Inference for biased transformation models*

Xuehu Zhu^{1,2}, Tao Wang³, Junlong Zhao⁴, and Lixing Zhu^{4,5,*}

¹ School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

² Qilu Securities Institute for Financial Studies, Shandong University, China

³ Department of Biostatistics, Shanghai Jiaotong University, Shanghai, China

⁴ School of Statistics, Beijing Normal University, Beijing, China

⁵ Department of Mathematics, Hong Kong Baptist University, Hong Kong

Abstract

Working regression models are often parsimonious for practical use and however may be biased. This is because either some strong signals to the response are not included in working models or too many weak signals are excluded in the modelling stage, which make cumulative bias. Thus, estimating consistently the parameters of interest in biased working models is then a challenge. This paper investigates the estimation problem for linear transformation models with three aims. First, to identify strong signals in the original full models, a sufficient dimension reduction approach is applied to transferring linear transformation models to pro forma linear models. This method can efficiently avoid high-dimensional nonparametric estimation for the unknown model transformation. Second, after identifying strong signals, a semiparametric re-modeling with some artificially constructed predictors is performed to correct model bias in working models. The construction procedure is introduced and a ridge ratio estimation is proposed to determine the

*Corresponding author. E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

number of these predictors. Third, root- n consistent estimators of the parameters in working models are defined and the asymptotic normality is proved. The performance of the new method is illustrated through simulation studies and a real data analysis.

Key words: Estimation consistency, Linear transformation models, Model bias correction, Non-sparse structure.

Running head: Inference for biased models.

1 Introduction

Consider the linear transformation model:

$$H(Y) = X^\top \beta + \epsilon, \tag{1}$$

where Y is the response variable, $H(\cdot)$ denotes an unknown monotone transformation function, $X = (X_1, X_2, \dots, X_p)^\top$ is the p -dimensional predictor vector having mean zero, $\beta = (\beta_1, \beta_2, \dots, \beta_p)^\top$ is the regression parameter vector of interest, and the error ϵ , which is independent of X , has a continuous distribution. As H is unknown, without loss of generality, we assume that the L_2 norm of β is one. This model, which circumvents the so-called curse of dimensionality, is widely used in statistical analysis.

In practical use, it is often not to contain all the predictors in a working model as when there are many predictors, we often remove those ‘weak’ signals for the response such that a parsimonious working model can be built. This is particularly the case in high-dimensional paradigms because to well estimate the parameter of interest β , variable selection is indispensable and hopefully only those ‘strong’ signals are retained in the working model. Here ‘strong’ signals are reflected by the magnitudes of the corresponding coefficients β_i . To have estimation consistency, model sparsity is an usual assumption. That is, a small portion of the components of β are assumed to be significantly non-zero and the others are zero. However, this is often hard to judge. In practice, it is often that many weak signals with small coefficients, but non-zero coefficients in the model are removed out from the regression function. When the number of these predictors is large, the cumulative bias of working model is often not negligible. Further the selection process would also unwittingly rule out some strong signals. All these cause the difficulty of having estimation consistency. Relevant references includes Leeb and Pötscher (2005) and

Zhang and Huang (2008) who have observed similar problems. For semiparametric or nonparametric modelling, parsimonious modelling is even more important, otherwise, consistent estimation is harder to achieve because nonparametric estimation would be involved. The linear transformation model is an example as the function $H(\cdot)$ is nonparametric.

Lin et al. (2016) investigated this issue for linear models. They proposed a bias correction method to achieve estimation consistency via a semiparametric remodeling. Their method highly relies on the linear model structure, and the two methods proposed by Lin et al. (2016), which select quasi-instrumental variables for the semiparametric remodeling, would be complex to implement and theoretically do not have the consistency of the estimated number of quasi-instrumental variables. For the linear transformation model under study, the problem is more difficult. First, as the model transformation function H is unknown and when X is high-dimensional, it is not a wise idea to use a nonparametric method to estimate the transformation function H and then the parameter of interest. To have an estimation procedure without estimating the transformation function H , our idea used in this paper is to apply a sufficient dimension reduction approach to transferring the original model to an equivalent pro forma ‘linear’ model. With this new formulation, estimating the transformation function H is completely averted. The key for this method to work is the equivalence of strong and weak signals between the original nonparametric transformation and the pro forma linear model. This will be confirmed in the next section. Therefore, for bias correction, we will work on the pro forma linear model and then, if necessary, establish a new transformation (maybe still nonparametric) to fit data. It is worth mentioning that unlike the classical linear model, given the selected predictors, the conditional expectation of the error in this pro forma ‘linear’ model is not centered, while it just has a zero expectation of the product of the error and the predictors. An idea in spirit similarly as that of Lin

et al. (2016) is used to construct some ‘new’ predictors such that a semiparametric re-modelling is possible to remove bias. However, as the linearity is just pro forma, the bias correction is more delicate than that in Lin et al. (2016). The detail will be elaborated in Section 3. Second, we suggest a ridge ratio estimation (RR) to estimate the number of those constructed predictors. The consistency will be proved.

We now formulate the model. Assume for simplicity that the first q predictors, represented by $\mathcal{I} = \{1, 2, \dots, q\}$, are relatively strong signals, and the last $p - q$ predictors are weak signals but not necessarily irrelevant to the response. We rewrite X and β as $X = (Z^\top, U^\top)^\top$ and $\beta = (\beta_{\mathcal{I}}^\top, \beta_{\mathcal{I}^c}^\top)^\top$, respectively, where $Z = (X_1, \dots, X_q)^\top$, $U = (X_{q+1}, \dots, X_p)^\top$ and $\mathcal{I}^c$ is the compliment of $\mathcal{I}$ in the complete set $\{1, \dots, p\}$. Correspondingly, the full model is rewritten as

$$H(Y) = Z^\top \beta_{\mathcal{I}} + \tilde{\epsilon}, \quad (2)$$

where $\tilde{\epsilon} = U^\top \beta_{\mathcal{I}^c} + \epsilon$ denotes the error term in this working model. A typical example is a working model that is based on a variable selection procedure such as the LASSO (Tibshirani, 1996). As commented above, when some strong signals are not included in the regression part or when the full model (1) is not sparse, the working model (2) could be biased:

$$E(\tilde{\epsilon}|Z) = \sum_{j \notin \mathcal{I}} \beta_j E(X_j|Z) \neq 0. \quad (3)$$

Consequently, it is of importance to obtain a consistent estimate of $\beta_{\mathcal{I}}$ or more precisely the normalized $\beta_{\mathcal{I}}$ as in the nonparametric transformation model, the scale of $\beta_{\mathcal{I}}$ is not identifiable. Without notational confusion, we still use $\beta_{\mathcal{I}}$ as its normalized one. The goal of the present paper is to construct a bias-corrected model and to obtain a root- n consistent estimate of $\beta_{\mathcal{I}}$. Our idea is to first transfer the linear transformation model to an equivalent pro forma ‘linear’ model such that

the modelling becomes easier without involving nonparametric estimation for the model transformation H . After that, we introduce a new method to conveniently determine the number of “new” predictors for the purpose of bias correction. The semiparametric re-modelling can then be implemented for the purpose of estimation consistency.

The rest of this paper is organized as follows. In Section 2, we transfer the original model as an equivalent pro forma linear model. We then give the definition of the index set of relatively strong signals in a non-sparse transformation model and the identifiability of the index set. In Section 3, after introducing a vector of ‘new’ predictors, a partially linear remodeling is implemented in which the parameters of interest are in the linear component when both the response and regression function are re-centered. The ridge ratio method to determine the number of ‘new’ predictors is also suggested in this section and the asymptotic properties are derived. It is noteworthy that the identifiability of the index set is mainly for theoretical purpose such that the selected model is non-random. For practical use, it is not necessary. This can be seen in Section 3. Numerical simulations are conducted to examine the performance of the new method in Section 4. All the proofs are postponed to the appendix.

2 Working model construction

2.1 An equivalent pro forma linear model

As the transformation function H is unknown, directly implementing variable selection procedure would involve nonparametric estimation for this function. This could again suffer from the dimensionality difficulty and computational instability.

Thus, we suggest the following approach to transform the model to be a pro forma ‘linear’ model.

Let $\Sigma_X = \text{Cov}(X)$, $T = \Sigma_X^{-1/2}X$ and $\alpha = \Sigma_X^{1/2}\beta$. Then we have $\beta^\top X = \alpha^\top T$. Let A_1 denote a $p \times (p-1)$ matrix such that $(A_1, \alpha/|\alpha|)$ is an orthogonal matrix. Feng et al. (2014) showed that under a mild regularity condition, we can get that for any function $h(\cdot)$ of the response Y , there exists some constant κ_h satisfying

$$\Sigma_X^{-1} \text{Cov}\{X, h(Y)\} = \kappa_h \beta. \quad (4)$$

When X is elliptically symmetric, the designed condition holds. Recently, Guan et al. (2016) found that this can be achieved for some mixture multivariate skew-elliptical distributions. Write $\gamma = \kappa_h \beta = (\gamma_1, \dots, \gamma_p)^\top$. As the transformation function H is unknown, we can assume that the L_2 -norm of β is one. Note that γ is proportional to β . Without loss of generality, we also assume that these two parameters are the same, otherwise, we can use the normalized parameter vector $\gamma/\|\gamma\|$. For any given function $h(\cdot)$, we arrive at a model that resembles the transformation model assuming $E\{h(Y)\} = 0$ as:

$$h(Y) = X^\top \gamma + e, \quad (5)$$

where $e = h(Y) - X^\top \gamma$ such that $E(eX) = 0$.

Note that this pro forma ‘linear’ model does not guarantee $E(e|X) = 0$, while it only has $E(eX) = 0$. Technically we need some more sophisticated skills to handle the estimation consistency. In Section 3, we will have a more detailed study for this issue. Regardless of this difficulty, its main advantage is that the model is free of the model transformation $H(\cdot)$ with no need of nonparametric estimation. Further, because γ is proportional to β and $h(\cdot)$ is given by user, we can simply use a penalized least squares method, e.g., the LASSO (Tibshirani, 1996), to select

the predictors that correspond to the strong signals in model (1). Specifically, the LASSO estimate is defined as

$$\hat{\gamma} = \arg \min_{\gamma} \left\{ \sum_{i=1}^n \{h(y_i) - x_i^\top \gamma\}^2 + \lambda \sum_{j=1}^p |\gamma_j| \right\}, \quad (6)$$

where $\{(x_i, y_i), i = 1, \dots, n\}$ is a random sample on (X, Y) and $\lambda = \lambda_n$ is a tuning parameter. See Wang et al. (2012) and Wang et al. (2015) for more details. In the sparse case and under certain regularity conditions, one can follow Zhang and Huang (2008) to show that, with a large probability, the LASSO estimate satisfies that

$$\sum_{j \in \mathcal{I}} |\hat{\gamma}_j - \gamma_j|^2 \leq M \frac{\log(p)s}{n},$$

where $\mathcal{I} = \{j : \gamma_j \neq 0\}$, the number s is the number of nonzero coefficients, and M is a positive constant.

To improve the performance of the LASSO, we can also adopt the post-LASSO estimation suggested by Belloni and Chernozhukov (2013): using the LASSO to estimate the index set $\mathcal{I}$ by $\hat{\mathcal{I}}$, and then defining the least squares estimate as

$$\hat{\gamma}_{\hat{\mathcal{I}}} = (\mathbf{X}_{\hat{\mathcal{I}}}^\top \mathbf{X}_{\hat{\mathcal{I}}})^{-1} (\mathbf{X}_{\hat{\mathcal{I}}})^\top h(y),$$

where $\mathbf{X}_{\hat{\mathcal{I}}}$ is the sub-matrix of $\mathbf{X} = (x_1, \dots, x_n)^\top$ with its column vectors according to $\hat{\mathcal{I}}$ and $h(y) = \{h(y_1), \dots, h(y_n)\}^\top$. In our notation, $\mathbf{X}_{\hat{\mathcal{I}}}$ is just $\mathbf{Z}$ with the index set $\hat{\mathcal{I}}$.

From the above description, we can see that after efficiently estimating γ and selecting the vector $\gamma_{\mathcal{I}} = (\gamma_1, \dots, \gamma_q)^\top$ that is associated with the strong signals, the original model becomes: for some constant c

$$H(Y) = cZ^\top \gamma_{\mathcal{I}} + \tilde{\epsilon}.$$

Again, as H is unknown, we can assume that $c = 1$ without loss of generality. Thus, this paper mainly focuses on identifying and estimating $\gamma_{\mathcal{I}}$. As described before, it is not possible to directly work on this model as $\tilde{\epsilon}$ would be biased and H is unknown. We shall work on the pro forma linear model (5) to identify $\gamma_{\mathcal{I}}$. Note that when both β and γ are normalized, we can simply consider $\beta = \gamma$. Thus, without notational confusion, we will still use $\beta_{\mathcal{I}}$ for $\gamma_{\mathcal{I}}$ throughout the rest of the paper.

Clearly, it is important to identify a post-LASSO selection-based working model. When the model is sparse, it has been intensively discussed in the literature. In this section, we only offer a very brief review on it. Under the non-sparse case that is considered in this paper, without notational confusion, we still use $\mathcal{I}$ denote the index set of relatively strong signals as $\mathcal{I} = \{1 \leq j \leq p, |\beta_j| \geq \tau_n\}$ for some positive constant τ_n . The predictors according to these indices are regarded as relatively strong signals. For $\beta_j \neq 0$ with $j \notin \mathcal{I}$, the corresponding predictors are regarded as weak signals comparatively. Assume that $\hat{\mathcal{I}} = \mathcal{I}$ as we will prove $\hat{\mathcal{I}} = \mathcal{I}$ with a probability going to one. With the aforementioned notation, the post-LASSO selection-based working model is then given by

$$h(Y) = Z^{\top} \beta_{\mathcal{I}} + \tilde{\epsilon}, \quad (7)$$

where $\tilde{\epsilon}$ satisfies

$$E(\tilde{\epsilon}|Z) = \sum_{j=q+1}^p E(X_j|Z)\beta_j.$$

Here, the given transformation function h selected by user is not the model transformation function H . As Feng et al. (2014) pointed out, the empirical distribution function $\hat{F}(\cdot)$ of Y can be a good choice. Thus, if Z and U are correlated, the post-selection-based model (7) tends to be biased. Consequently, the least squares post-LASSO estimate is not possible to be consistent to $\beta_{\mathcal{I}}$.

2.2 Identification of $\beta_{\mathcal{I}}$ when the model is not sparse

For sparse models, we can distinguish between strong and weak signals through zero and non-zero coefficients. However, this principle cannot be applied to nonsparse models as coefficients are not exactly zero even when the related predictors do not affect the response significantly. Thus, what we can distinguish is between ‘relatively strong’ and ‘relatively weak’ signals with relatively large and small coefficients. Without confusion, we will call them strong signals and weak signals throughout the rest of this paper. Our strategy is to use their convergence rates to separate them. Actually, this is also the case in practice. As we are using the LASSO to estimate the coefficients, we need to consider for what magnitude of the coefficients the LASSO works to pick out strong signals. Thus, the conditions and the definition of the index set $\mathcal{I}$ of strong signals are assumed and defined in non-sparse models. Afterwards, we prove the consistency that $\hat{\mathcal{I}} = \mathcal{I}$ with a probability going to one.

Introduce some notations first. Write $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p) \in R^{n \times p}$ with $\mathbf{X}_j$ being its j -th column. Without loss of generality, assume that each $\mathbf{X}_j$ is centered. Then the sample covariance matrix $\mathbf{C} = n^{-1} \mathbf{X}^\top \mathbf{X}$ can be expressed in a block-wise form as follows:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix},$$

where $\mathbf{C}_{ij} = n^{-1} \mathbf{X}_{(i)}^\top \mathbf{X}_{(j)}$ for $i, j = 1, 2$, $\mathbf{X}_{(1)} = \mathbf{Z} = (\mathbf{X}_1, \dots, \mathbf{X}_q)$ and $\mathbf{X}_{(2)} = \mathbf{U} = (\mathbf{X}_{q+1}, \dots, \mathbf{X}_p)$.

Assume that q is fixed. Let $\beta = (\beta_{\mathcal{I}}^\top, \beta_{\mathcal{I}^c}^\top)^\top \in R^p$, where $\beta_{\mathcal{I}} = (\beta_1, \dots, \beta_q)^\top \in R^q$ and $\beta_{\mathcal{I}^c} = (\beta_{q+1}, \dots, \beta_p)^\top \in R^{p-q}$ that are associated with relatively strong and weak signals respectively. Let $\tau_n = \min\{|\beta_j|, 1 \leq j \leq q\}$. To correctly select strong signals in non-sparse model, we assume the following conditions.

- C1. Assume that with a probability tending to 1, there exist some positive constants M_1 and M_2 such that $\frac{1}{n}\mathbf{X}_j^\top \mathbf{X}_j < M_1$ for $1 \leq j \leq p$ and $M_2^{-1} < a^\top \mathbf{C}_{11} a \leq M_2$ for any $\|a\|_2 = 1$;
- C2. Assume that the strong irrepresentable condition (Zhao and Yu, 2006) holds with a probability tending to 1. That is, with a probability tending to 1, there exists a positive constant vector c satisfying $|\mathbf{C}_{21}(\mathbf{C}_{11})^{-1}\text{sign}(\beta_{\mathcal{I}})| \leq \mathbf{1}_{p-q} - c$, where $\mathbf{1}_{p-q}$ is a $p - q$ vector of ones and the inequality holds element-wise.
- C3. $\|\beta_{\mathcal{I}^c}\|_1 \ll \lambda_n/n \ll \tau_n$, $\sqrt{n}\tau_n \rightarrow \infty$ and $\lambda_n/\sqrt{n} \rightarrow \infty$.

Condition C3 is critical for the success of the LASSO in non-sparse models, imposing the requirement on the strength of ‘strong signals’ and tuning parameter, such that the strong signals can be correctly identified. In Condition C3, we do not impose concrete order on λ_n/n . Therefore, λ_n/n can be large and consequently, $\|\beta_{\mathcal{I}^c}\|_1$ can be large. The key requirement is that τ_n should be larger than $\max\{\|\beta_{\mathcal{I}^c}\|_1, n^{-1/2}, \lambda_n/n\}$.

Condition C1 is commonly used in the literature. Condition C2 is the Strong irrepresentable condition for random design. The following proposition describes that Conditions C1 and C2 are reasonable. Assume that $X \sim N(0, \Sigma_X)$, where $\Sigma_X = (\sigma_{ij})$. Denote the largest and the smallest eigenvalues of Σ_X as $\lambda_{\max}(\Sigma_X)$ and $\lambda_{\min}(\Sigma_X)$, respectively.

Proposition 1. We have the following conclusions.

- (i) Assume that $0 < c_1^{-1} < \lambda_{\min}(\Sigma_X) \leq \lambda_{\max}(\Sigma_X) < c_1 < \infty$ for some constant $c_1 > 0$. If $\log(p)/n = o(1)$, Condition C1 holds.
- (ii) Under the condition of (i), assume further that $\max_{1 \leq i \neq j \leq p} \sigma_{ij} < (\min_{1 \leq i \leq p} \sigma_{ii})c_2/(2q -$

1) for some constant $0 < c_2 < 1$. Then Condition C2 holds with a probability tending to 1.

The following theorem confirms the expectation for the case in this paper that, with a probability approaching to one, the strong signals with large absolute values of coefficients are selected into the post-LASSO selection-based working model (7). Thus, in an asymptotic sense it is safe to regard the model (7) as a non-random model.

Theorem 1 *Assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent and identically distributed with $E(e^{2k}) < \infty$ for some positive integer k . Under the identifiability conditions C1-C3 and $p = o\{(\lambda_n/\sqrt{n})^{2k}\}$, $P(\hat{\mathcal{I}} = \mathcal{I}) \rightarrow 1$.*

3 Re-modeling and inference

3.1 Bias-correction

Because of the remodelling in (5), the model becomes linear and the methodology in Lin et al. (2016) can be employed although the linearity is just pro forma. Note that $P(\hat{\mathcal{I}} = \mathcal{I}) \rightarrow 1$. We then without loss of generality suppose

$$\Sigma_{Z,Z} = \text{Cov}(Z, Z) = I_{q \times q}, \quad (8)$$

where $I_{q \times q}$ is a $q \times q$ identity matrix. Since the model (7) is a low-dimensional transformation model with the predictors Z , the standardization on Z is feasible and does not affect the theoretical result. To correct the model bias, we define a new predictor vector V that is from the original predictors and is specified later. Let $\eta = \sum_{j=q+1}^p X_j \beta_j$, $g(V) = E(\eta|V)$ and $\xi(V) = \eta - g(V)$. Then, we can rewrite

the model (5) as

$$h(Y) = Z^\top \beta_{\mathcal{I}} + g(V) + \xi(V) + e. \quad (9)$$

If we can find a suitable V such that $E\{\xi(V)|Z, V\} = 0$, the model (5) is reduced to a partially linear model with linear component $Z^\top \beta_{\mathcal{I}}$ and nonparametric component $g(V)$, respectively.

3.2 Determination of V

1. Determination at the population level. As the technique is very similar to that in Lin et al. (2016), we only present a brief description in this subsection. First construct V by an affine transformation of the predictors. Let $U^* = (U_{k_1}, \dots, U_{k_d})^\top$ be d random variables selected from U , where $\|\text{corr}(U_{k_1}Z)\| \geq \dots \geq \|\text{corr}(U_{k_d}Z)\| > 0$ for some integer $d \geq 1$. Define

$$\tilde{Z} = (Z^\top, \tilde{U}^\top)^\top,$$

where $\tilde{U} = (\Sigma_{U^*, U^*} - \Sigma_{U^*, Z} \Sigma_{Z, U^*})^{-1/2} (U^* - \Sigma_{U^*, Z} Z)$, $\Sigma_{U^*, U^*} = \text{Cov}(U^*)$ and $\Sigma_{U^*, Z} = \text{Cov}(U^*, Z)$. Then one can easily see that $\tilde{U}$ and Z are statistically uncorrelated and that the covariance matrix of $\tilde{Z}$ is an identity matrix. We define $r \times (q + d)$ matrix with $r < q + d$ such that the constructed variable V has the formula as $V = A\tilde{Z}$.

To determine the favorable A , assume the following linearity condition:

$$E(U|C^\top \tilde{Z}) = E(U) + \Sigma_{U, \tilde{Z}} C (C^\top \Sigma_{\tilde{Z}, \tilde{Z}} C)^{-1} C^\top \{\tilde{Z} - E(\tilde{Z})\}, \quad (10)$$

for any given matrix C . This condition is well-known in the context of sufficient dimension reduction. Li (1991) pointed out that it is satisfied when X follows an elliptically contoured distribution. On the other hand, we note that this condition is not strong because Hall and Li (1993) showed that it holds approximately when the dimension p is high in the sense that $p \rightarrow \infty$ as $n \rightarrow \infty$. Thus, the linearity

condition is typically regarded as mild. Recently, Guan et al. (2016) proved that for some skewed mixture distributions, existing methods could also be available.

As Lin et al (2016) did, consider the spectral decomposition that leads to $\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}} = Q\Lambda Q^\top$, where Q is a $(q+d) \times (q+d)$ orthogonal matrix and Λ is a diagonal matrix whose first r diagonal elements are positive while the rest are zero. Let $Q = (Q_1, Q_2)$ and $Q_1^\top = (Q_{11}^\top, Q_{12}^\top)^\top$, where Q_1 is a $(q+d) \times r$ matrix and Q_{11} is a $q \times r$ matrix. The following proposition shows that when $A = Q_1^\top$, the model (9) is an unbiased semi-parametric model.

Proposition 2 *Under the linearity condition (10), assume that the eigenvalues of $Q_1^\top Q_1$ are not equal to 1 and $\Sigma_{Z,Z} = I_q$. If $A = Q_1^\top$, then model (9) satisfies $E\{\xi(V)|Z, V\} = 0$.*

In order to ensure that the solution of (17) is unique, we assume that the predictors V and U^* satisfy $E[\{Z - E(Z|V)\}\{Z - E(Z|V)\}^\top] > 0$. This condition is mild because V is a linear combination of Z and U^* . The justification of this proposition is described in the Appendix.

2. *Determination at the sample level.* Define an estimate of the matrix $\Sigma_{U,\tilde{Z}}^\top \Sigma_{U,\tilde{Z}}$ as

$$\widehat{\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}}} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \tilde{Z}_i U_i^\top U_j \tilde{Z}_j^\top.$$

For linear models, Lin et al. (2016) proved that if

$$E(U_k \tilde{Z}_j) \leq c, \quad 1 \leq k \leq p-q, 1 \leq j \leq q+d, \quad (11)$$

for some constant c , the estimate is consistent:

$$\frac{1}{p-q} \widehat{\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}}} - \frac{1}{p-q} \Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}} = O_p(n^{-1/2}). \quad (12)$$

This result can be similarly proved in our case and thus, the detail is omitted. Similarly as the case at the population level, the estimate $\frac{1}{p-q} \widehat{\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}}}$ can accordingly

have a spectral decomposition as $\hat{Q}\hat{\Lambda}\hat{Q}^\top$, where $\hat{Q}$ is a $(q+d) \times (q+d)$ orthogonal matrix, and $\hat{\Lambda}$ is a diagonal matrix whose first r diagonal elements are positive while the others are close to zero asymptotically. Let $\hat{Q} = (\hat{Q}_1, \hat{Q}_2)$ and $\hat{Q}_1^\top = (\hat{Q}_{11}^\top, \hat{Q}_{12}^\top)^\top$, where $\hat{Q}_1$ is a $(q+d) \times r$ matrix and $\hat{Q}_{11}$ is a $q \times r$ matrix. Then, we arrive at a sample version of V as

$$\hat{V} = \hat{Q}_1^\top \tilde{Z}. \quad (13)$$

3.3 Estimating the dimension r of V

It is obvious that the dimension r of V is unknown and needs to be estimated. Lin et al. (2016) advised a data-adaptive criterion that is not easy to implement and its estimation consistency is still in question. In this paper, we suggest a ridge ratio estimation as follows. Recall that $\frac{1}{p-q} \widehat{\Sigma_{\tilde{Z}, U}^\top \Sigma_{U, \tilde{Z}}} = \hat{Q}\hat{\Lambda}\hat{Q}^\top$, where $\hat{Q}$ is a $(q+d) \times (q+d)$ orthogonal matrix and $\hat{\Lambda}$ is a diagonal matrix. Write $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_1, \dots, \hat{\lambda}_{q+d})$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_{q+d} \geq 0$. The estimate is defined as:

$$\hat{r} = \arg \min_{1 \leq j \leq q+d-1} \left\{ \frac{\hat{\lambda}_{j+1} + c_n}{\hat{\lambda}_j + c_n} \right\}. \quad (14)$$

The rationale of this estimation is from the following justification. Suppose that there is a number K such that $\lambda_K > 0$ and for $j > K$, $\lambda_K = 0$. Thus, $\lambda_{j+1}/\lambda_j > 0$ for $j < K$, $\lambda_{K+1}/\lambda_K = 0$, and $\lambda_{j+1}/\lambda_j = 0/0$. To make the ratio be well defined, we consider the ratios with a ridge value $c > 0$: $(\lambda_{j+1} + c)/(\lambda_j + c)$. It is clear that K is the minimizer of the ratios. When all λ_j 's are replaced by the corresponding estimates, we need to have a sequence of ridge values c_n that converges to zero at a certain rate. The algorithm is motivated by Xia et al. (2015), Zhu et al. (2016a) and Zhu et al. (2016b) in different settings. The following proposition states the consistency.

Proposition 3 *Under the conditions in Theorem 1 and $\frac{c \log n}{\sqrt{n}} \leq c_n \rightarrow 0$ for some*

fixed constant $c > 0$, as $n \rightarrow \infty$, $\hat{r} = r$ with a probability going to one.

This proposition shows that the range of c_n is rather wide to ensure the consistency. In the simulations, we suggest a value of c_n as $5 \log n / \sqrt{n}$. The justification of this proposition is given in the appendix.

3.4 Estimation

Under the selection consistency of $\mathcal{I}$, we have shown that the working model (5) can be reformulated into a partially linear model as:

$$h(Y) = Z^\top \beta_{\mathcal{I}} + g(V) + \xi_1, \quad (15)$$

where $\xi_1 = \xi(V) + e$. Following the argument right below of (5), we still only have

$$E\{\xi_1 Z\} = E\{\xi(V)Z\} + E(eZ) = E[Z E\{\xi(V)|Z, V\}] = 0,$$

rather than $E\{\xi_1|Z\} = 0$ because, as stated before, $E\{e|Z\}$ may not be equal to 0. This is different from the case in Lin et al. (2016). To solve this problem such that we can well estimate the parameter $\beta_{\mathcal{I}}$ of interest, we consider the conditional expectation given V in both sides of the model (15):

$$E\{h(Y)|V\} = \{E(Z|V)\}^\top \beta_{\mathcal{I}} + g(V) + E\{\xi_1|V\}. \quad (16)$$

We then have

$$h(Y) - E\{h(Y)|V\} = \{Z - E(Z|V)\}^\top \beta_{\mathcal{I}} + \xi_2, \quad (17)$$

where $\xi_2 = \xi_1 - E\{\xi_1|V\}$ and $E\{\xi_2|V\} = 0$. This centralization step allows us to establish a semiparametric model like the classical one and then the estimation can be easier to implement.

When V is replaced by $\hat{V}$ and the relevant conditional expectations are estimated nonparametrically, we can have an approximate model as

$$\hat{h}(y_i) \approx \hat{z}_i^\top \hat{\beta}_{\hat{\mathcal{I}}} + \hat{\xi}_{2i}, \quad (18)$$

where

$$\begin{aligned}\hat{h}(y_i) &= h(y_i) - \frac{\sum_{k=1}^n h(y_i) L_H(\hat{v}_k - \hat{v}_i)}{\sum_{k=1}^n L_H(\hat{v}_k - \hat{v}_i)}, \\ \hat{z}_i &= z_i - \frac{\sum_{k=1}^n z_k L_H(\hat{v}_k - \hat{v}_i)}{\sum_{k=1}^n L_H(\hat{v}_k - \hat{v}_i)}, \\ \hat{\xi}_{2i} &= \xi_{1i} - \frac{\sum_{k=1}^n \xi_{1k} L_H(\hat{v}_k - \hat{v}_i)}{\sum_{k=1}^n L_H(\hat{v}_k - \hat{v}_i)},\end{aligned}$$

and L_H is a r -dimensional product kernel function in this paper:

$$L_H(u_1, u_2, \dots, u_r) = \frac{1}{h^r} K(u_1) K(u_2) \cdots K(u_r).$$

Thus, for the bias-corrected working model (17), an estimate of $\beta_{\mathcal{I}}$ is given by

$$\hat{\beta}_{\hat{\mathcal{I}}} = \hat{S}^{-1} \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{\top} \hat{h}(y_i), \quad (19)$$

where $\hat{S} = \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{\top} \hat{z}_i$.

In summary, the proposed estimation algorithm for the non-sparse transformation model has two main steps.

- Step 1. For a pre-determined function h , use the LASSO to estimate the set of relatively strong signals $\mathcal{I}$ by $\hat{\mathcal{I}}$. Practically, as argued in Feng et al. (2014), we often use the empirical distribution of Y , which can have a transformation-free property. The readers can refer to their paper.
- Step 2. Construct a vector V and build the bias-corrected working model (15), where the dimension of V is estimated by the ridge ratio estimate $\hat{r}$ defined in (14). Then $\beta_{\mathcal{I}}$ is estimated by (19).

3.5 Asymptotics

In order to study the asymptotic properties of the new estimate, some regularity conditions are assumed as follows.

- B1. The function $g(\cdot)$ is twice continuously differentiable;
- B2. The kernel function $K(\cdot)$ satisfies $\int K(u)du = 1$, $\int u^j K(u)du = 0$ for $j = 1, 2, \dots, m-1$ and $\int u^m K(u)du < \infty$;
- B3. The bandwidth $h = O_p(n^{-1/(2m+r)})$;
- B4. The constraint moment condition (11) holds.

It is noteworthy that Conditions B1-B3 are commonly used for estimating semi-parametric models, and Condition B4 is to guarantee the estimation consistency of the relative estimates in the above subsection. The result is stated in the following.

Theorem 2 *Under Conditions B1-B4 and the conditions in Theorem 1, we have that as $n \rightarrow \infty$, $\sqrt{n}(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})$ is asymptotically normal $N(0, S^{-1}\Sigma S^{-1})$, where $\Sigma = E[\{Z - E(Z|V)\}^\top \{Z - E(Z|V)\} \{e - E(e|V) + \xi(V)\}^2]$ and $S = Q_{12}^\top Q_{12}$.*

The justification of this theorem can be found in the appendix.

4 Numerical studies

4.1 Simulations

In this subsection, we use simulated data to assess the finite-sample performance of the proposed method. The Gaussian kernel function is used throughout the simulation study. For estimating the dimension r of V in Section 3.2, we adopt the ridge value $c_n = 5 \log n / \sqrt{n}$. The two criteria in Lin et al. (2016) for approximately constructing the variables V are also examined. These three methods are denoted respectively as RR-PLM-LASSO, LIN1-PLM-LASSO and LIN2-PLM-LASSO. For comparison purpose, we also examine the performance of the post-LASSO selection

least squares estimate (OLS-LASSO). According to our method, transferring the linear transformation model to the pro forma ‘linear’ model, OLS-LASSO can also be applicable to identify the variables in the working model.

For an estimate $\hat{\beta}_{\hat{\mathcal{I}}}$ of $\beta_{\hat{\mathcal{I}}}$, we measure its performance by computing the sample absolute correlation coefficient between the estimated predictor and the true one, $\text{ACC} = |\text{corr}(\mathbf{X}_{\hat{\mathcal{I}}}\hat{\beta}_{\hat{\mathcal{I}}}, \mathbf{X}_{\hat{\mathcal{I}}}\beta_{\hat{\mathcal{I}}})|$, the vector correlation coefficient, $\text{VCC} = \hat{\beta}_{\hat{\mathcal{I}}}^{s\top} \beta_{\hat{\mathcal{I}}}^s$, the mean squared error, $\text{MSE} = \|\hat{\beta}_{\hat{\mathcal{I}}}^s - \beta_{\hat{\mathcal{I}}}^s\|_2$, and the generalized mean squared error, $\text{GMSE} = (\hat{\beta}_{\hat{\mathcal{I}}}^s - \beta_{\hat{\mathcal{I}}}^s)^\top (n^{-1} \mathbf{X}_{\hat{\mathcal{I}}}^\top \mathbf{X}_{\hat{\mathcal{I}}}) (\hat{\beta}_{\hat{\mathcal{I}}}^s - \beta_{\hat{\mathcal{I}}}^s)$, where for a vector v , we use v^s to denote its normalized vector. The larger ACC and VCC are and the smaller MSE and GMSE are, the better a method is. We run each simulation example 200 times and the averages are reported in the tables.

Example 1. Consider the following linear model

$$Y = X^\top \beta + \epsilon,$$

where $I = \{1, 2, 3\}$ and $\beta_I = (1, 0.8, 0.8)^\top$. The rest of the coefficients are first generated from an uniform distribution $U(-0.1, 0.1)$ and then the negative ones are set to zero. Once the coefficient vector β is generated, we regard β_I as the relatively strong signals. The predictor vector X follows a multivariate normal distribution $N(u, \Sigma)$, where $u_j = 0, j \in I, u_j = 2, j \notin I$, and the covariance matrix is

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{p \times p}.$$

The error term $\epsilon \sim N(0, \sigma^2)$. To examine the disturbance impact from error and the correlation between the components of X , different values of σ and ρ are considered. In this example $n = 100$ and $p = 200$. Using this model, all competitors are well applicable.

Example 2. Consider the transformation model with a nonlinear link function:

$$H(Y) = X^\top \beta + \epsilon,$$

where

$$H^{-1}(Y) = \frac{\exp(Y)}{1 + \exp(Y)} + Y.$$

We use the same setup as in **Example 1**.

Clearly, the models in **Examples 1** and **2** are non-sparse. We summarize the simulation results in Tables 1 and 2. We can observe clearly that the OLS-LASSO performs poorly. Even when σ and ρ are small, the average values of MSE and GMES of the OLS-LASSO are alarmingly large. This confirms that the OLS-LASSO is inconsistent in non-sparse settings.

It is not difficult to find that the ridge ratio estimate outperforms those of Lin et al. (2016) in terms of both parameter estimation ($\hat{\beta}_{\hat{\mathcal{I}}}$) and predictor construction ($X_{\hat{\mathcal{I}}}^\top \hat{\beta}_{\hat{\mathcal{I}}}$). Generally, we can see that the ridge ratio estimation and the two criteria in Lin et al. (2016) for constructing the artificial variables, are uniformly superior to the OLS-LASSO in all the cases considered. However, it is also worthwhile to point out that the performance gets worse as ρ and σ increase. When ρ is very large, the LASSO can not consistently select the relatively strong signals, but our method still works better than the OLS-LASSO.

Example 3. It is expected that the new proposed method is still applicable to sparse models. To this end, we consider the same setting as in **Example 2**, except for a sparse structure with $\beta = (1, 0.8, 0.8, 0, \dots, 0)^\top$. The simulation results are shown in Table 3.

Two observations are noteworthy. First, our method performs well, with the sample absolute correlation coefficient and the vector correlation coefficient being larger

Table 1: Simulation results for Example 1 with the linear link function $H(\cdot)$ and the non-sparse structure. $X \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$ with the elements $\sigma_{ij} = I(i = j) + \rho I(i \neq j)$, and $\epsilon \sim N(0, \sigma^2)$.

	σ	ρ	RR-PLM-LASSO	LIN1-PLM-LASSO	LIN2-PLM-LASSO	OLS-LASSO
MSE	$\sigma = 0.1$	0.1	0.0375(0.0001)	0.0429(0.0003)	0.0489(0.0001)	0.0528(0.0002)
		0.2	0.0651(0.0007)	0.0670(0.0004)	0.0858(0.0012)	0.1116(0.0008)
		0.3	0.0948(0.0108)	0.1099(0.0024)	0.1022(0.0155)	0.1802(0.0063)
	$\sigma = 0.5$	0.1	0.0692(0.0016)	0.0902(0.0067)	0.0914(0.0055)	0.1179(0.0048)
		0.2	0.0940(0.0027)	0.0903(0.0044)	0.1043(0.0139)	0.1935(0.0068)
		0.3	0.1032(0.0034)	0.1814(0.0016)	0.1830(0.0136)	0.2638(0.0078)
GMSE	$\sigma = 0.1$	0.1	0.1610(0.0341)	0.1175(0.0180)	0.2567(0.1203)	0.3422(0.2174)
		0.2	0.3384(0.0777)	0.2093(0.0480)	0.4023(0.0982)	1.1450(1.3231)
		0.3	0.4366(0.2270)	0.6121(0.1435)	0.6485(0.4087)	2.1493(3.4126)
	$\sigma = 0.5$	0.1	0.4250(0.1036)	0.2122(0.0596)	0.4669(0.1142)	1.1776(1.3210)
		0.2	0.4221(0.1396)	0.4852(0.1066)	0.5584(0.2529)	2.5381(2.5819)
		0.3	0.4889(0.1393)	0.6421(0.1584)	0.6523(0.6548)	6.4116(8.7100)
ACC	$\sigma = 0.1$	0.1	0.9875	0.9780	0.9834	0.9740
		0.2	0.9819	0.9811	0.9687	0.9713
		0.3	0.9760	0.9831	0.9644	0.9553
	$\sigma = 0.5$	0.1	0.9752	0.9631	0.9698	0.9596
		0.2	0.9741	0.9742	0.9609	0.9505
		0.3	0.9787	0.9847	0.9564	0.9493
VCC	$\sigma = 0.1$	0.1	0.9185	0.9195	0.9212	0.9136
		0.2	0.9014	0.8860	0.9024	0.8791
		0.3	0.8921	0.8709	0.8846	0.8419
	$\sigma = 0.5$	0.1	0.8864	0.8973	0.9015	0.8771
		0.2	0.8881	0.8761	0.8863	0.8365
		0.3	0.9024	0.8676	0.8915	0.8112

than 0.99, and the MSE and GMSE being less than 0.03. Second, it is comparable to the OLS-LASSO. Nevertheless, the OLS-LASSO has a slightly better performance. One possible reason is that the two-stage estimation procedure involves more errors caused by nonparametric estimation.

Table 2: Simulation results for Example 2 with the nonlinear link function $H(\cdot)$ and the non-sparse structure. $X \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$ with the elements $\sigma_{ij} = I(i = j) + \rho I(i \neq j)$, and $\epsilon \sim N(0, \sigma^2)$.

	σ	ρ	RR-PLM-LASSO	LIN1-PLM-LASSO	LIN2-PLM-LASSO	OLS-LASSO
MSE	$\sigma = 0.1$	0.1	0.0485(0.0001)	0.0428(0.0001)	0.0475(0.0002)	0.0582(0.0001)
		0.2	0.0604(0.0019)	0.0597(0.0023)	0.0656(0.0051)	0.0819(0.0031)
		0.3	0.0857(0.0044)	0.0817(0.0036)	0.1070(0.0150)	0.1359(0.0066)
	$\sigma = 0.5$	0.1	0.0857(0.0026)	0.0822(0.0023)	0.1036(0.0050)	0.1051(0.0045)
		0.2	0.0912(0.0037)	0.0887(0.0029)	0.1057(0.0055)	0.1304(0.0050)
		0.5	0.1324(0.0057)	0.1583(0.0028)	0.1490(0.0196)	0.2327(0.0210)
GMSE	$\sigma = 0.1$	0.1	0.0919(0.0088)	0.0814(0.0079)	0.0934(0.0108)	0.1009(0.0129)
		0.2	0.1266(0.0227)	0.1372(0.0241)	0.1386(0.0371)	0.1712(0.0351)
		0.3	0.1893(0.0537)	0.2111(0.0679)	0.2481(0.1012)	0.4255(0.1512)
	$\sigma = 0.5$	0.1	0.2446(0.0475)	0.2388(0.0460)	0.2528(0.1236)	0.3456(0.1320)
		0.2	0.2571(0.0552)	0.2732(0.0587)	0.2876(0.0859)	0.4584(0.1229)
		0.3	0.3809(0.0976)	0.7863(0.1234)	0.4843(0.1655)	1.3704(0.5788)
ACC	$\sigma = 0.1$	0.1	0.9830	0.9854	0.9837	0.9784
		0.2	0.9821	0.9831	0.9808	0.9717
		0.3	0.9775	0.9782	0.9721	0.9641
	$\sigma = 0.5$	0.1	0.9697	0.9711	0.9649	0.9641
		0.2	0.9755	0.9762	0.9731	0.9615
		0.3	0.9716	0.9668	0.9681	0.9540
VCC	$\sigma = 0.1$	0.1	0.9132	0.9160	0.9137	0.9033
		0.2	0.9057	0.9060	0.8996	0.8949
		0.3	0.8886	0.8806	0.8729	0.8646
	$\sigma = 0.5$	0.1	0.9035	0.9048	0.9006	0.8839
		0.2	0.8986	0.9004	0.8967	0.8689
		0.3	0.8793	0.8664	0.8680	0.8291

Example 4. We now consider the well-known Box-Cox transformation model as:

$$Y^{(\lambda)} = c + X^\top \beta + \epsilon,$$

Table 3: Simulation results for Example 3 with the linear link function $H(\cdot)$ and the sparse structure. $X \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$ with the elements $\sigma_{ij} = I(i = j) + \rho I(i \neq j)$, and $\epsilon \sim N(0, \sigma^2)$.

	σ	ρ	RR-PLM-LASSO	LIN1-PLM-LASSO	LIN2-PLM-LASSO	OLS-LASSO
MSE	$\sigma = 0.1$	0.1	0.0005	0.0014	0.0007	0.0001
		0.2	0.0012	0.0033	0.0040	0.0001
		0.3	0.0013	0.0037	0.0029	0.0001
	$\sigma = 0.5$	0.1	0.0049	0.0056	0.0055	0.0036
		0.2	0.0058	0.0074	0.0071	0.0039
		0.3	0.0077	0.0134	0.0090	0.0071
GMSE	$\sigma = 0.1$	0.1	0.0004	0.0012	0.0006	0.0001
		0.2	0.0009	0.0027	0.0021	0.0001
		0.3	0.0010	0.0028	0.0035	0.0001
	$\sigma = 0.5$	0.1	0.0051	0.0061	0.0059	0.0044
		0.2	0.0067	0.0121	0.0088	0.0057
		0.3	0.0089	0.0280	0.0144	0.0073
ACC	$\sigma = 0.1$	0.1	0.9998	0.9995	0.9997	0.9999
		0.2	0.9996	0.9990	0.9987	0.9999
		0.3	0.9997	0.9991	0.9993	0.9999
	$\sigma = 0.5$	0.1	0.9982	0.9980	0.9980	0.9987
		0.2	0.9983	0.9979	0.9979	0.9989
		0.3	0.9979	0.9944	0.9969	0.9985
VCC	$\sigma = 0.1$	0.1	0.9997	0.9992	0.9996	0.9999
		0.2	0.9993	0.9983	0.9979	0.9999
		0.3	0.9993	0.9981	0.9985	0.9999
	$\sigma = 0.5$	0.1	0.9975	0.9971	0.9972	0.9981
		0.2	0.9970	0.9962	0.9964	0.9980
		0.3	0.9952	0.9903	0.9931	0.9964

where

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}.$$

Set $\lambda = 0.8$ and $c = 8$. We generate β, X and ϵ in the same way as in **Example 1**,

except that the covariance matrix of X takes the form

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho \\ \rho^2 & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{pmatrix}_{p \times p}.$$

The results are reported in Table 4. We present the results of our method, but use three aforementioned criteria to construct the artificial variables V . We then can examine what criterion can perform well. The comparison between Tables 1 - 2 and Table 4 demonstrates the following findings. The correlation structure of X has a slightly negative effect for our method in this non-sparse setting. But the ridge ratio estimation performs better than two criteria in Lin et al. (2016) to construct the artificial variables V , particularly when the correlation becomes larger.

In summary, our method is particularly useful when the true model is non-sparse, and it is competitive to the classical method in the sparse setting. Nevertheless, the estimation procedure is relatively more complex, because it involves nonparametric estimation. This price seems to be paid indispensably as for biased models, there are no good methods to handle them in the literature.

4.2 Real data example

For illustration, we analyze the well-known Baseball hitters' salary data set, which was originally published for the 1988 ASA Statistical Graphics and Computing Data Exposition. The data for information on salary consists of 255 samples with 16 independent predictors, which include the numbers of times at bat (X_1), hits (X_2), home runs (X_3), runs (X_4), runs batted in (X_5), walks (X_6) in 1986, the numbers of years in the major leagues (X_7), times at bat (X_8), hits (X_9), home runs (X_{10}),

runs (X_{11}), runs batted in (X_{12}), walks (X_{13}) during the players' entire career up to 1986; putouts (X_{14}), assists (X_{15}) and errors (X_{16}) in 1986. The logarithm of annual salary in 1987 is used to be the response variable (Y). Fit the data by using the following linear transformation model:

$$H(Y) = \sum_{i=1}^{16} X_i \beta_i + \epsilon, \quad (20)$$

where $\beta = (\beta_1, \dots, \beta_{16})^\top$, $\|\beta\| = 1$ and $X = (X_1, \dots, X_{16})$. By using the LASSO, $\tilde{X} = (X_2, X_6, X_7, X_{11})$ can be regarded as the strong signals. Then a working model is obtained as:

$$H(Y) = \tilde{X}^\top \beta_{\tilde{X}} + \eta. \quad (21)$$

Let $\hat{\beta}_{\tilde{X}}^{\text{LS}}$ and $\hat{\beta}_{\tilde{X}}^{\text{RR}}$ respectively represent the post-LASSO selection least squares estimator without bias correction and our estimator. The value of absolute difference $\text{MSE} = \|\hat{\beta}_{\tilde{X}}^{\text{RR}} - \hat{\beta}_{\tilde{X}}^{\text{LS}}\|_2 = 0.1704$ and the value of a multiple correlation coefficient like discrepancy $\text{VCC} = (\hat{\beta}_{\tilde{X}}^{\text{RR}})^\top \hat{\beta}_{\tilde{X}}^{\text{LS}} = 0.9147$. As there exists discrepancy between these two estimators, and we have proved the estimation consistency of our estimator, the discrepancy would suggest that the working model (21) exists bias and the least squares would not estimate $\beta_{\tilde{X}}$ well.

To further study the performance of our method, a larger artificial model is considered as a full model with additional predictors:

$$H(Y) = \sum_{i=1}^{16} \beta_i X_i + \sum_{i=17}^{24} \beta_i X_i + \sum_{i=25}^{32} \beta_i X_i + \sum_{i=33}^p \beta_i X_i + \epsilon, \quad (22)$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$, $\|\beta\| = 1$ for identifiability. To make correlation between the strong signals $\tilde{X}$ of real data set and the additional predictors $\{X_i\}_{i=17}^p$, the data set is generated as follows. $\{X_i\}_{i=1}^{16}$ are corresponding to the predictors in the model (20). $X_i = (X_2 + W_i)/2$ for $17 \leq i \leq 24$, with W_i being independently from the exponential distribution with mean 1, $X_i = (X_6 + W_i)/2$ for $25 \leq i \leq 32$, with

W_i being generated independently from the standard normal distribution, and X_i for $33 \leq i \leq p$ are generated independently from the standard normal distribution. Consider the different dimensions $p = 66, 116$, and 316 . Each experiment is repeated 100 times. When the LASSO is applied, we find that $\tilde{X} = (X_2, X_6, X_7, X_{11})$ are the relatively strong signals for $p = 66, 216$ and 316 although some other variables are also regarded as strong signals.

To compare the performance of the post-LASSO selection least squares estimator and our method, other than the above two criteria, MSE and VCC, two more criteria are the generalized absolute difference $\text{GMSE} = (\hat{\beta}_{\tilde{X}}^{\text{RR}} - \hat{\beta}_{\tilde{X}}^{\text{LS}})^\top (n^{-1} \mathbf{X}_{\tilde{X}}^\top \mathbf{X}_{\tilde{X}}) (\hat{\beta}_{\tilde{X}}^{\text{RR}} - \hat{\beta}_{\tilde{X}}^{\text{LS}})$ and absolute correlation coefficient between both estimated predictors $\text{ACC} = |\text{corr}(\mathbf{X}_{\tilde{X}} \hat{\beta}_{\tilde{X}}^{\text{RR}}, \mathbf{X}_{\tilde{X}} \hat{\beta}_{\tilde{X}}^{\text{LS}})|$, where $\hat{\beta}_{\tilde{X}}^{\text{RR}}$ and $\hat{\beta}_{\tilde{X}}^{\text{LS}}$ respectively denote our estimator and the post-LASSO selection least squares estimator.

To make the comparison reasonable, we also use the strong signals associated with the first four largest absolute values of coefficients selected by the LASSO, which equals the number of strong signals when the dimension $p = 16$. The results are presented in Tables 5 and 6 for different dimensions of p . We can then clearly observe that the discrepancy between these two estimators becomes larger when the dimension p gets larger. When we use the fixed number of predictors with the strong signals, the discrepancy between these two estimators exists, but the dimension p has slight impact for the performance of our method. The difference between these two estimators suggests the existence of bias when the working model is used and the bias correction should be useful.

To further check whether our method is robust against to the dimension in this example, we fix the strong signals $\tilde{X} = (X_2, X_6, X_7, X_{11})$ and compare our estimators under different dimension cases with our estimators under the case with the dimension $p = 16$. This can also check whether the estimation consistency or accuracy is

affected by the dimension of full model. We compute the four similar quantities as the above. The results are reported in Table 7. The discrepancy between $\hat{\beta}_p^{\text{RR}}$ and $\hat{\beta}_{16}^{\text{RR}}$ almost changes nothing as the dimension p increases, where $\hat{\beta}_p^{\text{RR}}$ denotes our estimator with the dimension p and $\hat{\mathcal{I}} = (2, 6, 7, 11)$.

5 Acknowledgments

Xuehu Zhu gratefully acknowledges financial support from National Natural Science Foundation of China (No.11601415, No.11571204) and China Postdoctoral Science Foundation (No.2016M590934). Lixing Zhu was supported by a grant from the University Grants Council of Hong Kong, Hong Kong, China. The work was also supported in part by National Natural Science Foundation of China (No.11401465). The authors thank the editor, the associate editor and two anonymous referees for their constructive comments and suggestions which led to a substantial improvement of an early manuscript.

6 Appendix

Proof of Proposition 1. (1) We first show Condition C1 holds. Let $\mathbf{C} = (\hat{\sigma}_{ij})$. Note that Σ_X has bounded eigenvalues and that $(\log p)/n \rightarrow 0$. By the tail probability of normal variable (Vershynin, 2010), it is easy to see that, for any $\epsilon > 0$,

$$P(\max_{1 \leq j \leq p} |\hat{\sigma}_{jj}/\sigma_{jj} - 1| > \epsilon) \leq p \max_{1 \leq j \leq p} P(|\hat{\sigma}_{jj}/\sigma_{jj} - 1| > \epsilon) \leq \exp(\log p - \delta_1 n \epsilon^2) \rightarrow 0,$$

where $\delta_1 > 0$ is a constant. Similarly, for any $\epsilon > 0$, we have

$$P(\max_{i \neq j} |\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon) \leq p(p-1) \max_{i \neq j} P(|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon) \leq \exp(2 \log p - \delta_2 n \epsilon^2) \rightarrow 0,$$

where $\delta_2 > 0$ is a constant. Altogether, we have

$$\max_{1 \leq i, j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| \xrightarrow{P} 1. \quad (23)$$

Let $\mathbf{X}_{(1)} = (X_1, \dots, X_q)^T$ is the subvector of X associated with the strong signals. Since q is fixed, then we have $\mathbf{C}_{11} \xrightarrow{p} \text{Cov}(\mathbf{X}_{(1)})$. Combining with the assumption on the eigenvalue of Σ_X , this result leads to Condition C1.

(2) Consider Condition C2. Let $\tilde{c}_{\min} = \min_{1 \leq i \leq p} \hat{\sigma}_{ii}$. The proof consists of two steps. In Step 1, we first show that if

$$c_{\max} := \max_{i \neq j} \hat{\sigma}_{ij} \leq \tilde{c}_{\min} c_2 / (2q - 1), \quad (24)$$

where $c_2 < 1$ is defined in (ii), then Strong irrepresentable condition (Zhao and Yu, 2006) holds. In Step 2, we show that under the conditions, (24) holds with a probability tending to 1.

Step 1. The proof is similar to that of Corollary 2 of Zhao and Yu (2006). Consider the first entry of $|\mathbf{C}_{21} \mathbf{C}_{11}^{-1} \text{sign}(\beta_{\mathcal{I}})|$, which takes the form $|\alpha^\top \mathbf{C}_{11}^{-1} \text{sign}(\beta_{\mathcal{I}})|$, where $\alpha^\top$ is the first row of $\mathbf{C}_{21}$. By (24), we have

$$|\alpha^\top \mathbf{C}_{11}^{-1} \text{sign}(\beta_{\mathcal{I}})| \leq \|\alpha\|_{\max} \|\text{sign}(\beta_{\mathcal{I}})\|_0 \cdot \frac{1}{f} \leq \frac{\tilde{c}_{\min} c q}{2q - 1} \cdot \frac{1}{f} \quad (25)$$

where f is the lower bound of the smallest eigenvalue of $\mathbf{C}_{11}$. For any unit vector $x = (x_1, \dots, x_q)^\top$, we have

$$\begin{aligned} e_1 > x^\top \mathbf{C}_{11} x &= \sum_{i=1}^q x_i^2 \hat{\sigma}_{ii} + \sum_{i \neq j} x_i \hat{\sigma}_{ij} x_j \\ &\geq \tilde{c}_{\min} - \sum_{i \neq j} |x_i| |\hat{\sigma}_{ij}| |x_j| \\ &\geq \tilde{c}_{\min} - \frac{\tilde{c}_{\min}}{2q - 1} \sum_{i \neq j} |x_i| |x_j| \\ &= \tilde{c}_{\min} - \frac{\tilde{c}_{\min}}{2q - 1} \left(\sum_{i,j} |x_i| |x_j| - 1 \right) \\ &\geq \tilde{c}_{\min} - \frac{\tilde{c}_{\min}(q - 1)}{2q - 1} \\ &= \tilde{c}_{\min} q / (2q - 1). \end{aligned}$$

Combining with (25), we have $|\alpha^\top \mathbf{C}_{11}^{-1} \text{sign}(\beta_{\mathcal{I}})| \leq c$. Then we have $|\mathbf{C}_{21} \mathbf{C}_{11}^{-1} \text{sign}(\beta_{\mathcal{I}})| < 1$. This completes the proof of Step 1.

Step 2. It is easy to derive the conclusion that (24) holds with a probability tending to 1 by (23) and the assumption that $\max_{1 \leq i \neq j \leq p} \sigma_{ij} < (\min_{1 \leq i \leq p} \sigma_{ii})c/(2q - 1)$. This completes the proof. $\square$

The proof of Theorem 1. Since Conditions C1 and C2 hold with probability tending to 1, or equivalently, $P(\mathcal{E}) \rightarrow 1$, where

$$\mathcal{E} = \left\{ \frac{1}{n} \mathbf{X}_j^\top \mathbf{X}_j < M_1, 1 \leq j \leq p; M_2^{-1} < a^\top \mathbf{C}_{11} a \leq M_2, \text{ for any } \|a\|_2 = 1; \right. \\ \left. |\mathbf{C}_{21} (\mathbf{C}_{11})^{-1} \text{sign}(\beta_{\mathcal{I}})| \leq \mathbf{1}_{p-q} - \nu \right\}.$$

We only need to discuss about the events $\mathcal{E}$.

We use the similar argument used to prove Proposition 1 in Zhao and Yu (2006). It is well known that for convex optimization problem, x^* is the optimal solution if and only if x^* satisfies the associated KKT condition. Since the objective function is strictly convex in Lasso, the estimate $\hat{\beta}$ is completely determined by the KKT condition, regardless the true parameter being sparse or non-sparse. Therefore, the main idea of Zhao and Yu (2006) can be used here. Of course, the KKT condition for non-sparse model is different from that for sparse model, with additional terms related with the nonsparse part as shown below. In the following argument, we mainly state the parts that are different from Zhao and Yu (2006) and for conciseness omit the details of the same parts.

Let $\mathbf{e}_n = (e_1, e_2, \dots, e_n)$, $\mathbf{Q}_{(1)} = \frac{1}{\sqrt{n}} \mathbf{X}_{(1)}^\top \mathbf{e}_n$, $\mathbf{Q}_{(2)} = \frac{1}{\sqrt{n}} \mathbf{X}_{(2)}^\top \mathbf{e}_n$. Similar to the argument of Zhao and Yu (2006), applying the KKT condition and the irrepresentable condition C2, we know that the sign consistency of the estimate $P(\hat{\beta} =_s \beta) \geq P(A_n \cap B_n)$ where " $=_s$ " stands for the same sign in the two sides of

the equation and

$$A_n = \{ |(\mathbf{C}_{11})^{-1} \mathbf{Q}_{(1)}| < \sqrt{n}(|\beta_{\mathcal{I}}| - \frac{\lambda_n}{2n} |(\mathbf{C}_{11})^{-1} \text{sign}(\beta_{\mathcal{I}})| - |(\mathbf{C}_{11})^{-1}(\mathbf{C}_{12})\beta_{\mathcal{I}^c}|) \},$$

$$B_n = \{ |\mathbf{C}_{21}(\mathbf{C}_{11})^{-1} \mathbf{Q}_{(1)} - \mathbf{Q}_{(2)}| < \frac{\lambda_n}{2\sqrt{n}} \nu - |\sqrt{n} \mathbf{C}_{22} \beta_{\mathcal{I}^c}| \}.$$

Note that, for sparse model with $\beta_{\mathcal{I}^c} = 0$, A_n and B_n are exactly identical to those defined in Zhao and Yu (2006). For non-sparse model, we have two additional terms $|(\mathbf{C}_{11})^{-1}(\mathbf{C}_{12})\beta_{\mathcal{I}^c}|$ and $|\sqrt{n} \mathbf{C}_{22} \beta_{\mathcal{I}^c}|$. We only need to show that both $P(A_n^c) \rightarrow 0$ and $P(B_n^c) \rightarrow 0$. It suffices to ensure that the right hand sides of the inequalities defined in the events A_n and B_n tend to ∞ .

For any $A = (a_{ij})$, let $|A|_{\max} = \max_{i,j} |a_{ij}|$. Recall that q is fixed. Denoting $D = \mathbf{C}_{11}^{-1} \mathbf{C}_{12}$, under the assumption C1, we see that $|D|_{\max} := d < \infty$ and consequently, $\|D_{i\cdot} \beta_{\mathcal{I}^c}\|_1 \leq d \|\beta_{\mathcal{I}^c}\|_1 \ll |\tau_n|$, for $1 \leq i \leq q$, where $D_{i\cdot}$ stands for the i th row of D . Therefore, $|(\mathbf{C}_{11})^{-1}(\mathbf{C}_{12})\beta_{\mathcal{I}^c}|$ is dominated by the term $\sqrt{n}|\beta_{\mathcal{I}}|$. Moreover, by Conditions C1 and C3, we see that $\frac{\lambda_n}{2n} |(\mathbf{C}_{11})^{-1} \text{sign}(\beta_{\mathcal{I}})|$ is dominated by $|\beta_{\mathcal{I}}|$. $\sqrt{n}|\beta_{\mathcal{I}}|$ is a dominant term in the right hand side of the inequality defined in the event A_n . Consequently, by the proof of Zhao and Yu (2006) and assumption C3, we known that

$$P(A_n^c \cap \mathcal{E}) \leq qP(|z_i^n| > \sqrt{n}\tau_n) = O\{(\sqrt{n}\tau_n)^{-k}\} \rightarrow 0,$$

where z_i is the element of the vector $\mathbf{C}_{11}^{-1} \mathbf{Q}_1$ (Zhao and Yu, 2006). Consequently, it holds that

$$P(A_n^c) \leq P(A_n^c \cap \mathcal{E}) + P(\mathcal{E}^c) \rightarrow 0.$$

Similarly, combining with Condition C3, we see that $\lambda_n \nu / (2\sqrt{n})$ is a dominant term in the right hand side of the inequality defined in B_n . Consequently, invoking the assumption $E(e^{2k}) < \infty$ and the same argument of Zhao and Yu (2006), we have $P(B_n^c \cap \mathcal{E}) = O(pn^k / \lambda_n^{2k}) \rightarrow 0$, which tends to zero by the assumption on p . Consequently, it holds that $P(B_n^c) \leq P(B_n^c \cap \mathcal{E}) + P(\mathcal{E}^c) \rightarrow 0$. $\square$

Proof of Proposition 2. The justification for this proposition is similar to that of Lemma 3.2 in Lin et al. (2016), thus we only give a sketch of the proof. The singular value decomposition of matrix in Banerjee and Roy (2014) yields that

$$\Sigma_{U,\tilde{Z}} = R \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} Q,$$

where R is a $(p-d) \times (p-d)$ orthogonal matrix, Q is a $(p+d) \times (p+d)$ orthogonal matrix, Λ_r is a diagonal matrix with the diagonal elements of Λ_r being the square roots of non-zero eigenvalues of $\Sigma_{U,\tilde{Z}}^\top \Sigma_{U,\tilde{Z}}$. Therefore, we have

$$\Sigma_{U,\tilde{Z}}^+ \Sigma_{U,\tilde{Z}} = Q \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q^\top = Q_1 Q_1^\top,$$

where $\Sigma_{U,\tilde{Z}}^+$ is the Moore-Penrose generalized inverse matrix of $\Sigma_{U,\tilde{Z}}$, $Q = (Q_1, Q_2)$ and $Q_1^\top Q_1 = I_r$ and $Q_1^\top = (Q_{11}^\top, Q_{12}^\top)$. Let $B = \begin{pmatrix} I_r & 0 \\ A_1 & A_2 \end{pmatrix}$ and $A = (A_1, A_2) = (Q_{11}^\top, Q_{12}^\top)$. Then we have

$$BB^\top = \begin{pmatrix} I_r & 0 \\ A_1 & A_2 \end{pmatrix} \begin{pmatrix} I_r & A_1^\top \\ 0 & A_2^\top \end{pmatrix} = \begin{pmatrix} I_r & A_1^\top \\ A_1 & A_2^\top \end{pmatrix}.$$

As the eigenvalues of $Q_{11}^\top Q_{11}$ are not equal to 1, $A_2 A_2^\top = I_r - A_1 A_1^\top$ is invertible.

We can achieve the inverse matrix of $BB^\top$ in the following form:

$$(BB^\top)^{-1} = \begin{pmatrix} I_r + A_1^\top (A_2 A_2^\top)^{-1} A_1 & -A_1^\top (A_2 A_2^\top)^{-1} \\ -(A_2 A_2^\top)^{-1} A_1 & (A_2 A_2^\top)^{-1} \end{pmatrix}.$$

Therefore, we can get the projection operator as:

$$P = B^\top (BB^\top)^{-1} B = \begin{pmatrix} I_r & 0 \\ 0 & A_2^\top (A_2 A_2^\top)^{-1} A_2 \end{pmatrix}. \quad (26)$$

It is clear that $AP = A$, $AA^\top = I_r$, $A^\top A = \Sigma_{U,\tilde{Z}}^+ \Sigma_{U,\tilde{Z}}$ and $\Sigma_{U,\tilde{Z}} \Sigma_{U,\tilde{Z}}^+ \Sigma_{U,\tilde{Z}} = \Sigma_{U,\tilde{Z}}$.

Combining the above results yields the following equation:

$$\Sigma_{U,\tilde{Z}} A^\top (AA^\top)^{-1} A \{\tilde{Z} - E(\tilde{Z})\} = \Sigma_{U,\tilde{Z}} P \{\tilde{Z} - E(\tilde{Z})\}. \quad (27)$$

Under the linearity condition (10), altogether Lemma A1 and Lemma A2 in Lin et al. (2016) with (26) and (27), we conclude the proof. $\square$

Proof of Proposition 3. From the equation (12), we have $\frac{1}{p-q}\widehat{\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}}} - \frac{1}{p-q}\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}} = O_p(n^{-1/2})$. As proved by Zhu and Ng (1995) or Zhu and Fang (1996), the root- n consistency of the eigenvalues of $\frac{1}{p-q}\widehat{\Sigma_{\tilde{Z},U}^\top \Sigma_{U,\tilde{Z}}}$ is retained, namely, $\hat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$ for $i = 1, \dots, q + d$.

Since for any $1 < l \leq r$, $\lambda_l > 0$, we have $\hat{\lambda}_l = \lambda_l + O_p(1/\sqrt{n})$. On the other hand, for any $r < l \leq q + d$, $\lambda_l = 0$, then we have $\hat{\lambda}_l = O_p(1/\sqrt{n})$. Thus, for any $l < r$, we have $\lambda_l > 0$, $\lambda_{l+1} > 0$ and

$$\begin{aligned} \frac{\hat{\lambda}_{(r+1)} + c_n}{\hat{\lambda}_r + c_n} - \frac{\hat{\lambda}_{(l+1)} + c_n}{\hat{\lambda}_l + c_n} &= \frac{\lambda_{r+1} + c_n + O_p(1/\sqrt{n})}{\lambda_r + c_n + O_p(1/\sqrt{n})} - \frac{\lambda_{l+1} + c_n + O_p(1/\sqrt{n})}{\lambda_l + c_n + O_p(1/\sqrt{n})} \\ &\rightarrow \frac{\lambda_{r+1}}{\lambda_r} - \frac{\lambda_{l+1}}{\lambda_l} = -\frac{\lambda_{l+1}}{\lambda_l} < 0. \end{aligned}$$

Further, due to the facts that $\lambda_r > 0$ and for any $l > r$, $\lambda_l = 0$, we can achieve

$$\begin{aligned} \frac{\hat{\lambda}_{(r+1)} + c_n}{\hat{\lambda}_r + c_n} - \frac{\hat{\lambda}_{(l+1)} + c_n}{\hat{\lambda}_l + c_n} &= \frac{c_n + O_p(1/\sqrt{n})}{\lambda_r + c_n + O_p(1/\sqrt{n})} - \frac{c_n + O_p(1/\sqrt{n})}{c_n + O_p(1/\sqrt{n})} \\ &= \frac{c_n + o_p(c_n)}{\lambda_r + c_n + O_p(1/\sqrt{n})} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)} \\ &\rightarrow -1 < 0. \end{aligned}$$

Therefor, altogether, we can conclude that $\hat{r} = r$ with a probability going to 1. $\square$

Proof of Theorem 2. To prove this theorem, we first prove that the equation $E[\{Z - E(Z|V)\}\xi_2(V)] = 0$ holds. By the application of Proposition 2, we can get the equation $X^\top \beta = Z^\top \beta_{\mathcal{I}} + g(V) + \xi(V)$, where $E\{\xi(V)|Z, V\} = 0$. Note that the random variable $\xi_1(V) = e + \xi(V)$ satisfies $E\{Z\xi_1(v)\} = E(Ze) + E\{Z\xi(V)\} = 0$ and $\xi_2(Z, V) = \xi_1(Z, V) - E\{\xi_1(Z, V)|V\} = e - E(e|V) + \xi(V)$. Thus, it is easy to see that some elementary calculation about conditional expectation can lead to the

following:

$$\begin{aligned} E\{Z\xi_2(Z, V)\} &= E\{Ze + Z\xi(V) - ZE(e|V)\} = E(Z[E\{\xi(V)|Z\}]) = 0, \\ E\{E(Z|V)\xi_2(Z, V)\} &= E\{Z\xi_2(Z, V)\} = 0. \end{aligned}$$

That is, we have derived that $E[\{Z - E(Z|V)\}\xi_2(V)] = 0$ holds. From the structure of the working model (18), we can derive an estimate of $\beta_{\mathcal{I}}$ as:

$$\hat{\beta}_{\mathcal{I}} = \hat{S}^{-1} \frac{1}{n} \sum_{k=1}^n \hat{z}_i^\top \hat{h}(y_i),$$

where $\hat{S} = \frac{1}{n} \sum_{k=1}^n \hat{z}_i^\top \hat{z}_i$. By applying the mean value theorem, there exists a matrix V^* between V and $\hat{V}$ satisfying that

$$\frac{1}{n} \sum_{k=1}^n L_H(\hat{v}_k - v) = \frac{1}{n} \sum_{k=1}^n L_H(v_k - v) + \frac{1}{n} \sum_{k=1}^n (\hat{v}_k - v_k)^\top L'_H(v_k^* - v),$$

where $L'_H(\cdot)$ denote the first derivative of $L_H(\cdot)$. Under the regularity conditions B1-B3, we have $\|\hat{v}_k - v_k\| = O_p(n^{-1/2})$ and thus $\frac{1}{n} \sum_{k=1}^n (\hat{v}_k - v_k)^\top L'_H(v_k^* - v) = O_p(n^{-1/2})$.

The application of the standard nonparametric technique (see, e.g. Härdle et al., 2000) yields that

$$\frac{1}{n} \sum_{k=1}^n L_H(v_k - v) - f_V(v) = O_p\left(h^m + \frac{1}{\sqrt{nh^r}}\right),$$

and then

$$\frac{1}{n} \sum_{k=1}^n L_H(\hat{v}_k - v) - f_V(v) = O_p\left(h^m + \frac{1}{\sqrt{nh^r}}\right) + O_p(n^{-1/2}),$$

where f_V denotes the density function of V . We can also similarly justify

$$\frac{1}{n} \sum_{k=1}^n z_k L_H(\hat{v}_k - v) - \frac{1}{n} \sum_{k=1}^n z_k L_H(v_k - v) = O_p(n^{-1/2}).$$

Therefore, we have

$$\hat{h}(y_i) = \tilde{h}(y_i) + O_p(n^{-1/2}), \quad \hat{z}_i = \tilde{z}_i + O_p(n^{-1/2}), \quad \hat{\xi}_{2i} = \tilde{\xi}_{2i} + O_p(n^{-1/2}),$$

where

$$\begin{aligned}\tilde{h}(y_i) &= h(y_i) - \frac{\sum_{k=1}^n h(y_i) L_H(v_k - v_i)}{\sum_{k=1}^n L_H(v_k - v_i)}; \quad \tilde{g}(v_i) = g(v_i) - \frac{\sum_{k=1}^n g(v_i) L_H(v_k - v_i)}{\sum_{k=1}^n L_H(v_k - v_i)}; \\ \tilde{z}_i &= z_i - \frac{\sum_{k=1}^n z_k L_H(v_k - v_i)}{\sum_{k=1}^n L_H(v_k - v_i)}; \quad \tilde{\xi}_{1i} = \xi_{1i} - \frac{\sum_{k=1}^n \xi_{1i} L_H(v_k - v_i)}{\sum_{k=1}^n L_H(v_k - v_i)}.\end{aligned}$$

Therefore, we can get

$$\hat{\beta}_{\tilde{\mathcal{I}}} - \beta_{\mathcal{I}} = \tilde{S}^{-1} \frac{1}{n} \sum_{i=1}^n \{\tilde{z}_i \tilde{g}(v_i) + \tilde{z}_i \tilde{\xi}_{2i}\} \{1 + O_p(h^m + \frac{1}{\sqrt{nh^r}}) + O_p(n^{-1/2})\} + o_p(n^{-1/2}),$$

where $\tilde{S} = \frac{1}{n} \sum_{k=1}^n \tilde{z}_i^\top \tilde{z}_i$. Combining the above conclusions yields $\hat{S} - \tilde{S} = O_p(n^{-1/2})$.

To complete the justification of this theorem, we also prove the following three facts for $j = 1, \dots, p$:

- (i) $H_{1j} = 1/\sqrt{n} \sum_{i=1}^n \tilde{z}_i \tilde{g}(v_i) = o_p(1)$;
- (ii) $H_{2j} = 1/\sqrt{n} \sum_{i=1}^n \tilde{z}_{ij} \sum_{k=1}^n \omega_{nk}(v_i) \tilde{\xi}_{2i} = o_p(1)$;
- (iii) $H_{2j} = 1/\sqrt{n} \sum_{i=1}^n \tilde{z}_i \tilde{\xi}_{2i} \xrightarrow{d} N(0, \Sigma)$,

where $\omega_{in}(v) = K(\frac{v-v_i}{h}) / \sum_{j=1, j \neq i}^n K(\frac{v_j-v_i}{h})$ and $\tilde{g}(v_i) = \sum_{j=1, j \neq i}^n \omega_{in}(v_i) \{g(v_i) - g(v_j)\}$. Following the same way as the proof of Theorem 2.1.2 in Härdle et al. (2000), we can easily make (i) and (ii) be proved. Because of $E[\{Z - E(Z|V)\} \xi_2] = 0$, the central limit theorem yields that as $n \rightarrow \infty$,

$$1/\sqrt{n} \sum_{i=1}^n \tilde{z}_i \tilde{\xi}_{2i} \xrightarrow{d} N(0, \Sigma).$$

where $\Sigma = E[\{Z - E(Z|V)\}^\top \{Z - E(Z|V)\} \{e - E(e|V) + \xi(V)\}^2]$. Similarly as the proof for Lemma A.2 in Härdle et al. (2000), we can deduce that $\tilde{S} = \frac{1}{n} \sum_{k=1}^n \tilde{z}_i^\top \tilde{z}_i$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \tilde{z}_i^\top \tilde{z}_i = S,$$

where $S = E\{Z - E(Z|V)\}\{Z - E(Z|V)\}^\top$.

Since $Q_1 Q_1 = I_{r \times r}$ and the condition about the eigenvalues of $Q_{11}^\top Q_{11}$ in Proposition 2 implies that $Q_{12}^\top Q_{12}$ is invertible, it is easy to verify that $S = Q_{12}^\top Q_{12}$. By Slutsky's theorem, we therefore get that as $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}}) \xrightarrow{d} N(0, S^{-1} \Sigma S^{-1}).$$

The proof is completed. □

References

- Banerjee, S. and Roy, A. (2014). *Linear algebra and matrix analysis for statistics*. CRC Press.
- Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521-547.
- Feng, Z., Wang, T. and Zhu, L. X. (2014). Transformation-based estimation. *Computational Statistics & Data Analysis*, **78**, 186-205.
- Guan, Y., Xie, C. L. and Zhu, L. X. (2016). Sufficient dimension reduction with mixture multivariate skew elliptical distributions, *Statistica Sinica*, Accepted
- Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* , **21**, 867-889.
- Härdle, W., Liang, H., and Gao, J. T. (2000). *Partially Linear Models*. Physica-Verlag, Heidelberg.
- Leeb, H. and Pötscher, B. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**, 21-59.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of American Statistical Association*, **86**, 316-327.
- Lin, L., Zhu, L. X. and Gai, Y. J. (2016). Inference for biased models: A quasi-instrumental variable approach. *Journal of Multivariate Analysis*, **145**, 22-36
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267-288.

- Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- Wang, T., Xu, P. and Zhu, L. X. (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, **109**, 221-235.
- Wang, T., Zhang, J., Liang, H. and Zhu, L. X. (2015). Estimation of a groupwise additive multiple-index model and its applications. *Statistica Sinica*, **25**, 551-566.
- Xia, Q., Xu, W. L. and Zhu, L. X. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, **25**, 1025-1044.
- Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567-1594.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning Research*, **7**, 2541-2563.
- Zhu, L. X. and Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *The Annals of Statistics*, **24**, 1053-1067.
- Zhu, L. X. and Ng, K. W. (1995). Asymptotics for sliced inverse regression. *Statistica Sinica*, **5**, 727-736.
- Zhu, X. H., Chen, F., Guo, X. and Zhu, L. X. (2016a). Heteroscedasticity testing for regression models: A dimension reduction-based model adaptive approach. *Computational Statistics & Data Analysis*, **103**, 263-283.
- Zhu, X. H., Guo, X. and Zhu, L. X. (2016b). An adaptive-to-model test for partially parametric single-index models. *Statistics and Computing*, accepted. <http://link.springer.com/article/10.1007/s11222-016-9680-z>

Table 4: Simulation results for Example 4 with the non-sparse Box-Cox transformation model. $X \sim N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p \times p}$ with the elements $\sigma_{ij} = \rho^{|i-j|}$, and $\epsilon \sim N(0, \sigma^2)$.

	σ	ρ	RR-PLM-LASSO	LIN1-PLM-LASSO	LIN2-PLM-LASSO
MSE	$\sigma = 0.1$	0.1	0.0084	0.0333	0.0096
		0.3	0.0118	0.0394	0.0129
		0.5	0.0201	0.0768	0.0276
		0.7	0.0752	0.2111	0.1315
	$\sigma = 0.5$	0.1	0.0116	0.0390	0.0128
		0.3	0.0193	0.0652	0.0208
		0.5	0.0241	0.0886	0.0402
		0.7	0.0807	0.2304	0.1347
	$\sigma = 1.0$	0.1	0.0084	0.0313	0.0094
		0.3	0.0113	0.0319	0.0117
		0.5	0.0143	0.0442	0.0178
		0.7	0.0917	0.1280	0.2785
GSE	$\sigma = 0.1$	0.1	0.0150	0.0380	0.0176
		0.3	0.0214	0.0564	0.0248
		0.5	0.0285	0.0679	0.0481
		0.7	0.0803	0.1694	0.2126
	$\sigma = 0.5$	0.1	0.9967	0.9847	0.9962
		0.3	0.9979	0.9896	0.9977
		0.5	0.9975	0.9875	0.9962
		0.7	0.9972	0.9878	0.9930
	$\sigma = 1.0$	0.1	0.9967	0.9847	0.9962
		0.3	0.9959	0.9822	0.9955
		0.5	0.9964	0.9845	0.9935
		0.7	0.9915	0.9654	0.9824
VCC	$\sigma = 0.1$	0.1	0.9330	0.9193	0.9324
		0.3	0.9298	0.9160	0.9293
		0.5	0.9283	0.9000	0.9246
		0.7	0.9220	0.8728	0.8991
	$\sigma = 0.5$	0.1	0.9330	0.9193	0.9324
		0.3	0.9195	0.8966	0.9188
		0.5	0.9294	0.8971	0.9213
		0.7	0.8946	0.8198	0.8676

Table 5: Simulation results with all strong signals selected by the LASSO for real data example.

p	MSE	VCC	GMSE	ACC
66	0.5406	0.7296	0.4549	0.8372
116	0.6172	0.6913	0.5649	0.8001
316	0.8875	0.5562	0.8240	0.6969

Table 6: Simulation results with the four strong signals that associated with the first four largest absolute values of coefficients by the LASSO in real data example.

p	MSE	VCC	GMSE	ACC
66	0.5815	0.7092	0.5004	0.8202
116	0.5831	0.7084	0.4781	0.8220
316	0.5195	0.7402	0.4261	0.8495

Table 7: Simulation results with the fixed strong signals $\tilde{X} = (X_2, X_6, X_7, X_{11})$. Let $\widetilde{\text{MSE}} = \|\hat{\beta}_p^{\text{RR}} - \hat{\beta}_{16}^{\text{RR}}\|_2$, $\widetilde{\text{VCC}} = (\hat{\beta}_p^{\text{RR}})^\top \hat{\beta}_{16}^{\text{RR}}$, $\widetilde{\text{GMSE}} = (\hat{\beta}_p^{\text{RR}} - \hat{\beta}_{16}^{\text{RR}})^\top (n^{-1} \mathbf{X}_{\hat{\mathcal{I}}}^\top \mathbf{X}_{\hat{\mathcal{I}}})(\hat{\beta}_p^{\text{RR}} - \hat{\beta}_{16}^{\text{RR}})$ and $\widetilde{\text{ACC}} = |\text{corr}(\mathbf{X}_{\hat{\mathcal{I}}} \hat{\beta}_p^{\text{RR}}, \mathbf{X}_{\hat{\mathcal{I}}} \hat{\beta}_{16}^{\text{RR}})|$, where $\hat{\beta}_p^{\text{RR}}$ denote our estimator with the dimension p and $\hat{\mathcal{I}} = (2, 6, 7, 11)$.

p	$\widetilde{\text{MSE}}$	$\widetilde{\text{VCC}}$	$\widetilde{\text{GMSE}}$	$\widetilde{\text{ACC}}$
66	0.0985	0.9507	0.0135	0.9966
116	0.0701	0.9649	0.0119	0.9972
316	0.0814	0.9592	0.0148	0.9968