delta}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{R}\boldsymbol{\Sigma}_0 \mathrm{E}(Z_2^{-2}) \\
& = && (p_2-2)^2 \boldsymbol{J}_0 \mathrm{E}(Z_1^{-2}) + (p_2-2)^2 \boldsymbol{\gamma}\boldsymbol{\gamma}^\top \mathrm{E}(Z_2^{-2}), \text{where } Z_2 = \chi^2_{p_2+4,\alpha}(\Delta), \\
\text{3rd term} & = && -2 \lim_{n\to\infty} \mathrm{E}\left(n(p_2-2) \hat{\Lambda}^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \right) \\
& = && -2(p_2-2) \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \lim_{n\to\infty} \mathrm{E}\left(\hat{\Lambda}^{-1} \sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \right) \\
& = && -2(p_2-2) \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \\
& \times && \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \hat{\Lambda}^{-1} \right) (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{R}\boldsymbol{\Sigma}_0 \\
& + && 2(p_2-2) \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \\
& \times && \lim_{n\to\infty} \mathrm{E}\left(\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \hat{\Lambda}^{-1} \right) \sqrt{n}(\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r})^\top (\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \boldsymbol{R}\boldsymbol{\Sigma}_0 \\
& = && -2(p_2-2) \boldsymbol{J}_0 \mathrm{E}(Z_1^{-1}) - 2(p_2-2) \boldsymbol{\gamma}\boldsymbol{\gamma}^\top \mathrm{E}\left(Z_2^{-2} - Z_1^{-1} \right), \text{by Lemma 1.}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) & = && \mathrm{Var}(\hat{\boldsymbol{\beta}}) + (p_2-2)\boldsymbol{J}_0 \left((p_2-2)\mathrm{E}(Z_1^{-2}) - 2\mathrm{E}(Z_1^{-1}) \right) \\
& + && (p_2-2)\boldsymbol{\gamma}\boldsymbol{\gamma}^\top \left((p_2-2)\mathrm{E}(Z_2^{-2}) - 2\mathrm{E}(Z_2^{-1} - Z_1^{-1}) \right), \text{and}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{AR}(\hat{\boldsymbol{\beta}}^{SE};\boldsymbol{Q}) &= \mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) + (p_2-2)\mathrm{trace}(\boldsymbol{Q}\boldsymbol{J}_0)\left((p_2-2)\mathrm{E}(Z_1^{-2}) - 2\mathrm{E}(Z_1^{-1})\right) \\
&+ (p_2-2)\left((p_2-2)\mathrm{E}(Z_2^{-2}) - 2\mathrm{E}(Z_2^{-1} - Z_1^{-1})\right)\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}.
\end{aligned}
$$

e. AR of positive shrinkage estimator $\hat{\boldsymbol{\beta}}^{PSE}$

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{PSE}) &= \lim_{n\to\infty}\mathrm{E}\left(n(\hat{\boldsymbol{\beta}}^{PSE}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{PSE}-\boldsymbol{\beta})^\top\right) \\
&= \mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) + \lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})^2(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top I(\hat{\Lambda}<p_2-2)\right) \\
&- 2\lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\left((\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(1-(p_2-2)\hat{\Lambda}^{-1})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})\right)^\top\right. \\
&\times \left. I(\hat{\Lambda}<p_2-2)\right) \\
&= \mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) - \lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})^2(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top I(\hat{\Lambda}<p_2-2)\right) \\
&- 2\lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})I(\hat{\Lambda}<p_2-2)(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\right), \text{where} \\
\text{2nd term} &= \mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) - \lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})^2(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top I(\hat{\Lambda}<p_2-2)\right) \\
&= -\boldsymbol{J}_0\mathrm{E}\left((1-(p_2-2)Z_1^{-1})^2 I(Z_1<p_2-2)\right) \\
&- \boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathrm{E}\left((1-(p_2-2)Z_2^{-1})^2 I(Z_2<p_2-2)\right), \text{ and} \\
\text{3rd term} &= -2\lim_{n\to\infty}\mathrm{E}\left(n(1-(p_2-2)\hat{\Lambda}^{-1})I(\hat{\Lambda}<p_2-2)(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\right) \\
&= -2\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1} \\
&\times \lim_{n\to\infty}\mathrm{E}\left((1-(p_2-2)\hat{\Lambda}^{-1})I(\hat{\Lambda}<p_2-2)\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) \\
&= 2\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathrm{E}\left((1-(p_2-2)Z_1^{-1})I(Z_1<p_2-2)\right), \text{by Lemma 1.}
\end{aligned}
$$

Therefore,

$$
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{PSE}) = \mathrm{Var}(\hat{\boldsymbol{\beta}}^{SE}) - \boldsymbol{J}_0\mathrm{E}\left((1-(p_2-2)Z_1^{-1})^2 I(Z_1<p_2-2)\right)
$$

$$
\begin{aligned}
& - \quad \boldsymbol{\gamma}\boldsymbol{\gamma}^\top \mathrm{E}\left((1-(p_2-2)Z_2^{-1})^2 I(Z_2 < p_2 - 2)\right) \\
& + \quad 2\boldsymbol{\gamma}\boldsymbol{\gamma}^\top \mathrm{E}\left((1-(p_2-2)Z_1^{-1}) I(Z_1 < p_2 - 2)\right), \text{and} \\
\mathrm{AR}(\hat{\boldsymbol{\beta}}^{PSE};\boldsymbol{Q}) \quad = & \quad \mathrm{AR}(\hat{\boldsymbol{\beta}}^{SE};\boldsymbol{Q}) - \mathrm{trace}\,(\boldsymbol{Q}\boldsymbol{J}_0)\,\mathrm{E}\left((1-(p_2-2)Z_1^{-1})^2 I(Z_1 < p_2 - 2)\right) \\
& - \quad \mathrm{E}\left((1-(p_2-2)Z_2^{-1})^2 I(Z_2 < p_2 - 2)\right)\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma} \\
& + \quad 2\mathrm{E}\left((1-(p_2-2)Z_1^{-1}) I(Z_1 < p_2 - 2)\right)\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}.
\end{aligned}
$$

Remarks:

- If $\boldsymbol{\delta} = \boldsymbol{0}$ (i.e the non-centrality parameter $\Delta = 0$), the ARs of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are $\mathrm{trace}\,(\boldsymbol{Q}\boldsymbol{\Sigma}_0)$ and $\mathrm{AR}(\hat{\boldsymbol{\beta}};\boldsymbol{Q}) - \mathrm{trace}\,(\boldsymbol{Q}\boldsymbol{J}_0)$, respectively.
- The AR of $\hat{\boldsymbol{\beta}}$ is constant while the AR of $\tilde{\boldsymbol{\beta}}$ becomes unbounded as $\boldsymbol{\delta}$ in $\boldsymbol{\gamma} = -\boldsymbol{\Sigma}_0\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Sigma}_0\boldsymbol{R}^\top)^{-1}\boldsymbol{\delta}$ away from the null vector.
- $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$ have the lower risks than $\hat{\boldsymbol{\beta}}$ when $\Delta = 0$. However, for any intermediate value of Δ, AR of $\hat{\boldsymbol{\beta}}^{PT}$ is larger than AR of $\hat{\boldsymbol{\beta}}$.
- If $\Delta \to \infty$, the AR of $\hat{\boldsymbol{\beta}}^{PT} \to \hat{\boldsymbol{\beta}}$.
- As $\Delta > 0$, ARs of $\hat{\boldsymbol{\beta}}^{SE}$ and $\hat{\boldsymbol{\beta}}^{PSE}$ increase. However, both estimators outperform $\hat{\boldsymbol{\beta}}$ and are admissible estimators when compared to $\hat{\boldsymbol{\beta}}$.
- $\hat{\boldsymbol{\beta}}^{PSE}$ is asymptotically superior to $\hat{\boldsymbol{\beta}}^{SE}$ in the entire parameter space induced by Δ.

The comparison of performance among estimators will be explored numerically through a simulation study with different number of insignificant covariates and sample sizes in next section (3.5).

3.5 Simulation Study

In this section we conduct simulation studies in various settings to evaluate the performance of the proposed pretest and shrinkage estimators, in estimating the parametric part of GPLM to compare with the unrestricted GPLM estimator.

Our simulations are based on sample sizes $n = 200$, 250, 300, and 350. We consider a logistic partially linear model and generate data from this model

$$\text{logit}(\boldsymbol{y}_i = 1|\boldsymbol{x}_i, \boldsymbol{t}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + m_{\boldsymbol{\beta}}(t_i), \quad i = 1, 2, \cdots, n, \tag{3.9}$$

where

- the parametric covariates $\boldsymbol{x}_i$ are drawn from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{I}_p$,
- t_i is generated from the uniform distribution $[-1, 1]$ with $\boldsymbol{x}$ and $\boldsymbol{t}$ are independent, and
- $m_{\boldsymbol{\beta}}(t)$ is assumed as $m_{\boldsymbol{\beta}}(t) = 1.5 \times sin(2\pi t)$.

To investigate the performance of the pretest and shrinkage estimators with respect to the unrestricted GPLM estimator, we first define three models as follows.

a. Simulation model

We consider model (3.9) as our simulation model and assume two sets of true values of coefficients $\boldsymbol{\beta}_{sim}$ for the parametric part of model (3.9),

- $\boldsymbol{\beta}_{sim} = (1.5, -2.1, -1.25, 3, -0.75)$,
- $\boldsymbol{\beta}_{sim} = (1.5, -2.1, -1.25, 3, -0.75, 1.45, 0.3)$.

Given the values of the parameters, the design matrix, and the nonparametric part $m_{\boldsymbol{\beta}}(t)$, we then generate binary responses from this simulation model.

b. Unrestricted model

The unrestricted model is defined as the logistic partially linear model (3.9), with $\boldsymbol{\beta}_U = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. In this model $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{sim}$ and $\boldsymbol{\beta}_2$ is a $p_2 \times 1$ vector of zeros, that is, insignificant covariates. Thus, the true values of the parameters are $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{sim}$, where $\boldsymbol{\beta}_1$ is either $(1.5, -2.1, -1.25, 3, -0.75)$ or $(1.5, -2.1, -1.25, 3, -0.75, 1.45, 0.3)$ and $\boldsymbol{\beta}_2 = \mathbf{0}_{p_2\times 1}$.

With this set up, the covariates related to $\boldsymbol{\beta}_2$ turn out not to be statistically significant with the response, that means these covariates may not make any contribution in the presence of other significant covariates. We consider several scenarios: $p_1 = 5$ and 7; and $p_2 = 3$, 5, 8, 10, 13 and 15.

c. Restricted model

The restricted model is just the unrestricted model subject to the constraint $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$. The restricted estimator will be based on an $\boldsymbol{R}$ with $H_0 : \boldsymbol{R\beta} = \boldsymbol{0}$ where $\boldsymbol{R} = \left[\boldsymbol{0}_{p_2\times(p-p_2)}, \boldsymbol{I}_{p_2}\right]$, $\boldsymbol{r} = \boldsymbol{0}_{p_2\times 1}$, where $\boldsymbol{0}_{p_2\times 1}$ is an $p_2 \times 1$ zero vector and $\boldsymbol{\beta}_R = (\boldsymbol{\beta}_{1R}^\top, \boldsymbol{\beta}_{2R}^\top)^\top$. The dimension of $\boldsymbol{\beta}_{1R}$ and $\boldsymbol{\beta}_{2R}$ are $p_1 \times 1$ and $p_2 \times 1$, respectively, such that $p = p_1 + p_2$.

With the restriction $\boldsymbol{R\beta} = \boldsymbol{0}$, the restricted GPLM is not substantially different from the unrestricted GPLM. In this case, $\Delta = ||\boldsymbol{\beta} - \boldsymbol{\beta}_R||^2 = 0$, where $|| \cdot ||$ denotes the Euclidian norm.

To evaluate the behavior of the pretest and shrinkage estimators when the restricted model is significantly different from the unrestricted model, we assume $\boldsymbol{\beta}_{2R} = (\sqrt{d}, 0, 0, \ldots, 0)'$ so that $\Delta = ||\boldsymbol{\beta} - \boldsymbol{\beta}_R||^2 = d$, where d is a positive constant. Here Δ is the difference between the unrestricted and the restricted model according to the local alternative.

The numerical performance of the pretest and shrinkage estimators is evaluated under both $H_0 : \Delta = 0$ and $H_a : \Delta = d$ for $0 < d \leq 2.0$.

Based on the above setups of three defined models, the responses were generated using different Δ values, where $\Delta = (0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0)$ to study the effect of n, p_2 and Δ on restricted, pretest and shrinkage estimators in terms of bias and mean squared error.

We used 1000 replications for each scenario since the result did not change significantly with any increase in the number of replication. Generalized Speckman method were applied to estimate parameters. In this simulation and the application to credit score data in next section, we consider the Epanechnikov kernel function $K(u) = \frac{15}{16}(1-u^2)^2 I_{(|u|\leq 1)}$. The weight matrix $\boldsymbol{Q}$ in the quadratic loss function is Section (3.4.4) was set to $\boldsymbol{I}_{p\times p}$. We then calculate $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}^{PT}$, $\hat{\boldsymbol{\beta}}^{SE}$, and $\hat{\boldsymbol{\beta}}^{PSE}$ numerically for each simulated data set.

Bandwidth selection is quite important to visualize the distributions. For the estimation of nonparametric part $m_{\boldsymbol{\beta}}(t) = 1.5 \times sin(2\pi t)$ of our procedure, we used the Scott's rule of thumb to compute the bandwidth $h = 1.06 * \hat{\sigma} n^{-1/5}$ which satisfies the conditions of Theorem 1.

The quadratic bias, mean squared error (MSE), and relative mean squared error (RMSE) are used to evaluate the performance of any proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{g}}$ with respect to UG ($\hat{\boldsymbol{\beta}}$). The relative $\mathrm{RMSE}_{\mathrm{g}}$ is defined as

$$\mathrm{RMSE}_{\mathrm{g}} = \frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{g}})},$$

where $\mathrm{g} = 1, 2, 3, 4$ denote the relative MSE of the RG, PT, SE and PSE, respectively with respect to UG.

The MSE for $\mathrm{g} = 1, 2, 3, 4$ is calculated based on the formula

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{g}}) = \mathrm{trace}[\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{g}})] + ||\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{g}})||^2.$$

The trace of a covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{g}}$ and the average of $||\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{g}})||^2$ are calculated from 1000 simulated data sets to compute $\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{g}})$.

Note: When $\text{RMSE}_{\text{g}} > 1$, it indicates either the RG, PT, SE or PSE is better or has lower risk than the UG.

3.5.1 Quadratic Biases of UG, RG, PT, SE and PSE when $\Delta \geq 0$

The quadratic bias (QB) for each estimator is calculated as the difference between the mean of the estimates obtained from the 1000 replications and the true value of the parameter. As it is a scaler quantity, we can compare all estimators of logistic partially linear model for varying sample sizes $n = 250$ to $n = 350$.

The QBs of UG, RG, PT, SE and PSE when $\Delta \geq 0$ are given in Theorem 2, it shows that the non-centrality parameter Δ is common in the QB expressions except for UG, that is, we mainly compare biases for different values of Δ when the sample size varies.

The plots of QBs of the UG, RG, PT, SE and PSE with $\alpha = 0.05$ are provided in Figure 3.1 with different values of Δ. Some findings from these plots are as follows.

- For all values of Δ, the QB of the UG is approximately 0. However, due to sampling fluctuation, the simulated QB is not exactly zero.

- The QB of RG is unbounded as Δ increases and tends to ∞ as $\Delta \to \infty$.
- The QB of PT is a function of Δ and the significance level α. In this simulation study, we provided the result of PT for $\alpha = 0.05$.
- The QB of PT approaches to that of the UG as Δ increases.
- The SE and PSE are biased but are bounded in Δ and the PSE has lower or equal QB than the SE.
- As Δ increases, more sampling fluctuations occur in the QBs of all estimators for $n = 250$ as compared to the QBs for $n = 350$.

3.5.2 RMSEs of UG, RG, PT, SE and PSE when $\Delta \geq 0$

The RMSE results from the simulation study are presented in Figure 3.8 and Table 3.1 when the number of significant covariates $p_1 = 5$ and 7 along with the different number of insignificant covariates p_2.

Note: RMSEs for RG, PT, SE and PSE are presented by the dash (in green), dotted (in gold), dotdash (in blue), and longdash (in red) lines, respectively, and the curves are split up for sample sizes $n = 250$ and $n = 350$ by the filled circle and filled triangle point-up, respectively. The dotted pink line is the benchmark line for UG.

Summary of findings from Figure 3.8 and Table 3.1

- RMSEs of all the estimators relative to UG are highest when Δ is 0, n is fixed and p_2 varies, subject to the random sampling fluctuations.

- RG consistently outperforms all other estimators at and near the null hypothesis because of its unbiasedness.

- As Δ increases, the risk of the RG increases and becomes unbounded, that is, the RMSE of the RG decreases and approaches zero. Therefore, if RG is nearly correctly specified, it is an optimal estimator over the entire parameter space.

- When $\Delta = 0$ and p_2 values range from small to medium, PT outperforms the PSE and SE. However, the performance of PT diminishes compared to PSE and SE as p_2 increases. Therefore, for large p_2, PT underperforms SE and PSE.

- For example, when $n = 250$ and 350, we see that PSE outperforms SE (in terms of RMSE) for Δ close to zero (see Figures 3.8(a) and (f)). Nevertheless, SE and PSE are superior to UG in terms of RMSE for all the combinations of p_2 and n.

- For fixed n, SE and PSE gain the highest RMSEs when p_2 is large (e.g., $p_2 \geq 13$; see the curves with crosses at each data point) and Δ is zero.

It is interesting to see that the improvement due to shrinkage depends on the value of p_2 relative to n.

- As an example, note that $p_2 = 5$ relative to $n = 250$ is larger than $p_2 = 5$ relative to $n = 300$; a comparison between the RMSE in Table 3.1 indicates more improvement due to shrinkage for $p_2 = 5$ with $n = 250$ as compared to $p_2 = 5$ with $n = 300$. Similarly, $p_2 = 5$ with $n = 250$ produces more accurate estimates than $p_2 = 5$ with $n = 350$.

- For small-sized Δ (e.g $0 \leq \Delta < 0.5$), the PSE outperforms the SE. However, when Δ is large, the RMSE of PSE quickly converges to SE ($\hat{\boldsymbol{\beta}}^{PSE} = \hat{\boldsymbol{\beta}}^{SE}$). Thus, shrinkage is the most beneficial when there is no substantial difference between the unrestricted and the restricted GPLM.

- The improvement due to shrinkage depends on the value of p_2 relative to n. In general, shrinkage is useful when p_2 is large relative to n and $\Delta = 0$.

To explore the effect of p_2 and n, we also make a comparison between the MSE curves in Figure 3.15. For fixed n, we see that the larger the value of p_2, the less accurate the estimates are, a well known result in the real data analysis. We also see that as n increases with p_2 fixed, the accuracy of the estimates increases (e.g., compare Figures 3.15((a)-(f)).

We did not report the simulation results for backfitting algorithm as the

Table 3.1: RMSE of RG, PT, SE, and PSE with respect to UG when $\Delta = 0$. Here, $p_1 = 5, 7$ and $n = 250, 300, 350$.

	$p_2 = 5$	$p_2 = 10$	$p_2 = 15$	$p_2 = 3$	$p_2 = 8$	$p_2 = 13$
Estimators	$n = 250$, $p_1 = 5$			$n = 250$, $p_1 = 7$		
RG	1.81	3.03	5.47	1.41	2.31	4.33
PT	1.53	2.06	2.19	1.27	1.69	1.94
SE	1.34	1.98	2.78	1.10	1.63	2.24
PSE	1.40	2.05	2.86	1.13	1.68	2.27
Estimators	$n = 300$, $p_1 = 5$			$n = 300$, $p_1 = 7$		
RG	1.73	2.63	4.73	1.37	2.23	3.51
PT	1.49	2.00	2.20	1.26	1.63	2.13
SE	1.34	1.92	2.68	1.09	1.61	2.30
PSE	1.38	1.99	2.80	1.12	1.65	2.35
Estimators	$n = 350$, $p_1 = 5$			$n = 200$, $p_1 = 7$		
RG	1.67	2.59	3.97	1.37	2.03	3.17
PT	1.46	2.13	2.37	1.27	1.72	1.92
SE	1.32	1.90	2.57	1.06	1.57	2.12
PSE	1.37	1.98	2.69	1.12	1.61	2.18

quadratic biases and MSEs, and RMSE of this algorithm are very similar to generalized Speckman method.

In summary, our simulation study provides the following features:

(a) the PT, SE and PSE have uniformly lower risk than the UG across the entire parameter space;

(b) shrinkage is most useful when (i) there is no substantial difference between the unrestricted and the restricted GPLM ($\Delta \approx 0$), and (ii) the number of auxiliary covariates (p_2) is large relative to the sample size

n; and

(c) PSE outperforms SE when (i) $\Delta \approx 0$, and (ii) p_2 is large relative to n.

3.6 Application to Credit Scoring Data

Credit scores are markers that allow lenders and financial institutions to check a person's reliability for paying off the debt in time. Thus credit scoring data are quite important in the risk assessment process. We apply our proposed estimation methods to one such real credit scoring data.

The German credit scoring data set (available at https://archive.ics.uci.-edu/ml/datasets/ statlog + (german+credit+data)) contains observations on 20 socioeconomic variables for 1000 individuals, on the basis of which they have been classified as good or bad credit risks. All individuals belong to the same bank. The response variable, `Creditability` in the dataset corresponds to the risk label, 0 has been classified as bad credit risk (300 cases) and 1 has been classified as good credit risk (700 cases).

Our objective here is to construct a logistic partially regression model and to apply proposed methods to estimates the regression parameter efficiently that can be used to determine if a new applicant is in good or bad credit risk situation based on a set of socioeconomic variables. For fitting logistic regression, we merge classes of several categorical predictors because of insufficient number of observations in each category and we only used 10 out of 20 covariates. The following Table (3.2) presents nine categorial and one continuous covariates with their classes which are either numeric or categorical in nature.

Table 3.2: Descriptions of the selected variables of credit scoring dataset

Name of variable	Description	Value
1. Account.Balance	Status of existing checking account	No account (0) Having account (1)
2. Savings.Stocks	Savings account or stocks	No (0) Yes (1)
3. Length.Employment	Current employment period	Unemployed/ $<$ 1 year (0) 1 year or more (1)
4. Type.Apartment	Type of apartment	Not rented (0) Rented (1)
5. Purpose	Purpose of Credit	For tangible items (0) Others (1)
6. Payment.Status	Payment Status	No loan balance (0) Some outstanding loan (1)
7. No.Credits	Number of existing credits at the bank	One (0) Two or more (1)
8. Instalment	Installment rate in % of disposable income	Less than 20% (0) 20% or more (1)
9. Occupation	Occupation	Unskilled (0) Skilled (1)
10. Age	Age of applicant	Number of years

We initially take the unrestricted logistic partially regression model and fit this model using the Epanechnikov kernel

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 \text{ Account.Balance} + \beta_2 \text{ Savings.Stocks} \\
&+ \beta_3 \text{ Length.Employment} + \beta_4 \text{ Type.Apartment} + \beta_5 \text{ Purpose} \\
&+ \beta_6 \text{ Payment.Status} + \beta_7 \text{ No.Credits} + \beta_8 \text{ Instalment} \\
&+ \beta_9 \text{ Occupation} + \text{f(age)}.
\end{aligned}
$$

The backward elimination procedure based on AIC and residual deviance criteria are used to select a reduced GPLM that contains only the significant covariates from the unrestricted GPLM.

To find a restricted model which fits the data adequately, where the observed value does not differ so much from the predicted value, we look at the deviance residuals. We also look at the AIC value for the fitted model. We find that the deviance residual and AIC value for the unrestricted model are 1082.21 and 1112.4, respectively and also for restricted model are 1089.3 and 1114.1, respectively. This suggest that the restricted model is a useful and parsimonious approximation to the unrestricted model.

This procedure selects five significant predictors from the unrestricted GPLM: Account.Balance, Savings.Stocks, Length.Employment, Type.Apartment, and Purpose. We assume that $\boldsymbol{\beta}_2 = (\beta_6, \beta_7, \beta_8, \beta_9)^\top$ and fit the restricted model including only the significant predictors (Account.Balance, Savings.Stocks, Length.Employment, Type.Apartment, and Purpose) subject to the restriction $\boldsymbol{\beta}_2 = \mathbf{0}$. We use Scott's rule method (Scott, 1992) to select average optimal bandwidth of 7.53.

Table 3.3: Credit scoring data analysis – unrestricted, restricted, pretest, shrinkage and positive shrinkage estimates of the GPLM parameters and their standard errors

Coefficients	UG (SE)	RG (SE)	PT (BSE)	SE (BSE)	PSE (BSE)
β_1 (Account.Balance)	1.00(0.162)	1.03(0.161)	0.88(0.215)	0.87(0.214)	0.87(0.214)
β_2 (Savings.Stocks)	0.55(0.164)	0.52(0.161)	0.42(0.217)	0.42(0.216)	0.42(0.215)
β_3 (Length.Employment)	0.48(0.173)	0.47(0.171)	0.28(0.234)	0.28(0.234)	0.28(0.233)
β_4 (Type.Apartment)	0.51(0.164)	0.50(0.161)	0.26(0.218)	0.26(0.218)	0.26(0.218)
β_5 (Purpose)	−0.80(0.155)	−0.75(0.152)	−0.68(0.201)	−0.68(0.198)	−0.68(0.198)
β_6 (Payment.Status)	0.04(0.190)				
β_7 (No.Credits)	0.13(0.195)				
β_8 (Instalment)	0.36(0.152)				
β_9 (Occupation)	−0.22(0.184)				

Since we have one dataset, we use case resampling bootstrap method to calculate the estimates and standard errors of pretest, shrinkage and positive shrinkage estimators. In the bootstrap, we draw 1000 new samples (of size 800) from the data matrix $(\boldsymbol{y}_i^*, \boldsymbol{x}_i^*, T_i^*)$. We then refit the GPLM using these data based on the method described in Section 2. We compute the bootstrap parameter estimates using 1000 bootstrap samples. Table 3.3 presents the mean of the estimates obtained by the resampling bootstrap method for the credit scoring dataset and the bootstrap standard errors (SE: standard error and BSE: bootstrap standard error) of the estimated coefficients. The RMSEs of RG, PT, SE, and PSE with respect to UG are 1.43, 1.00, 1.14 and 1.15, respectively. The results are consistent with the simulation study and our theoretical findings provide recommendations about the usefulness of the proposed pretest and shrinkage estimators.

3.7 Summary

We have applied the maximum likelihood method with generalized Speckman algorithm to estimate the regression parameters of GPLM and named this as unrestricted estimate, UG. We also estimate the parameters when some of them are restricted to a subspace and named this as restricted estimate, RG. We study the relative risk dominance of the pretest and shrinkage estimators which are defined based on the UG and RG. We derive the expressions of biases and risks and used a Monte Carlo simulation study to calculate the numerical biases, mean squared errors and risks (inverse of relative mean squared error) of the estimators.

Our simulation studies show that the restricted estimator offer a numerically superior performance compared to the unrestricted, pretest and shrinkage estimators near the null hypothesis $\boldsymbol{R\beta} = \boldsymbol{r}$, but this estimator performs poorly when the restriction is seriously violated. The risk of the pretest estimator is lower than that of the UG (or higher relative MSE with respect to the UG) at and near the restriction in the simulation study. Our simulation study also shows that the shrinkage estimators have smaller mean squared error than the unrestricted estimators in terms of MSE for large region of parameter space even when there exist omitted significant predictor in the specified model. Under alternative hypothesis, it shows that the relative MSEs of PT, SE and PSE converge to one.

We applied the proposed estimation method to the credit scoring data. We calculated the RMSEs of RG, PT, SE and PSE with respect to UG based on the bootstrap resampling method as we cannot calculate RMSE based on one data set. It shows that RG, SE, and PSE perform better than the UG but PT does not show good performance for this dataset.

To summarize, it shows that simulation and real data results justify the better performance of shrinkage estimators in terms of higher accuracy and lower variability in the estimation of regression parameters for GPLM. It is of great interest to study the proposed method for the longitudinal data and when the number of covariates grows with the sample size.

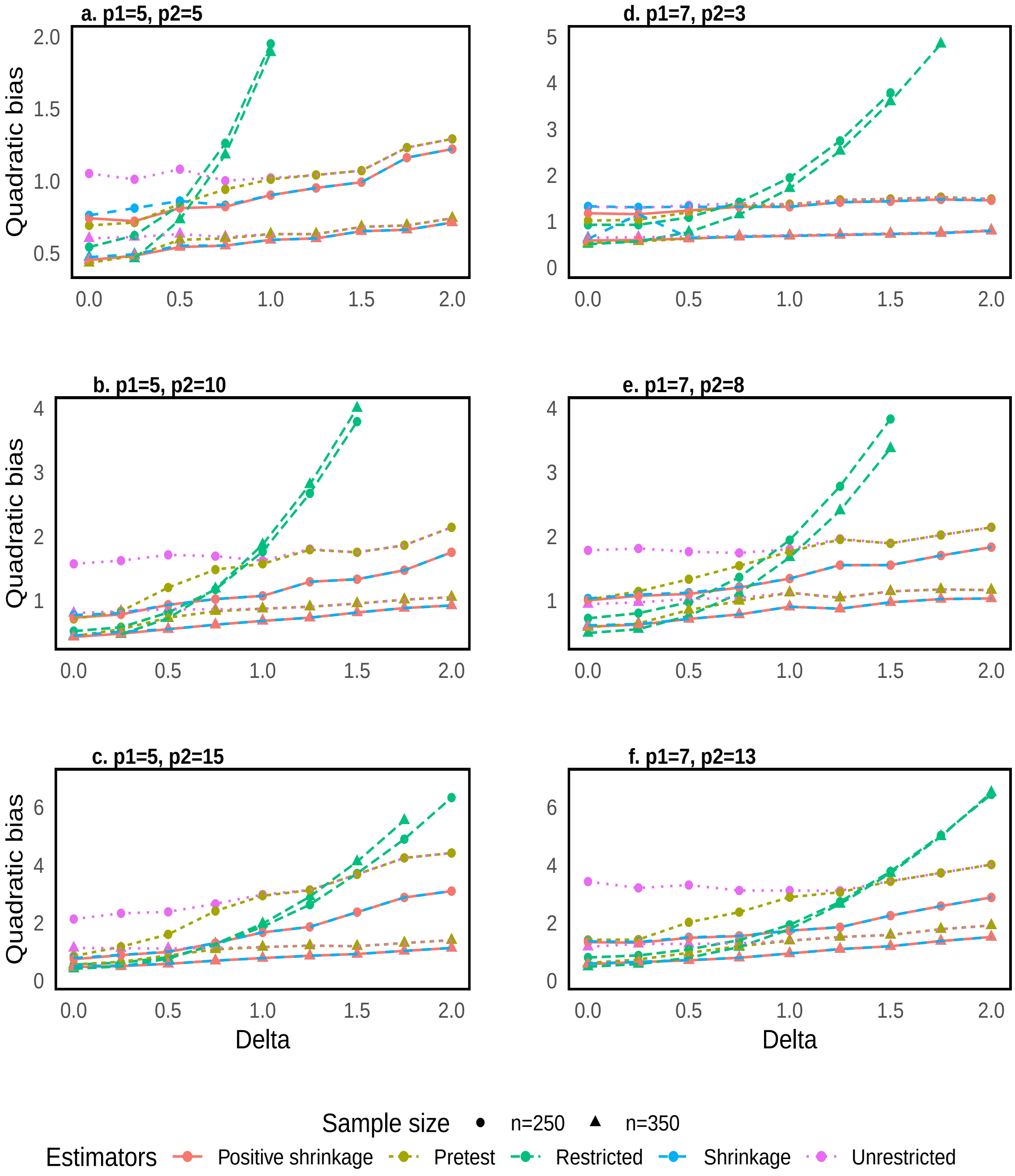

Figure 3.1: Simulated QB curves of proposed estimators in GPLM for $n = 250, 350$ and $p_2 = 3, 5, 8, 10, 13, 15$.

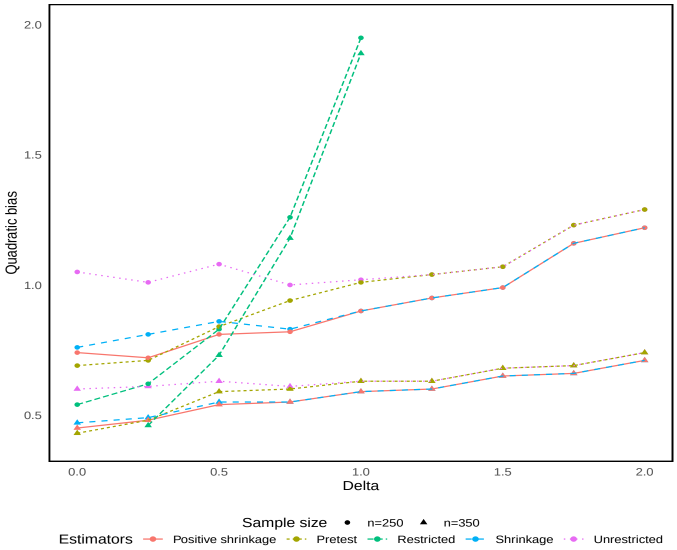

Figure 3.2: Simulated QB curves of proposed estimators in GPLM for $n =$ 250, 350 and $p = 10, p_1 = 5, p_2 = 5$.

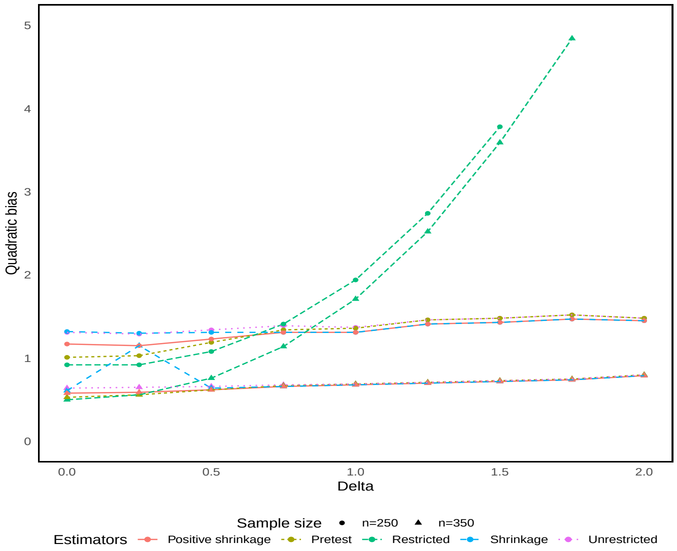

Figure 3.3: Simulated QB curves of proposed estimators in GPLM for $n =$ 250, 350 and $p = 10, p_1 = 7, p_2 = 3$.

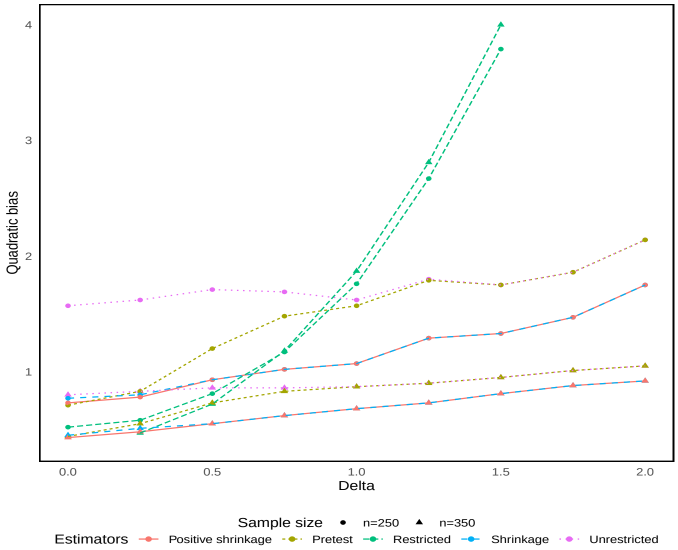

Figure 3.4: Simulated QB curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 5, p_2 = 10$.

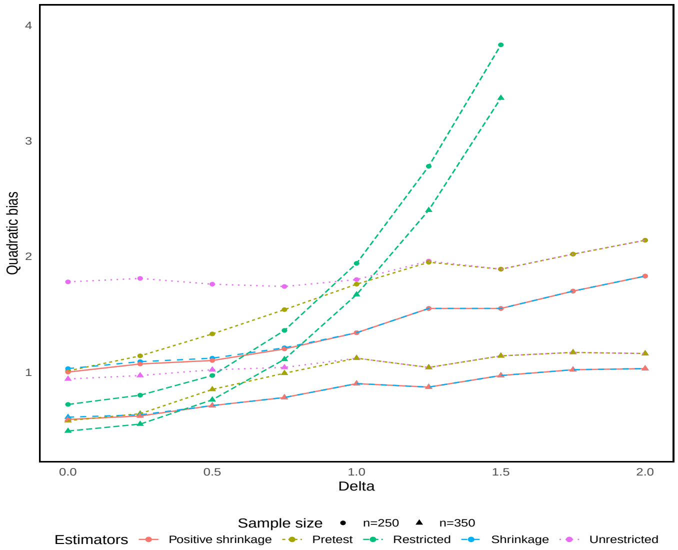

Figure 3.5: Simulated QB curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 7, p_2 = 08$.

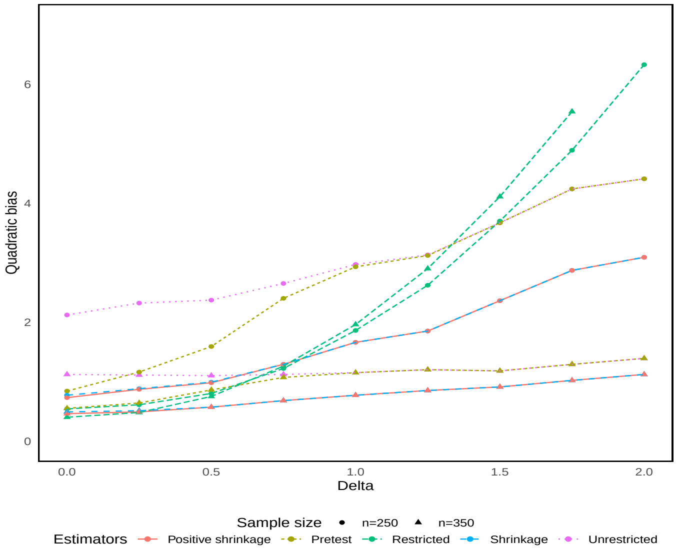

Figure 3.6: Simulated QB curves of proposed estimators in GPLM for $n =$ 250, 350 and $p = 20, p_1 = 5, p_2 = 15$.

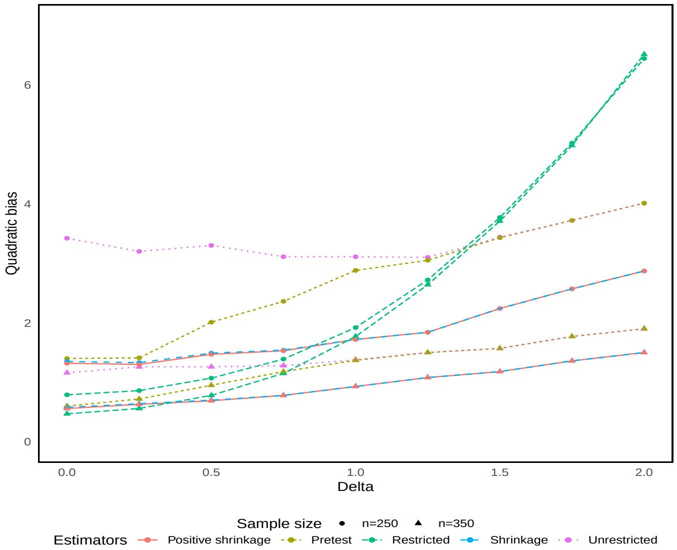

Figure 3.7: Simulated QB curves of proposed estimators in GPLM for $n =$ 250, 350 and $p = 20, p_1 = 7, p_2 = 13$.

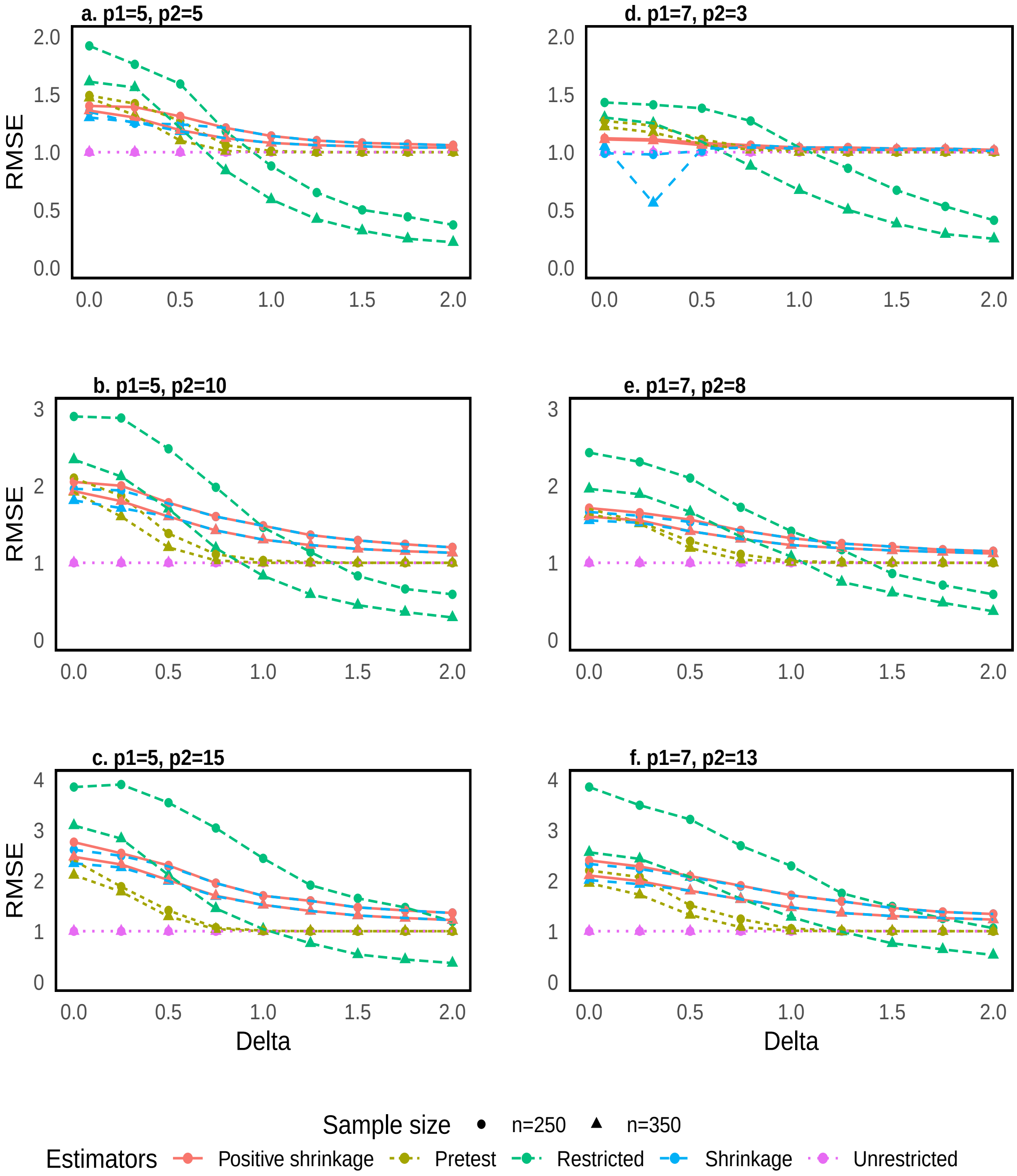

Figure 3.8: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p_2 = 3, 5, 8, 10, 13, 15$.

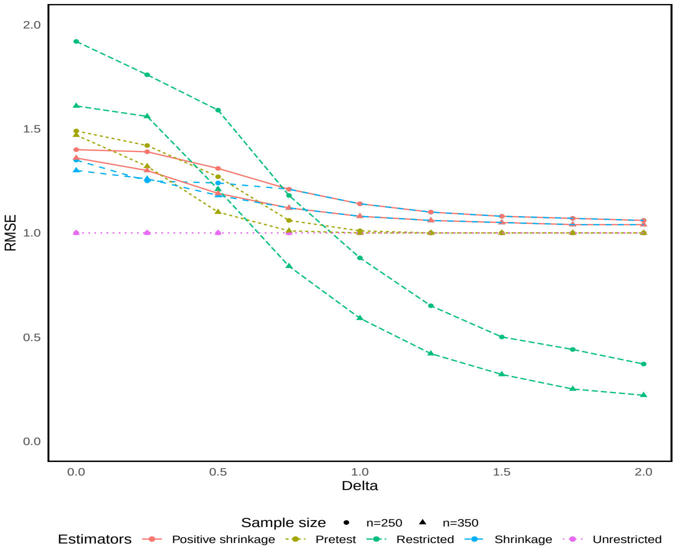

Figure 3.9: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 10, p_1 = 5, p_2 = 5$.

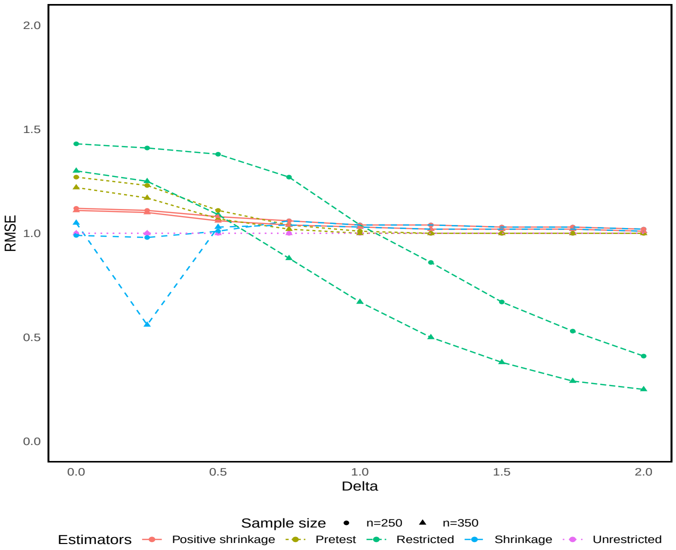

Figure 3.10: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 10, p_1 = 7, p_2 = 3$.

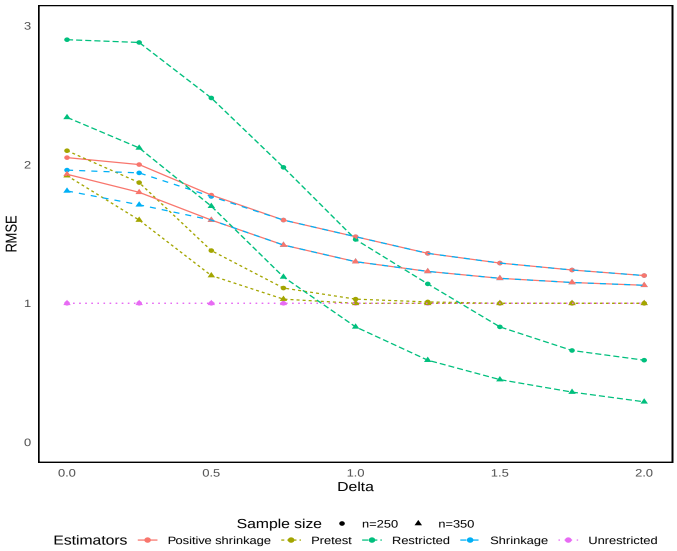

Figure 3.11: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 5, p_2 = 10$.

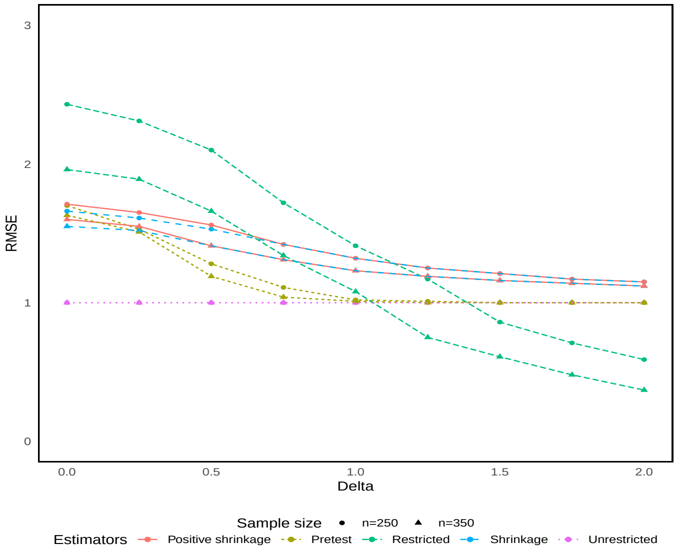

Figure 3.12: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 7, p_2 = 08$.

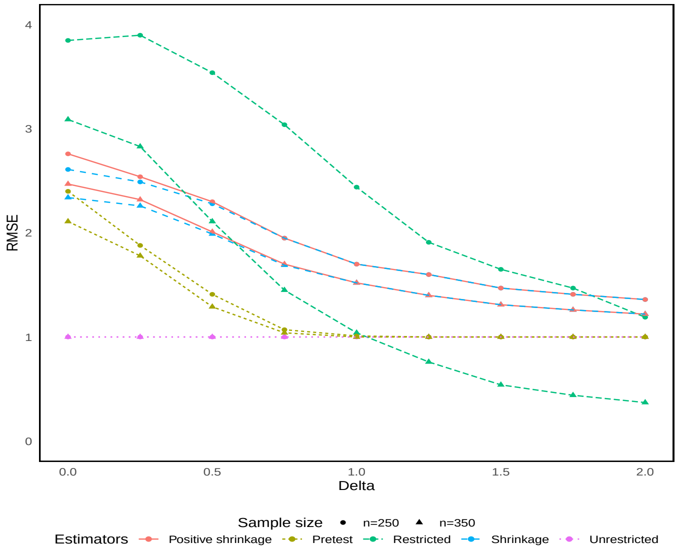

Figure 3.13: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 20, p_1 = 5, p_2 = 15$.

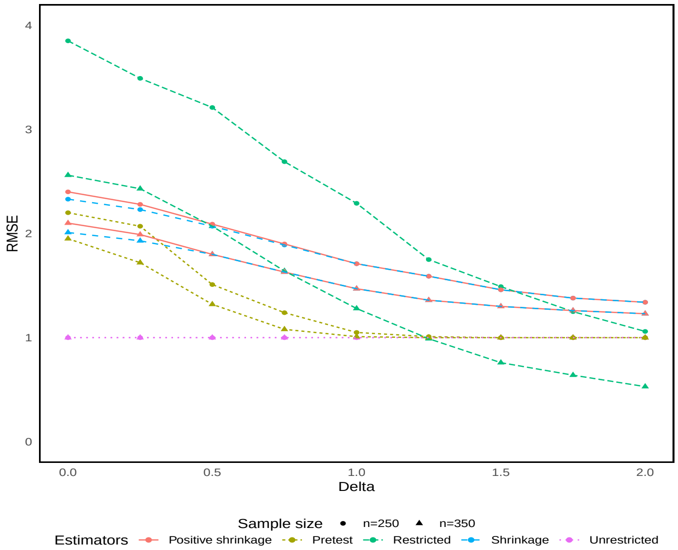

Figure 3.14: Simulated RMSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 20, p_1 = 7, p_2 = 13$.

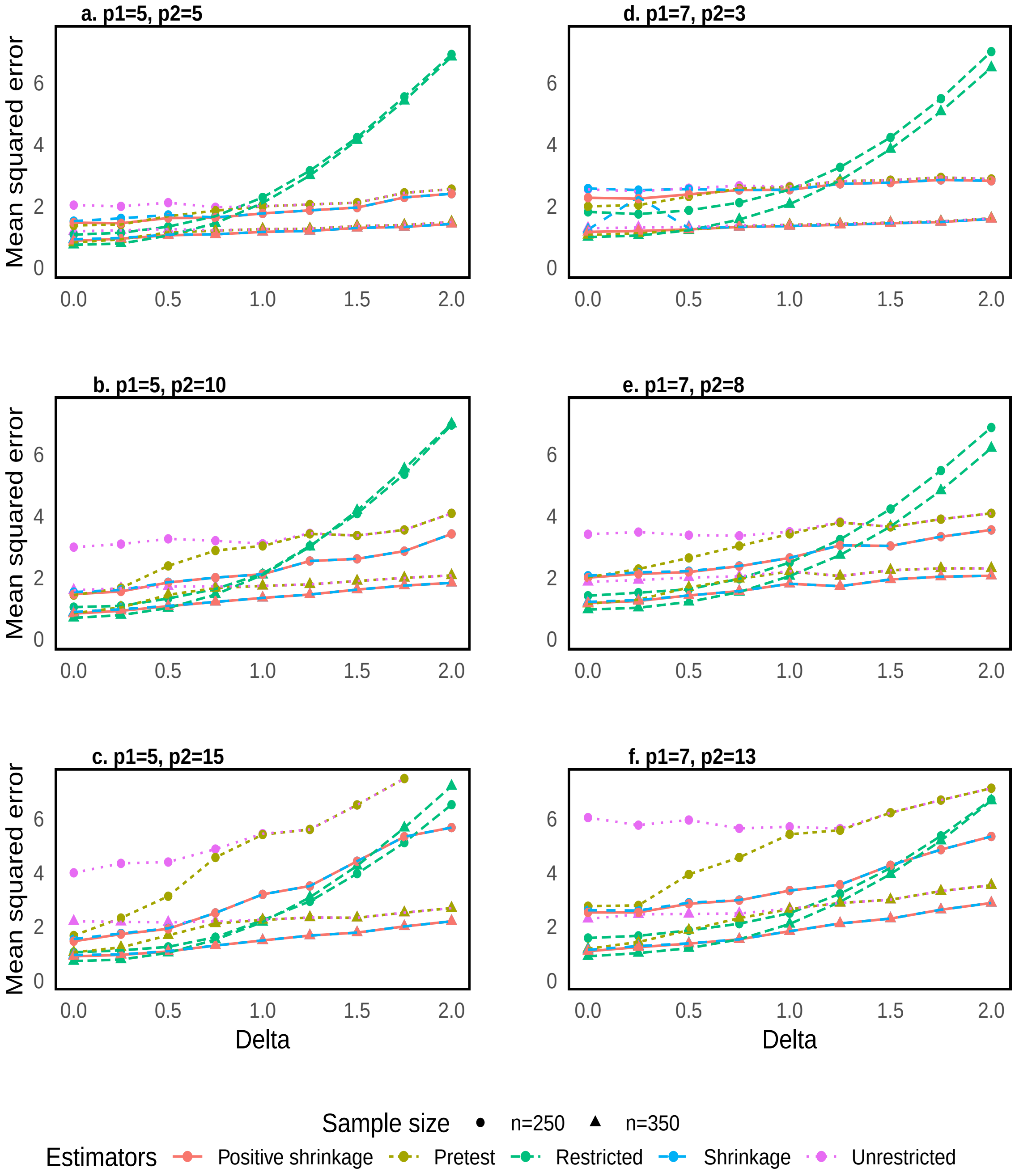

Figure 3.15: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p_2 = 3, 5, 8, 10, 13, 15$.

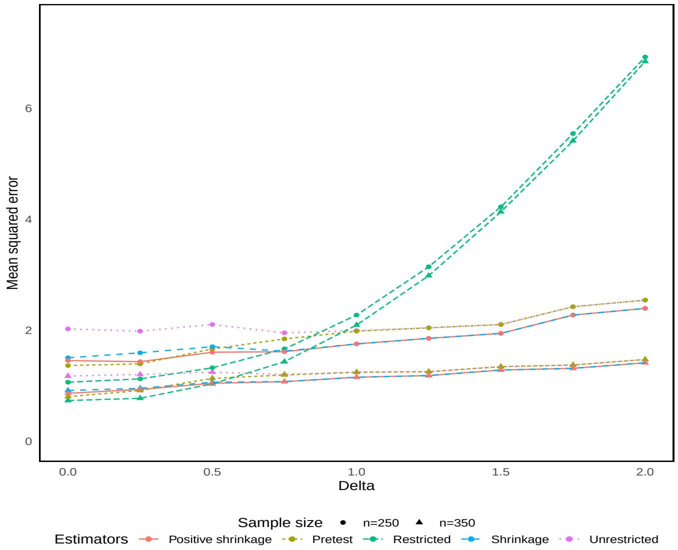

Figure 3.16: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 10, p_1 = 5, p_2 = 5$.

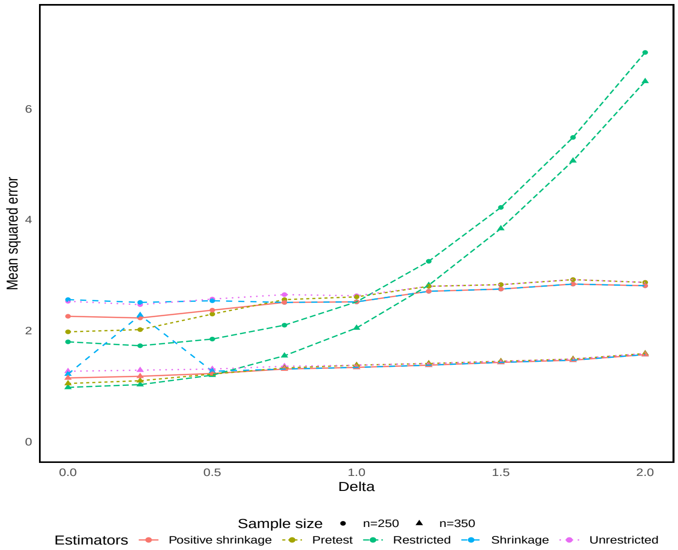

Figure 3.17: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 10, p_1 = 7, p_2 = 3$.

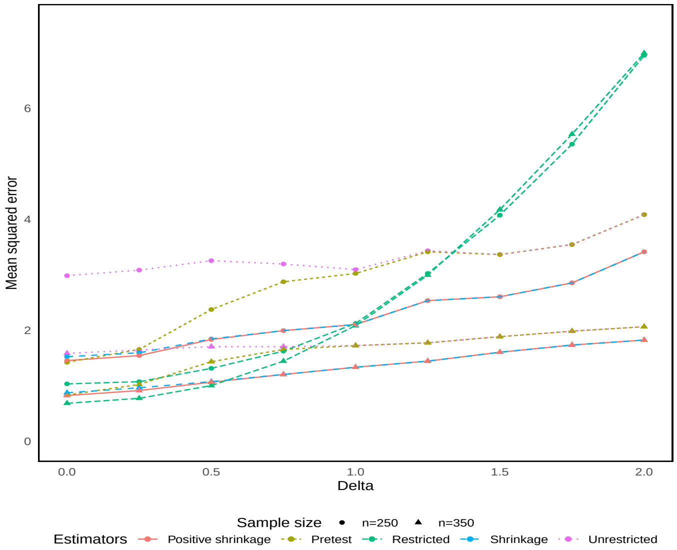

Figure 3.18: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 5, p_2 = 10$.

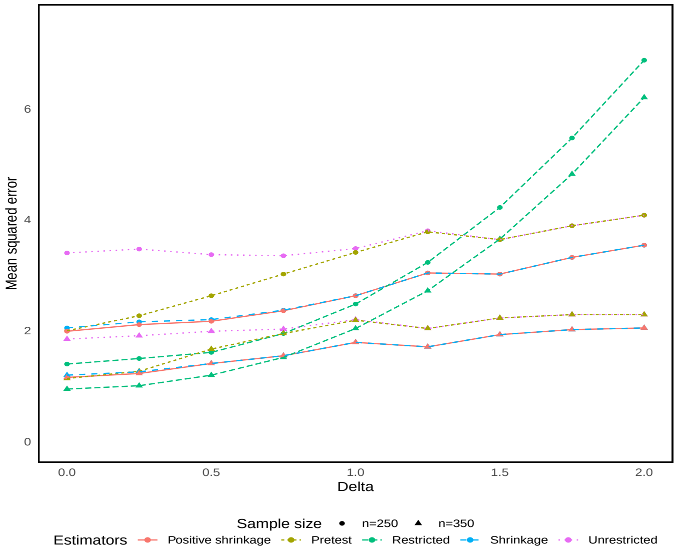

Figure 3.19: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 15, p_1 = 7, p_2 = 08$.

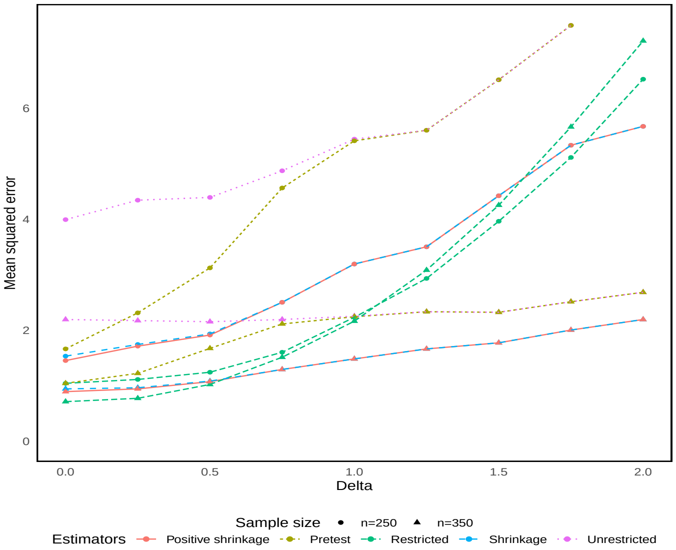

Figure 3.20: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 20, p_1 = 5, p_2 = 15$.

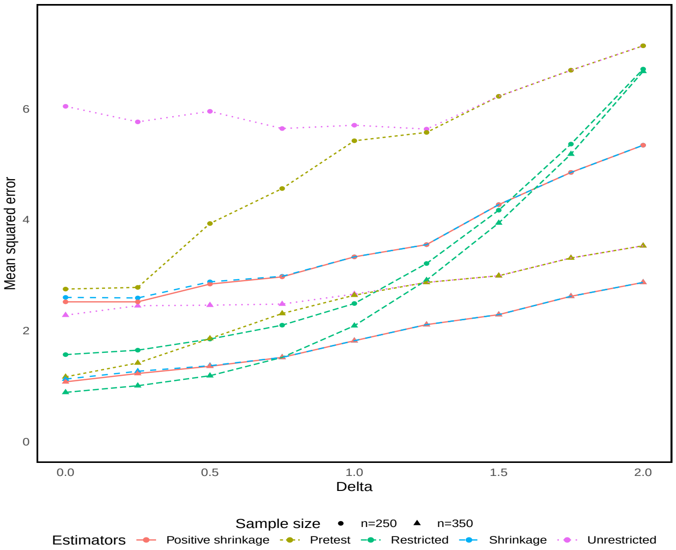

Figure 3.21: Simulated MSE curves of proposed estimators in GPLM for $n = 250$, 350 and $p = 20, p_1 = 7, p_2 = 13$.

Chapter 4

Optimal Design for Pretest and Shrinkage Estimators

4.1 Introduction

We often forget that a crucial step in statistical inference is the design stage. A research design is carefully used for data collection to obtain good estimation of the model parameters. A carefully designed study can provide valid, precise and efficient inference for the estimators at minimal time and cost. This fact motivated us to further improve our pretest and shrinkage estimators using optimal design theory.

In this chapter, we will first construct an optimal design according to some criterion of interest. Given an optimal design, we generate our data to obtain our pretest and shrinkage estimators. Our studies evidently show that the proposed estimators using optimal design theory outperform the estimators without using optimal design.

As our work is based on optimal design, we start with a general description of optimal design theory and some optimal design concepts such as design definition, design measure, variance-function, information matrix, different criterion functions. We also discuss some important properties. Later we will describe different types of design, requirements for a good design, and a class of optimization problems that may be needed according to the different types of designs. We will determine the optimality conditions according to our optimization problem and consider some algorithms to construct the optimal designs.

4.1.1 Experimental Design

Let y be the response variable. The models used to consider the problem of selecting a design have the form as

$$y \sim \pi(y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma),$$

where

- $\pi(.)$ is a probability model;
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^\top$ are design variables, $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$ with $\mathcal{X}$ being the design space;
- $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^T$ are unknown parameters. The true value of $\boldsymbol{\theta}$ belongs to a set $\Theta \in \mathbb{R}^k$.
- σ is an unkown fixed nuisance parameter.

In most applications, the design space $\mathcal{X}$ is compact. For each $\boldsymbol{x} \in \mathcal{X}$, an experiment is performed and the outcome is a random variable $y(\boldsymbol{x})$, where $\text{Var}(y(\boldsymbol{x})) = \sigma^2$.

A value for $\boldsymbol{x}$ is selected first from the design space to obtain an observation on the response y. Given that $\boldsymbol{x}$ can be set to any chosen value in $\mathcal{X}$, a natural question comes in mind is at what values of $\boldsymbol{x}$ should observations on y be taken so that we obtain a best inference for all or some of the parameters $\boldsymbol{\theta}$. Suppose that we take n observations. Such a best selection of the values of $\boldsymbol{x}$ or how to allocate the n observations to the elements of the design space is known as an optimal design.

4.1.2 Estimation of Parameters

Generally the mode of inference is decided first. Suppose this is point estimation of the parameters. The solution considered for this case will hold for other modes of inference as well.

It is important to choose n values ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$) to yield the "best" point estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Given the availability of methods in obtaining the esitmator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\theta}}$ be an unbiased estimator for $\boldsymbol{\theta}$ in which the components $\hat{\theta}_j$ are correlated, the $k \times k$ dispersion matrix of $\hat{\boldsymbol{\theta}}$ about $\boldsymbol{\theta}$ is defined by

$$D(\hat{\boldsymbol{\theta}}) = E([\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^\top).$$

Here $D(\hat{\boldsymbol{\theta}})$ contains information about the accuracy of the parameter estimators not only in its diagonal elements (which measure the mean squared deviation of $\hat{\theta}_j$ from θ_j), but also in the off-diagonal cross product elements. Therefore, the smaller is $D(\hat{\boldsymbol{\theta}})$ the better is the accuracy of the parameter estimators.

Let y_i be the observation obtained at $\boldsymbol{x}_i$, the linear models can be written as

$$E(y_i) = \boldsymbol{v}_i^\top \boldsymbol{\theta}, \tag{4.1}$$

where $\boldsymbol{v}_i = (f_1(\boldsymbol{x}_i), f_2(\boldsymbol{x}_i), \ldots, f_k(\boldsymbol{x}_i))^\top$ for $i = 1, 2, \ldots, n$. There will be several equalities between the $\boldsymbol{x}_i$'s and more than one observation is taken at the same value of $\boldsymbol{x}$.

Suppose that $y_1, y_2, \ldots, y_n$ are independent random variables with equal variance σ^2, y_i's then satisfy the linear model given as

$$E(Y) = X\boldsymbol{\theta}, \qquad D(Y) = \sigma^2 I_{n\times n}, \tag{4.2}$$

where

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$,
- $\boldsymbol{X}$ is an $n \times k$ design matrix whose $(i, j)th$ element is $f_j(\boldsymbol{x}_i)$
- $I_{n\times n}$ is an $n \times n$ identity matrix,
- $D(\boldsymbol{y})$ is the dispersion matrix of $\boldsymbol{y}$.

Least squares estimators are usually a conventional choice for the standard linear model (4.2) and have the property of being the best linear unbiased estimators (BLUE). The estimators are the solution of the following equation

$$(\boldsymbol{X}^\top \boldsymbol{X})\,\hat{\boldsymbol{\theta}} \;=\; \boldsymbol{X}^\top \boldsymbol{y}. \tag{4.3}$$

Here, $\boldsymbol{X}^\top\boldsymbol{X}$ is a $k \times k$ information matrix of $\boldsymbol{\theta}$. The larger the $\boldsymbol{X}^\top\boldsymbol{X}$ is, the greater the information is in the experiment. However, if all the parameters $\boldsymbol{\theta}$ are of interest, we select $\boldsymbol{x}$ in such a way that the $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular. In this case, the equation (4.3) will give a unique solution which has the form as

$$\hat{\boldsymbol{\theta}} \;=\; (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}. \tag{4.4}$$

The expectation and dipersion matrix of the estimator $\hat{\boldsymbol{\theta}}$ are $\boldsymbol{\theta}$ and $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, respectively. The predicted value of the response at $\boldsymbol{x}$ is

$$\hat{Y}(\boldsymbol{x}) = \boldsymbol{f}^\top(\boldsymbol{x})\hat{\boldsymbol{\theta}}$$

where $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}),\, f_2(\boldsymbol{x}),\, \ldots,\, f_k(\boldsymbol{x}))^T$.

Remarks:

1. As we can see the dispersion matrix of $\hat{\boldsymbol{\theta}}$ does not depend on $\boldsymbol{\theta}$ but depends proportionally on the parameter σ^2. Therefore, $\{\boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \ldots,\, \boldsymbol{x}_n\}$ needs to be selected in such way to make $D(\hat{\boldsymbol{\theta}})$ as small as possible or to make $(\boldsymbol{X}^\top \boldsymbol{X})$ large in some sense.

2. The explicit form of the expected value of $y(\boldsymbol{x})$ can be written as $E(y|\boldsymbol{v}) = \boldsymbol{v}^\top\boldsymbol{\theta}$, where $\boldsymbol{v} = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))^T$ for $\boldsymbol{v} \in \mathcal{V}$; and $\{\boldsymbol{v} \in \mathbb{R}^k \;:\; \boldsymbol{v} = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))^T,\; \boldsymbol{x} \in \mathcal{X}\}$. In this case, choosing a vector $\boldsymbol{x}$ from the design space $\mathcal{X}$ is same as choosing a k-vector $\boldsymbol{v}$ in the k-dimensional space $\mathcal{V} = \boldsymbol{f}(\mathcal{X})$. Hence $\mathcal{V}$ is the image

under f of the original design space $\mathcal{X}$. Thus, $\mathcal{V}$ is also an induced design space.

4.1.3 Discrete Design Space

The original design space $\mathcal{X}$ is continuous. To express the design problem more precisely, we consider discretization. After the discretization of $\mathcal{X}$, we can deal with the induced design space $\mathcal{V}$ which consists of J distinct vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_J$. At this point, typically $\mathcal{V}$ is a discrete design space. A value for $\boldsymbol{v}$ must be chosen from the J elements of the space $\mathcal{V}$ to obtain an observation on y.

We now need to find the answer to two questions. First, at which of the points $\boldsymbol{v}_j$ should observations be taken? Second, how many observations should be taken at these points to obtain "best" least squares estimators of the parameters?

There are two options to solve this problem; one is by means of an exact design problem and the other one is by an approximate design problem.

4.1.4 Exact Design Problem

Suppose we observe n observations. We must decide how many of these, say n_j, to take at $\boldsymbol{v}_j$, $\sum\limits_{j=1}^{J} n_j = n$. Let $\boldsymbol{n} = (n_1, n_2, \ldots, n_J)^\top$, the matrix $(\boldsymbol{X}^\top \boldsymbol{X})$ can be expressed as

$$\boldsymbol{X}^\top \boldsymbol{X} = M(\boldsymbol{n}) = \sum_{j=1}^{J} n_j \boldsymbol{v}_j \boldsymbol{v}_j^\top = \boldsymbol{V} \boldsymbol{N} \boldsymbol{V}^\top$$

where $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_J]$ and $\boldsymbol{N} = \operatorname{diag}(n_1, n_2, \ldots, n_J)$.

Our problem is to choose $\boldsymbol{n}$ in such way to make $M(\boldsymbol{n})$ big in some sense. Due to integer programming problem, n_j's are required to be integer. In this context, this is known as an exact design problem. However, as the theory of calculus cannot be used to identify optimal solutions, integer programming problems are usually difficult to solve even without any additional constraints. Fortunately, there is a convenient way to solve this problem, which is given in the following.

4.1.5 Approximate Design Problem

To overcome the limitation of interger programming problem, approximate design is preferrable. Consider $\boldsymbol{M}(\boldsymbol{n}) = n\boldsymbol{M}(p)$ where $\boldsymbol{M}(p)$ can be expressed as

$$\boldsymbol{M}(p) = \sum_{j=1}^{J} p_j \, \boldsymbol{v}_j \, \boldsymbol{v}_j^\top = \boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top.$$

In this form, $\boldsymbol{P} = \text{diag}(p_1, \, p_2, \, \ldots, \, p_J)$ with $p_j{=}\frac{n_j}{n}$ for $j = 1, 2, \ldots, J$. Here, p_j is the proportion of observations taken at $\boldsymbol{v_j}$, $p_j \geq 0$, and $\sum\limits_{j=1}^{J} p_j = 1$ and $p = (p_1, \, p_2, \, \ldots, \, p_J)$ gives the resultant distribution on the induced design space $\mathcal{V}$.

The matrix $\boldsymbol{M}(p)$ is known as the information matrix. Thus the problem becomes choosing p in such way to make $\boldsymbol{M}(p)$ large subject to $p_j{=}n_j/n$. Relaxing the latter design to $p_j \geq 0$ and $\sum\limits_{j=1}^{J} p_j = 1$ will give an approximate design problem. This is a simpler or more flexible problem to solve. Another advantage is that it is not much visibly different from the original.

4.1.6 Design Measure and the Variance Function

Consider p as a defining probability distribution on $\mathcal{V}$ we have

$$\boldsymbol{M}(p) \quad = \quad E_p[\boldsymbol{v}\,\boldsymbol{v}^\top] \quad \text{where} \quad P(\boldsymbol{v} = \boldsymbol{v}_j) = p_j. \tag{4.5}$$

Since the information matrix $\boldsymbol{M}(p)$ is symmetric and nonnegative definite, a design can be defined by a set of weights or probabilities p_j, where p_j is assigned to $\boldsymbol{v}_j \in \mathcal{V}$. In this design, the weight p_j may be set to 0.

The notation p is referred as the vector $(p_1,\, p_2,\, \ldots,\, p_J)$ and also as a probability distribution on $\mathcal{V}$ that induces a distribution or measure on the original design space $\mathcal{X}$. We have the full form as

$$p \;=\; \left\{ \begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_J \\ p_1 & p_2 & \ldots & p_J \end{array} \right\} \tag{4.6}$$

where the first line gives the design points with p_j the associated design weights. Note that $\sum\limits_{j=1}^{J} p_j = 1$ and $0 \leq p_j \leq 1$ for all j.

The support of a design measure p in the design space $\mathcal{V}$ is defined to be those vertices $\boldsymbol{v}_j$ with nonzero weights under p, and it is given by:

$$Supp(p) \;=\; \{\boldsymbol{v_j} \in \mathcal{V} \,:\, p_j > 0,\, j = 1, 2,\, \ldots,\, J\}$$

An optimal design, say p^* often existes in such way that $Supp(p^*)$ is a strict subset of $\mathcal{V}$. The standardized variance of the predicted response at $\boldsymbol{x}$ for the design (4.6) is given by

$$d(\boldsymbol{x},\, p) \;=\; \boldsymbol{f}^\top(\boldsymbol{x})\, \boldsymbol{M}^{-1}(p)\, \boldsymbol{f}(\boldsymbol{x}), \tag{4.7}$$

where $\boldsymbol{M}(p)$ is the information matrix. This standardized variance plays an important role in our optimization problem.

4.2 Design Criteria

In statistical modeling, the first objective is to obtain the good estimation of the parameters. Good estimation is usually defined by a variety of criteria. The criterion function, say ϕ, can be expressed in terms of the information matrix $\boldsymbol{M}(p)$. We have

$$\phi(p) = \psi\{\boldsymbol{M}(p)\}$$

where $\boldsymbol{M}(p) = \sum_{j=1}^{J} p_j \, \boldsymbol{v}_j \, \boldsymbol{v}_j^\top = \boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top$.

It may be possible to obtain the best inference for all or some of the unknown parameters $\boldsymbol{\theta} \in \Theta$ by making $\boldsymbol{M}(p)$ large. Therefore, serveral ways to maximize the real valued function $\phi(p) = \psi\{\boldsymbol{M}(p)\}$ may be considered to make $\boldsymbol{M}(p)$ large.

The criterion defined by the function ϕ is usually called ϕ-optimality and the design which maximizes $\phi(p)$ is called a ϕ-optimal design. Variety of criteria has been studied in the literature. Some possible criteria in the context of our estimation strategy in pretest and shrinkage estimation are D-optimality, A-optimality, G-optimality, linear optimality and c-optimality. In this research, we will generate our data for pretest and shrinkage estimation based on the D-optimality criterion. This criterion is also the most important design criterion in applications.

4.2.1 D-optimality

In D-optimality, we maximize the determinant of the information matrix $\boldsymbol{M}(p)$ or the logarithm of its determinant $log\,det\,\{\boldsymbol{M}(p)\}$. The criterion function being maximized is defined as

$$\phi_D(p) = \psi_D\{\boldsymbol{M}(p)\} = log\,det\{\boldsymbol{M}(p)\} = \; -log\,det\{\boldsymbol{M}^{-1}(p)\}. \tag{4.8}$$

Maximizing the determinant of the information matrix is equivalent to minimizing the determinant of the covariance matrix of the parameter estimators (the reciprocity property of the covariance matrix and the information matrix). Therefore, in D-optimality, we minimize the generalized variance of the parameter estimators.

There is also an interesting link between the D-optimal design and the standardized variance of the predicted response. Suppose we have a design variable x for a given model and let p^* be the D-optimal design.

Kiefer and Wolfowitz (1960) show that maximizing the D-optimal criterion is equivalent to

$$\underset{p}{inf} \;\; \underset{x}{sup} \;\; d(x,p) \;\; = \;\; \underset{x}{sup} \;\; d(x,p^*) = k$$

where $d(x,p) = \boldsymbol{f}^{\top}(x)\,\boldsymbol{M}^{-1}(p)\,\boldsymbol{f}(x)$ is the standardized variance of the predicted response and k is the number of parameters.

Various motivations for D-optimality extend beyond to the idea of point estimation and joint inference of the parameters $\boldsymbol{\theta}$. If the error terms are assumed to follow a normal distribution, the general form of the joint confidence region for $\boldsymbol{\theta} \in \Theta$ is described by an ellipsoid

$$\{\boldsymbol{\theta} \;:\; (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\top} \, \boldsymbol{M}(p) \, (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \;\leq\; c\}, \tag{4.9}$$

where c is a critical value and $\hat{\boldsymbol{\theta}}$ is the least squares estimator or the maximum likelihood estimator of $\boldsymbol{\theta}$. In D-optimality criterion, $\boldsymbol{M}(p)$ is choosen to make the volume of the ellipsoid as small as possible since this volume is proportional to $[det\{\boldsymbol{M}(p)\}]^{-\frac{1}{2}}$. The value of $[log\, det\{\boldsymbol{M}(p)\}]$ is finite if and only if $\boldsymbol{M}(p)$ is non-singular (i.e. when all the unknown parameters are estimable).

On the other hand, the above D-optimality criterion can be expressed in terms of the eigenvalues of $\boldsymbol{M}(p)$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of $\boldsymbol{M}(p)$, the eigenvalues of $\boldsymbol{M}^{-1}(p)$ are $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_k$. These eigenvalues are proportional to the squares of the lengths of the axes of the confidence ellipsoid. Thus we see that the D-optimal design minimizes the product of the eigenvalues of $\boldsymbol{M}^{-1}(p)$: $\prod_{i=1}^{k} 1/\lambda_i$.

This is the most extensively studied of all design criteria. The references include Kiefer (1959), Fedorov (1972), Silvey (1980), Berger and Wong (2009), Atkinson et al. (2007), Shah and Sinha (1989), Pukelsheim (1993), Mandal and Torsney (2006), Mandal et al. (2005), Torsney and Mandal (2001), Torsney (1983), and Torsney (1988).

4.2.2 An Important Property of D-optimality

The D-optimality criterion $[\phi_D(p)]$ has several useful properties. We will discuss one of the notable properties of this criterion. Since the D-optimality criterion is a concave function of the positive definite symmetric matrices, whenever the criterion function ϕ_D is finite, it is differentiable and its first partial derivatives are given by

$$\frac{\partial \phi_D}{\partial p_j} \;=\; \boldsymbol{v}_j^\top \boldsymbol{M}^{-1}(p)\, \boldsymbol{v}_j. \tag{4.10}$$

Property:

The criterion function ϕ_D is invariant under a non-singular linear transformation of $\mathcal{V}$. That is, the D-optimal design is invariant under linear transformation of the scale of the independent variables.

Proof:

Recall $\boldsymbol{M}(p)$: $\boldsymbol{M}(p) = \boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top$. Suppose $\mathcal{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_J]$ is the transformation of $\mathcal{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_J]$ under the linear transformation $\boldsymbol{\omega}_j = \boldsymbol{A}\boldsymbol{v}_j$, where $\boldsymbol{A}$ is a $k \times k$ non-singular matrix, the information matrix of a design after assigning weight p_j to $\boldsymbol{\omega}_j$ has the form as

$$\boldsymbol{M}_\omega(p) = \boldsymbol{\mathcal{W}} \boldsymbol{P} \boldsymbol{\mathcal{W}}^\top = \boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top \boldsymbol{A}^\top.$$

The criterion function $\phi_D\{M_\omega(p)\}$ then can be obtained as follows.

$$\begin{aligned} \phi_D\{\boldsymbol{M}_\omega(p)\} &= log\, det\{\boldsymbol{M}_\omega(p)\} \\ &= log\, det\{\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top \boldsymbol{A}^\top\} \\ &= log\,[\,det\{\boldsymbol{V}\boldsymbol{P}\boldsymbol{V}^\top\} \times det\{\boldsymbol{A}\}^2\,] \\ &= log\, det\{\boldsymbol{M}(p)\} + log\, det\{\boldsymbol{A}\}^2 \\ &= \phi_D\{\boldsymbol{M}(p)\} + \text{c}, \quad \text{where c is a constant.} \end{aligned}$$

4.2.3 Relative Efficiency

To compare different designs, we consider the relative efficiency which is a function or measure that enables us to compare the efficiencies of two designs. Let p be a design of any given model of k parameters and let p^* be the D-optimal design, the relative efficiency of the design p with respect to the D-optimal design p^* (i.e., D-efficiency of the design p) is defined as

$$D_{eff} = \left\{ \frac{det\, \boldsymbol{M}(p)}{det\, \boldsymbol{M}(p^*)} \right\}^{1/k}. \tag{4.11}$$

Note that taking the kth root of the ratio of the determinants will give us an efficiency measure that is proportional to design size, irrespective of the dimension of the model.

4.2.4 Other Criteria

As mentioned in previous subsection, apart from D-optimality there are also several other choices of criteria of interest, such as: A-optimality, G-optimality, linear optimality and c-optimality. Following is a summary of criterion functions of c-optimality, linear optimality, and D_A-optimality.

1. c-optimality: $\psi\{\boldsymbol{M}(p)\}$ = -$\boldsymbol{c}^\top \boldsymbol{M}^{-1}(p)\boldsymbol{c}$.
2. Linear optimality: $\psi\{\boldsymbol{M}(p)\}$ = -Trace $(A\boldsymbol{M}^{-1}(p)A^\top)$.
3. D_A-optimality: $\psi\{\boldsymbol{M}(p)\}$ = -logdet$(A\boldsymbol{M}^{-1}(p)A^\top)$.

Here, $\boldsymbol{c}$ and $\boldsymbol{A}$ are respectively a given vector and $s \times k$ matrix where $s < k$. It is noted that c-optimality is appropriate if the interest is only about $\boldsymbol{c}^\top\boldsymbol{\theta}$. However, if the interest is in the inference for $\boldsymbol{A\theta}$, linear optimality or D_A-optimality will be more suitable.

4.3 Optimality Conditions

Our goal is to obtain an optimal design based on a criterion function. This is equivalent to choosing the proportion p_j of observations, taken at x_j to ensure good estimation of $\boldsymbol{\theta}$ by optimizing some criterion. Once the criterion function is defined, we need to determine the optimality conditions based on the constraints of the proposed optimization problem. For any of the optimality criteria as listed in previous subsection, the criterion function $\phi(p)$ will be optimized subject to the constraints $p_j \geq 0$ and $\sum\limits_{j=1}^{J} p_j = 1$. This optimization problem is considered as a general problem.

Various problems in statistics require the calculation of such probability distributions or measures and optimal regression design is an example. Other examples are parameter estimation, adaptive design, and stratified sampling. Generally, to find an optimizing distribution, we need to determine optimality conditions for a given class of optimization problems.

4.3.1 Examples of the general problems

To find an optimizing distribution, says p^*, of the above general problem, recall that p can be referred as a probability distribution on both the induced design space $\mathcal{V}$ and the original design space $\mathcal{X}$.

We express our optimization problem in two different formats - one in terms of the design p and the other in terms of the information matrix $\boldsymbol{M}(p)$.

Optimization problem one in terms of the design p:
Maximize a criterion $\phi(p)$ over $\mathcal{P} \equiv \big\{p = (p_1, p_2, \ldots, p_J) : p_j \geq 0, \sum_{j=1}^{J} p_j = 1\big\}$. This problem can be viewed as a constrained optimization problem. The full constraint region is a closed bounded convex set.

Optimization problem two in terms of the design $\boldsymbol{M}(p)$:
Maximize $\psi(x)$ over the convex hull (of the points $G(\boldsymbol{v}_1), \ldots, G(\boldsymbol{v}_J)$)

$$\mathcal{CH}\{\mathcal{G}(\mathcal{V})\} \;=\; \big\{x = x(p) = \sum_{j=1}^{J} p_j\, G(\boldsymbol{v}_j) \;:\; p = (p_1,\, p_2,\, \ldots, p_J) \in \mathcal{P}\big\},$$

where $G(.)$ is a given one-to-one function and $\mathcal{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J\}$ is a known set of vector (or matrix) vertices of fixed dimension.

Alternatively, assuming the value $G(\boldsymbol{v}_j)$ with probability p_j, we can express $x(p)$ as $x(p) = E_p[G(\boldsymbol{v})]$, where $G(\boldsymbol{v})$ is a random variable. So the previous problem can be solved for

$$\phi(p) \;=\; \psi\{E_p[G(\boldsymbol{v})]\},\;\; x \;=\; E_p[G(\boldsymbol{v})] \;=\; \sum_{j=1}^{J} p_j G(\boldsymbol{v}_j). \qquad (4.12)$$

An example of either the above two problems is clearly a general optimal linear regression design problem. However, as with our design problem, a generalization of the second problem would be to seek a probability measure which is defined on $\mathcal{V}$ to maximize a function $\phi(p)$.

When we consider optimality conditions for an optimization problem, we could approach in two ways. First, we could find an optimizing p^* directly. Second, we can first find an x^* which maximizes $\psi(x)$ over $\mathcal{CH}\{\mathcal{G}(\mathcal{V})\}$ and then find a p^* such that $x(p^*) = x^*$. However, we will consider the former one as the principal approach. This approach requires conditions explicitly to define an optimizing p^*.

4.3.2 Directional derivatives:

The emphasis in finding the optimizing distribution is on a differential calculus approach. One of the tools playing an important simplifying role in the differential calculus of optimization is the directional derivative of Whittle (1973). We define optimality conditions in terms of point to point directional derivatives of a criterion function $\phi(p)$ at a point p in the direction of another point q. The directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q is defined as

$$F_\phi\{p, q\} = \lim_{\varepsilon \to 0+} \frac{\phi\{(1-\varepsilon)p + \varepsilon q\} - \phi(p)}{\varepsilon}. \tag{4.13}$$

Here the criterion function $\phi(.)$ can be any function with no constraints on p. The derivative of $F_\phi\{p,\, q\}$ exists even if $\phi(.)$ is not differentiable. However, if $\phi(.)$ is differentiable, then (4.13) becomes

$$F_\phi(p, q) = (q-p)^\top \frac{\partial \phi}{\partial p} = \sum_{j=1}^{J} (q_j - p_j)\, d_j, \text{ where } d_j = \partial\phi/\partial p_j,\ j = 1, 2, \ldots, J,$$

(Mandal and Torsney, 2006).

Let F_j denote the vertex directional derivative of $\phi(.)$ at p. To find the directional derivatives towards the vertices of the feasible region, we first need to simplify $F_\phi(p, q)$ by taking q as e_j where e_j is the j^{th} unit vector in $\mathbb{R}^J$. Thus F_j can be expressed as

$$F_j = F_\phi(p, e_j) = \frac{\partial \phi}{\partial p_j} - \sum_{i=1}^{J} p_i \frac{\partial \phi}{\partial p_i} = d_j - \sum_{i=1}^{J} p_i d_i.$$

Remarks:

- If $\phi(.)$ is differentiable at an optimizing distribution p^*, the first-order conditions for $\phi(p^*)$ to be a local maximum of $\phi(.)$ in the feasible region of the problem are

$$F_j^* = F_\phi\{p^*, e_j\} \begin{cases} = 0 & \text{for} \quad p_j^* > 0 \\ \leq 0 & \text{for} \quad p_j^* = 0. \end{cases} \tag{4.14}$$

- If the criterion $\phi(.)$ is concave on the feasible region, the first-order condition (4.14) is both necessary and sufficient for optimality. This result is known as the general equivalence theorem in optimal design (Kiefer, 1974).

4.4 A Class of Algorithms

Typically it is not possible to evaluate an optimal solution p^* explicitly. Therefore, in order to determine the optimal weights, we often require an algorithm. A class of multiplicative algorithms which neatly satisfy the constraints of the general problem has the form as given by

$$p_j^{(r+1)} \quad \propto \quad p_j^{(r)} f(d_j^{(r)}), \tag{4.15}$$

where $d_j^{(r)} = \partial\phi/\partial p_j|_{p=p^{(r)}}$ and the function $f(.)$ satisfies certain conditions and may depend on a free positive parameter δ. In view of the conditions for optimality, a solution to our maximization problem is a fixed point of the iteration but can also be the solutions to the problem when any subset of weights is set to zero.

Torsney (1977) first proposed this type of iteration which requires derivatives to be positive by taking $f(d) = d^\delta$ with $\delta > 0$. Subsequent empirical studies include Silvey et al. (1978) which is a study of the choice of δ

when $f(d) = d^{\delta}$, $\delta > 0$. Torsney (1988) considered $f(d) = e^{\delta d}$ in a variety of applications. Mandal and Torsney (2006) explored systematic choices of $f(.)$. Torsney and Mandal (2001) and Mandal et al. (2005) considered constrained optimal design problems. Mandal et al. (2017) considered the logistic cumulative density function as a choice of $f(.)$ to optimize a non-standard criterion. Titterington (1976) did a monotonicity proof for D-optimality. Torsney (1983) studied monotonicity of some particular values of δ for some criterion $\phi(p)$. Torsney (1983) established a sufficient condition for monotonicity when the criterion is homogeneous of certain degree and also proved this condition to hold for design criteria such as c- and A-optimal criteria. In other cases the value $\delta = 1$ can be shown to yield an expectation–maximization (EM) algorithm. This algorithm is known to be monotonic and convergent (see Dempster et al. (1977)). Fedorov (1972) and Wynn (1972) considered vertex direction algorithms. These are useful when many weights are zero at the optimum. At the other case, constrained steepest ascent or Newton type iterations are appropriate (see Wu (1978) and Atwood (1980)).

We maximize a criterion function $\phi(p)$ subject to the basic constraints on the design weights: $p_j \geq 0$, $j = 1, 2, \ldots, J$ and $\sum_{j=1}^{J} p_j = 1$. The iteration (4.19) neatly submits to these basic constraints. Convergence of this algorithm could be slow if we do not choose the function $f(.)$ and its arguments in an objective way. Thus the goal is to find appropriate choice(s) of these, that is, to develop strategies for better convergence of the algorithm for constructing designs that optimize standard regression design criteria. Following this the general form of the algorithm is given below. The iteration at the $(r+1)^{\text{th}}$ step is

$$p_j^{(r+1)} \quad = \quad \frac{p_j^{(r)} f(x_j^{(r)},\, \delta)}{\sum\limits_{j=1}^{J} p_j^{(r)} f(x_j^{(r)},\, \delta)}\,, \tag{4.16}$$

where x can be taken as the partial derivatives and $f(x,\,\delta)$ is a positive and strictly increasing function in x and δ is a free positive parameter.

4.5 Properties of the Algorithm

The choice of $f(.)$ plays an important role in the convergence of the algorithm. We will study some important properties of the directional derivatives to make a proper choice of $f(.)$.

1. Under the conditions imposed on $f(.)$, algorithm (4.16) guarantees $F_\phi\{p^{(r)},\, p^{(r+1)}\} \geq 0$ where $F_\phi\{p^{(r)},\, p^{(r+1)}\}$ is the directional derivative of $\phi(.)$ at the current iteration $p^{(r)}$ in the direction of the next iteration $p^{(r+1)}$.

2. Let $supp(p) = \{\boldsymbol{v}_j \in \mathcal{V} :\ p_j > 0\}$ denote the support of the distribution p. Under the above iteration, we have $supp(p^{(r+1)}) \subseteq supp(p^{(r)})$, but

some weights can converge to zero.

3. If $p^{(r)} = p^*$, then $p^{(r+1)} = p^*$ so that $F_\phi\{p^{(r)},\ p^{(r+1)}\} = 0$, with p^* being a fixed point of the iteration.

4. The partial derivatives d_j corresponding to nonzero p_j^* must share a common value.

In next section, we will construct our optimal designs, generate data, obtain pretest, shrinkage, and positive shrinkage estimators and compare their performances.

4.6 Simulation Study

To see the finite-sample performance of pretest, shrinkage, and positive shrinkage estimators with and without using optimal design technique in the pre-modeling stage, we conduct Monte Carlo simulation studies under the various scenarios. The data are generated from the following multiple linear regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{4.17}$$

where the covariates $\boldsymbol{x}_i$ are generated from a uniform distribution $\mathbf{U}(-1, 1)$ without applying the optimal design technique. The D-optimality criterion are applied when we used optimal design for generating covariates from the interval $[-1, 1]$. In this case we get the optimal values of the covariates from the interval $[-1, -1]$ using optimal design instead of taking random covariate values from $\mathbf{U}(-1, 1)$. The errors ε_i are generated from standard normal distribution. For each scenario, we calculate the relative mean squared error for sample size $n = 128$, 192, 256, and 320 and repeat each simulation 1000 times.

In this simulation, we have partitioned the coefficient $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are $p_1 \times 1$ and $p_2 \times 1$ vectors, respectively. We consider three sets of true $\boldsymbol{\theta}_1$ values ($p_1 = 1$, 2, and 3), that is, $\boldsymbol{\theta}_1 = 1.5$, $\boldsymbol{\theta}_1 = (1.5, -2.1)$, and $\boldsymbol{\theta}_1 = (1.5, -2.1, -1.25)$ and also the number of corresponding inactive covariates p_2 in the model are 5, 3 and 2, respectively.

To explore the behavior of the pretest and shrinkage estimators, we define $\Delta = \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\|$, where $\|.\|$ is the Euclidean norm, $\boldsymbol{\theta}^0 = (\boldsymbol{\theta}_1^\top, \mathbf{0}^\top)^\top$, and $\mathbf{0}$ is a zero vector with different dimensions p_2. Note that Δ is the difference between the unrestricted and restricted models and the performance of the pretest and shrinkage estimators is assessed when $H_0 : \Delta = 0$ and $H_a : \Delta = d_\Delta$ for $0 \leq d_\Delta \leq 2$. For each simulation set up, 1000 simulation runs are conducted using $\Delta = (0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0)$. For each simulated data set, the relative mean squared errors of restricted (RE), pretest (PT), shrinkage (SE), and positive shrinkage (PSE) with respect to unrestricted (UE) are calculated when the covariates $\boldsymbol{x}_i$ are generated from

$\mathbf{U}(-1,1)$ and the interval $[-1,1]$ using optimal design technique. Quadratic bias and mean squared error (MSE) of the estimators are also calculated for each simulated data set with and without optimal design settings.

4.6.1 Covariates generated using D-optimal criterion

We consider $m = 6$ and rewrite the model (4.17) as follows.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \varepsilon = \boldsymbol{v}_x^\top \boldsymbol{\theta} + \varepsilon, \quad (4.18)$$

where $\boldsymbol{v}_x = (1, x_1, x_2, x_3, x_4, x_5, x_6)^\top$ and $\boldsymbol{\theta}_x = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$.

In order to find the optimal design, we first need to discretize the design space. An ideal discretization would be some form of uniform grid on a continuous design space. In practice the discretization that is used is the image under the regression function of a uniform grid on the design space. Let this induced design space be $\mathcal{V}$ with J points or vertices. In our model, we have a standardized continuous design space. Therefore, the design space consists of $J = (5)^6 = 15625$ combinations of $(x_1, x_2, x_3, x_4, x_5, x_6)$. Let pr be the arbitrary weight of each combination, the design can be written as

$$\boldsymbol{\zeta} = \left\{ \begin{array}{cccc} (x_{11}, x_{21}, \cdots, x_{61}) & (x_{11}, x_{21}, \cdots, x_{62}) & \cdots & (x_{15}, x_{25}, \cdots, x_{65}) \\ \mathrm{pr}_1 & \mathrm{pr}_2 & \cdots & \mathrm{pr}_{15625} \end{array} \right\}$$

where the first row gives the values of combinations of $(x_1, x_2, x_3, x_4, x_5, x_6)$ respectively while the second row gives the corresponding proportions or weights. In our design, the variables $\mathrm{pr}_1, \mathrm{pr}_2, \ldots, \mathrm{pr}_J$ must be nonnegative and sum to 1. An iteration which neatly submits to these and enjoys some properties is the following algorithm:

$$\mathrm{pr}_j^{(r+1)} = \mathrm{pr}_j^{(r)} f(x_j^{(r)}, \delta) \Bigg/ \sum_{i=1}^{J} \mathrm{pr}_i^{(r)} f(x_i^{(r)}, \delta) \tag{4.19}$$

where $x_j^{(r)} = d_j^{(r)}$, $d_j^{(r)} = \frac{\partial \phi}{\partial \mathrm{pr}_j}\Big|_{\mathrm{pr}=\mathrm{pr}^{(r)}}$ (partial derivatives at r^{th} iteration, i.e., at $\mathrm{pr} = \mathrm{pr}^{(r)}$). The function $f(x, \delta)$ must be positive and strictly increasing in the first argument x. Its second argument δ is a free positive parameter. We will see that it is actually a tuning parameter for the convergence of the algorithm.

We apply the above optimization algorithm by assigning initial weights $\mathrm{pr}_j^0 =$ 1/J for $j = 1, 2, \cdots$, J. Let $f(x) = x^{\delta}$, where $x = d$ and d is the partial derivatives of D-optimality criterion. We use $\delta = 1.9$ in the above algorithm. According to the design criterion, the partial and directional derivatives are calculated and then we can run the algorithms to obtain the optimal design points $\boldsymbol{\zeta}$ and weights until it converges.

Here F_j's are the vertex directional derivatives of the criterion function. In our case, the criterion function is the D-optimality criterion. In order to calculate the directional derivatives, we first need to calculate the partial derivatives of the criterion function.

The final optimal design has 64 combinations of $(x_1, x_2, x_3, x_4, x_5, x_6)$ where values for each x_i is -1 or 1 and the corresponding optimal weight for each combination is 0.015625. The results are summarized in the following table.

Table 4.1: Optimal weights with the corresponding values of $x_1, x_2, x_3, x_4, x_5, x_6$

x_1	x_2	x_3	x_4	x_5	x_6	pr	x_1	x_2	x_3	x_4	x_5	x_6	pr
-1	-1	-1	-1	-1	-1	0.015625	1	-1	-1	-1	-1	-1	0.015625
-1	1	-1	-1	-1	-1	0.015625	1	1	-1	-1	-1	-1	0.015625
-1	-1	1	-1	-1	-1	0.015625	1	-1	1	-1	-1	-1	0.015625
-1	1	1	-1	-1	-1	0.015625	1	1	1	-1	-1	-1	0.015625
-1	-1	-1	1	-1	-1	0.015625	1	-1	-1	1	-1	-1	0.015625
-1	1	-1	1	-1	-1	0.015625	1	1	-1	1	-1	-1	0.015625
-1	-1	1	1	-1	-1	0.015625	1	-1	1	1	-1	-1	0.015625
-1	1	1	1	-1	-1	0.015625	1	1	1	1	-1	-1	0.015625
-1	-1	-1	-1	1	-1	0.015625	1	-1	-1	-1	1	-1	0.015625
-1	1	-1	-1	1	-1	0.015625	1	1	-1	-1	1	-1	0.015625
-1	-1	1	-1	1	-1	0.015625	1	-1	1	-1	1	-1	0.015625
-1	1	1	-1	1	-1	0.015625	1	1	1	-1	1	-1	0.015625
-1	-1	-1	1	1	-1	0.015625	1	-1	-1	1	1	-1	0.015625
-1	1	-1	1	1	-1	0.015625	1	1	-1	1	1	-1	0.015625
-1	-1	1	1	1	-1	0.015625	1	-1	1	1	1	-1	0.015625

Continued on next page

Table 4.1 – *Continued from previous page*

x_1	x_2	x_3	x_4	x_5	x_6	pr	x_1	x_2	x_3	x_4	x_5	x_6	pr
-1	1	1	1	1	-1	0.015625	1	1	1	1	1	-1	0.015625
-1	-1	-1	-1	-1	1	0.015625	1	-1	-1	-1	-1	1	0.015625
-1	1	-1	-1	-1	1	0.015625	1	1	-1	-1	-1	1	0.015625
-1	-1	1	-1	-1	1	0.015625	1	-1	1	-1	-1	1	0.015625
-1	1	1	-1	-1	1	0.015625	1	1	1	-1	-1	1	0.015625
-1	-1	-1	1	-1	1	0.015625	1	-1	-1	1	-1	1	0.015625
-1	1	-1	1	-1	1	0.015625	1	1	-1	1	-1	1	0.015625
-1	-1	1	1	-1	1	0.015625	1	-1	1	1	-1	1	0.015625
-1	1	1	1	-1	1	0.015625	1	1	1	1	-1	1	0.015625
-1	-1	-1	-1	1	1	0.015625	1	-1	-1	-1	1	1	0.015625
-1	1	-1	-1	1	1	0.015625	1	1	-1	-1	1	1	0.015625
-1	-1	1	-1	1	1	0.015625	1	-1	1	-1	1	1	0.015625
-1	1	1	-1	1	1	0.015625	1	1	1	-1	1	1	0.015625
-1	-1	-1	1	1	1	0.015625	1	-1	-1	1	1	1	0.015625
-1	1	-1	1	1	1	0.015625	1	1	-1	1	1	1	0.015625
-1	-1	1	1	1	1	0.015625	1	-1	1	1	1	1	0.015625
-1	1	1	1	1	1	0.015625	1	1	1	1	1	1	0.015625

We compare the proposed estimators using quadratic bias (QB) and mean squared error (MSE). Figures 4.1 - 4.8 show the QB curves and Figures 4.15 - 4.22 show the MSE curves of the estimators for different values of Δ and different combination of inactive covariates p_2. These figures also provide the comparison of quadratic biases and MSE among different sample sizes with and without optimal design settings.

The key findings of QB and MSE for proposed estimators in both settings are summarized below when $\Delta \geq 0$. First we start with without-optimal-design technique:

1. All figures show that the quadratic bias and MSE of restricted estimator (RE) is lower than the pretest and shrinkage estimators (SE and PSE) at and near $\Delta = 0$ and this is obvious because of its unbiasedness.

2. The RE is consistently outperforming all other estimators near the null hypothesis. However, as Δ increases, the quadratic bias and MSE of the RE increases and becomes unbounded. Therefore, if RE is nearly correctly specified, it is the optimal estimator over entire parameter space. Because of the nature of RE, this estimator heavily depends on the validity of H_0.

3. Since Δ is the deviation from the null hypothesis value, it is clear that one cannot go wrong with the use of shrinkage estimators despite $\Delta > 0$, that is, in worst case the risk of pretest and shrinkage estimators

will be same as UE. If $\Delta = 0$ then pretest and shrinkage estimators have lower bias and MSE compared to the UE.

4. The pretest, shrinkage and positive shrinkage estimators have uniformly lower risk than the UE across the entire parameter space.

In summary, RE performs better than PT, SE and PSE but these three estimators are not as sensitive to an increase in Δ as compared to the RE. Thus the shrinkage strategy is preferable when the number of inactive predictors is three and greater.

Next we start comparing QB and MSE when applying optimal design technique:

1. The QB and MSE curves in Figures 4.1 - 4.8 and Figures 4.15 - 4.22 show that the QB and MSE curves when applying optimal design are always under the curves when without applying optimal design. It gives evidence that optimal design points and weights are helping to reduce the bias and MSE of pretest, shrinkage, positive shrinkage estimators.
2. It supports another great strong point that even with smaller sample size (n=128 compared to n=320), the proposed estimators with optimal design setup outperform the unrestricted estimator.

In summary it shows that the larger the sample size is, the lower the bias and MSE are. However, cost will increase when sample size increases. The differences of QB and MSE curves of the estimators with and without using optimal design are significant.

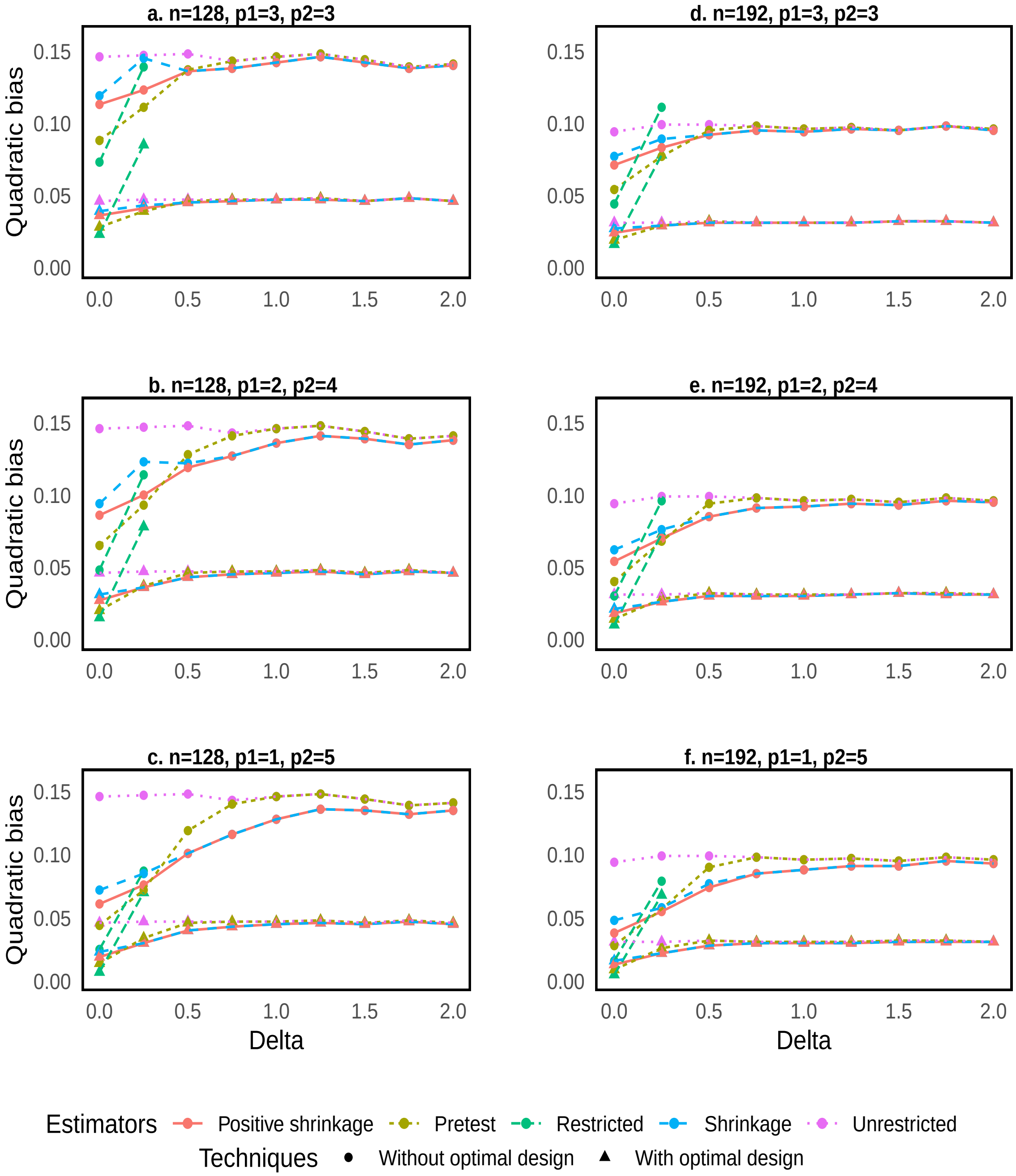

Figure 4.1: Simulated QB curves of proposed estimators with and without optimal design for $n = 128, 192$ and $p_2 = 3, 4, 5$

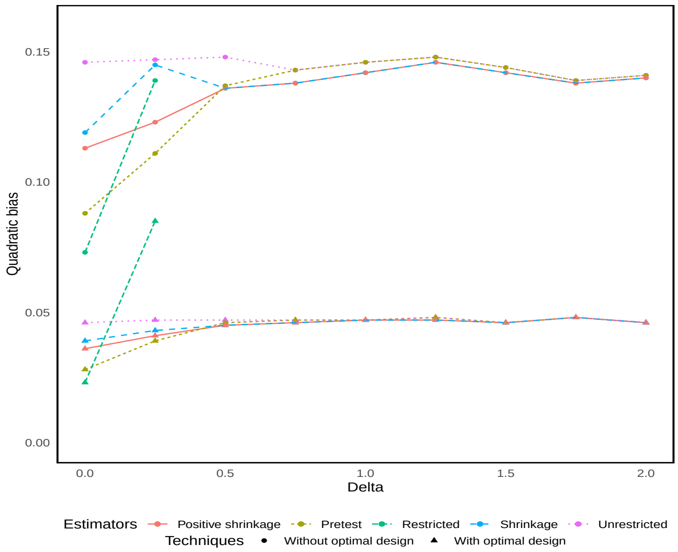

Figure 4.2: Simulated QB curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 3, p_2 = 3$.

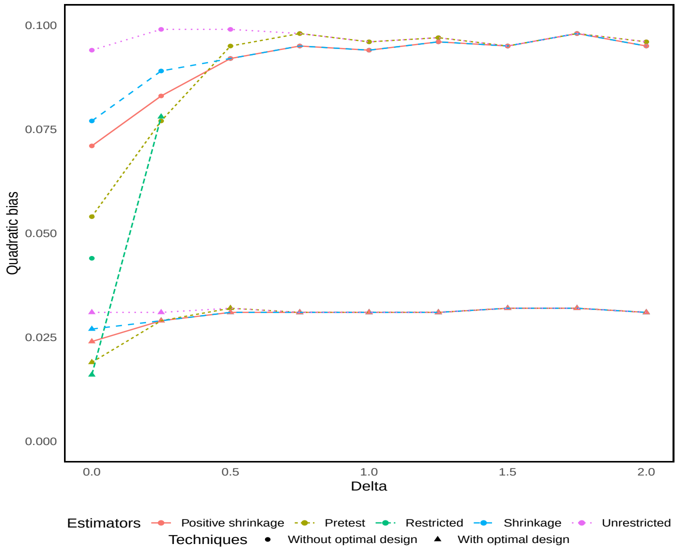

Figure 4.3: Simulated QB curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 3, p_2 = 3$.

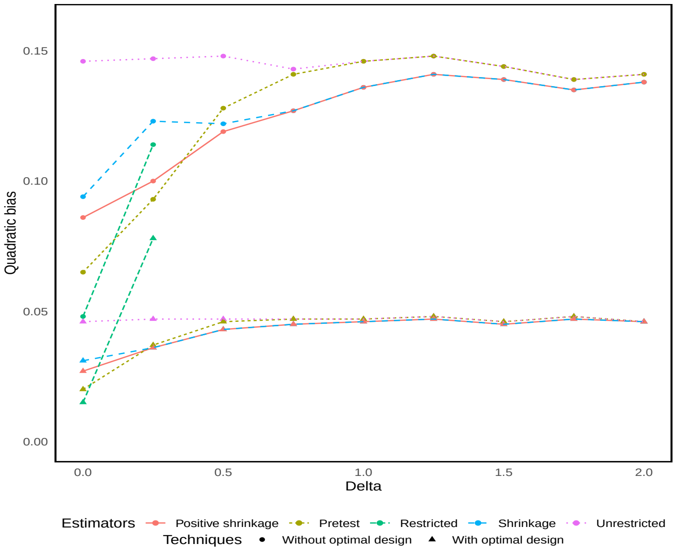

Figure 4.4: Simulated QB curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 2, p_2 = 4$.

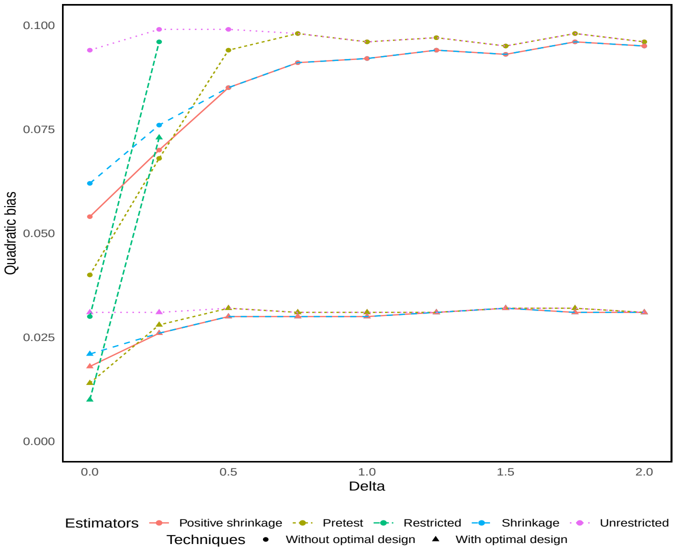

Figure 4.5: Simulated QB curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 2, p_2 = 4$.

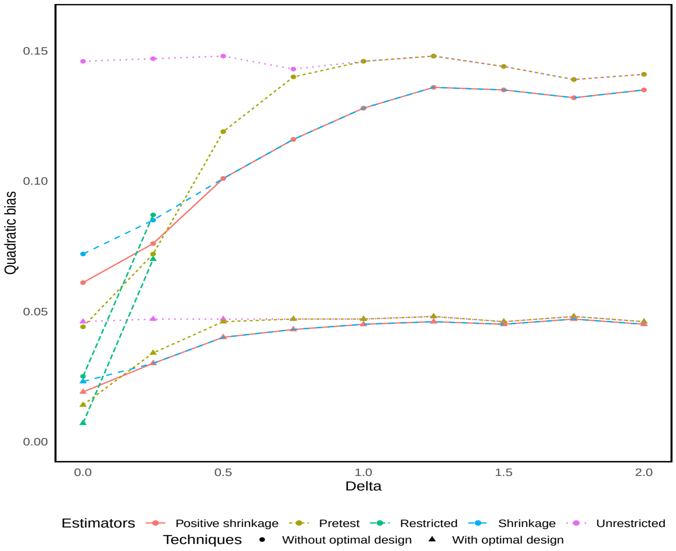

Figure 4.6: Simulated QB curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 1, p_2 = 5$.

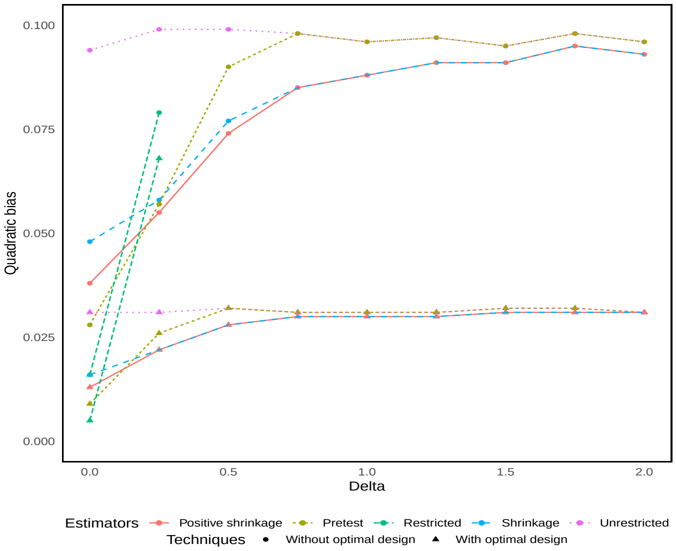

Figure 4.7: Simulated QB curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 1, p_2 = 5$.

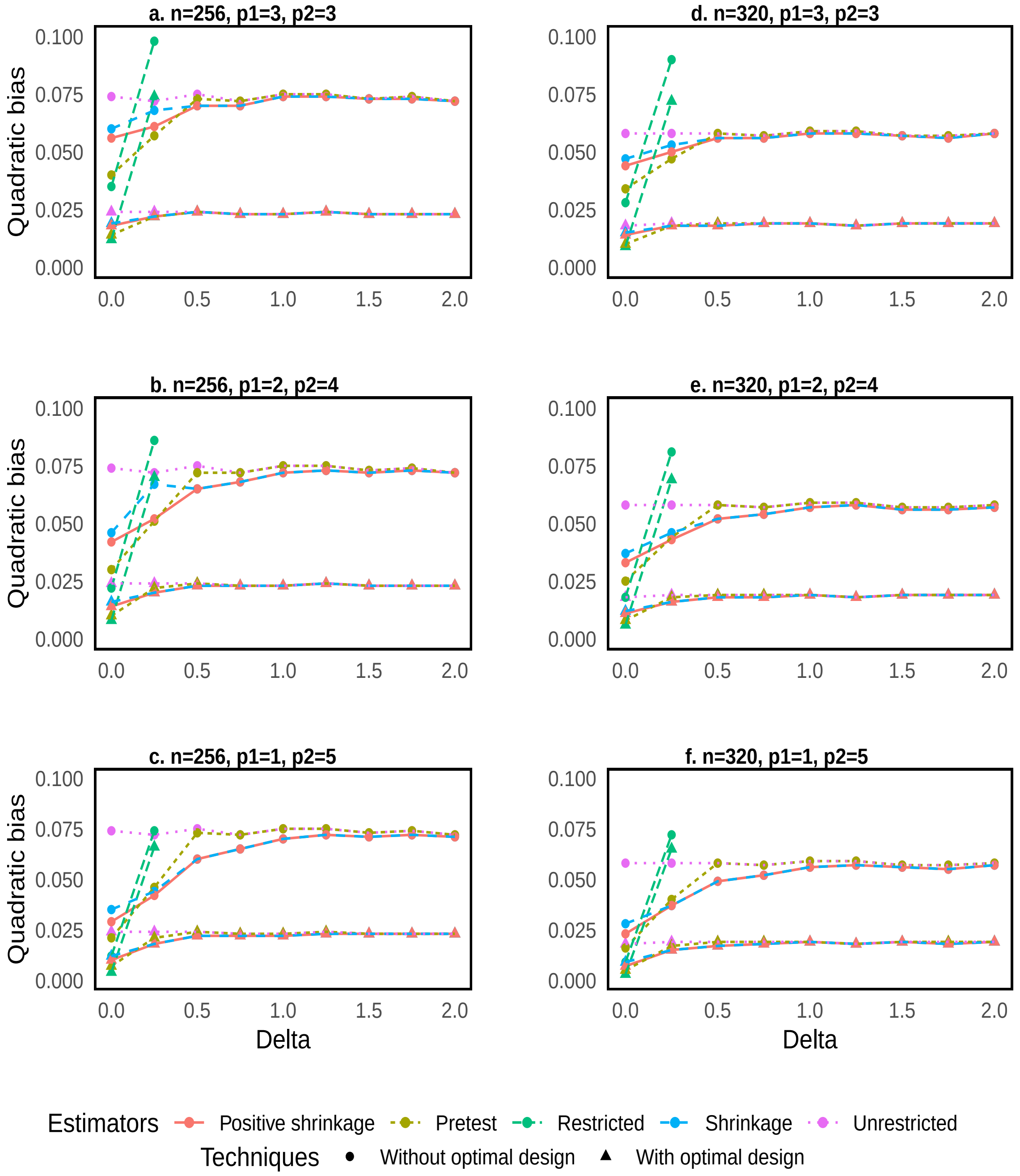

Figure 4.8: Simulated QB curves of proposed estimators with and without optimal design for $n = 256, 320$ and $p_2 = 3, 4, 5$

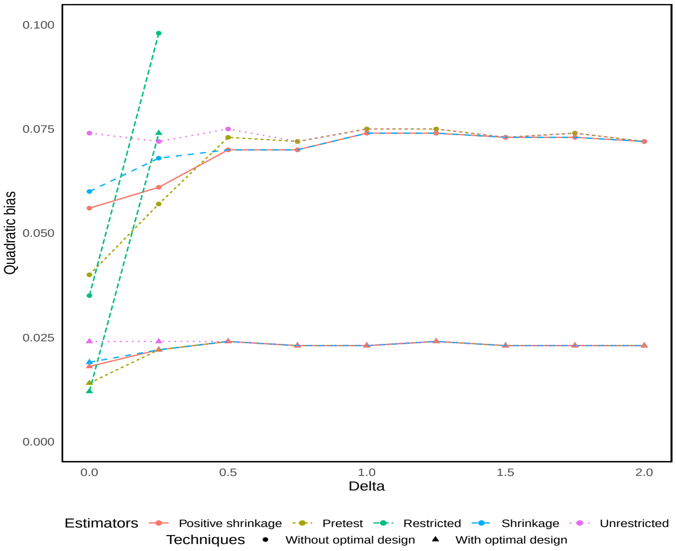

Figure 4.9: Simulated QB curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 3, p_2 = 3$.

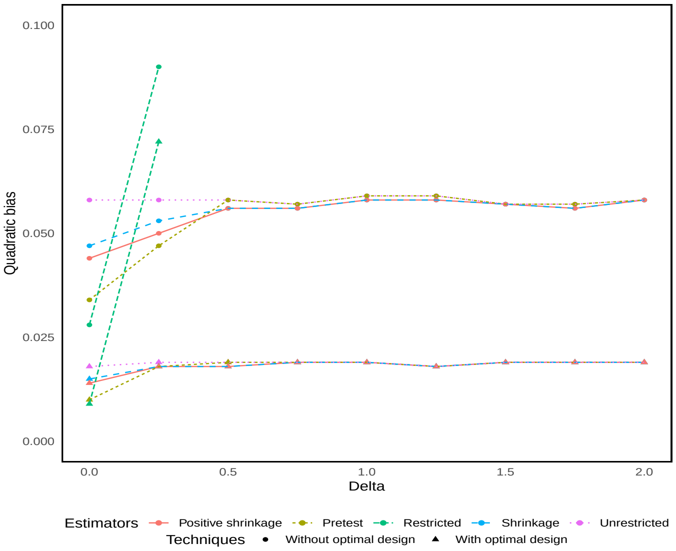

Figure 4.10: Simulated QB curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 3, p_2 = 3$.

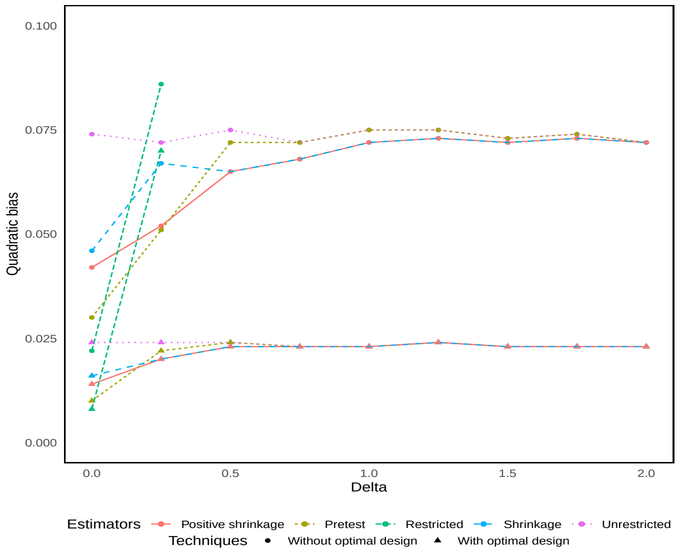

Figure 4.11: Simulated QB curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 2, p_2 = 4$.

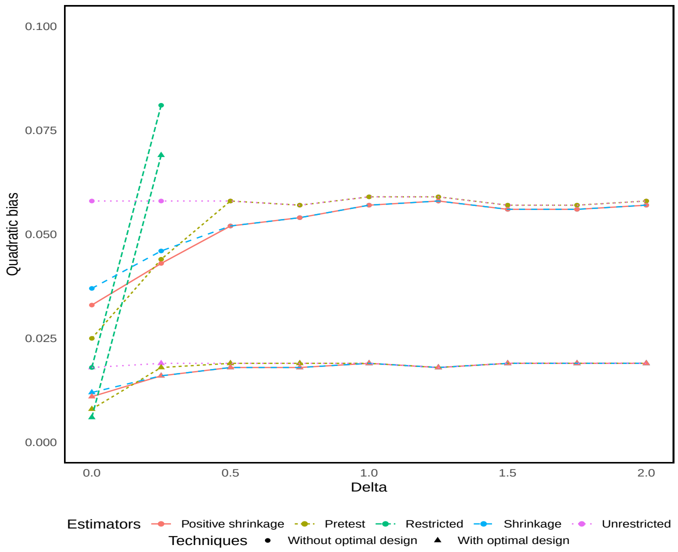

Figure 4.12: Simulated QB curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 2, p_2 = 4$.

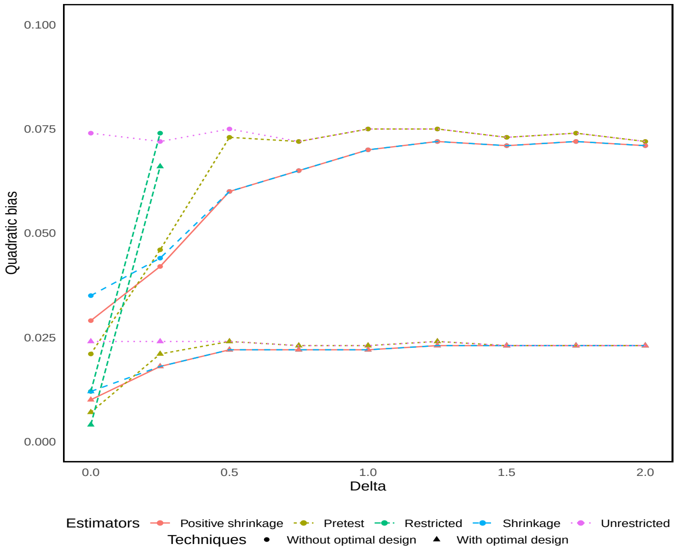

Figure 4.13: Simulated QB curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 1, p_2 = 5$.

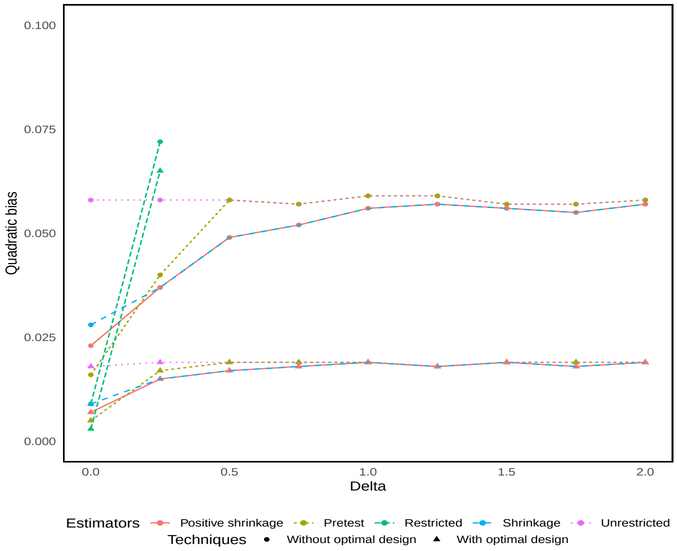

Figure 4.14: Simulated QB curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 1, p_2 = 5$.

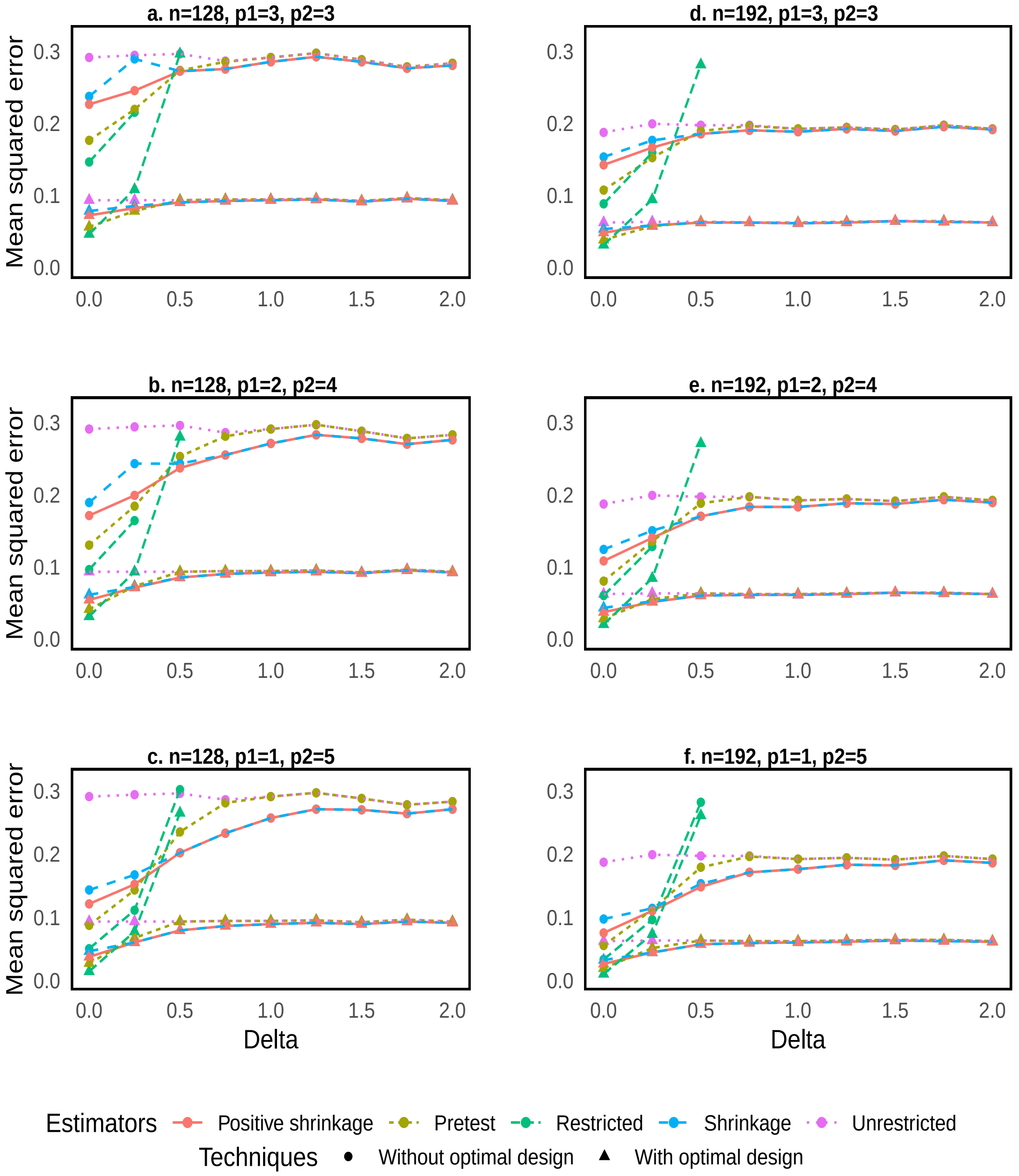

Figure 4.15: Simulated MSE curves of proposed estimators with and without optimal design for for $n = 128, 192$ and $p_2 = 3, 4, 5$

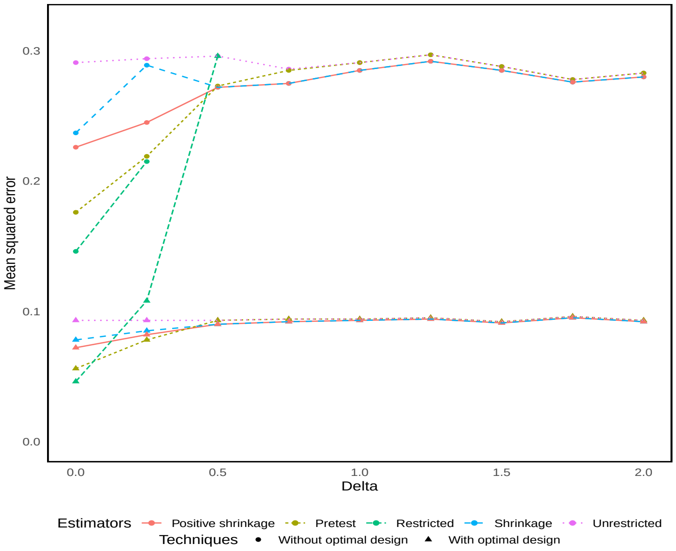

Figure 4.16: Simulated MSE curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 3, p_2 = 3$.

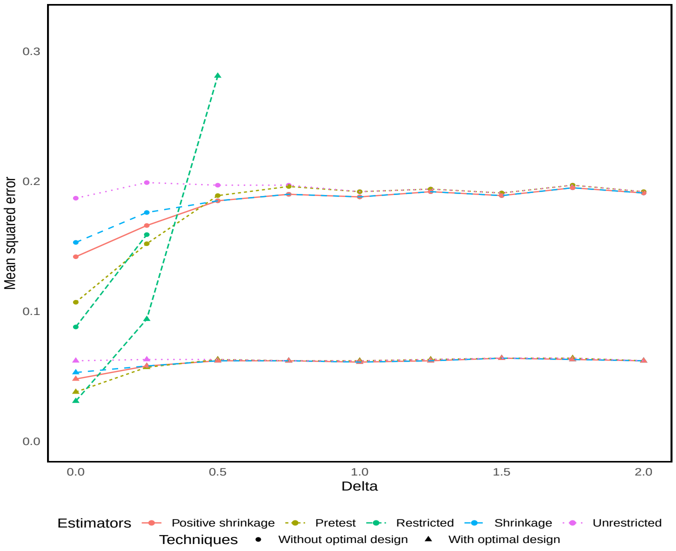

Figure 4.17: Simulated MSE curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 3, p_2 = 3$.

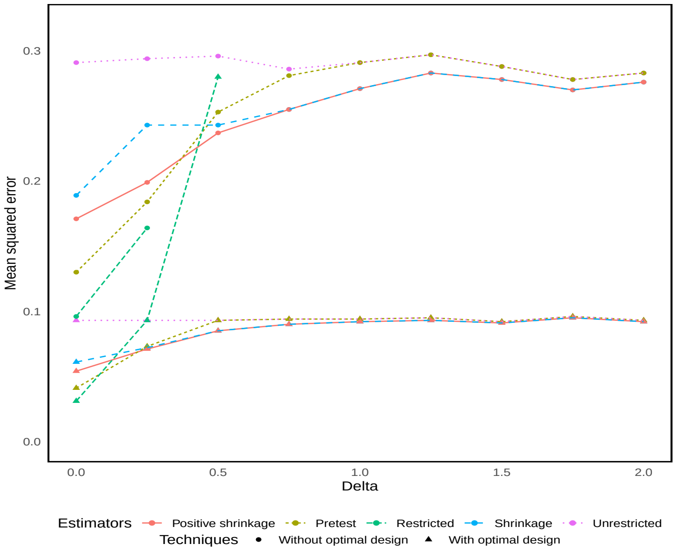

Figure 4.18: Simulated MSE curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 2, p_2 = 4$.

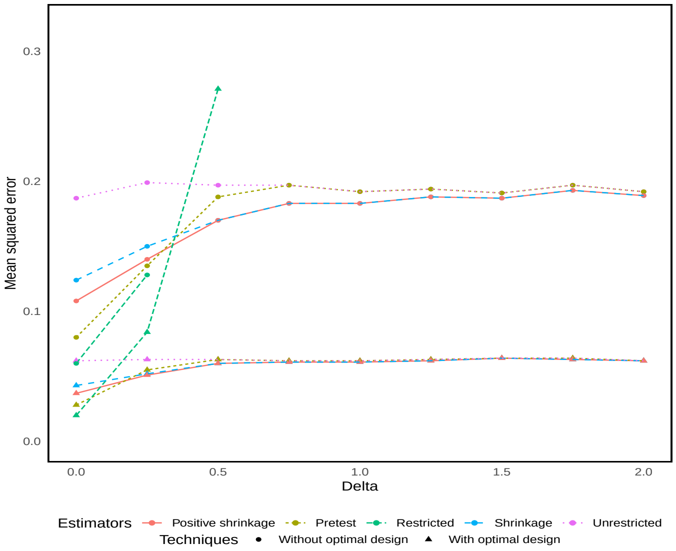

Figure 4.19: Simulated MSE curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 2, p_2 = 4$.

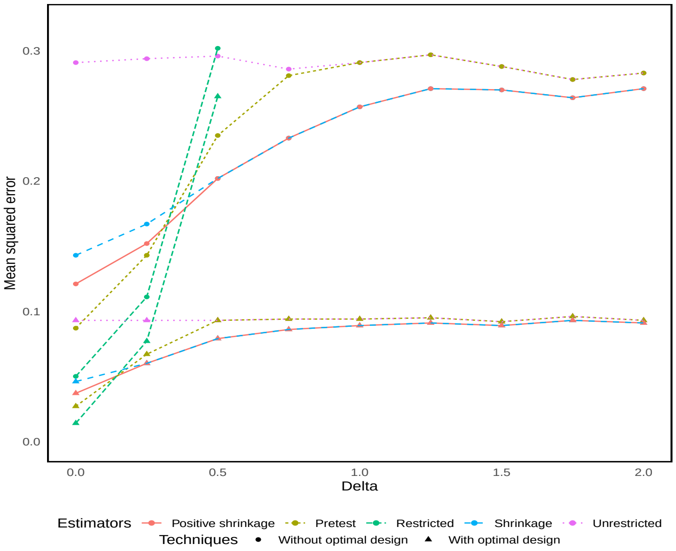

Figure 4.20: Simulated MSE curves of proposed estimators with and without optimal design for $n = 128$ and $p = 6, p_1 = 1, p_2 = 5$.

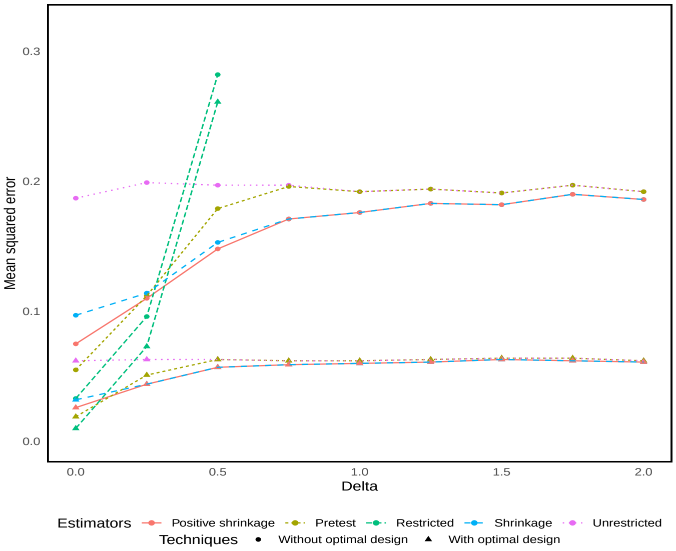

Figure 4.21: Simulated MSE curves of proposed estimators with and without optimal design for $n = 192$ and $p = 6, p_1 = 1, p_2 = 5$.

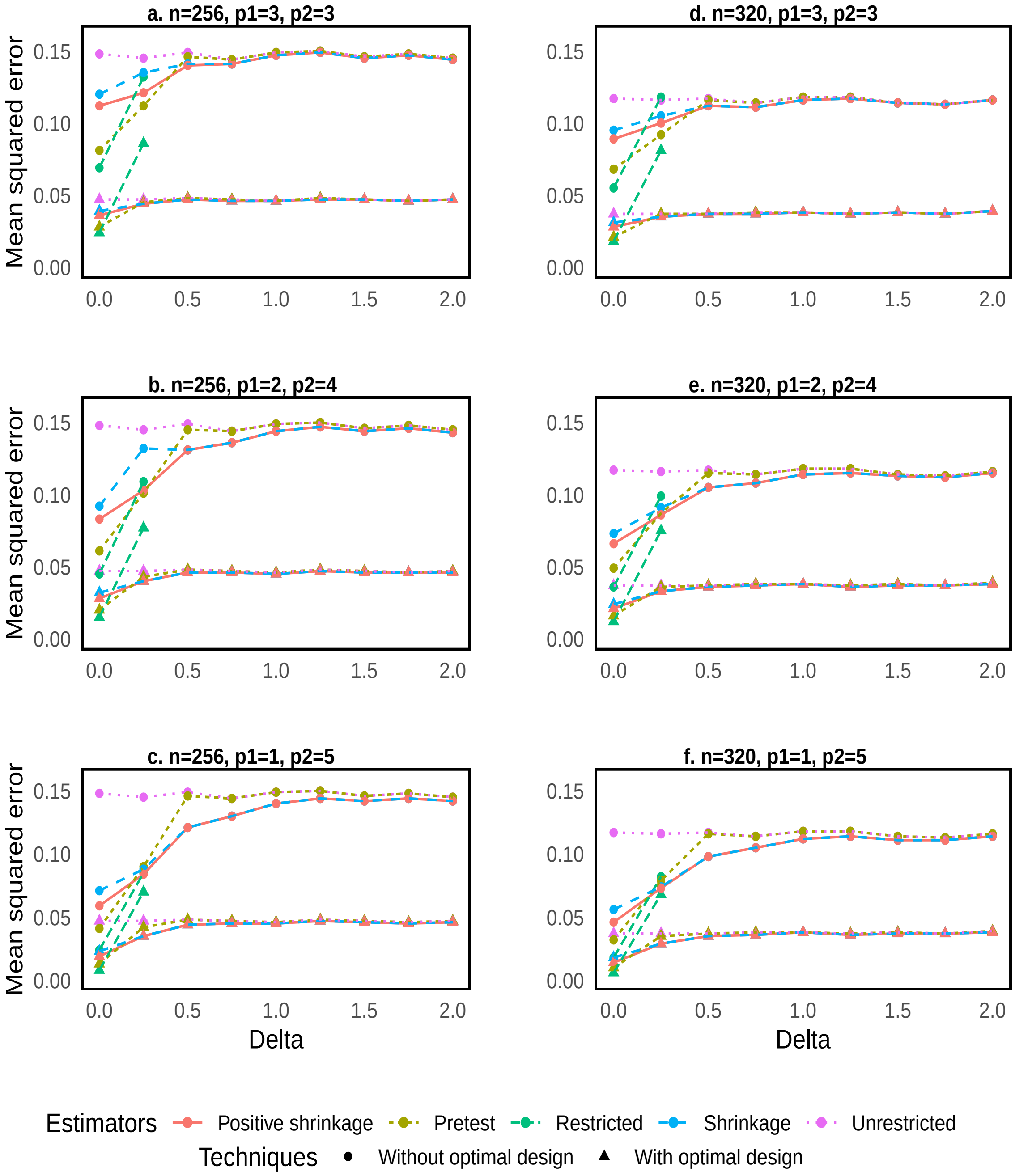

Figure 4.22: Simulated MSE curves of proposed estimators with and without optimal design for $n = 256, 320$ and $p_2 = 3, 4, 5$

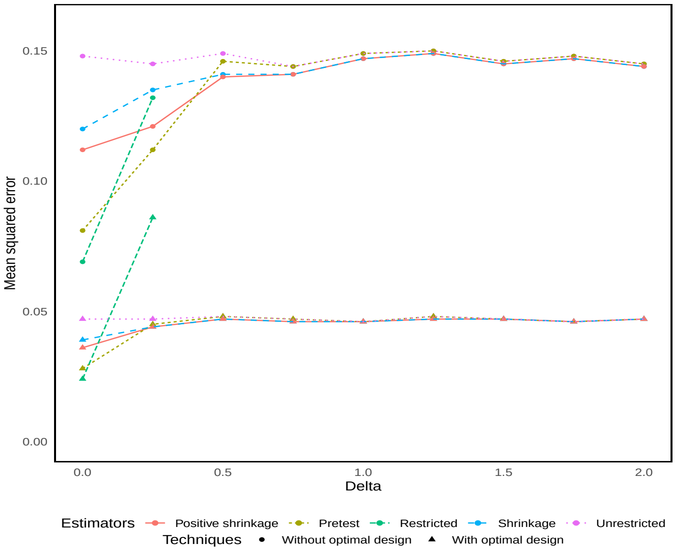

Figure 4.23: Simulated MSE curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 3, p_2 = 3$.

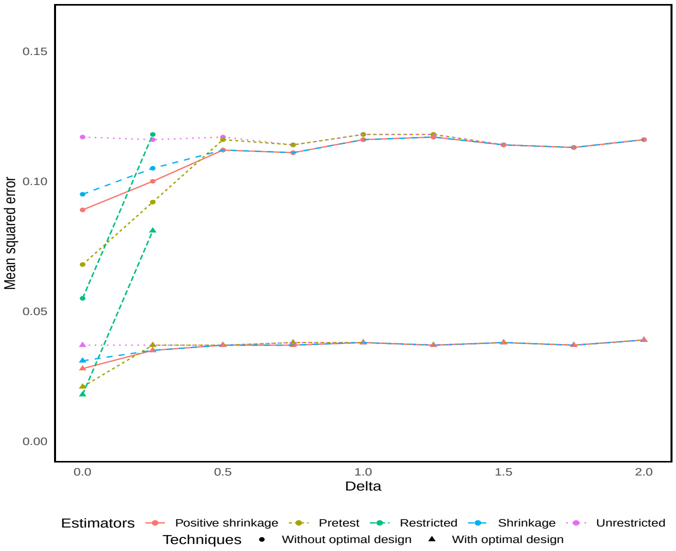

Figure 4.24: Simulated MSE curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 3, p_2 = 3$.

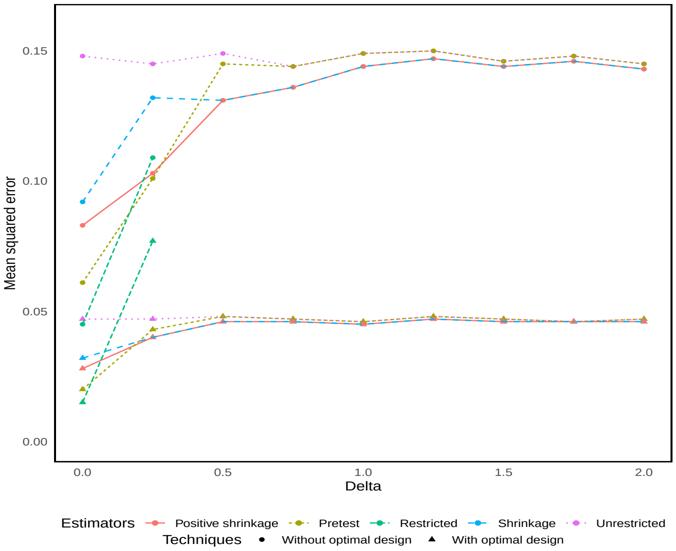

Figure 4.25: Simulated MSE curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 2, p_2 = 4$.

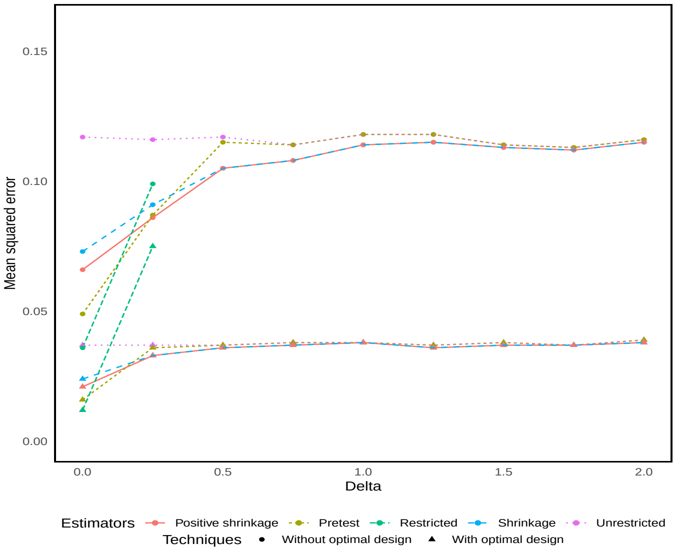

Figure 4.26: Simulated MSE curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 2, p_2 = 4$.

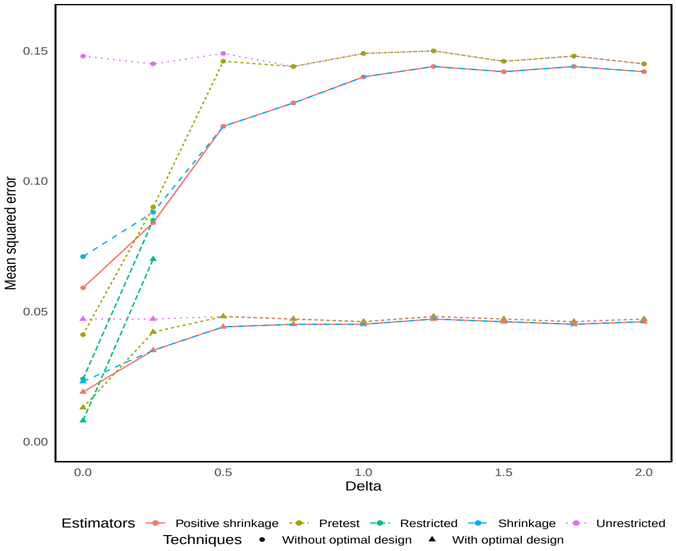

Figure 4.27: Simulated MSE curves of proposed estimators with and without optimal design for $n = 256$ and $p = 6, p_1 = 1, p_2 = 5$.

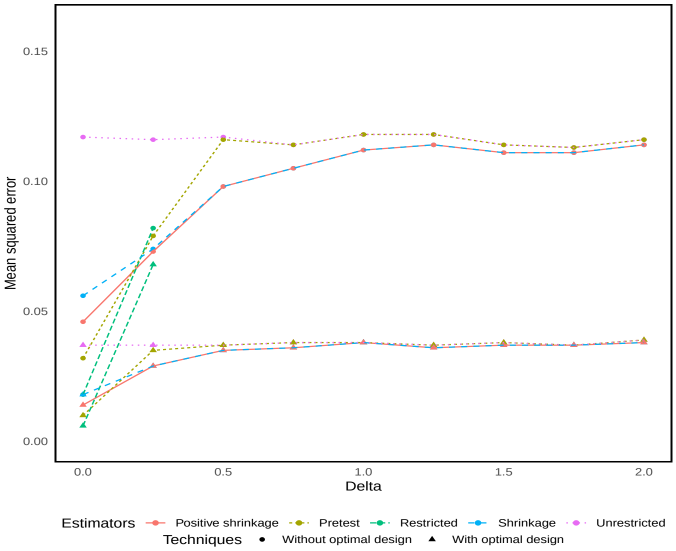

Figure 4.28: Simulated MSE curves of proposed estimators with and without optimal design for $n = 320$ and $p = 6, p_1 = 1, p_2 = 5$.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis we have tried to solve two new estimation problems. We are not aware of any literature in which these two problems have been addressed.

In particular, we have addressed the problems in developing pretest and shrinkage estimators for generalized partial linear models (GPLM) and extending shrinkage estimators using optimal design theory in multiple linear regression. In each problem, a detailed Monte Carlo simulation study is conducted to examine the performance of the proposed pretest and shrinkage estimators. Application of the pretest and shrinkage estimators have been demonstrated in generalized partial linear regression models with a real credit scoring data.

In Chapters 2 and 3, we have applied the maximum likelihood method with generalized Speckman algorithm to estimate the regression parameters of GPLM and named this as unrestricted estimate, UG. We also estimate the parameters when some of them are restricted to a subspace and named this as restricted estimate, RG. We also developed pretest and shrinkage estimators. We study the relative risk dominance of the pretest and shrinkage estimators which are defined based on the UG and RG by deriving mathematically their asymptotic risks and biases. We derive the expressions of biases and risks. We used a Monte Carlo simulation study to calculate the numerical biases, mean squared errors and risks (inverse of relative mean squared error) of the estimators. Our simulation studies show that the restricted estimators offer numerically superior performance compared to the unrestricted, pretest and shrinkage estimators near the null hypothesis $\boldsymbol{R\beta} = \boldsymbol{r}$, but this performs poorly when the restriction is seriously violated. The risk of the pretest estimator is lower than that of the UG (or higher relative MSE with respect to the UG) at and near the restriction in the simulation study. Our simulation study also shows that the shrinkage estimators have smaller mean squared error than the shrinkage estimators in terms of MSE for large region of parameter space even when there exist omitted significant predictor in the specified model. Under alternative hypothesis, it shows that the relative

MSEs of PT, SE and PSE converges to one.

In Chapter 4, we extend the proposed pretest and shrinkage estimators incorporating with optimal design theory, specifically the most popular D-optimality criterion. The combinations of numerical inputs or level of values are obtained using the optimization algorithm under D-optimality criterion. Optimal weights with the corresponding sizes of each combination of covariates are obtained. For this purpose we used a class of multiplicative algorithms that are indexed by a function. This function has to be positive and increasing. The function may depend on a free positive parameter. This algorithm satisfies the basic constraints of our optimization problem and possesses many nice properties. We also determined the optimality conditions using directional derivatives. We estimated the regression parameters of multiple linear model using maximum likelihood estimation and named it unrestricted estimator. The restricted maximum likelohood estimator of mulitple linear regression then was obtained under constraints as that of applying for GPLM in Chapter 2. Then, we combined unrestricted and restricted estimators optimally to obtain pretest and shrinkage estimators. A Monte Carlo simulation was conducted to investigate the performance of proposed estimators while applying pre-modeling optimal design theory. Simulation studies show that estimators with applying optimal design theory in advance will far outperform than the regular proposed pretest and shrinkage estimators.

5.2 Future Work

The focus of this thesis was to develop pretest and shrinkage estimators to generalized partial linder models for independent data and hybrid of shrinkage and optimal design has only applied to multiple linear regression model. A possible future work would be to extend our proposed estimators to genealized partially linear models for longitudinal data. Furthermore, the joint work of shrinkage estimators and optimal design can be expanded to more complicated models.

It is of great interest to study the proposed method for the longitudinal data and when the number of predictor grows with the sample size (e.g. high-dimensional data) in which pre-planning optimal design theory may help in minimizing cost and time. Another possible work could be to construct an optimal design subject to a given cost (budget), then accordingly we can develop the shrinkage estimators.

Bibliography

Ahmad, I., Leelahanon, S., and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *The Annals of Statistics*, 33:258–283.

Ahmed, S. E. (2014). *Penalty, shrinkage and pretest strategies: variable selection and estimation*. Springer International Publishing, New York, USA.

Ahmed, S. E., Doksum, K., Hossain, S., and You, J. (2007). Shrinkage, pretest and LASSO estimators in partially linear models. *Australian and New Zealand Journal of Statistics*, 49(4):461–471.

Ahmed, S. E. and Fallahpour, S. (2012). Shrinkage estimation strategy in quasi-likelihood models. *Statistics and Probability Letters*, 82:2170–2179.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Clarendon Press, Oxford.

Atwood, C. L. (1980). Convergent design sequences for sufficiently regular optimality criteria, II: singular case. *Annals of Statistics*, 8:894–912.

Bancroft, T. (1944). On biases in estimation due to the use of preliminary test of significance. *The Annals of Mathematical Statistics*, 15:190–204.

Battauz, M. and Bellio, R. (2021). Shrinkage estimation of the three-parameter logistic model. *British Journal of Mathematical & Statistical Psychology*. https://doi.org/10.1111/bmsp.12241.

Berger, M. and Wong, W. (2009). *An introduction to optimal designs for social and biomedical research*. John Wiley and Sons, Chichester.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore.

Boente, G., Cao, R., Manteiga, W. G., and Rodriguez, D. (2016). Testing in generalized partially linear models: A robust approach author links open overlay panel. *Statistics & Probability Letters*, 28(3):531–549.

Boente, G., He, X., Zhou, J., et al. (2006). Robust estimates in generalized partially linear models. *Annals of Statistics*, 34(6):2856–2878.

Boente, G. and Rodriguez, D. (2010). Robust inference in generalized partially linear models. *Computational Statistics and Data Analysis*, 54(12):2942–2966.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals Statistics*, 17:453–555.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489.

Chen, H. and Shiau, J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference*, 27:187–201.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Dunn, p. K. and Smyth, G. K. (2018). *Generalized linear models With examples in R.* Springer, New York, NY.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales. *Journal of American Statistical Association*, 81:310–320.

Fedorov, V. (1972). *Theory of optimal experiments.* Academic Press, New York.

Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage Estimation (1st ed.).* Switzerland.

Härdle, W., Liang, H., and Gao, J. (2012a). *Partially linear models*. Springer Science and Business Media, Berlin.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2012b). *Nonparametric and semiparametric Models*. Springer Science and Business Media, New York.

Hastie, T. J., Tibshirani, R., and Wainwright, M. J. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, Boca Raton.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall, New York.

Hossain, S., Ahmed, S. E., and Doksum, K. A. (2015). Shrinkage, pretest, and penalty estimators in generalized linear models. *Statistical Methodology*, 24:52–68.

Hossain, S., Ahmed, S. E., Yi, Y., and Chen, B. (2016). Shrinkage and pretest estimators for longitudinal data analysis under partially linear models. *Journal of Nonparametric Statistics*, 28:531–549.

Hossain, S., Doksum, K. A., and Ahmed, S. E. (2009). Positive-part shrinkage and absolute penalty estimators in partially linear models. *Linear Algebra and its Applications*, 430:2749–2761.

Hossain, S. and Lac, L. A. (2021). Optimal shrinkage estimations in partially linear single-index models for binary longitudinal data. *TEST*. https://doi.org/10.1007/s11749-021-00753-3.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *Journal of the Royal Statistical Society, Series B*, 21:272–319.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2:849–879.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Leng, C., Liang, H., and Martinson, N. (2011). Parametric variable selection in generalized partially linear models with an application to assess condom use by HIV-infected patients. *Statistics in Medicine*, 30(12):2015–2027.

Li, R. and Nie, L. (2008). Efficient statistical inference procedures for partially nonlinear models and their applications. *Biometrics*, 64(3):904–911.

Liang, H. (2008). Generalized partially linear models with missing covariates. *Journal of Multivariate Analysis*, 99:880–895.

Mandal, S., Arabi Belaghi, R., Mahmoudi, A., and Aminnejad, M. (2019). Stein-type shrinkage estimators in gamma regression model with application to prostate cancer data. *Statistics in Medicine*, 38:4310–4322.

Mandal, S. and Torsney, B. (2006). Construction of optimal designs using a clustering approach. *Journal of Statistical Planning and Inference*, 136(3):1120–1134.

Mandal, S., Torsney, B., and Carriere, K. (2005). Constructing optimal designs with constraints. *Journal of Statistical Planning and Inference*, 128:609–621.

Mandal, S., Torsney, B., and Chowdhury, M. (2017). Optimal designs for minimizing covariances among parameter estimators in a linear model. *Australian & New Zealand Journal of Statistics*, 59(3):255–273.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, vol. 37 of monographs on statistics and applied probability, second edition.* Chapman and Hall, London.

Müller, M. (2001). Estimation and testing in generalized partial linear models – a comparative study. *Statistics and Computing*, 11:299–309.

Ni, X., Zhang, H., and Zhang, D. (2009). Automatic model selection for partially linear model. *Journal of Multivariate Analysis*, 100:2100–2111.

Nkurunziza, S. and Chen, F. (2013). On extension of some identities for the bias and risk functions in elliptically contoured distributions. *Journal of Multivariate Analysis*, 122:190–201.

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Applied Statistics*, 40:133–141.

Pukelsheim, F. (1993). *Optimal design of experiments*. Wiley Series in Probability and Mathematical Statistics, New York.

Raheem, S. E., Ahmed, S. E., and Doksum, K. A. (2012). Absolute penalty and shrinkage estimation in partially linear models. *Computational Statistics and Data Analysis*, 56(4):874–891.

Rahman, J., Luo, S., Fan, Y., and Liu, X. (2020). Semiparametric efficient inferences for generalised partially linear models. *Journal of Nonparametric Statistics*, 32(3):704–724.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley and Sons, New York.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511.

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models. *The Annals of Statistics*, 20:1768–1802.

Shah, K. and Sinha, B. (1989). *Theory of optimal designs. Lecture Notes in Statistics*, volume 54. Springer-Verlag.

Silvey, S. (1980). *Optimal Design*. Chapman and Hall, London.

Silvey, S. D., Titterington, D. M., and Torsney, B. (1978). An algorithm for optimal designs on a finite design space. *Communications in Statistics - Theory and Methods*, 7(14):1379–1389.

Speckman, P. E. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society. Series B*, 50(3):413–436.

Titterington, D. M. (1976). Algorithms for computing *D*-optimal designs on a finite design space. *Proc. 1976 Conf. on Information Sciences and Systems*, pages 213–216.

Torsney, B. (1977). Contribution to discussion of 'maximum likelihood estimation via the em algorithm' by dempster et al. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:26–27.

Torsney, B. (1983). *A moment inequality and monotonicity of an algorithm. In: Fiacco A.V., Kortanek K.O. (eds) Semi-infinite programming and applications. Lecture notes in economics and mathematical systems*, volume 215. Springer, Berlin, Heidelberg.

Torsney, B. (1988). Computing optimizing distributions with applications in design, estimation and image processing. In Dodge, Y., Fedorov, V. V., and Wynn, H. P., editors, *Optimal design and analysis of experiments*, chapter 15, pages 361–370. Elsevier Science Publishers B. V., North Holland.

Torsney, B. and Mandal, S. (2001). *Construction of constrained optimal designs* Optimum design 2000, volume 215. Kluwer Academic Publishers.

Vaart, A. (1998). *Asymptotic Statistics* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, Cambridge.

Wellner, J. A., Klaassen, C. A. J., and Ritov, Y. (2006). Semiparametric models: a review of progress since bkrw (1993). *In: Fan, J.; Koul, HL., editors. Frontiers in Statistics*, pages 25–44.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35:123–130.

Wu, C. F. J. (1978). Some iterative procedures for generating nonsingular optimal designs. *Communications in Statistics – Theory and Methods*, 14:1399–1412.

Wynn, H. P. (1972). Results in the theory and construction of *D*-optimum experimental designs (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):133–147.

Xu, J. and Yang, H. (2012). On the preliminary test backfitting and Speckman estimators in partially linear models and numerical comparisons. *Communications in Statistics–Simulation and Computation*, 41:327–341.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97:1042–1054.