ORIGINAL PAPER

Stein Estimation and Statistical Shrinkage Methods

Stein-rule M-estimation in sparse partially linear models

Enayetur Raheem[1] · **S. Ejaz Ahmed[2]** · **Shuangzhe Liu[3]**

Abstract

We propose and investigate the statistical properties of shrinkage M-estimators based on Stein-rule estimation for partially linear models under the assumption of sparsity. We are mainly interested in estimating regression coefficients parameter sub-vector with strong signals when the sparsity assumption may or may not hold. Thus, we consider two models, one including all the predictors, leading to a full (unrestricted, or over-fitted) model estimation; and the other with only a few influential predictors, resulting in a submodel (restricted, or under-fitted model) estimation problem. Generally speaking, submodel estimators perform better than full model estimators, when the assumption of sparsity is nearly correct. However, a small departure from this assumption makes submodel estimators biased and inefficient, questioning its applicability for practical reason. On the other hand, the full model estimators may not be desirable due to interpretability and higher estimation errors, specially when a large number of predictors are included in the model. For this reason, we propose shrinkage strategies which combine both full model and submodel estimators in an optimal way. The asymptotic properties of the suggested estimators have been studied both analytically and numerically. The asymptotic bias and risk of the estimators are derived in closed form. In addition, a simulation study is conducted to examine the performance of the estimators in practical settings when sparsity assumption may or may not hold. Our simulation results consolidate the theoretical properties of the estimators.

✉ Shuangzhe Liu
  shuangzhe.liu@canberra.edu.au

  Enayetur Raheem
  sraheem1@nccu.edu

  S. Ejaz Ahmed
  sahmed5@brocku.ca

[1] North Carolina Central University, Durham, NC, USA

[2] Brock University, St. Catharines, ON, Canada

[3] University of Canberra, Canberra, ACT, Australia

1 Introduction

We study robust shrinkage M-estimation in partially linear models (PLM) with a scaled error term. Ahmed et al. (2006) considered robust shrinkage-estimation of regression parameters when it is a priori suspected that the regression parameters could be restricted to a linear subspace. They studied the asymptotic properties of variants of Stein-rule M-estimators. For some insights on shrinkage estimation strategies, we refer to Ahmed and Fallahpour (2012), Ahmed and Raheem (2012), Raheem et al. (2012), Ahmed (2014), Ma et al. (2014), Ahmed et al. (2016), Sun et al. (2020) and Opoku et al. (2021). Recently, Ahmed et al. (2023) extended shrinkage strategies to high-dimension when the sparsity assumption cannot be judiciously justified. Maruyama et al. (2023) presented both classical and recent shrinkage estimation developments including results of admissibility of generalized Bayes estimators in the presence of a nuisance scale parameter.

Application of robust statistical techniques, including M-estimation and related approaches, to address modeling and prediction challenges are found in various domains, e.g. industrial data modeling (Zhou et al., 2020), disease incidence prediction (Susanti et al., 2020), and ratio-type estimation (Rather et al., 2022). These techniques are valuable for handling data imperfections, outliers, and non-normality.

Related works also include those by Arashi et al. (2014). They discussed the improvement of preliminary tests and Stein-rule Liu estimators specifically tailored for the ill-conditioned elliptical linear regression model. The authors focused on addressing the challenges posed by data that exhibit multicollinearity or non-normality, providing more accurate estimations in such scenarios. Norouzirad and Arashi (2019) discussed the use of preliminary tests and Stein-type shrinkage ridge estimators in the context of robust regression. They explored methods for improving the robustness and accuracy of regression models, particularly when dealing with outliers or influential data points. Recently, Shih et al. (2021) proposed a robust ridge M-estimators, incorporating pretests and Stein-rule shrinkage techniques for estimating the intercept term in regression models. They aimed to enhance the robustness of ridge regression when dealing with outliers and influential observations. On the other hand, in addition to the original results on Huber type M-estimation, a review of recent results and applications was given by e.g. Farcomeni and Ventura (2012).

Generally speaking, a PLM is more flexible than a linear model since it includes a nonlinear component along with the linear components. A PLM may provide a better alternative to the classical linear regression model in a situation where one or more predictors have nonlinear relationship with the response variable. Robust regression models are designed to overcome some of the limitations of classical linear regression in a host of scenarios. For example, least squares regression is highly sensitive to outliers, and is subject to underlying assumptions. Any violation of these assumptions may have serious impact on the validity of the fitted model.

In this paper, we extend the available works to a PLM, and develop shrinkage M-estimators. We construct shrinkage M-estimators of regression parameters under the sparsity assumption. Our analytical and numerical results establish the superiority of shrinkage M-estimators over full model and submodel M-estimators. We focus on shrinkage M-estimation of regression coefficients in a PLM when sparsity assumption may or may not hold. In our setup, the nonparametric part is estimated by the kernel-based method.

1.1 Statement of the problem

Consider a PLM of the form

$$Y = X\beta + g(T) + \sigma e, \tag{1.1}$$

where $Y = (y_1, y_2, \ldots, y_n)^\top$ is an n-response vector, $X = (x_1^\top, x_2^\top, \ldots, x_n^\top)^\top$ is the $n \times p$ design matrix with x_i's as known row p-vectors, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the p-vector of regression parameters, $g(T) = (g(t_1), g(t_2), \ldots, g(t_n))^\top$ is an unknown real-valued function, $e = (e_1, e_2, \ldots, e_n)^\top$ is the n-vector of random errors with mean $E(e) = 0$, e_i's are independent and identically distributed (iid) random variables having a continuous distribution, F, free from any unknown scale parameter $\sigma > 0$, where the $()^\top$ denotes transpose of a vector or a matrix. In passing, we would like to remark here that without loss of generality the intercept term is not included in establishing the asymptotic properties of the estimators.

Under the assumption of the sparsity, the data matrix X can be partitioned as $X = (X_1 : X_2)$ with $\beta = (\beta_1^\top, \beta_2^\top)^\top$, where X_1 and X_2 are $n \times p_1$ and $n \times p_2$ submatrices of predictors with strong signals and no signals, respectively. Thus, the model can be rewritten as

$$Y = X_1\beta_1 + X_2\beta_2 + g(T) + \sigma e, \quad p = p_1 + p_2 < n.$$

Under the assumption of sparsity, that is, $\beta_2 = 0$, then we have the submodel, under-fitted or restricted model as follows:

$$Y = X_1\beta_1 + g(T) + \sigma e, \quad p_1 < n,$$

and the remaining discussion follows.

In practice, the submodel can be readily obtained by applying a suitable variable selection method to full model. We are primarily interested in estimating β_1, when β_2 may or may not be a null vector. In other words, when practitioners may not be certain that model is fully sparse or not. In an effort to help data analyst, we propose shrinkage M-estimates based on Stein-rule to improve the performance of the under-fitted model estimators.

The remainder of the paper is organized as follows. In Sect. 2, we define a kernel-based least-squares (LS) estimators and discuss a two-step procedure to estimate the nonparametric function in a PLM. In Sect. 3, we propose Stein-rule shrinkage

M-estimators. Asymptotic properties of the estimators are presented in Sect. 4. The expressions for asymptotic bias and risk for the estimators are derived in Sect. 5. Monte Carlo simulation results are conducted in Sect. 6. Our concluding remarks are made in Sect. 7. Finally, our derivations of the theoretical results are included in the appendix.

2 Proposed LS estimation

In this section, we propose our robust LS estimation method with a two-step procedure.

Again, consider a PLM of the form

$$Y = X\boldsymbol{\beta} + g(T) + \sigma\boldsymbol{e}, \tag{2.1}$$

We first linearize (2.1) by estimating $g(\cdot)$ using kernel smoothing. We then confine ourselves to the estimation of $\boldsymbol{\beta}$ based on the partial residuals which attains the usual parametric convergence rate $n^{-1/2}$ without under-smoothing the nonparametric component $g(\cdot)$; see e.g. Speckman (1988).

Now, we describe the estimation process. We assume $\left\{y_i, \boldsymbol{x}_i^\top, t_i; i = 1, 2, \ldots, n\right\}$ satisfy (2.1). If $\boldsymbol{\beta}$ is the true parameter, then by $E(\boldsymbol{e}_i) = 0$, we have

$$g(t_i) = E(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}), \quad i = 1, 2, \ldots, n.$$

A natural nonparametric estimator of $g(\cdot)$ given $\boldsymbol{\beta}$ is

$$g^*(t, \boldsymbol{\beta}) = \sum_{i=1}^n W_{ni}(t)(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where

$$W_{ni}(t) = \frac{K((t_i - t)/h)}{\sum_{j=1}^n K((t_j - t)/h)}, \tag{2.2}$$

with $K(\cdot)$ being a kernel function which is a non-negative function integrable on $\mathfrak{R}$, and h being a bandwidth parameter. We need to make the assumptions as outlined in Appendix B.

Now, we define the conditional expectations

$$\gamma_0(t) = E(y|T = t), \text{ and}$$
$$\boldsymbol{\gamma}(t) = (\gamma_1(t), \gamma_2(t), \ldots, \gamma_n(t))^\top,$$

where $\gamma_j(t) = E(\boldsymbol{x}_j|T = t)$.

We estimate $\boldsymbol{\beta}$ using

$$\hat{\boldsymbol{\beta}} = \arg\min SS(\boldsymbol{\beta}) = (\widetilde{X}^\top \widetilde{X})^{-1} \widetilde{X}^\top \widetilde{Y}, \tag{2.3}$$

with

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} - g^*(t_i, \boldsymbol{\beta}) \right)^2 = \sum_{i=1}^{n} (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2, \qquad (2.4)$$

where $\widetilde{\boldsymbol{Y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^\top$, $\widetilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n)^\top$, $\tilde{y}_i = y_i - \gamma_0(t)$, and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\gamma}(t)$ for $i = 1, 2, \ldots, n$.

The conditional expectations $\gamma_0(t)$ and $\boldsymbol{\gamma}(t)$ are obtained using a classical nonparametric approach through

$$\hat{\gamma}_0(t) = \sum_{i=1}^{n} W_{ni}(t) y_i, \text{ and}$$

$$\hat{\gamma}_j(t) = \sum_{i=1}^{n} W_{ni}(t) x_{ij},$$

where $W_{ni}(t)$ is defined in (2.2). Clearly, once we obtain the estimates $\hat{\gamma}_0(t)$ and $\hat{\gamma}_j(t)$, they can be plugged into (2.4) prior to the estimation of $\boldsymbol{\beta}$.

The above procedure was independently proposed by Denby (1986) and Speckman (1988). A similar approach was taken by Ahmed et al. (2007) in estimating the nonparametric component in a PLM.

We obtain the robust M-estimators of the parameters of a PLM using a two-step procedure as follows:

Step 1 We first estimate $\gamma_0(t)$ and $\gamma_j(t)$ through kernel smoothing as described above. We denote the estimates by $\hat{\gamma}_0(t)$ and $\hat{\gamma}_j(t)$, respectively.
Step 2 The estimates in Step 1 are then plugged into (2.2). So the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained by regressing the residuals $y_i - \hat{\gamma}_0(t)$ and $\boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t)$ using a robust procedure. We denote the residuals as $\hat{r}_i = y_i - \hat{\gamma}_0(t)$ and $\boldsymbol{u}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)$.

Consistency and asymptotic normality of the estimators can be found in Appendix Section B.1 and the reference therein.

3 Proposed shrinkage M-estimation strategies

In this section, we propose our full model and submodel estimators, and formulate a test statistic which has asymptotically a non-central χ^2 distribution.

Let $\hat{\boldsymbol{\beta}}_1^{\text{RM}}$ be the restricted estimator of $\boldsymbol{\beta}_1$ where $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and $\hat{\beta}_1^{\text{UM}}$ be the unrestricted estimator of β_1 when $\boldsymbol{\beta}_2$ may not be a null vector. Following Ahmed (2014), a Stein-type M-estimator (SM), $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$ of $\boldsymbol{\beta}_1$ can be defined as

$$\hat{\boldsymbol{\beta}}_1^{\text{SM}} = \hat{\boldsymbol{\beta}}_1^{\text{RM}} + (\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}}) \left\{ 1 - \kappa \psi_n^{-1} \right\}, \quad \text{for } \kappa = p_2 - 2 \text{ and } p_2 \geq 3,$$

where ψ_n is a distance statistic defined later in (3.7). To avoid the over-shrinkage problem, the positive-rule Stein-type M-estimator (SM+) has the form

$$\hat{\boldsymbol{\beta}}_1^{\text{SM+}} = \hat{\boldsymbol{\beta}}_1^{\text{RM}} + (\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}}) \left\{1 - \kappa \psi_n^{-1}\right\}^+, \quad \text{for } p_2 \geq 3,$$

where $z^+ = \max(0, z)$.

3.1 Full model and submodel estimation strategies for $\hat{\boldsymbol{\beta}}_1$

For a suitable absolutely continuous function $\rho : \mathfrak{R} \rightarrow \mathfrak{R}$, with derivative ϕ, an M-estimator of $\boldsymbol{\beta}$ is defined as a solution of the minimization

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}). \tag{3.1}$$

Generally, an M-estimator is regression-equivariant, i.e.,

$$\boldsymbol{M}_n(c\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{a}) = c\boldsymbol{M}_n(\boldsymbol{Y}) + c\,\boldsymbol{a}, \quad \text{for } \boldsymbol{a} \in \mathfrak{R}_p,$$

and robustness depends on the choice of $\rho(\cdot)$. But it is generally not scale-equivariant. That is, it may not satisfy

$$\boldsymbol{M}_n(c\boldsymbol{Y}) = c\boldsymbol{M}_n(\boldsymbol{Y}), \quad \text{for } c > 0.$$

To have the estimators scale and regression equivariant, we need to studentize them. The studentized M-estimator is defined as as solution of the minimization

$$\min_{\beta \in \mathfrak{R}^p} \sum_{i=1}^{n} \rho\left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}}{S_n}\right), \tag{3.2}$$

where $S_n = S_n(\boldsymbol{Y}) \geq 0$ is an appropriate scale statistic that is regression equivariant and scale equivariant, i.e.,

$$S_n(c(\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{a})) = cS_n(\boldsymbol{Y}), \quad \text{for } \boldsymbol{a} \in \mathfrak{R}_p \text{ and } c > 0.$$

According to Jurečcková and Sen (1996), the minimization in (3.2) should be supplemented by a rule how to define $\boldsymbol{M}_n$ in the case when $S_n(\boldsymbol{Y}) = 0$. However, in general, this happens with probability zero, and the specific rule does not affect the asymptotic properties of $\boldsymbol{M}_n$. There are additional regularity conditions needed with (3.2), which we present in Appendix A. Further details may be found in Jurečcková and Sen (Jurečcková and Sen (1996), page 217).

Now, in an effort to define the M-estimator for $\boldsymbol{\beta}_1$, we define $C = X^\top X$ with $X = (X1 : X2)$ as follows:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} X_1^\top X1 & X_1^\top X2 \\ X_2^\top X1 & X_2^\top X2 \end{pmatrix}$$

Also, we define

$$C_{22.1} = C_{22} - C_{21} C_{11}^{-1} C_{12}.$$

Note that, if $C_{21} = 0$, then $C_{22.1} = C_{22}$. Otherwise, $C_{22} - C_{22.1}$ is positive semi-definite. We assume that C and $C_{22.1}$ are positive.

A studentized unrestricted M-estimator of $\boldsymbol{\beta}$ is defined as a solution of (3.2). Let us denote it by

$$\hat{\boldsymbol{\beta}}^{\mathrm{UM}} = \left(\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} \right)^\top, \left(\hat{\boldsymbol{\beta}}_2^{\mathrm{UM}} \right)^\top \right)^\top.$$

A studentized restricted M-estimator of $\boldsymbol{\beta}_1$ is obtained by minimizing

$$\min_{\boldsymbol{\beta}_1 \in \mathfrak{R}^{p_1}} \sum_{i=1}^{n} \rho \left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}^\top \boldsymbol{\beta}_1}{S_n} \right), \tag{3.3}$$

where S_n is regression-invariant so is not affected by the restricted environment. Since $\rho(\cdot)$ is assumed to have derivative $\phi(\cdot)$, we rewrite $\hat{\boldsymbol{\beta}}^{\mathrm{UM}}$ as a solution of

$$\boldsymbol{M}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i \, \phi \left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}}{S_n} \right) = \boldsymbol{0}. \tag{3.4}$$

In other words,

$$\boldsymbol{M}_n(\hat{\boldsymbol{\beta}}^{\mathrm{UM}}) = \boldsymbol{0}.$$

Similarly, $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ is a solution of

$$\boldsymbol{M}_{n_1}(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_{i1} \, \phi \left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}^\top \boldsymbol{\theta}_1}{S_n} \right) = \boldsymbol{0}. \tag{3.5}$$

Now, let

$$\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}} = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_{i2} \, \phi \left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}^\top \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}}{S_n} \right). \tag{3.6}$$

Note that $\boldsymbol{M}_n$ is a $(p_1 + p_2)$-vector, $\boldsymbol{M}_{n_1}$ is a p_1-vector and $\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}$ is a p_2-vector.

3.2 Test statistic

Following (Jurečcková and Sen (1996), Sect. 10.2) a suitable test statistic can be formulated as follows:

$$\psi_n = \frac{\left[\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}\right]^{\top} \boldsymbol{C}_{22.1}^{-1} \left[\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}\right]}{\hat{\sigma}_{\Phi_n}}, \tag{3.7}$$

where

$$\hat{\sigma}_{\Phi_n}^2 = (n - p_2)^{-1} \sum_{i=1}^{n} \phi^2 \left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}^{\top} \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}}{S_n}\right). \tag{3.8}$$

Directly applying the Lemma 5.5.1 in (Jurečcková and Sen (1996), page 220), it can be shown that under the sparsity assumption, that is, $\boldsymbol{\beta}_2$ is a null vector

$$\psi_n \xrightarrow{d} \chi_{p_2}^2.$$

Further, under (local) alternative hypothesis ψ_n has a non-central χ^2 distribution.

It is to be mentioned here that unlike LS estimators, M-estimators are not linear. Even if the distribution function F is normal, the finite sample distribution theory of M-estimators is not simple. Asymptotic methods Jurečcková and Sen (1996) have been used to overcome this difficulty.

4 Asymptotic properties of the estimators

In this section, we establish the asymptotic properties of the estimators. This facilitates in finding the asymptotic distributional bias (ADB), asymptotic distributional quadratic bias (ADQB), and asymptotic distributional quadratic risk (ADQR) of the estimator of the regression parameter vector $\boldsymbol{\beta}_1$.

Under the assumed regularity conditions

$$\lim_{n \to \infty} \frac{C_n}{n} = \boldsymbol{Q}, \tag{4.1}$$

where

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix},$$

it is known that under non-sparsity assumption, that is under local alternatives, $\boldsymbol{\beta}_2 \neq \boldsymbol{0}$,

$$\frac{\psi_n}{n} \to \gamma(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2; \boldsymbol{Q}) > 0, \quad \text{as } n \to \infty,$$

such that the shrinkage factor $\kappa \psi_n^{-1} = \mathcal{O}_p(n^{-1})$. This implies, asymptotically, there is no shrinkage effect. Therefore, to obtain meaningful asymptotics, we consider a class of local alternatives, $\{K_n\}$, given by

$$K_n : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2n} = \frac{\boldsymbol{\omega}}{\sqrt{n}}, \tag{4.2}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_{p_2})^\top \in \mathfrak{R}^{p_2}$ is a fixed vector and $||\boldsymbol{\omega}|| < \infty$, so that the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ reduces to $H_0 : \boldsymbol{\omega} = \boldsymbol{0}$.

For an estimator $\boldsymbol{\beta}_1^*$ and a positive-definite matrix $\boldsymbol{W}$, we define the loss function of the form

$$L(\boldsymbol{\beta}_1^*; \boldsymbol{\beta}_1) = n(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)^\top \boldsymbol{W} (\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1).$$

Thus, the risk function is defined as follows:

$$\begin{aligned}
R[(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1); \boldsymbol{W}] &= n E[(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)^\top \boldsymbol{W} (\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)] \\
&= n \operatorname{tr}[\boldsymbol{W} \{E(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)^\top\}] \\
&= \operatorname{tr}(\boldsymbol{W} \boldsymbol{\Omega}^*),
\end{aligned} \tag{4.3}$$

where tr denotes the trace operator and $\boldsymbol{\Omega}^*$ is the covariance matrix of $\sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)$. Whenever $\lim_{n \to \infty} \hat{\boldsymbol{\Omega}}_n^* = \hat{\boldsymbol{\Omega}}^*$ exists, the asymptotic risk is defined by

$$R_n(\boldsymbol{\beta}_{1n}^*, \boldsymbol{\beta}_1; \boldsymbol{W}) \to R(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1; \boldsymbol{W}) = \operatorname{tr}(\boldsymbol{W} \hat{\boldsymbol{\Omega}}^*).$$

Consider the asymptotic cumulative distribution function (cdf) of $\sqrt{n}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_1)$ under $\{K_n\}$ exists, and is defined as

$$G(\boldsymbol{y}) = P\left[\lim_{n \to \infty} \sqrt{n}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_1) \le \boldsymbol{y}\right].$$

This is known as the asymptotic distribution function (ADF) of $\boldsymbol{\beta}_1^*$. Suppose that $G_n \to G$ at all points of continuity as $n \to \infty$, and let $\hat{\boldsymbol{\Omega}}^*$ be the covariance matrix of G. Then the ADR is defined as

$$R(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1; \boldsymbol{W}) = \operatorname{tr}(\boldsymbol{W} \boldsymbol{\Omega}_G^*).$$

As noted in Ahmed et al. (2006), if $G_n \to G$ in second moment, then ADR is the asymptotic risk. However, this is a stronger mode of convergence, and is hard to analytically prove for shrinkage M-estimators. Therefore, they suggested using asymptotic distributional risk.

Now let

$$\boldsymbol{\Gamma} = \int \int \cdots \int \boldsymbol{y} \boldsymbol{y}^\top dG(\boldsymbol{y})$$

be the dispersion matrix which is obtained from ADF. The asymptotic distributional quadratic risk (ADQR) may be defined as

$$R(\boldsymbol{\beta}_1^*; \boldsymbol{\beta}_1) = \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}), \tag{4.4}$$

where $\boldsymbol{\Gamma}$ is the asymptotic distributional mean squared error (ADMSE) of the estimators.

To establish the asymptotic properties of the estimators, we present two important theorems.

Theorem 1 *Consider an absolutely continuous function $f(\cdot)$ with derivative $f'(\cdot)$ which exists everywhere, and finite Fisher information*

$$I(f) = \int_R \left(\frac{-f'(x)}{f(x)}\right)^2 dF(x) < \infty.$$

Under $\{K_n\}$ and the assumed regularity conditions, ψ_n has asymptotically a non-central chi-square distribution with non-centrality parameter $\Delta = \boldsymbol{\omega}^\top \boldsymbol{Q}_{22.1}\boldsymbol{\omega}\gamma^{-2}$, where

$$\gamma^2 = \frac{\int_R \phi^2(y)\mathrm{d}F(y)}{\int_R \phi(x)[-f'(x)/f(x)]^2\mathrm{d}F(x)}, \tag{4.5}$$

and $\phi(\cdot)$ is defined in ([3.4](#3.4)) or Appendix A.

Theorem 2 *We have, under the assumed regularity conditions, as $n \to \infty$*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{UM} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}, \gamma^2 \boldsymbol{Q}^{-1}). \tag{4.6}$$

Proofs of these theorems are available in Jurečková and Sen ([1996](#1996)).

5 Asymptotic bias and risk of the estimators

In this section, we present the asymptotic distribution, bias and risk results for each of our estimators. We also compare their risk performances.

Theorem 3 *Under the local alternative K_n and the assumed regularity conditions, we have as $n \to \infty$*

(i) $\eta_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{UM} - \boldsymbol{\beta}_1) \xrightarrow{d} N(\boldsymbol{0}, \gamma^2 \boldsymbol{Q}_{11.2}^{-1})$,

(ii) $\eta_2 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{UM} - \hat{\boldsymbol{\beta}}_1^{RM}) \xrightarrow{d} N(\boldsymbol{\delta}, \boldsymbol{\Sigma}^*)$, $\quad \boldsymbol{\delta} = -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}$,

(iii) $\eta_3 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{RM} - \boldsymbol{\beta}_1) \xrightarrow{d} N(-\boldsymbol{\delta}, \boldsymbol{\Omega}^*)$, $\quad \boldsymbol{\Omega}^* = \gamma^2 \boldsymbol{Q}_{11}^{-1}$.

We have, under $\{K_n\}$

$$\sqrt{n}\left((\hat{\boldsymbol{\beta}}_1^{UM} - \boldsymbol{\beta}_1)^\top, (\hat{\boldsymbol{\beta}}_2^{UM} - n^{-\frac{1}{2}}\boldsymbol{\omega})^\top\right)^\top \xrightarrow{d} N(\boldsymbol{0}, \gamma^2 \boldsymbol{Q}^{-1}),$$

where $\boldsymbol{Q}$ is partitioned as in (4.1).

Also, we have the joint distributions as follows:

$$
\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\delta} \end{pmatrix}, \begin{pmatrix} \gamma^2 \boldsymbol{Q}_{11.2}^{-1} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}^* \end{pmatrix} \right]
$$
$$
\begin{pmatrix} \eta_2 \\ \eta_3 \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{\delta} \\ -\boldsymbol{\delta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}^* \end{pmatrix} \right].
$$

The proof for this theorem is given in Appendix C.

5.1 Asymptotic bias of the estimators

The asymptotic distributional bias (ADB) of an estimator $\boldsymbol{\beta}^*$ is defined as

$$
\text{ADB}(\boldsymbol{\beta}^*) = E \left\{ \lim_{n \to \infty} n^{\frac{1}{2}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}) \right\}.
$$

Theorem 4 *Under the assumed regularity conditions and the stated theorems above, and under $\{K_n\}$, the ADB of the estimators are as follows:*

$$
ADB(\hat{\boldsymbol{\beta}}_1^{UM}) = \boldsymbol{0}
$$
$$
ADB(\hat{\boldsymbol{\beta}}_1^{RM}) = -\boldsymbol{\delta}
$$
$$
ADB(\hat{\boldsymbol{\beta}}_1^{SM}) = \kappa \boldsymbol{\delta} E \left\{ \chi_{p_2+2}^{-2}(\Delta) \right\}
$$
$$
ADB(\hat{\boldsymbol{\beta}}_1^{SM+}) = ADB(\hat{\boldsymbol{\beta}}_1^{SM}) - \boldsymbol{\delta} \left[H_{p_2+2}(\kappa, \Delta) - E \left\{ \kappa \chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) < \kappa) \right\} \right],
$$

where $E\left\{ \chi_a^{-2}(\Delta) \right\}$ is the expected value of an inverse of a non-central χ^2 random variable with a degrees of freedom and non-centrality parameter Δ, and $H_a(y, \Delta)$ is the cdf of the a non-central χ^2 random variable with a degrees of freedom and non-centrality parameter Δ.

The proof for this theorem is given in Appendix D.

Let us define the asymptotic distributional quadratic bias (ADQB) of an estimator $\boldsymbol{\beta^*}$ of $\boldsymbol{\beta}_1$ by

$$
ADQB(\boldsymbol{\beta}^*) = [ADB(\boldsymbol{\beta}^*)]^\top \boldsymbol{\Sigma}^{-1} [ADB(\boldsymbol{\beta}^*)],
$$

where $\boldsymbol{\Sigma}$ is the dispersion matrix of $\hat{\boldsymbol{\beta}}_1^{UM}$ as $n \to \infty$. In our case, the dispersion matrix is $\boldsymbol{Q}_{11}$. Thus, the asymptotic quadratic distributional bias of the estimators are given below.

$$
ADQB(\hat{\boldsymbol{\beta}}_1^{UM}) = \boldsymbol{0},
$$
$$
ADQR(\hat{\boldsymbol{\beta}}_1^{RM}) = \boldsymbol{\delta}^\top \boldsymbol{Q}_{11}^{-1} \boldsymbol{\delta}
$$

$$\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) = \kappa^2 \boldsymbol{\delta}^\top \boldsymbol{Q}_{11}^{-1} \boldsymbol{\delta} \left[E \left\{ \chi_{p_2+2}^{-2}(\Delta) \right\} \right]^2$$

$$\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) = \boldsymbol{\delta}^\top \boldsymbol{Q}_{11}^{-1} \boldsymbol{\delta} \left[H_{p_2+2}(\kappa, \Delta) - E \left\{ \kappa \chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) < \kappa) \right\} \right].$$

The above expression reveal that, as expected, the unrestricted estimator of $\boldsymbol{\beta}_1$ is asymptotically unbiased. On the other, the bias function of restricted estimator is a function of the sparsity parameter (non-centrality parameter Δ), so under the sparsity assumption, the estimator is asymptotically unbiased. However, it is an unbounded function of Δ, not a desirable property.

It can be seen that both shrinkage estimators are also function of Δ, more importantly, they are bounded function of the non-centrality parameter. The magnitude of bias increases as Δ increases and then converges to zero as $\Delta \to \infty$. As expected the bias curve of positive-part shrinkage estimator is below or equal the curve the curve of the shrinkage estimators.

The bias is a function of MSE (risk), so onward we focus on the risk properties of the estimators.

5.2 Asymptotic risk and risk performance of the estimators

In Appendix E, we present the derivation of the expressions for asymptotic distributional mean square error (ADMSE), and consequentially the risk expressions of the respective estimators.

From the ADMSE and ADQR results in Appendix E, we see clearly that the risk of the classical unrestricted estimator is independent of the sparsity assumption, so its risk take a constant value of $\text{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}))$. On the other hand, the risk of the restricted estimator depends on the sparsity assumption, and when the assumption is nearly correct then $R(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) \leq R(\hat{\boldsymbol{\beta}}_1^{\text{UM}})$ and a strict inequality will hold for some values in the parameter space induced by the sparsity parameter. However, beyond this small interval in the parameter space, the unrestricted estimator will dominate the restricted estimator. As a matter of fact of the the restricted estimator risk is unbounded function of sparsity parameter, a undesirable property.

Interestingly, but not surprisingly, both shrinkage estimators are superior to benchmark estimators in the entire parameter space. For a suitable choice of $\boldsymbol{W}$, it can be verified that $R(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) \leq R(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) \leq R(\hat{\boldsymbol{\beta}}_1^{\text{UM}})$, and the strict inequality will hold for some values in the parameter space. Thus, the shrinkage estimators dominate the classical M-estimator. Further, the shrinkage estimators will outperform the restricted estimator except in a small interval where sparsity assumption may hold. Thus, we recommend the use of the shrinkage estimators, they are in closed form and free from any tuning parameter.

6 Simulation studies

In this section, we conduct a simulation study to appraise the performance of in practical setting and to quantify the relative behavior of the estimators. We perform Monte Carlo simulation experiments to examine the quadratic risk performance of the proposed estimators. We simulate the response from the following model:

$$y_i = \sum_{l=1}^{p_1} x_{il}\beta_l + \sum_{m=p_1+1}^{p} x_{im}\beta_m + \sin(4\pi t_i) + \varepsilon_i \tag{6.1}$$

where β_l is a $p_1 \times 1$ vector and β_m is $p_2 \times 1$ vector of parameters with $p = p_1 + p_2$, and $\varepsilon_i \sim N(0, 1)$, $i = 1, \ldots, n$. Furthermore, $x_{i1} = (\zeta_{i1}^{(1)})^2 + \zeta_i^{(1)} + \xi_{i1}$, $x_{i2} = (\zeta_{i2}^{(1)})^2 + \zeta_i^{(1)} + 2\xi_{i2}$, $x_{is} = (\zeta_{is}^{(1)})^2 + \zeta_i^{(1)}$ with $\zeta_{is}^{(1)} \sim N(0, 1)$, $\zeta_i^{(1)} \sim N(0, 1)$, $\xi_{i1} \sim$ Bernoulli(0.35) and $\xi_{i2} \sim$ Bernoulli(0.35) for all $s = 3, \ldots, p$.

We are interested in testing the assumption of the sparsity in the form of statistical hypothesis $H_0 : (\beta_{p_1+1}, \beta_{p_1+2}, \ldots, \beta_{p_1+p_2}) = \mathbf{0}$. Our aim is to estimate $\boldsymbol{\beta}_1$ when sparsity assumption may or may not be true. We partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)$. Each realization was repeated 5000 times to obtain stable results. For each realization, we calculated the MSE of the estimators.

We define $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1^\top, 0)^\top$ and $|| \cdot ||$ is the Euclidean norm. In addition, Δ^* and S_n were estimated by median absolute deviation (MAD). To determine the behavior of the estimators for $\Delta^* > 0$, further data sets were generated from those distributions under the alternative hypothesis.

6.1 Error distributions

In an effort to evaluate the performance of the proposed estimators numerically, we perform a simulation study. We generate data from four different error distributions, namely the standard normal, contaminated normal, standard logistic distribution, and standard Laplace distribution, respectively.

The cumulative distribution function

$$F(x) = \lambda N(0, \omega^2) + (1 - \lambda)N(0, 1) \tag{6.2}$$

was used to generate the standard normal and contaminated normal errors, where λ is the parameter indicating whether the standard normal or its contaminated version is returned. We consider $\lambda = 0$ and $\lambda = 0.9$, respectively. Indeed, for $\lambda = 0$ we get the standard normal errors, while for $\lambda = 0.9$, with $\omega^2 \neq 1$, we obtain the scale contaminated normal errors.

The standard logistic distribution has cdf

$$F(x) = \frac{1}{1 + e^{-x}}, \quad x \in \mathfrak{R}. \tag{6.3}$$

The standard Laplace distribution has cdf

$$F(x) = \frac{1}{2} \left[1 + \text{sign}(x)(1 - e^{-|x|}) \right], \quad x \in \mathfrak{R}. \tag{6.4}$$

6.2 Relative risk comparison

The risk performance of an estimator of $\boldsymbol{\beta}_1$ was measured by comparing its MSE with that of the unrestricted M-estimator. We numerically calculated the relative MSE (RMSE) of the proposed estimators $\hat{\boldsymbol{\beta}}_1^{\text{RM}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SM+}}$ to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1^{\text{UM}}$, given by

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} : \hat{\boldsymbol{\beta}}_1^{*}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{UM}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{*})}, \tag{6.5}$$

where $\hat{\boldsymbol{\beta}}_1^{*}$ is one of the proposed estimators. The amount by which an RMSE is larger than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^{*}$ over $\hat{\boldsymbol{\beta}}_1^{\text{UM}}$; see also Fig. 1.

We compute the RMSE values for $n = 30, 50$ and various configurations of (p_1, p_2) based on Huber's $\rho-$function. Our results are presented in Fig. 1 and Tables 1–4.

Figure 1 shows the RMSE values of various M-estimators. Here, Δ^* indicates the correctness of the submodel under sparsity assumption. Thus, $\Delta^* > 0$ quantify the degree of deviation from the assumed model. Figure 1 clearly shows that the restricted estimator is the best when Δ^* is close to the origin. However, the restricted estimator become inefficient and the RMSE goes below 1 very quickly as Δ^* deviates from zero. The RMSE of restricted estimator is depicted by the dashed line in Fig. 1. In the simulation study, the restricted estimator shows similar behaviour for all the error distributions considered in this study.

Tables 1–4 portrayed similar characteristic of the estimators. Both shrinkage estimators dominate the classical M-estimator, and positive-rule shrinkage estimator (SM+) dominates the shrinkage estimator. As for example, Table 1 presents the RMSEs for $(p_1, p_2) = (3, 5)$ and $n = 30$. For the standard normal error, the gain in risk for the positive-rule shrinkage M-estimator is 3.161 times that of the classical M-estimator provided that the model specification is correct (i.e., $\Delta^* = 0$). For the same configuration, when the error distribution is the standard Laplace, the gain in risk for SM+ is 2.273 times that of unrestricted estimator. Interestingly, for the large dimensional case $(p_1, p_2) = (5, 20)$ in Table 3, the gain is much higher with the value 7.325 and
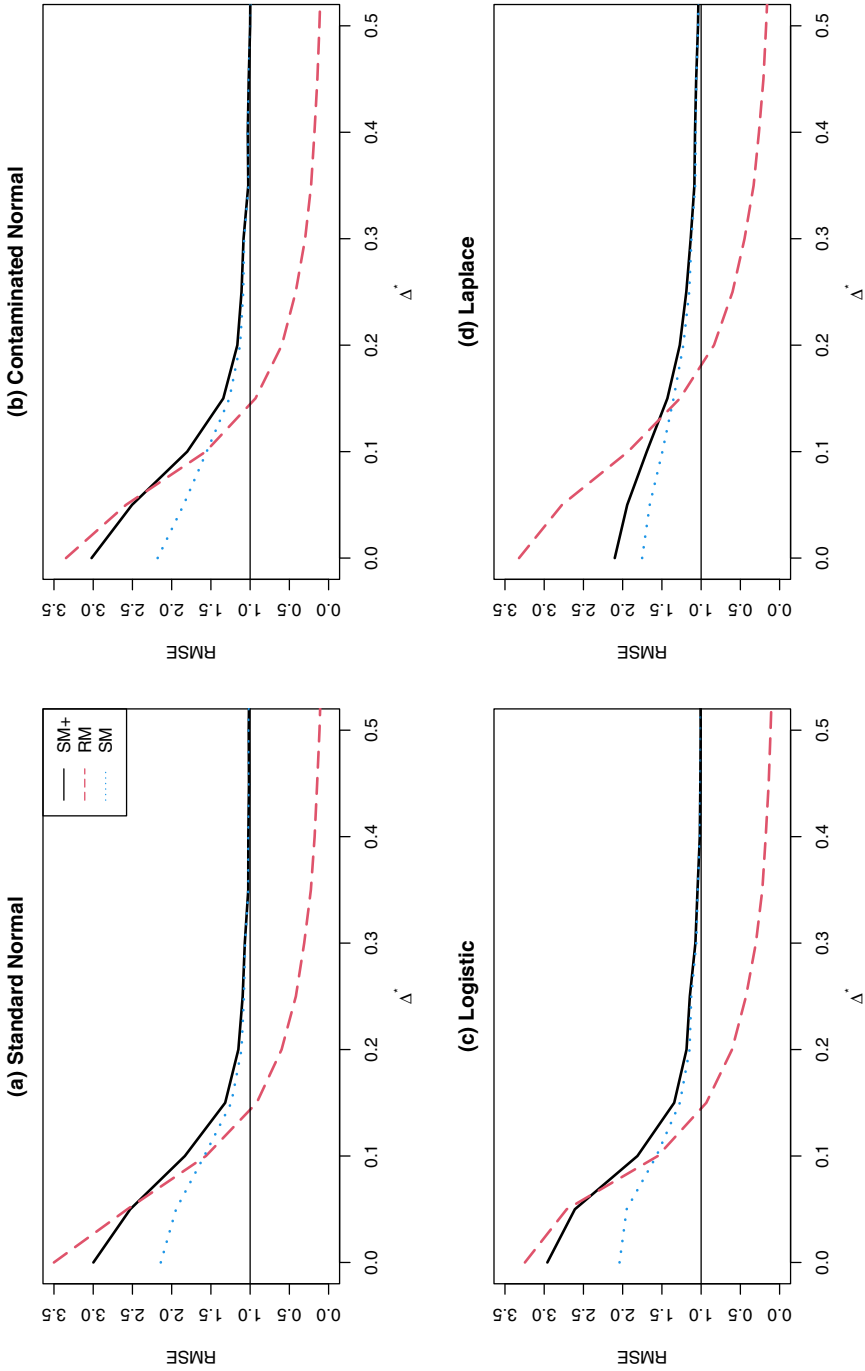
Fig. 1 RMSE values for RM, SM, and SM+ estimators with respect to the unrestricted M-estimator for $(p_1, p_2) = (3, 4)$, $n = 50$, when Huber's ρ-function is considered. Specially, RMSE = 1 when the estimators perform equivalently

Table 1 RMSE values for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (3, 5)$, $n = 30$, based on Huber's ρ−function for different error distributions

Error	Δ^*	$\hat{\beta}_1^{RM}$	$\hat{\beta}_1^{SM}$	$\hat{\beta}_1^{SM+}$
Standard	0.00	3.695	2.035	3.161
Normal	0.05	3.472	2.084	3.224
	0.10	2.386	1.758	2.515
	0.15	1.619	1.496	1.835
	0.20	1.065	1.216	1.479
	0.25	0.814	1.202	1.261
	1.00	0.060	0.995	0.997
Contaminated	0.00	3.867	2.163	3.195
Normal	0.05	2.842	1.733	2.621
	0.10	2.250	1.707	2.352
	0.15	1.476	1.492	1.839
	0.20	1.100	1.288	1.473
	0.25	0.737	1.082	1.163
	1.00	0.060	1.017	1.017
Standard	0.00	3.532	1.921	2.991
Logistic	0.05	3.288	1.922	3.004
	0.10	2.400	1.846	2.434
	0.15	1.656	1.551	1.853
	0.20	1.129	1.323	1.464
	0.25	0.758	1.158	1.208
	1.00	0.062	0.996	0.996
Standard	0.00	3.853	1.895	2.273
Laplace	0.05	3.628	1.743	2.056
	0.10	2.719	1.597	1.974
	0.15	2.179	1.426	1.753
	0.20	1.418	1.329	1.538
	0.25	1.090	1.273	1.360
	1.00	0.093	1.014	1.016

4.200, respectively, demonstrating the applicability, power and beauty of the Stein-rule estimators in high-dimensional cases.

In closing, our numerical results strongly corroborate the theoretical properties of the suggested estimators.

7 Concluding remarks

In this paper, the shrinkage M-estimation strategies in the context of a partially linear regression model are developed. The statistical properties of shrinkage and positive-rule shrinkage M-estimators are investigated when the sparsity assumption may or may not hold. The expressions for bias and risk of the estimators are presented in closed

Table 2 RMSE values for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (3, 9)$, $n = 50$, based on Huber's $\rho-$function for different error distributions

Error	Δ^*	$\hat{\beta}_1^{RM}$	$\hat{\beta}_1^{SM}$	$\hat{\beta}_1^{SM+}$
Standard	0.00	5.552	3.607	5.462
Normal	0.05	4.269	3.098	4.407
	0.10	2.574	2.322	2.919
	0.15	1.601	1.827	2.027
	0.20	1.085	1.503	1.583
	0.25	0.726	1.340	1.370
	1.00	0.051	0.994	0.994
Contaminated	0.00	5.443	3.576	5.301
Normal	0.05	4.457	3.009	4.510
	0.10	2.755	2.442	3.096
	0.15	1.668	1.788	2.028
	0.20	1.022	1.417	1.500
	0.25	0.750	1.313	1.353
	1.00	0.051	1.001	1.001
Standard	0.00	5.702	3.364	5.410
Logistic	0.05	4.422	3.316	4.459
	0.10	2.641	2.364	2.927
	0.15	1.666	1.941	2.139
	0.20	1.040	1.475	1.582
	0.25	0.710	1.338	1.374
	1.00	0.049	1.009	1.009
Standard	0.00	5.827	2.734	3.256
Laplace	0.05	5.187	2.489	2.989
	0.10	3.351	2.333	2.624
	0.15	2.332	2.142	2.242
	0.20	1.462	1.692	1.783
	0.25	1.039	1.565	1.606
	1.00	0.075	1.021	1.021

form. The relative performance of the estimators is critically examined, the positive-rule shrinkage estimator is found to perform better than the unrestricted estimator. Further, it outshines the restricted estimator except in small interval when the submodel at the hand assumed to be to nearly true model.

In the simulation study, we numerically compute relative mean squared errors of the restricted-M, shrinkage-M, and positive-rule shrinkage M-estimators compared to the unrestricted M-estimator. Four different error distributions are considered to study the performance of the proposed estimators. Our numerical provides support for the positive-rule shrinkage estimators under varying degrees of model misidentification, as well. The submodel restricted M-estimator outperforms all other estimators when

Table 3 RMSE values for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (5, 9)$, $n = 50$, based on Huber's $\rho-$function for different error distributions

Error	Δ^*	$\hat{\beta}_1^{RM}$	$\hat{\beta}_1^{SM}$	$\hat{\beta}_1^{SM+}$
Standard	0.00	3.838	2.772	3.705
Normal	0.05	3.202	2.438	3.179
	0.10	2.494	2.168	2.641
	0.15	1.638	1.708	1.923
	0.20	1.097	1.453	1.531
	0.25	0.778	1.201	1.260
	1.00	0.062	1.016	1.016
Contaminated	0.00	3.653	2.592	3.472
Normal	0.05	3.172	2.328	3.161
	0.10	2.391	2.035	2.530
	0.15	1.632	1.748	1.986
	0.20	1.123	1.453	1.550
	0.25	0.797	1.244	1.297
	1.00	0.064	1.028	1.028
Standard	0.00	3.740	2.730	3.678
Logistic	0.05	3.299	2.554	3.298
	0.10	2.424	2.125	2.544
	0.15	1.596	1.690	1.931
	0.20	1.077	1.383	1.480
	0.25	0.769	1.283	1.326
	1.00	0.059	1.002	1.002
Standard	0.00	3.919	2.125	2.626
Laplace	0.05	3.872	2.158	2.461
	0.10	2.613	1.923	2.207
	0.15	2.015	1.798	1.895
	0.20	1.519	1.637	1.703
	0.25	1.157	1.464	1.528
	1.00	0.099	1.022	1.022

there is sparsity. However, a small departure from this condition makes the restricted very inefficient, questioning its applicability for practical purposes. We suggest to use positive-rule shrinkage M-estimators due to its performance in the entire parameter space.

More importantly, the performance positive-rule shrinkage M-estimators is noticeable when p_2 is large, this work can be extended to high-dimensional cases, we refer to Ahmed et. al. (2023). We plan to study such extensions in a separate communication.

Table 4 RMSE values for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (5, 20)$, $n = 50$, based on Huber's $\rho-$function for different error distributions

Error	Δ^*	$\hat{\beta}_1^{RM}$	$\hat{\beta}_1^{SM}$	$\hat{\beta}_1^{SM+}$
Standard	0.00	7.469	5.415	7.328
Normal	0.05	6.034	4.502	6.145
	0.10	3.809	3.343	3.992
	0.15	2.230	2.487	2.727
	0.20	1.437	1.901	1.985
	0.25	1.019	1.638	1.672
	1.00	0.072	1.037	1.037
Contaminated	0.00	7.900	5.809	7.974
Normal	0.05	6.115	4.864	6.171
	0.10	3.593	3.111	3.820
	0.15	2.295	2.446	2.726
	0.20	1.491	1.967	2.054
	0.25	1.005	1.613	1.636
	1.00	0.073	1.028	1.028
Standard	0.00	7.366	5.569	7.238
Logistic	0.05	6.085	4.701	6.042
	0.10	3.767	3.391	4.100
	0.15	2.296	2.494	2.714
	0.20	1.482	1.920	2.004
	0.25	1.018	1.589	1.624
	1.00	0.072	1.025	1.025
Standard	0.00	8.550	3.841	4.200
Laplace	0.05	7.369	3.424	3.817
	0.10	4.973	3.262	3.457
	0.15	3.116	2.510	2.622
	0.20	2.149	2.266	2.342
	0.25	1.473	1.938	1.945
	1.00	0.112	1.054	1.054

Declarations

Conflicts of interest The authors declare no conflict of interest.

Appendix

A. Regularity conditions

Here, we list the regularity conditions needed for the minimization problem in (3.2). Detailed discussions about these conditions can be found in (Jurečcková and Sen (1996), p. 217–218).

For the studentized M-estimators, consider that $\phi = \rho'$ can be decomposed as

$$\phi = \phi_1 + \phi_2 + \phi_3, \tag{A.1}$$

where ϕ_1 is an absolutely continuous function with absolutely continuous derivative, ϕ_2 is a continuous piecewise linear function that is constant in a neighbourhood of $\pm\infty$, and ϕ_3 is a non-decreasing step function.

The following conditions are imposed on (3.2).

RC1 $S_n(Y)$ is regression invariant and scale equivariant, $S_n > 0$ a.s., and

$$\sqrt{n}(S_n - S) = O_p(1)$$

for some functional $S = S(F) > 0$.

RC2 The function $h(t) = \int \rho((z-t)/S) \mathrm{d}F(z)$ has the unique minimum at $t = 0$.

RC3 For some $\delta > 0$ and $\eta > 1$,

$$\int_{-\infty}^{\infty} \left\{ |z| \sup_{|u| \le \delta} \sup_{|v| \le \delta} \left| \phi_1'' \left(\frac{e^{-v}(z+u)}{S} \right) \right| \right\}^{\eta} \mathrm{d}F(z) < \infty$$

and

$$\int_{-\infty}^{\infty} \left\{ |z|^2 \sup_{|u| \le \delta} \left| \frac{\phi_1''(z+u)}{S} \right| \right\}^{\eta} \mathrm{d}F(z) < \infty,$$

where $\phi_1'(z) = \frac{d}{dz}\phi_1(z)$ and $\phi_1''(z) = \frac{d^2}{dz^2}\phi_1(z)$.

RC4 ϕ_3 is a continuous, piecewise linear function with knots at $\mu_1, \ldots, \mu_k$, which is constant in a neighborhood of $\pm\infty$. Hence the derivative ϕ_3' of ϕ_3 is a step function

$$\phi_3'(z) = \alpha_\nu \quad \text{for } \mu_\nu < z < \mu_{\nu+1}, \nu = 0, 1, \ldots, k,$$

where $\alpha_0, \alpha_1, \ldots, \alpha_k \in \mathfrak{R}_1$, $\alpha_0 = \alpha_k = 0$ and $\infty = \mu_0 < \mu_1 < \cdots < \mu_k < \mu_{k+1} = \infty$. Further, we assume that $f(z) = \frac{\mathrm{d}F(z)}{dz}$ is bounded in neighbourhood of S_{μ_j}, $j = 1, 2, \ldots, k$.

RC5 $\phi_3(z) = \lambda_\nu$ for $q_\nu < z \le q_{\nu+1}$, $\nu = 1, 2, \ldots, m$ where $-\infty = q_0 < q_1 < \cdots < q_m < q_{m+1} = \infty$, $-\infty < \lambda_0 < \lambda_1 < \cdots < \lambda_m < \infty$. We further assume that $f'(z)$ and $f''(z)$ are bounded in the neighbourhood of S_{q_j}, $j = 1, 2, \ldots, m$.

B. Assumptions

Assumption B.1 The function $g(\cdot)$ satisfies the Lipschitz condition of order 1 on $[0, 1]$.

Assumption B.2 The probability weight functions $W_{ni}(\cdot)$ satisfy

a) $\max_{1 \leq i \leq n} \sum_{j=1}^{n} W_{ni}(t_j) = \mathcal{O}(1)$,

b) $\max_{1 \leq i, \, j \leq n} \sum_{j=1}^{n} W_{ni}(t_j) = \mathcal{O}(n^{-2/3})$,

c) $\max_{1 \leq j \leq n} \sum_{i=1}^{n} W_{ni}(t_j) I(|t_i - t_j| > c_n) = \mathcal{O}(d_n)$, where I is the indicator function, c_n satisfies $\limsup_{n \to \infty} nc_n^3$, and d_n satisfies $\limsup_{n \to \infty} nd_n^3 < \infty$.

Remark 1 The usual polynomial and trigonometric functions satisfy Assumption B.1.

Remark 2 Under regular conditions, the Nadaraya-Watson kernel weights, Priestley and Chao kernel weights, locally linear weights and Gasser-Müller kernel weights satisfy Assumption B.2. If we consider the pdf of $U[-1, 1]$ as the kernel function as

$$K(t) = \frac{1}{2} I_{[-1,1]}(t),$$

with $t_i = \frac{i}{n}$, and the bandwidth $cn^{-1/3}$ where c is constant, then the Priestley and Chao kernel weights satisfy Assumption B.2, and the weights are

$$W_{ni}(t) = \frac{1}{2cn^{\frac{2}{3}}} \left(\left| t - \frac{i}{n} \right| \leq cn^{-\frac{1}{3}} \right)^{(t)}.$$

For a detailed discussion on the assumptions above, see Ahmed et al. (2007).

B.1 Consistency and Asymptotic Normality

Now, we denote the random vector $(R(T), \boldsymbol{U}(T)^\top)^\top$ with the same distribution as $(r_i, \boldsymbol{u}_i^\top)^\top$.

Consistency of the regression parameters in a semi-parametric model has been proved in great detail in Bianco and Boente (2004). We omit the details but present only the set of assumptions, lemma, and theorem which are needed for proving asymptotic normality and consistency of the estimators.

Let $\tilde{\rho}$ and $\widetilde{W}$ be score and weight functions, respectively. The asymptotic distribution of $\boldsymbol{\beta}$ is defined as a solution of

$$\sum_{i=1}^{n} \tilde{\rho} \left(\frac{\hat{r}_i - \hat{\boldsymbol{u}}_i \hat{\boldsymbol{\beta}}}{s_n} \right) \widetilde{W}(\|\hat{\boldsymbol{u}}_i\|) \hat{\boldsymbol{u}}_i = 0, \tag{B.1}$$

where $\hat{r}_i = y_i - \hat{\gamma}_0(t)$, $\boldsymbol{u}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)$, and s_n is an estimate of the residual scale.

To derive the asymptotic distribution of $\boldsymbol{\beta}$ we must have t_i in a compact set, so without loss of generality, we assume that $t_i \in [0, 1]$. We need the following set of assumptions. See Bianco and Boente (2004) for details.

A1 $\tilde{\rho}$ is odd, bounded, continuous, and twice differentiable with bounded derivative $\tilde{\rho}'$ and $\tilde{\rho}''$ such that $\phi_1(t) = t\tilde{\rho}'(t)$ and $\phi_2(t) = t\tilde{\rho}''(t)$ are bounded.

A2 $E(\tilde{W}(||\boldsymbol{U}(T)||)||\boldsymbol{U}(T)||^2) < \infty$ and the matrix

$$\boldsymbol{A} = E\left(\tilde{\rho}'\left(\frac{R(T) - \boldsymbol{U}(T)^\top \boldsymbol{\beta}}{\sigma}\right)\tilde{W}(||\boldsymbol{U}(T)||)\boldsymbol{U}(T)\boldsymbol{U}(T)^\top\right)$$

is nonzero.

A3 $\tilde{W}(u) = \tilde{\rho}_1(u)u^{-1} > 0$ is a bounded function which satisfies the Lipschitz condition of order 1. Further, $\tilde{\rho}_1$ is bounded with bounded derivative.

A4 $E(\tilde{W}(||\boldsymbol{U}(T)||)\boldsymbol{U}(T)|T = t) = 0$ for almost all t.

A5 The functions $\boldsymbol{x}_j(t), 0 \le j \le p$ are continuous in $[0, 1]$ with continuous first derivative.

Remark 3 According to Robinson ([1988](#)), condition A2 is needed so that no element of X can be predictable by T. A2 guarantees that there is no multicollinearity in the columns of $X - \tilde{X}_j(T)$. In other words, X has to be free from multicollinearity. Also, condition A5 is a standard requirement in kernel estimation in semi-parametric models in order to guarantee asymptotic normality.

Lemma B.1 *Let $(y_i, \boldsymbol{x}_i^\top, t_i)^\top, 1 \le i \le n$ be independent random vectors satisfying ([2.1](#)) with e_i independent of $(\boldsymbol{x}_i^\top, t_i)^\top$. Assume that t_i are random variable with $t_i \in [0, 1]$. Denote $(R(T), \boldsymbol{U}(T)^\top)^\top$ a random vector with the same distribution as*

$$(r_i, \boldsymbol{u}_i^\top)^\top = (y_i - \hat{\gamma}_0(t_i), [\boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)]^\top)^\top.$$

Further, let $\hat{\boldsymbol{\gamma}}_j(t_i), 0 \le j \le p$ be the estimates of $\gamma_j(t_i)$ such that

$$\sum_{t \in [0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{p} 0, \quad 0 \le j \le p.$$

If $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and $s_n \xrightarrow{p} \sigma$, then under the stated assumptions A1-A3, $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$, where $\boldsymbol{A}$ is defined in A2, and

$$\boldsymbol{A}_n = n^{-1}\sum_{i=1}^{n}\tilde{\rho}'\left(\frac{\hat{r}_i - \hat{\boldsymbol{u}}_i^\top \hat{\boldsymbol{\beta}}}{s_n}\right)\tilde{W}(||\hat{\boldsymbol{u}}_i||)\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^\top,$$

where $\xrightarrow{p}$ denotes convergence in probability.

Proof The proof is available in the appendix of Bianco and Boente ([2004](#)).

Theorem B.1 *Let $(y_i, \boldsymbol{x}_i^\top, t_i)^\top, 1 \le i \le n$ be independent random vectors satisfying ([2.1](#)) with e_i independent of $(\boldsymbol{x}_i^\top, t_i)^\top$. Assume that t_i are random variables with $t_{i_n} \in [0, 1]$. Denote $(R(T), \boldsymbol{U}(T)^\top)^\top$ a random vector with the same distribution as*

$$(r_i, \boldsymbol{u}_i^\top)^\top = (y_i - \gamma_0(t_i), (x_i - \boldsymbol{\gamma}(t_i))^\top)^\top.$$

Further, let $\hat{\gamma}_j(t)$, $0 \leq j \leq p$ be estimates of $\gamma_j(t)$ such that first derivative of $\hat{\gamma}_j(t)$ exists and is continuous, and

$$n^{1/4} \sum_{t \in [0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{P} 0, \quad 0 \leq j \leq p, \qquad (B.2)$$

$$\sum_{t \in [0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{P} 0, \quad 0 \leq j \leq p. \qquad (B.3)$$

Then, if $s_n \xrightarrow{p} \sigma$, under A1-A5,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{Q})$$

with $\boldsymbol{Q} = \boldsymbol{A}^{-1} \boldsymbol{\Sigma} (\boldsymbol{A}^{-1})^\top$, where $\boldsymbol{A}$ is defined in A2 and

$$\boldsymbol{\Sigma} = \sigma^2 E \left(\tilde{\rho}^2 \left(\frac{R(T) - \boldsymbol{U}(T)^\top \boldsymbol{\beta}}{\sigma} \right) \widetilde{W}^2(\|\boldsymbol{U}(T)\|) \boldsymbol{U}(T) \boldsymbol{U}(T)^\top \right).$$

Proof The proof is available in Bianco and Boente (2004).

C. Proof for Theorem 5.1

For Theorem 5.1, we derive $\boldsymbol{\Sigma}_{12}$ as follows:

$$
\begin{aligned}
\boldsymbol{\Sigma}_{12} &= Cov(\eta_1, \eta_2) \\
&= Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}}) \\
&= Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{UM}}) - Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{RM}}) \\
&= Var(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) - Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{RM}}) \\
&= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{RM}}),
\end{aligned}
$$

where

$$
\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{RM}}) &= Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_1^{\text{UM}} + \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \hat{\boldsymbol{\beta}}_2^{\text{UM}}) \\
&= Var(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) + Cov(\hat{\boldsymbol{\beta}}_1^{\text{UM}}, \hat{\boldsymbol{\beta}}_2^{\text{UM}})[\boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}]^\top \\
&= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} + \gamma^2 \boldsymbol{Q}_{12} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{12} &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - \gamma^2 \boldsymbol{Q}_{12} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1} \\
&= -\gamma^2 \boldsymbol{Q}_{12} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1}
\end{aligned}
$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}^* &= \boldsymbol{\Omega}^* - \gamma^2 \boldsymbol{Q}_{11.2}^{-1} + \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{21} \\
&= \gamma^2 (\boldsymbol{Q}_{11}^{-1} - \boldsymbol{Q}_{11.2}^{-1} - \boldsymbol{Q}_{12} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1} - \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \boldsymbol{Q}_{21}).
\end{aligned}$$

D. Proof for Theorem 5.2

We present our proof as follows: Obviously, $\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) = 0$ and

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) &= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{RM}} - \boldsymbol{\beta}_1) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} + \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \hat{\boldsymbol{\beta}}_2^{\text{UM}} - \boldsymbol{\beta}_1) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \boldsymbol{\beta}_1) \right\} + E\left\{ \lim_{n\to\infty} \sqrt{n}(\boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \hat{\boldsymbol{\beta}}_2^{\text{UM}}) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n} \left(\boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \hat{\boldsymbol{\beta}}_2^{\text{UM}} \right) \right\} \\
&= \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12} \boldsymbol{\omega} \\
&= -\boldsymbol{\delta}.
\end{aligned}$$

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) &= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM}} - \boldsymbol{\beta}_1) \right\} \\
&= E\left\{ \lim_{n\to\infty} \left(\sqrt{n}\hat{\boldsymbol{\beta}}_1^{\text{SM}} - \sqrt{n}\boldsymbol{\beta}_1 \right) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})(-\kappa \psi_n^{-1}) \right\} \\
&= -\kappa E\left\{ \eta_2 \psi_n^{-1} \right\} \\
&= -\kappa(-\boldsymbol{\delta}) E\left\{ \chi_{p_2+2}^{-2}(\Delta) \right\} \\
&= \kappa \boldsymbol{\delta} E\left\{ \chi_{p_2+2}^{-2}(\Delta) \right\}.
\end{aligned}$$

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) &= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}} - \boldsymbol{\beta}_1) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}} - \boldsymbol{\beta}_1) - \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})(1 - \kappa \psi_n^{-1}) I(\psi_n < \kappa) \right\} \\
&= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM}} - \boldsymbol{\beta}_1) \right\} - E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})(1 - \kappa \psi_n^{-1}) I(\psi_n < \kappa) \right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - E\left\{ \eta_2 (1 - \kappa \psi_n^{-1}) I(\psi_n < \kappa) \right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - \boldsymbol{\delta} E\left\{ \left(1 - \kappa \chi_{p_2+2}^{-2}(\Delta^2) \right) I \left(\chi_{p_2+2}^2(\Delta^2) < \kappa \right) \right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - \boldsymbol{\delta} E\left\{ I \left(\chi_{p_2+2}^2(\Delta) \right) < \kappa \right\} - \boldsymbol{\delta} E\left\{ \kappa \chi_{p_2+2}^{-2}(\Delta) I \left(\chi_{p_2+2}^2(\Delta) < \kappa \right) \right\}
\end{aligned}$$

$$= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - \delta\left[H_{p_2+2}(\kappa, \Delta) - E\left\{\kappa\chi_{p_2+2}^{-2}(\Delta)I\left(\chi_{p_2+2}^2(\Delta) < \kappa\right)\right\}\right],$$

where $I()$ denotes an indicator function.

E. Derivation of Asymptotic Risk of the Estimators

Let us denote the ADMSE by $\boldsymbol{\Gamma}$, and then the expressions are listed as follows:

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) = \gamma^2 \boldsymbol{Q}_{11.2}^{-1}$$

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) = \gamma^2 \boldsymbol{Q}_{11}^{-1} + \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}^{\top} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1}$$

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) = \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - 2\kappa\left[E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \boldsymbol{\delta}\boldsymbol{\delta}^{\top} E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right.$$

$$\left. - \boldsymbol{\delta}\boldsymbol{\delta}^{\top} E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right]$$

$$+ \kappa^2\left[\boldsymbol{\Sigma}^* E(\chi_{p_2+2}^{-4}(\Delta)) + \boldsymbol{\delta}\boldsymbol{\delta}^{\top} E(\chi_{p_2+4}^{-4}(\Delta))\right].$$

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{SM}+}) = \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - 2\boldsymbol{\Sigma}_{21}E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$- 2\boldsymbol{\delta}\boldsymbol{\delta}^{\top}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}E(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}$$

$$+ 2\boldsymbol{\delta}\boldsymbol{\delta}^{\top} E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$+ \boldsymbol{\Sigma}^* E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))^2 I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$+ \boldsymbol{\delta}\boldsymbol{\delta}^{\top} E\left\{(1 - \chi_{p_2+4}^{-2}(\Delta))^2 I(\chi_{p_2+4}^2(\Delta) < \kappa)\right\}.$$

Proof

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) = E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \boldsymbol{\beta}_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \boldsymbol{\beta}_1)^{\top}\right\}$$

$$= E\{\eta_1\eta_1^{\top}\}$$

$$= \{Cov(\eta_1\eta_1^{\top}) + E(\eta_1)E(\eta_1)^{\top}\}$$

$$= Var(\eta_1)$$

$$= \gamma^2 \boldsymbol{Q}_{11.2}^{-1}.$$

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) = E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{RM}} - \boldsymbol{\beta}_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{RM}} - \boldsymbol{\beta}_1)^{\top}\right\}$$

$$= E\{\eta_3\eta_3^{\top}\}$$

$$= Cov(\eta_3, \eta_3^{\top}) + E(\eta_3)E(\eta_3)^{\top}$$

$$= Var(\eta_3) + E(\eta_3)E(\eta_3)^{\top}$$

$$= \gamma^2 \boldsymbol{Q}_{11}^{-1} + \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}^{\top} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1}.$$

$$
\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \beta_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \beta_1)^{\top}\right\} \\
&= E\left\{\lim_{n\to\infty} n\left[(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \beta_1) - (\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\kappa\psi_n^{-1}\right]\right. \\
&\qquad \left.\left[(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \beta_1) - (\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\kappa\psi_n^{-1}\right]^{\top}\right\} \\
&= E\left\{[\eta_1 - \eta_2\kappa\psi_n^{-1}][\eta_1 - \eta_2\kappa\psi_n^{-1}]^{\top}\right\} \\
&= E\left\{\eta_1\eta_1^{\top} - 2\kappa\psi_n^{-1}\eta_2\eta_1^{\top} + \kappa^2\psi_n^{-2}\eta_2\eta_2^{\top}\right\}. \quad\quad (A)
\end{aligned}
$$

Now

$$
\begin{aligned}
E\left\{\psi_n^{-1}\eta_2\eta_1^{\top}\right\} &= E\left\{E(\eta_2\eta_1^{\top}\psi_n^{-1}|\eta_2)\right\} = E\left\{\eta_2 E(\eta_1^{\top}\psi_n^{-1}|\eta_2)\right\} \\
&= E\left\{\eta_2\left[0 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{*-1}(\eta_2 - \delta)\right]^{\top}\psi_n^{-1}\right\} \\
&= E\left\{\eta_2(\eta_2 - \delta)^{\top}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{12}^{\top}\psi_n^{-1}\right\} \\
&= E\left\{\eta_2\eta_2^{\top}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\psi_n^{-1}\right\} - E\left\{\eta_2\delta^{\top}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\psi_n^{-1}\right\} \\
&= \left[Var(\eta_2)E(\chi_{p_2+2}^{-2}(\Delta)) + E(\eta_2)E(\eta_2)^{\top}E(\chi_{p_2+4}^{-2}(\Delta))\right]\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - E(\eta_2)\delta^{\top}E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&= \left[\boldsymbol{\Sigma}^{*}E(\chi_{p_2+2}^{-2}(\Delta)) + \delta\delta^{\top}E(\chi_{p_2+4}^{-2}(\Delta))\right]\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - \delta\delta^{\top}E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&= E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \delta\delta^{\top}E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - \delta\delta^{\top}E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}.
\end{aligned}
$$

By substituting $E\{\psi_n^{-1}\eta_2\eta_2^{\top}\}$ in (A), we get

$$
\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= E\{\eta_1\eta_1^{\top}\} - 2\kappa E\left\{\psi_n^{-1}\eta_2\eta_1^{\top}\right\} + \kappa E\left\{\psi_n^{-2}\eta_2\eta_2^{\top}\right\} \\
&= Var(\eta_1) - 2\kappa\left[E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \delta\delta^{\top}E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right. \\
&\quad \left. - \delta\delta^{\top}E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right] \\
&\quad + \kappa^2\left\{Var(\eta_2)E(\chi_{p_2+2}^{-4}(\Delta)) + E(\eta_2)E(\eta_2)^{\top})E(\chi_{p_2+4}^{-4}(\Delta))\right\} \\
&= \gamma^2 Q_{11.2}^{-1} - 2\kappa\left[E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \delta\delta^{\top}E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right. \\
&\quad \left. - \delta\delta^{\top}E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right] \\
&\quad + \kappa^2\left[\boldsymbol{\Sigma}^{*}E(\chi_{p_2+2}^{-4}(\Delta)) + \delta\delta^{\top}E(\chi_{p_2+4}^{-4}(\Delta))\right].
\end{aligned}
$$

$$\Gamma(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) = E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\text{SM+}} - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^{\text{SM+}} - \boldsymbol{\beta}_1)^\top\right\}$$

$$= \Gamma(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - 2E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})(\hat{\boldsymbol{\beta}}^{\text{UM}} - \boldsymbol{\beta}_1)^\top (1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$+ E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \hat{\boldsymbol{\beta}}_1^{\text{RM}})^\top (1 - \kappa\psi_n^{-1})^2 I(\psi_n < \kappa)\right\}$$

$$= \Gamma(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - 2E\left\{\eta_2\eta_1^\top (1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$+ E\left\{\eta_2\eta_2^\top (1 - \kappa\psi_n^{-1})^2 I(\psi_n < \kappa)\right\}. \tag{B}$$

By using the rule of conditional expectation, we obtain

$$E\left\{\eta_2\eta_2^\top (1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$= E\left[\eta_2 E\left\{\eta_1^\top (1 - \kappa\psi_n^{-1}I(\psi_n < \kappa)\right\} |\eta_2\right]$$

$$= E\left[\eta_2\left\{0 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{*-1}(\eta_2 - \boldsymbol{\delta})\right\}^\top (1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right]$$

$$= E\left\{\eta_2(\eta_2 - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}(1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$= E\left\{\eta_2\eta_2^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}(1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$- E\left\{\eta_2\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}(1 - \kappa\psi_n^{-1})I(\psi_n < \kappa)\right\}$$

$$= \left\{Var(\eta_2)E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right.$$

$$\left. + \boldsymbol{\delta}\boldsymbol{\delta}^\top E(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right\}$$

$$- \left\{\boldsymbol{\delta}\boldsymbol{\delta}^\top E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)\right\}.$$

Substituting the above in (B), we get

$$\Gamma(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) = \Gamma(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - 2\boldsymbol{\Sigma}_{21}E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$- 2\boldsymbol{\delta}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}E(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}$$

$$+ 2\boldsymbol{\delta}\boldsymbol{\delta}^\top E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$+ \boldsymbol{\Sigma}^* E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))^2 I(\chi_{p_2+2}^2(\Delta) < \kappa)$$

$$+ \boldsymbol{\delta}\boldsymbol{\delta}^\top E\left\{(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))^2 I(\chi_{p_2+4}^2(\Delta) < \kappa)\right\}.$$

Using the definition in (4.4), we have the ADQR expressions as follows:

$$R(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) = \text{tr}(\boldsymbol{W}\Gamma(\hat{\boldsymbol{\beta}}_1^{\text{UM}}))$$

$$= \text{tr}(\boldsymbol{W}\gamma^2\boldsymbol{Q}_{11.2}^{-1}),$$

$$R(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) = \text{tr}(\boldsymbol{W}\Gamma(\hat{\boldsymbol{\beta}}_1^{\text{RM}}))$$

$$= \text{tr}(\boldsymbol{W}\gamma^2 \boldsymbol{Q}_{11}^{-1}) + \text{tr}(\boldsymbol{W}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}^\top \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}),$$

$$R(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) = \text{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}))$$

$$= R(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) - 2\kappa E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\Sigma}_{21})$$

$$- 2\kappa E\left\{\chi_{p_2+4}^{-2}(\Delta)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21})$$

$$+ 2\kappa E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21})$$

$$+ \kappa^2 E\left\{\chi_{p_2+2}^{-4}(\Delta)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^*)$$

$$+ \kappa^2 E\left\{\chi_{p_2+4}^{-2}(\Delta)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}^\top),$$

$$R(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) = \text{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}))$$

$$= R(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - 2E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)\text{tr}(\boldsymbol{W}\boldsymbol{\Sigma}_{21})$$

$$- 2E(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\text{tr}(\boldsymbol{W}\boldsymbol{\delta}^\top\boldsymbol{\delta}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21})$$

$$+ 2E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\text{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}^\top)$$

$$+ E\left\{(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))^2 I(\chi_{p_2+2}^2(\Delta) < \kappa)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^*)$$

$$+ E\left\{(1 - \kappa\chi_{p_2+4}^{-2}(\Delta))^2 I(\chi_{p_2+4}^2(\Delta) < \kappa)\right\}\text{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}^\top).$$

References

Ahmed, S. E. (2014). *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*. New York, USA: Springer.

Ahmed, S. E., Ahmed, F., & Yüzbaşı, B. (2023). *Post-Shrinkage Strategies in Statistical and Machine Learning for High Dimensional Data*. Boca Raton, USA: CRC Press.

Ahmed, S. E., Doksum, K. A., Hossain, S., & You, J. (2007). Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian & New Zealand Journal of Statistics, 49*, 435–454.

Ahmed, S. E., & Fallahpour, S. (2012). Shrinkage estimation strategy in quasi-likelihood models. *Statistics & Probability Letters, 82*(12), 2170–2179.

Ahmed, S. E., Hussein, A. A., & Sen, P. K. (2006). Risk comparison of some shrinkage M-estimators in linear models. *Nonparametric Statistics, 18*, 401–415.

Ahmed, S. E., & Raheem, S. M. E. (2012). Shrinkage and absolute penalty estimation in linear regression models. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(6), 541–553.

Ahmed, S. E., & Yüzbaşı, B. (2016). Big data analytics: integrating penalty strategies. *International Journal of Management Science and Engineering Management, 11*(2), 105–115.

Arashi, M., Kibria, B. G., Norouzirad, M., & Nadarajah, S. (2014). Improved preliminary test and Stein-rule Liu estimators for the ill-conditioned elliptical linear regression model. *Journal of Multivariate Analysis, 126*, 53–74.

Bianco, A., & Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *Journal of Statistical Planning and Inference, 122*(1–2), 229–252.

Denby, L. (1986). Smooth regression functions. Statistical Research Report, 26. AT&T Bell Laboratories, Murray Hill.

Farcomeni, A., & Ventura, L. (2012). An overview of robust methods in medical research. *Statistical Methods in Medical Research, 21*(2), 111–133.

Jurečková, J., & Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. New York: Wiley.

Ma, T., Liu, S., & Ahmed, S. E. (2014). Shrinkage estimation for the mean of the inverse Gaussian population. *Metrika, 77*, 733–752.

Maruyama, Y., Kubokawa, T., & Strawderman, W. E. (2023). *Stein Estimation*. Singapore: Springer.

Norouzirad, M., & Arashi, M. (2019). Preliminary test and Stein-type shrinkage ridge estimators in robust regression. *Statistical Papers, 60*, 1849–1882.

Opoku, E. A., Ahmed, S. E., & Nathoo, F. S. (2021). Sparse estimation strategies in linear mixed effect models for high-dimensional data application. *Entropy, 23*(10), 1348.

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

Rather, K. U. I., Koçyiğit, E. G., Onyango, R., & Kadilar, C. (2022). Improved regression in ratio type estimators based on robust M-estimation. *PLoS ONE, 17*(12), e0278868.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica, 56*, 931–954.

Shih, J. H., Lin, T. Y., Jimichi, M., & Emura, T. (2021). Robust ridge M-estimators with pretest and Stein-rule shrinkage for an intercept term. *Japanese Journal of Statistics and Data Science, 4*, 107–150.

Speckman, P. (1988). Kernel smoothing in partial linear models. Journal of the Royal Statistical Society. *Series. B, 50*, 413–437.

Sun, R., Ma, T., & Liu, S. (2020). Portfolio selection: shrinking the time-varying inverse conditional covariance matrix. *Statistical Papers, 61*, 2583–2604.

Susanti, Y., Qona'ah, N., Ferawati, K., & Qumillaila, C. (2020). Prediction modeling of annual parasite incidence (API) of Malaria in Indonesia using Robust regression of M-estimation and S-estimation. *AIP Conference Proceedings, 2296*, 020100. https://doi.org/10.1063/5.0037417

Zhou, P., Xie, J., Li, W., Wang, H., & Chai, T. (2020). Robust neural networks with random weights based on generalized M-estimation and PLS for imperfect industrial data modeling. *Control Engineering Practice, 105*, 104633.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.