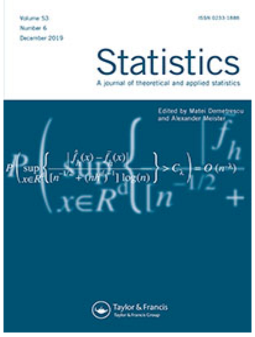

Taylor & Francis
Taylor & Francis Group

Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Principal varying coefficient estimator for high-dimensional models

Weihua Zhao, Fode Zhang, Xuejun Wang, Rui Li & Heng Lian

To cite this article: Weihua Zhao, Fode Zhang, Xuejun Wang, Rui Li & Heng Lian (2019) Principal varying coefficient estimator for high-dimensional models, Statistics, 53:6, 1234-1250, DOI: 10.1080/02331888.2019.1663521

To link to this article: https://doi.org/10.1080/02331888.2019.1663521

Published online: 16 Sep 2019.

STATISTICS
2019, VOL. 53, NO. 6, 1234–1250
https://doi.org/10.1080/02331888.2019.1663521

Taylor & Francis
Taylor & Francis Group

Check for updates

Principal varying coefficient estimator for high-dimensional models

Weihua Zhao[a], Fode Zhang[b], Xuejun Wang[c], Rui Li[d] and Heng Lian[e]

[a]School of Science, Nantong University, Nantong, People's Republic of China; [b]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, People's Republic of China; [c]School of Mathematical Sciences, Anhui University, Hefei, People's Republic of China; [d]School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, People's Republic of China; [e]Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong

ABSTRACT

We consider principal varying coefficient models in the high-dimensional setting, combined with variable selection, to reduce the effective number of parameters in semiparametric modelling. The estimation is based on B-splines approach. For the unpenalized estimator, we establish non-asymptotic bounds of the estimator and then establish the (asymptotic) local oracle property of the penalized estimator, as well as non-asymptotic error bounds. Monte Carlo studies reveal the favourable performance of the estimator and an application on a real dataset is presented.

ARTICLE HISTORY
Received 5 January 2019
Accepted 30 August 2019

KEYWORDS
Asymptotic properties; B-splines; sub-Gaussian distribution; ultra-high dimensionality

1. Introduction

Varying coefficient model (VCM) is often used when there exists an 'index variable' modulating the effect of the predictors in a standard linear regression model. This class of models was pioneered in [1,2], with the former focusing on cross-sectional data and the latter on more general time series data. More specifically, the model can be formulated as

$$Y_i = \mathbf{g}^{\mathrm{T}}(T_i)\mathbf{X}_i + \epsilon_i, \tag{1}$$

where $\mathbf{g}(t) = (g_1(t), \ldots, g_p(t))^{\mathrm{T}}$ are the coefficients to be estimated and T is usually called an index variable in this context. We assume the observations Y_i, T_i, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}, i = 1, \ldots, n$ are independent and identically distributed. For more efficient estimation, [3] proposed a two-step local linear estimator that can attain optimal convergence rate which may be different for different functions. Many works including [4–8] demonstrated the wide applicability of the varying coefficient model.

Although in varying coefficient models, the index variable T can be multivariate, in practice one almost always uses only a one-dimensional index variable to avoid the curse of dimensionality. In spite of this, estimating p one-dimensional nonparametric functions can still result in a large effective number of parameters to estimate, especially when p is

CONTACT Rui Li liruishanghai@hotmail.com School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, People's Republic of China

relatively large. In this paper, we propose ways to reduce the effective number of parameters to make the estimation more feasible and more efficient.

One aspect of our estimation approach is that it is based on the popular assumption of sparsity and uses a penalty to take advantage of this assumption. Penalized variable selection was pioneered in [9] and many improvements were proposed including [10–12]. More related to our work here, [13–16] all used penalized variable selection for varying coefficient models. The more important aspect of our estimation approach is motivated by the need to incorporate some kind of dimension reduction, which is similar to principal component analysis. More specifically, assume we use the decomposition $\mathrm{Cov}(\mathbf{g}(T)) = \widetilde{\mathbf{U}}\mathbf{D}\widetilde{\mathbf{U}}^{\mathrm{T}}$, where $\mathbf{D} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, this implies $\mathbf{g}(T) - E[\mathbf{g}(T)] = \widetilde{\mathbf{U}}\widetilde{\boldsymbol{\gamma}}(T)$, with $\widetilde{\boldsymbol{\gamma}}(T) = \widetilde{\mathbf{U}}^{\mathrm{T}}(\mathbf{g}(T) - E[\mathbf{g}(T)])$. In other words, we have performed an orthogonalization procedure so that the covariance of $\widetilde{\boldsymbol{\gamma}}(U)) = \mathbf{D}$ have orthogonal components. This step merely performs a transformation without any achieved dimension reduction. The principal varying coefficient model is motivated by ideas similar to principal component analysis. That is, when many eigenvalues are zero or approximately zero, we can ignore the small eigenvalues. In other words, we assume $\mathbf{g}(T) - E[\mathbf{g}(T)] = \mathbf{U}\boldsymbol{\gamma}(T)$, with $\mathbf{U}$ being the first r columns of $\widetilde{\mathbf{U}}$ and $\boldsymbol{\gamma}(T)$ being the first r components of $\widetilde{\boldsymbol{\gamma}}(T)$, respectively. Equivalently, we assume the p functions $\mathbf{g}(t)$ can be written as linear combinations of components of $\boldsymbol{\gamma}$ which means we now have a smaller number of functions to estimation.

Principal VCM was pioneered in [17] who proposed to fit the model using kernel methods. However, their study is limited to fixed dimensional models, while reducing the number of nonparametric functions is of higher importance in high-dimensional setting. Motivated by this, we hereby consider unpenalized principal VCM when $p = o(n)$ (diverging with n), and penalized principal VCM when $p \gg n$. We use polynomial splines here because of its computational convenience. This computational aspect has been promoted in [18–20]. Zhao et al. [21] also used splines for latent function estimation, which is however limited to the fixed-dimensional case. For unpenalized principal VCM, we establish *non-asymptotic* bounds for the estimator, and then show that the estimator is a local minimizer of the penalized objective function (so-called weak oracle property).

He et al. [22] considered multi-response varying coefficient models also using the principal functions. Compared to their work, here we provide also non-asymptotic analysis of the penalized estimator for the *global minimizer* and also propose a new heuristic algorithm that is much faster. For technical reasons, the non-asymptotic bound is only derived for the case a sparsity constraint is added to the optimization problem. Even so, our theoretical results are non-trivial since it involves nonconvex penalties while most previous non-asymptotic bounds are derived for convex penalties.

The rest of the article is organized as follows. In Section 2, estimation methodology for principal VCM is laid out and the non-asymptotic probabilistic bound for the mean squared errors of our estimator in estimation. Penalized variable selection is incorporated in Section 3 and the weak oracle property and non-asymptotic bound are presented. Section 4 contains our simulation studies and an analysis of the genome-wide association data. We conclude with a discussion in Section 5. The technical details are mostly contained in the Appendix.

2. High-dimensional principal varying coefficient models

We are given a sample of independent and identically distributed (i.i.d.) observations $(Y_i, T_i, \mathbf{X}_i), i = 1, \dots, n$, satisfying

$$Y_i = \mathbf{g}^{\mathrm{T}}(T_i)\mathbf{X}_i + \epsilon_i$$

with $E(\epsilon_i|\mathbf{X}_i, T_i) = 0$. As discussed previously, we have $\mathbf{g}(T) = \boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\gamma}(T)$ where $\boldsymbol{\alpha} = E[\mathbf{g}(T)], \mathbf{U} \in \mathcal{R}^{p\times r}$ and we impose the orthogonality condition $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}, E[\boldsymbol{\gamma}(T)] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\gamma}(T))$ is a diagonal matrix with entries on the diagonal linear ordered from large to small. Here $\boldsymbol{\alpha}, \mathbf{U}, \boldsymbol{\gamma}(.)$ are all parameters that we need to estimate in order to fit the model. Using our parametrization, if furthermore we assume that the first $r + 1$ eigenvalues are different from each other, $\mathbf{U}, \boldsymbol{\gamma}(.)$ can be identified. However, since we are only interested in the quantities $\mathbf{g}(T)$, we do not try to enforce identifiability. Unidentifiability will not cause problem when we only try to estimate $\mathbf{g}$. Furthermore, this will avoid the orthogonality constraint which is theoretically difficult to deal with. Similarly, we can remove the intercept and at the same time add the constant function as one component of $\boldsymbol{\gamma}$ so we can write $\mathbf{g}(T) = \mathbf{U}\boldsymbol{\gamma}(T)$ for some $\mathbf{U} \in \mathcal{R}^{p\times r}$. This also means we will remove the assumption $E[\boldsymbol{\gamma}(T)] = \mathbf{0}$.

The support of T is assumed to be a compact interval. For simplicity we just assume it is $[0, 1]$. Our estimation strategy for function estimation is based on polynomial splines. Let $t_0 = 0 < t_1 < \cdots < t_{K'} < 1 = t_{K'+1}$ be a sequence of internal knots which partition $[0, 1]$ as $[t_k, t_{k+1}), k = 0, \dots, K'$. Assume we use polynomial splines with order s. Given the knots, we can construct a a B-spline basis $\{B_1(x), \dots, B_K(x)\}$ with $K = K' + s$. For theoretical convenience, the B-spline basis is normalized to satisfy $\sum_{k=1}^{K} B_k(x) = \sqrt{K}$. In implementation, any scaling in place of $\sqrt{K}$ can be used.

Denote $\mathbf{B}(\cdot) = (B_1(\cdot), \dots, B_K(\cdot))^{\mathrm{T}}$. Equipped with the spline basis, writing $g_j(\cdot) \approx \mathbf{B}^{\mathrm{T}}(\cdot)\boldsymbol{\theta}_j$, our estimation procedure is carried out by minimizing

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\mathbf{B}^{\mathrm{T}}(T_i)\boldsymbol{\theta}_j X_{ij}\right)^2.$$

For principal varying coefficient estimator with a given r, writing $\gamma_j(\cdot) \approx \mathbf{B}^{\mathrm{T}}(\cdot)\mathbf{c}_j$, and $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_r)^{\mathrm{T}}$, we minimize

$$\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\mathbf{UCB}(T_i)\right)^2,$$

over $\mathbf{U} \in \mathcal{R}^{p\times r}$, $\mathbf{C} \in \mathcal{R}^{r\times K}$. Since $\mathbf{UC}$ is a $p \times K$ matrix with rank bounded by r, by regarding $\mathbf{UC}$ as a single quantity, the problem can be rewritten in a constrained form as

$$\min_{\mathrm{rank}(\boldsymbol{\Theta})\le r}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{B}(T_i))^2, \tag{2}$$

where we write $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)_{K\times p}$, and we denote the solution by $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}_1, \dots, \widehat{\boldsymbol{\theta}}_p)$ and then the estimated functions in the VCM are $\widehat{g}_j(\cdot) = \mathbf{B}^{\mathrm{T}}(\cdot)\widehat{\boldsymbol{\theta}}_j$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}}, \mathbf{Z}_{ij} =$

$X_{ij}\mathbf{B}(T_i)$, $\mathbf{Z}_j = (\mathbf{Z}_{1j}, \dots, \mathbf{Z}_{nj})^{\mathrm{T}} \in R^{n\times K}$, $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_p)$. Then the objective function can be written as $\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}\|^2$ where $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$ and $\mathrm{vec}(\cdot)$ is the vectorization operator that stacks the columns of a matrix into a vector.

Careful readers will have noticed that problem (2) is similar to the reduced rank regression model [23]. However, the difference is that there is only one single response in standard reduced-rank regression. Furthermore, the appearance of $\boldsymbol{\Theta}$ in (2) in between $\mathbf{X}$ and $\mathbf{B}$ makes the form differ from the standard reduced-rank regression which leads to significant differences in proof.

We now turn to our theoretical results. We use the following conditions.

(A1) The density of the index variable T is bounded away from zero and infinity.

(A2) The eigenvalues of $\sum_i \mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ are nonzero.

(A3) The error ϵ is sub-Gaussian, that is $E[\exp(t\epsilon)] \le \exp(\sigma^2 t^2/2)$ for all $t > 0$, where σ is a positive constant.

Assumption (A1) is standard in the literature for varying coefficient models. The assumption (A2) is mild. In fact, under some natural assumptions, the eigenvalues of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ are bounded and bounded away from zero, with probability approaching one (see Lemma A.1). Sub-Gaussianity of error in (A3) is used to obtain our non-asymptotic bound on the stochastic term as in Wei et al. [16].

Let $\boldsymbol{\theta}_{0j}$ be the spline coefficients in some (arbitrary) spline approximation of g_j and define $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^{\mathrm{T}}, \dots, \boldsymbol{\theta}_{0p}^{\mathrm{T}})^{\mathrm{T}}$. $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of a matrix, respectively. In the rest of the paper C denotes a generic positive constant which may assume different values even on the same line.

Theorem 2.1: *Under conditions* (A1)–(A3), *for any* $t > 0$, *and any* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \dots, \boldsymbol{\theta}_p^{\mathrm{T}})^{\mathrm{T}}$ *with* $\mathrm{rank}((\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)) \le r$,

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 \le C(\|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + \|\mathbf{R}\|^2) + C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}(r(K+p-r)+t),$$

with probability at least $1 - C\exp(-Ct)$, *where* $\mathbf{R} = (R_1, \dots, R_n)^{\mathrm{T}}$ *and* $R_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{g}(T_i) - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}_0^{\mathrm{T}}\mathbf{B}(T_i)$ *is the spline approximation error. Furthermore,*

$$E\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 \le CE[\|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + \|\mathbf{R}\|^2] + C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}r(K+p-r). \qquad (3)$$

Remark 2.1: The theorem does not assume that $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{01}, \dots, \boldsymbol{\theta}_{0p})$ has rank bounded by r. When $\boldsymbol{\Theta}_0$ has rank bounded by r, we can choose $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ which leads to $\|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| = 0$. In general, the term $\|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| = 0$ is small if $\boldsymbol{\Theta}_0$ is close to a rank r matrix.

Remark 2.2: Under mild assumptions, the eigenvalues of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ are bounded and bounded away from zero (see Lemma A.1 in the Appendix), with probability approaching one. Furthermore, it is easy to see that if we assume the following:

(A4) The functions g_j are in the Hölder space of order $d > 1$. That is $|g_j^{(m)}(x) - g_j^{(m)}(y)| \le C|x-y|^r$ for $d = m + r$ and m is the largest integer strictly smaller than d, where $g_j^{(m)}$ is the mth derivative of g_j. The order of the spline is larger than $d + 1/2$.

then $\|\mathbf{R}\|^2 = O_p(npK^{-2d})$. In this case, if $\text{rank}(\boldsymbol{\Theta}_0) \leq r$, we have $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 = O_p(pK^{-2d} + r(K + p - r)/n)$. The two terms can be regarded as squared bias and variance, respectively.

Remark 2.3: The term $r(K + p - r)$ is the degrees of freedom, or the effective number of parameters, for $\boldsymbol{\Theta}$. This can be justified by our proof in terms of covering number. Informally, the number of parameters in reduced rank regression can be counted as follows. Assuming the first r columns of $\boldsymbol{\Theta}$ are linearly independent, we have the parametrization $\boldsymbol{\Theta} = (\mathbf{B}_1, \mathbf{B}_1\mathbf{B}_2^{\mathrm{T}})$ for a rank r matrix $\boldsymbol{\Theta}$, where $\mathbf{B}_1$ is a $K \times r$ matrix and $\mathbf{B}_2$ is a $(p - r) \times r$ matrix, and the effective number of parameters is then $r(K + p - r)$.

In the above discussions, r is assumed to be given. We can determine r in a data-driven manner using Bayesian information criterion (BIC [24,25]):

$$\text{BIC}(r) := \log\left(\|\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\theta}}\|^2\right) + (2n)^{-1}(\log n)(r(p + K - r)),$$

where $\widehat{\boldsymbol{\Theta}}$ is the estimate using a given value of r.

3. Variable selection in PVCM

Dimension reduction alone is not sufficient if the number of predictors is relatively large. We can further add penalized variable selection to take advantage of sparsity assumption. Penalization can be used to identify significant functions within the p components $g_1, \ldots, g_p$. More concretely, we can define the sparse estimator $\widetilde{\boldsymbol{\theta}}$ by

$$\min_{\text{rank}(\boldsymbol{\Theta}) \leq r} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_j\|), \tag{4}$$

where λ is a tuning parameter controlling sparsity. Among the many possible penalty functions that has been proposed in the literature, here for specificity we use the SCAD penalty [10] defined by its first derivative

$$p'_\lambda(x) = \lambda\left\{I(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a - 1)\lambda}I(x > \lambda)\right\},$$

with $a > 2$ and $p_\lambda(0) = 0$. a can be treated as another tuning parameter, but as is usually done for simplicity we will set $a = 3.7$ which is often used. There are other possibilities in the choice of penalty. For example, minimax concave penalty [26] or adaptive lasso [12] can also be used and in general the view is that there is no dominant penalty that is always better and the choice is often based on taste and convenience of each individual researcher.

Under our sparsity assumption, we assume that $g_1, \ldots, g_s$ are nonzero for easy of presentation. The following result is called the local oracle property by Fan and Lv [27]. It means the estimator using only the first s components of $\mathbf{X}_i$, is a local minimizer of (4).

We further state the following additional conditions:

(B1) $n\lambda > 2\hat{\xi}$ with probability tending to 1, where

$$\hat{\xi} = \max_j \|\mathbf{Z}_j^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\theta}}^o)\|.$$

(B2) $\|\widehat{\boldsymbol{\theta}}_j^o\| > a\lambda$, for $j \leq s$.

In the following, we write the oracle estimator as $\widehat{\boldsymbol{\theta}}^o = ((\widehat{\boldsymbol{\theta}}^o_{(1)})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\theta}}^o_{(1)}$ is obtained by (2) using only the first s components of $\mathbf{X}_i$.

Theorem 3.1: *Let $\mathcal{A}(\lambda)$ be the set of local minimizers of* (4) *for a given λ. Under assumptions* (B1) *and* (B2), *we have*

$$P(\widehat{\boldsymbol{\theta}}^o \in \mathcal{A}(\lambda)) \to 1$$

as $n \to \infty$.

Remark 3.1: Conditions (B1) and (B2) are non-primitive assumptions, but for the convenience of proof. More primitive assumptions can be used to replace these. For example, under assumptions as mentioned in Remark 2.1, if $rank(\boldsymbol{\Theta}_0) \le r$, then $\|\mathbf{Z}\widehat{\boldsymbol{\theta}}^o - \mathbf{Z}\boldsymbol{\theta}_0\|^2 = O_p(r(K+s) + \|\mathbf{R}\|^2)$ and $\|\widehat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}_0\|^2 = O_p(r(K+s)/n + \|\mathbf{R}\|^2/n)$ if $\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}_{(1)}\mathbf{Z}_{(1)}/n)$ is bounded away from zero, where $\mathbf{Z}_{(1)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_s)$. Thus we can derive that (B1) and (B2) would be implied by $\lambda >> \sqrt{K\log(Kp)/n} + \sqrt{r(K+s)/n} + \|\mathbf{R}\|/\sqrt{n}$ and $\|\boldsymbol{\theta}_{0j}\| >> a\lambda$ for $j \le s$.

In fact, we note $\hat{\xi} = \max_j \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{e}\| + \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}(\widehat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}_0)\|$. If the errors are sub-Gaussian, we have

$$\max_j \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{e}\| = \sqrt{K}\max_{k,j}\left|\sum_i X_{ij}B_k(T_i)\epsilon_i\right| = O_p(\sqrt{nK\log(Kp)}),$$

using Lemma 2.2.2 of [28]. This implies

$$\begin{aligned}\hat{\xi} &= O_p(\sqrt{nK\log(Kp)} + \sqrt{n}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}_0)\|)\\ &= O_p(\sqrt{nK\log(Kp)} + \sqrt{nr(K+s)} + \sqrt{n}\|\mathbf{R}\|).\end{aligned}$$

Furthermore, it is easy to see that $\|\boldsymbol{\theta}_{0j}\| >> a\lambda$ and $\|\widehat{\boldsymbol{\theta}}^o_j - \boldsymbol{\theta}_{0j}\| = o_p(\lambda)$ implies (B2).

Based on the remarks above, we can derive the next Corollary.

Corollary 3.1: *Suppose* (A1)–(A3) *holds. Then, $\widehat{\boldsymbol{\theta}}^o$ is a local minimizer of* (4) *with probability tending to* 1 *if*

$$\lambda >> \sqrt{K\log(Kp)/n} + \sqrt{r(K+s)/n} + \|\mathbf{R}\|/\sqrt{n}$$

and

$$\lambda << \min_{j\le s}\|\boldsymbol{\theta}_{0j}\|,$$

provided the eigenvalues of $\mathbf{Z}^{\mathrm{T}}_{(1)}\mathbf{Z}_{(1)}/n$ are bounded and bounded away from zero with probability approaching one, and $rank(\boldsymbol{\Theta}_0) \le r$.

Remark 3.2: The conditions on λ implicitly constrains the dimension p. In fact, for λ to exist that satisfies the conditions, we must impose $\sqrt{K\log(Kp)/n} << \min_{j\le s}\|\boldsymbol{\theta}_{0j}\|$. Assuming $\min_{j\le s}\|\boldsymbol{\theta}_{0j}\|$ is bounded, this translate to $\log p << n/K$.

The above result only shows that the oracle estimator is a local minimizer of (4). In the following, we present non-asymptotic error bounds for the *global minimizer* of (4) similar to Theorem 2.1. The result does not require assumptions on the minimum signal size as in Corollary 3.1. However, it requires a stronger eigenvalue condition for $\mathbf{Z}$ and also we need a slight change of the minimization problem to

$$\min_{\operatorname{rank}(\boldsymbol{\Theta})\leq r, s(\boldsymbol{\theta})\leq s_{\max}} \|\mathbf{Y}-\mathbf{Z}\boldsymbol{\theta}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_j\|), \tag{5}$$

where $s(\boldsymbol{\theta})$ is the number of nonzero $\boldsymbol{\theta}_j$ in $\boldsymbol{\theta}$ (equivalently the number of nonzero columns of $\boldsymbol{\Theta}$). That is, an additional sparsity constraint is added for technical reasons in the proof. We require that $s_{\max}$ is an upper bound of true s, and both are allowed to diverge with n. Currently, we do not know how to show the bound for the global minimizer if this constraint is removed. In practice, if we choose λ such that the estimator is sparse enough, such a constraint will not be binding and thus does not change the estimator. Thus in implementation we still ignore such a constraint. Although unsatisfactory, such a sparsity constraint was also used in [29,30] to make the proofs feasible.

We now require the following assumption in place of (A2). The assumption is a sparse eigenvalue condition often used in penalized variable selection modelling.

(A2′) Let $S \subset \{1,\ldots,p\}$ and let $\mathbf{Z}_S$ denote the columns of $\mathbf{Z}$ associated with predictors in S. We assume eigevalues of $\mathbf{Z}_S^{\mathrm{T}}\mathbf{Z}_S/n$ are uniformaly bounded and bounded away from zero over the set $\{S : |S| \leq 2s_{\max}\}$.

Theorem 3.2: *Under conditions* (A1), (A2') *and* (A3), *with* $\lambda = C\sqrt{(r+\log p)/n}$ *for sufficiently large* $C > 0$, *we have for* $\widehat{\boldsymbol{\theta}}$ *the minimizer of* (5) *and any* $\boldsymbol{\theta}$ *satisfying* $\operatorname{rank}(\boldsymbol{\Theta}) \leq r$, $s(\boldsymbol{\theta}) \leq s_{\max}$,

$$E\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \leq CE[\|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + \|\mathbf{R}\|^2] + Cr(K+s(\boldsymbol{\theta})-r) + Cs(\boldsymbol{\theta})\log p.$$

Note that compared to Theorem 2.1, *the extra term* $s(\boldsymbol{\theta})\log p$ *can roughly be attributed to high-dimensionality. By taking* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ *when* $\operatorname{rank}(\boldsymbol{\Theta}_0) \leq r$ *and* $s(\boldsymbol{\theta}_0) = s \leq s_{\max}$, *we obtain*

$$E\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \leq CE[\|\mathbf{R}\|^2] + Cr(K+s-r) + Cs\log p.$$

In this paper, we use the cubic splines, which is probably the most popular choice in the literature. Furthermore, to ease the computational burden, we fix $K = 6$. One could select K using cross-validation or some information criterion but it would increase the computational burden with no appreciable numerical advantages in our experience. For complicated models, using a fixed K is not uncommon, and this is the case, for example, in [31,32]. This choice of K is small enough to avoid overfitting in typical problems with sample size not too small, and big enough to flexibly approximate many smooth functions accurately.

Finally, we can choose the tuning parameter λ, together with r, using the Bayesian Information Criterion

$$\operatorname{BIC}(r,\lambda) := \log\left(\|\mathbf{Y}-\mathbf{Z}\widetilde{\boldsymbol{\theta}}\|^2\right) + (2n)^{-1}(\log n)(r(K+\widetilde{s}-r)). \tag{6}$$

In the above, $\widetilde{\boldsymbol{\theta}}$ is the estimate based on (4), and $\widetilde{s}$ is the estimate for s. These estimates are obtained with the given tuning parameters (r,λ).

4. Numerical results

4.1. Algorithm

In this section, we give some details in implementing the penalized estimate in (4). To compute, given r, we firstly write $\boldsymbol{\Theta}^{\mathrm{T}} = \mathbf{D}\mathbf{E}^{\mathrm{T}}$ for $\mathbf{D} \in \mathcal{R}^{p\times r}$, $\mathbf{E} \in \mathcal{R}^{K\times r}$. We also put the constraint that $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$. With this orthogonal constraint, $\|\boldsymbol{\theta}_j\|$ is equal to $\|\mathbf{d}_j\|$, where $\mathbf{d}_j$ is the jth row of $\mathbf{D}$. Then we can state the estimation problem in the following form:

$$\min_{\mathbf{D},\mathbf{E}} \left\{ \sum_i (Y_i - ((\mathbf{B}^{\mathrm{T}}(T_i)\mathbf{E}) \otimes \mathbf{X}_i^{\mathrm{T}})\mathrm{vec}(\mathbf{D}))^2 + n \sum_j p_\lambda(\|\mathbf{d}_j\|) \right\}.$$

The numerical procedure to solve the above is simply to minimize over $\mathbf{D}$ with $\mathbf{E}$ fixed, and vice versa, done iteratively till convergence.

Given $\mathbf{E}$, minimizing over $\mathbf{D}$ ($\mathrm{vec}(\mathbf{D})$)

$$\min_{\mathbf{D}} \left\{ \sum_i (Y_i - ((\mathbf{B}^{\mathrm{T}}(T_i)\mathbf{E}) \otimes \mathbf{X}_i^{\mathrm{T}})\mathrm{vec}(\mathbf{D}))^2 + n \sum_j p_\lambda(\|\mathbf{d}_j\|) \right\} \tag{7}$$

is the same as the group SCAD problem of least square, which can be implemented by the group coordinate descent algorithm [33].

Given $\mathbf{D}$, minimizing over $\mathbf{E}$ is minimizing

$$\sum_i (Y_i - (\mathbf{B}^{\mathrm{T}}(T_i) \otimes (\mathbf{X}_i^{\mathrm{T}}\mathbf{D}))\mathrm{vec}(\mathbf{E}^{\mathrm{T}}))^2 \tag{8}$$

with the constraint $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$. It seems very hard to take into account the constraint $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$ directly. Thus we use a more heuristic approach. First the minimization problem is solved ignoring the constraint to get $\mathbf{E}$, which obviously has a closed-form solution since the problem is quadratic. Then we modify $\mathbf{E}$ so that it satisfies the constraint. To carry out this step, we first write $\mathbf{E} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ based on SVD, where $\mathbf{U}$ is $K \times r$ and $\mathbf{V}$ is $r \times r$ ($r \leq K$), and then set $\mathbf{E}$ to be $\mathbf{U}\mathbf{V}^{\mathrm{T}}$. This simple strategy works well in our experience.

For given r and tuning parameter λ, we update (7) and (8) alternatively until the difference of loss function ($\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{B}(T_i))^2$) between two iterations is smaller than the given precision (< 0.001). Then, we use the BIC given in (6) to select the optimal number of principal components and the optimal tuning parameter by the two-dimensional grid method. Simulation studies show that the above strategy works well. He et al. [22] used a more principled approach by optimization on the Stiefel manifold combined with manifold gradient descent. Their algorithm is much more complicated and much slower, although with some limited convergence guarantees. For example, when $n = 400$, it takes about 5 h to complete our simulation below using the proposed algorithm, while the algorithm of [22] takes about 10 times longer (by private communication, we learned that [22] has to use a cluster with multiple nodes in order to complete their simulations), with almost no differences in the produced results.

4.2. Simulation study

We carry out some simulation studies which investigates both the estimation error and rank selection error for our model. The data are generated from the following model:

$$Y_i = \mathbf{g}^{\mathrm{T}}(T_i)\mathbf{X}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbf{g}(u) = \mathbf{B}\boldsymbol{\gamma}(u)$, $\boldsymbol{\gamma}(u) = (2\sin(2\pi u), 3\exp(u-0.5))^{\mathrm{T}}$, $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)$ with $\mathbf{b}_1 = (\underbrace{1,\ldots,1}_{5}, \underbrace{-1,\ldots,-1}_{5}, \underbrace{0,\ldots,0}_{p-10})^{\mathrm{T}}$ and $\mathbf{b}_2 = (\underbrace{0,\ldots,0}_{5}, \underbrace{1,\ldots,1}_{5}, \underbrace{0,\ldots,0}_{p-10})^{\mathrm{T}}$. The covariates $\mathbf{X}_i$ are generated from a multivariate Gaussian distribution with the standard normal distribution as their marginal distributions and $\mathrm{Cov}(X_{ij_1}, X_{ij_2}) = 0.8^{|j_1-j_2|}$ for $1 \le j_1, j_2 \le p$, and T_i is simulated from $U[0,1]$, and the random error e_i follows the standard normal distribution. We focus on the sample size $n = 200, 300, 400$ and dimensionality $p = 100, 200$, *and* 400.

We note that the generating model have two latent functions, and we will also look at the performances of BIC (6) as p changes. To this end, we report the proportion of correctly identifying the number of principal components and use the 'FNR' (false negative rate) and 'FPR' (false positive rate) to evaluate the selection performance of non-parametric functions, where

$$\mathrm{FNR} = \frac{\#(\text{number of truly nonzero functions estimated as zeros})}{\#(\text{number of truly nonzero functions})}$$

and

$$\mathrm{FPR} = \frac{\#(\text{number of truly zero functions estimated as nonzeros})}{\#(\text{number of truly zero functions})}.$$

Besides, we report the average absolute error (AAE) for varying coefficient functions

$$\mathrm{AAFE} = \frac{1}{n}\sum_{i=1}^{n} |\widehat{\mathbf{g}}(T_i) - \mathbf{g}_0(T_i)|,$$

which indicates the overall error, where $|\mathbf{a}| = (|a_1| + \cdots + |a_p|)/p$ for any vector $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$. The simulation results over 200 replications are shown in Table 1. For comparisons, we also report the corresponding results of the ordinary varying coefficient model without dimensional reduction method [16] implemented by the R package grpreg [34].

From Table 1, we can see that the proportions of correctly identifying the number of principal components are very high in all cases. Moreover, in terms of FNR and FPR, our proposed principal component estimate can simultaneously select important varying coefficient functions and remove redundant covariates with high accuracy rate. On the other hand, the FNRs are larger than 0.2 for the ordinary varying coefficient model, which means that the VCM without dimensional reduction often misses important variables under our simulation settings. As a result, the AAE of our proposed model is significantly smaller than the corresponding result of ordinary VCM. In summary, the proposed penalized principal estimation algorithm is satisfactory for both estimation and selection performance.

Table 1. Estimation and selection results of simulation studies based on the 200 replications.

	p	Correct r	FNR		FPR		AAE	
			PVCM	VCM	PVCM	VCM	PVCM	VCM
$n = 200$	100	98%	0.0000	0.2725	0.0005	0.0006	0.0086	0.1144
	200	96%	0.0045	0.2850	0.0008	0.0009	0.0118	0.0623
	400	92%	0.0750	0.3150	0.0009	0.0009	0.0186	0.0293
$n = 300$	100	99%	0.0000	0.2460	0.0003	0.0000	0.0074	0.0972
	200	98%	0.0000	0.2560	0.0005	0.0003	0.0020	0.0258
	400	98%	0.0015	0.2855	0.0008	0.0005	0.0018	0.0249
$n = 400$	100	100%	0.0000	0.2260	0.0002	0.0000	0.0062	0.0895
	200	99%	0.0000	0.2450	0.0006	0.0002	0.0034	0.0459
	400	98%	0.0010	0.2725	0.0006	0.0006	0.0015	0.0228

Table 2. Estimation and selection results for genome-wide association data.

	r_{opt}	λ_{opt}	nonzero functions	fitting error
PVCM	4	0.0037	19	0.0664
VCM	–	0.0149	16	0.0926

Table 3. AAPEs for PVCM and VCM based on 200 splits of the data.

Size of		AAPE	
Training set	Test set	PVCM	VCM
200	127	0.2407	0.2490
250	77	0.1253	0.1891
300	27	0.1127	0.1806

4.3. *Empirical application*

The method is applied to a genome-wide association data. There are 54,241 single nucleotide polymorphisms (SNPs) in this dataset and we analyse it by the principal varying coefficient models. We try to select significant SNPs ($\mathbf{X}$) as predictors, with the indicator variable being 'birth weight'. The response variable we want to predict is the 'trait weaning weight'. The data can be accessed from http://www.ncbi.nlm.nih.gov/gds.

For this data, the three genotypes are coded as (1, 1), (0, 1) and (0, 0). Since the dimension of the predictor is excessive, an independence screening method [35] is first applied we only retain 500 SNPs before fitting our model. Finally the sample size in our data is 327 with 500 SNPs used. With $p = 500$ the real dimension of the problem in implementation using splines is $Kp = 3000$. This already makes the computational burden very heavy and thus we do not try higher dimensions. We found that four principal components are identified and there are 19 SNPs selected as important variables. These results and the absolute fitting error ($\frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$) have been shown in Table 2, where we also list the corresponding results for ordinary VCM. It is clear that PVCM outperforms VCM in terms of absolute fitting error for genome-wide association data.

Furthermore, we look at the performance of PVCM in terms of prediction error, in comparison to the standard VCM without dimension reduction. For this purpose, we randomly split the entire dataset into two parts. One part is used to fit the model and the other for calculating the prediction error. Table 3 reported the average absolute prediction errors (AAPE) based on 200 splits of the data. Different sizes of the partition are used as shown in the table. We see that PVCM exhibits better performance compared to standard VCM.

5. Conclusion

In this paper, we combine ideas from penalization and reduced rank regression for high-dimensional varying coefficient models. The proposed method can make the number of coefficient functions to be estimated smaller and in our numerical studies we show that it can obviously reduce the errors in model fitting and in prediction.

We established local oracle property of the penalized estimator in this paper. We also derived oracle bounds for the global estimator. However, one unsatisfactory aspect in our theory is that we have not been able to say anything about the estimator returned by the algorithm since our theory does not promise the computed solution is a local minimizer, let alone a global minimizer. On the other hand, our Monte Carlo studies suggest that the algorithm works satisfactorily in practice.

Acknowledgments

The authors sincerely thank the Editor Professor Meister, an Associate Editor and two reviewers for their insightful comments that greatly improved the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Li Rui's research was supported by National Social Science Foundation of China (No.17BTJ025). The research of Heng Lian is supported by City University of Hong Kong Start-up grant 7200521, by Hong Kong RGC general research fund 11301718 and 11300519, by Project 11871411 from NSFC and by the Shenzhen Research Institute, City University of Hong Kong. The research of Fode Zhang is partially supported by the Fundamental Research Funds for the Central Universities (No. JBK1901053, JBK1806002 and JBK140507) of China. The research of Xuejun Wang is partially supported by NSFC (11671012). The research of Weihua Zhao is partially supported by National Social Science Fund (15BTJ027).

References

[1] Chen R, Tsay RS. Functional-coefficient autoregressive models. J Am Stat Assoc. 1993;88(421): 298–308.
[2] Hastie T, Tibshirani R. Varying-coefficient models. J R Stat Soc Ser B-Methodol. 1993;55(4):757–796.
[3] Fan JQ, Zhang WY. Statistical estimation in varying coefficient models. Ann Stat. 1999;27(5):1491–1518.
[4] Fan JQ, Zhang JT. Two-step estimation of functional linear models with applications to longitudinal data. J R Stat Soc Ser B-Stat Methodol. 2000;62:303–322.
[5] Fan JQ, Zhang W. Statistical methods with varying coefficient models. Stat Interface. 2008;1:179–195.
[6] Hoover DR, Rice JA, Wu CO, et al. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998;85(4):809–822.

[7] Huang JHZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika. 2002;89(1):111–128.

[8] Xia Y, Zhang W, Tong H. Efficient estimation for semivarying-coefficient models. Biometrika. 2004;91:661–681.

[9] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B-Methodol. 1996;58(1):267–288.

[10] Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.

[11] Huang J, Horowitz JL, Ma SG. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Stat. 2008;36(2):587–613.

[12] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429.

[13] Lian H. Variable selection for high-dimensional generalized varying-coefficient models. Stat Sin. 2012b;22:1563–1588.

[14] Wang HS, Xia YC. Shrinkage estimation of the varying coefficient model. J Am Stat Assoc. 2009;104(486):747–757.

[15] Wang LF, Li HZ, Huang JHZ. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J Am Stat Assoc. 2008;103(484):1556–1569.

[16] Wei F, Huang J, Li HZ. Variable selection and estimation in high-dimensional varying-coefficient models. Stat Sin. 2011;21:1515–1540.

[17] Jiang Q, Wang H, Xia Y, et al. On a principal varying coefficient model. J Am Stat Assoc. 2013;108(501):228–236.

[18] Li Q. Efficient estimation of additive partially linear models. Int Econ Rev. 2000;41(4): 1073–1092.

[19] Wang L, Liu X, Liang H, et al. Estimation and variable selection for generalized additive partial linear models. Ann Stat. 2011;39(4):1827–1851.

[20] Wang L, Xue L, Qu A, et al. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. Ann Stat. 2014;42(2):592–624.

[21] Zhao W, Jiang X, Lian H. A principal varying-coefficient model for quantile regression: joint variable selection and dimension reduction. Comput Stat Data Anal. 2018;127:269–280.

[22] He K, Lian H, Ma S, et al. Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. J Am Stat Assoc. 2018;113:746–754.

[23] Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann Math Stat. 1951;29:327–351.

[24] Horowitz JL, Lee S. Nonparametric estimation of an additive quantile regression model. J Am Stat Assoc. 2005;100(472):1238–1249.

[25] Lian H. A note on the consistency of Schwarz's criterion in linear quantile regression with the scad penalty. Stat Probab Lett. 2012a;82(7):1224–1228.

[26] Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.

[27] Fan J, Lv J. Non-concave penalized likelihood with np-dimensionality. IEEE Trans Inf Theory. 2011;57(8):5467–5484.

[28] van der Vaart AW, Wellner JA. Weak convergence and empirical processes. New York: Springer Verlag; 1996.

[29] Fan Y, Lv J. Asymptotic equivalence of regularization methods in thresholded parameter space. J Am Stat Assoc. 2013;108(503):1044–1061.

[30] Zheng Z, Fan Y, Lv J. High dimensional thresholded regression and shrinkage effect. J R Stat Soc Ser B (Stat Methodol). 2014;76(3):627–649.

[31] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Am Stat Assoc. 2011;106(494):544–557.

[32] Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. Ann Stat. 2010;38(4):2282–2313.

[33] Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput. 2015;25(3):173–187.

[34] Breheny P, Zeng Y, Regularization paths for regression models with grouped covariates. R package version 3.1-4; 2018.
[35] Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. J Am Stat Assoc. 2014;109(507):1270–1284.
[36] Huang JHZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Stat Sin. 2004;14(3):763–788.
[37] Szarek SJ. Nets of grassmann manifold and orthogonal group. In: Proceedings of research workshop on Banach space theory, Iowa City, IA, Vol. 169; 1982. p. 185.
[38] She Y. Thresholding-based iterative selection procedures for model selection and shrinkage. Electron J Stat. 2009;3:384–415.
[39] She Y. Selective factor extraction in high dimensions. Biometrika. 2017;104:97–110.

Appendix

Here we present a lemma that guarantees the eigenvalues of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ are bounded and bounded away from zero with probability approaching one. The lemma is basically Lemma A.2 of [36] and thus the proof is omitted.

Lemma A.1: *Assume* (i) *the density of T is continuous and bounded and bounded away from zero*; (2) *The eigenvalues of $E[\mathbf{X}\mathbf{X}^{\mathrm{T}}|T=t]$ resides between two fixed constants, uniformly on* $[0,1]$; (iii) $Kp\log(Kp)/n = o(1)$. *Then with probability approaching 1, the eigenvalues of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ are also between two fixed constants.*

Proof of Theorem 2.1: Given a $\boldsymbol{\Theta} \in R^{K\times p}$ with rank at most r, and $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$, we have

$$\|\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\theta}}\|^2 \le \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}\|^2.$$

Working out the squares, we obtain

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 \le \|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + 2\langle \mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \mathbf{e} + \mathbf{R}\rangle,$$

where $\mathbf{e} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$.

Let $\Gamma = \{\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\theta}/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})} : \boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_p^{\mathrm{T}})^{\mathrm{T}}, \|\boldsymbol{\theta}\| = 1, \mathrm{rank}((\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)) \le r\}$. We first show that the covering entropy $\log N(\epsilon, \Gamma, l_2) \le r(K+p-r)\log(C/\epsilon)$. In fact, for $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\theta}/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})} \in \Gamma$, we can write $\boldsymbol{\Theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) = \mathbf{D}\mathbf{A}^{\mathrm{T}}$, $\mathbf{D} \in R^{K\times r}$, $\mathbf{A} \in R^{p\times r}$, $\|\mathbf{D}\|_F = 1$, $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$. The covering number, under the Frobenius norm for $\mathbf{D}$ satisfying these assumptions is $(C/\epsilon)^{Kr}$. For the covering number of $\mathbf{A}$ satisfying the orthogonal condition, we use the distance $d(\mathbf{A}_1, \mathbf{A}_2) = \|\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}\|_{op}$, and then by the result of [37] the covering number has a bound given by $(C/\epsilon)^{r(p-r)}$. Using that

$$\begin{aligned}\|\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{D}_2\mathbf{A}_2^{\mathrm{T}}\|_F &\le \|\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{D}_1\mathbf{A}_1\mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}\|_F + \|\mathbf{D}_1\mathbf{A}_1\mathbf{A}_2\mathbf{A}_2^{\mathrm{T}} - \mathbf{D}_2\mathbf{A}_2^{\mathrm{T}}\|_F \\ &\le \|\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}}\|_F\|\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}\|_{op} + \|\mathbf{D}_1\mathbf{A}_1\mathbf{A}_2 - \mathbf{D}_2\|_F,\end{aligned}$$

we have $\log N(\epsilon, \Gamma, l_2) \le r(K+p-r)\log(C/\epsilon)$.

Furthermore, since ϵ_i is sub-Gaussian, we have

$$E\exp\{t\langle \mathbf{e}, \boldsymbol{\eta}\rangle\} \le \exp\{Ct^2\|\boldsymbol{\eta}\|^2\}.$$

Using Dudley's integral entropy bound, we get

$$P\left(\sup_{\boldsymbol{\eta}\in\Gamma}\langle \mathbf{e}, \boldsymbol{\eta}\rangle > Ct + C\int_0^{\sigma}\sqrt{r(K+p-r)\log\left(\frac{C}{\epsilon}\right)}\,\mathrm{d}\epsilon\right) \le C\exp(-Ct^2).$$

Since $\int_0^\sigma \sqrt{r(K+p-r)\log(\frac{C}{\epsilon})}\,\mathrm{d}\epsilon \asymp \sqrt{r(K+p-r)}$, we get

$$\begin{aligned}\langle \mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}, \mathbf{e}\rangle &\le C\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|(t+\sqrt{r(K+p-r)})\\ &\le \frac{C}{\sqrt{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|(t+\sqrt{r(K+p-r)}),\end{aligned}$$

with probability at least $1-C\exp(-Ct^2)$.

Trivially, we have $\langle \mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}), \mathbf{R}\rangle \le \|\mathbf{R}\|\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|$. Thus with probability at least $1-C\exp(-Ct^2)$, we have

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \le \|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + C\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|\left(\|\mathbf{R}\| + \frac{\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}}{\sqrt{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}}(\sqrt{r(K+p-r)}+t)\right).$$

Using Cauchy-Schwarz inequality and that $\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\| \le \|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\| + \|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|$, we get

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \le C\|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + C\|\mathbf{R}\|^2 + C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}(r(K+p-r)+t^2),$$

with probability at least $1-C\exp(-Ct^2)$, this proved the first statement.

Letting $A := \|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 - C\|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 - C\|\mathbf{R}\|^2 - C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}r(K+p-r)$, the above shows $P(A \ge C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}t) \le C\exp(-Ct)$ and thus

$$E[A^2] = \int_0^\infty P(A^2 > t)\,\mathrm{d}t \le \int_0^\infty C\exp\left(-C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}\sqrt{t}\right)\mathrm{d}t \le C\left(\frac{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}\right)^2 \le C.$$

Thus

$$E\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \le CE[\|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + \|\mathbf{R}\|^2] + C\frac{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}r(K+p-r)+C.$$

To see (3), we only need to note that it is trivially true if $r=0$ (since in this case $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} = \mathbf{0}$), and when $r \ge 1$, obviously the last term (constant C) in the displayed equation above can be omitted. ■

Proof of Theorem 3.1: Recall that we define the oracle estimator as $\widehat{\boldsymbol{\theta}}^o = ((\widehat{\boldsymbol{\theta}}^o_{(1)})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\theta}}^o_{(1)}$ is obtained by (2) using only the first s components of $\mathbf{X}_i$. Since $\|\widehat{\boldsymbol{\theta}}^o_j\| > a\lambda$ for $j \le s$ there is a neighbourhood of $\widehat{\boldsymbol{\theta}}^o_{(1)}$ such that $\|\boldsymbol{\theta}_{(1)j}\| > a\lambda$ for any $\boldsymbol{\theta}_{(1)}$ in this neighbourhood, with rank of $\mathrm{vec}^{-1}(\boldsymbol{\theta}_{(1)})$ bounded by r, where we use vec^{-1} to denote the operation of rearranging of a Kp-vector into a $K \times p$ matrix, which implies $\sum_{j=1}^s p_\lambda(\|\widehat{\boldsymbol{\theta}}^o_j\|) = \sum_{j=1}^s p_\lambda(\|\boldsymbol{\theta}_{(1)j}\|)$. Thus, using that $\widehat{\boldsymbol{\theta}}^o$ is the oracle estimator, we have

$$\|\mathbf{Y}-\mathbf{Z}\widehat{\boldsymbol{\theta}}^o\|^2 + n\sum_{j=1}^p p_\lambda(\|\widehat{\boldsymbol{\theta}}^o_j\|) \le \|\mathbf{Y}-\mathbf{Z}\boldsymbol{\theta}\|^2 + n\sum_{j=1}^p p_\lambda(\|\boldsymbol{\theta}_j\|),$$

for any $\boldsymbol{\theta}$ of the form $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathrm{T}}_{(1)}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\theta}_{(1)}$ in a small neighbourhood of $\widehat{\boldsymbol{\theta}}^o_{(1)}$ and $\mathrm{rank}(\mathrm{vec}^{-1}(\boldsymbol{\theta}_{(1)})) \le r$.

What is left to show is that if $c > 0$ is small enough, for any $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathrm{T}}_{(1)}, \boldsymbol{\theta}^{\mathrm{T}}_{(2)})^{\mathrm{T}}$ that satisfies $\mathrm{rank}(\mathrm{vec}^{-1}(\boldsymbol{\theta})) \le r$, $\|\boldsymbol{\theta}_{(1)} - \widehat{\boldsymbol{\theta}}^o_{(1)}\| \le c$, $\|\boldsymbol{\theta}_{(2)}\| \le c$, we have

$$\|\mathbf{Y}-\mathbf{Z}\widetilde{\boldsymbol{\theta}}\|^2 + n\sum_{j=1}^p p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|) \le \|\mathbf{Y}-\mathbf{Z}\boldsymbol{\theta}\|^2 + n\sum_{j=1}^p p_\lambda(\|\boldsymbol{\theta}_j\|), \tag{A1}$$

where $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^{\mathrm{T}}_{(1)}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$.

The difference between the two sides above is

$$\|\mathbf{Z}_{(2)}\boldsymbol{\theta}_{(2)}\|^2 - 2\langle \mathbf{Y} - \mathbf{Z}_{(1)}\boldsymbol{\theta}_{(1)}, \mathbf{Z}_{(2)}\boldsymbol{\theta}_{(2)}\rangle + n\sum_{j=s+1}^{p} p_\lambda(\|\boldsymbol{\theta}_j\|),$$

where $\mathbf{Z}_{(1)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_s)$ and $\mathbf{Z}_{(2)} = (\mathbf{Z}_{s+1}, \ldots, \mathbf{Z}_p)$. We have $\|\mathbf{Z}_{(2)}\boldsymbol{\theta}_{(2)}\|^2 \leq \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\boldsymbol{\theta}_{(2)}\|^2$. Furthermore,

$$\begin{aligned} |\langle \mathbf{Y} - \mathbf{Z}_{(1)}\boldsymbol{\theta}_{(1)}, \mathbf{Z}_{(2)}\boldsymbol{\theta}_{(2)}\rangle| &\leq |\langle \mathbf{Z}_{(2)}^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}_{(1)}\widehat{\boldsymbol{\theta}}_{(1)}^{o}), \boldsymbol{\theta}_{(2)}\rangle| + |\langle \mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(1)}(\widehat{\boldsymbol{\theta}}_{(1)}^{o} - \boldsymbol{\theta}_{(1)}), \boldsymbol{\theta}_{(2)}\rangle| \\ &\leq \widehat{\xi}\sum_{j=s+1}^{p}\|\boldsymbol{\theta}_j\| + \lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\boldsymbol{\theta}}_{(1)}^{o} - \boldsymbol{\theta}_{(1)}\|\|\boldsymbol{\theta}_{(2)}\|, \end{aligned}$$

where we use $\lambda_{\max}(.)$ also to denote the top singular value of a (non-symmetric) matrix. If c is small enough, $p_\lambda(\|\boldsymbol{\theta}_j\|) = \lambda\|\boldsymbol{\theta}_j\|$. Thus

$$\begin{aligned} &\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\boldsymbol{\theta}_j\|) - \|\mathbf{Y} - \mathbf{Z}\widetilde{\boldsymbol{\theta}}\|^2 - n\sum_{j=1}^{p} p_\lambda(\|\widetilde{\boldsymbol{\theta}}_j\|) \\ &\geq (n\lambda - 2\hat{\xi})\sum_{j>s}\|\boldsymbol{\theta}_j\| - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\boldsymbol{\theta}}_{(1)}^{o} - \boldsymbol{\theta}_{(1)}\|\|\boldsymbol{\theta}_{(2)}\| - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\boldsymbol{\theta}_{(2)}\|^2 \\ &\geq (n\lambda - 2\hat{\xi})\|\boldsymbol{\theta}_{(2)}\| - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\boldsymbol{\theta}}_{(1)}^{o} - \boldsymbol{\theta}_{(1)}\|\|\boldsymbol{\theta}_{(2)}\| - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\boldsymbol{\theta}_{(2)}\|^2, \end{aligned}$$

where the last inequality is due to $\sum_{j>s}\|\boldsymbol{\theta}_j\| \geq \|\boldsymbol{\theta}_{(2)}\|$. Thus, if $n\lambda > 2\hat{\xi}$, and c is sufficiently small, the displayed will be positive with probability approaching 1, which establishes (A1). ■

Proof of Theorem 3.2: The proof is motivated by the results in [38,39]. As in the proof of Theorem 2.1, for any fixed $S \subset \{1, \ldots, p\}$ with $|S| \leq 2s_{\max}$, we have, with probability at least $1 - Ce^{-Ct^2}$, uniformly for all $\boldsymbol{\delta} = \mathrm{vec}(\boldsymbol{\Delta})$ with $\mathrm{rank}(\boldsymbol{\Delta}) \leq r$,

$$\langle \mathbf{Z}_S\boldsymbol{\delta}, \mathbf{e}\rangle \leq C\|\mathbf{Z}_S\boldsymbol{\delta}\|(t + \sqrt{r(K + m - r)}), \tag{A2}$$

where $m = |S \cap \{j : \boldsymbol{\delta}_j \neq 0\}|$. Next we make the bound uniform over $\{S : |S| \leq 2s_{\max}\}$. We will show that, with probability at most Ce^{-Ct},

$$\begin{aligned} \inf_{\substack{S,\boldsymbol{\delta}:|S|\leq 2s_{\max} \\ \boldsymbol{\delta}=\mathrm{vec}(\boldsymbol{\Delta}),\mathrm{rank}(\boldsymbol{\Delta})\leq r}} &\langle \mathbf{Z}_S\boldsymbol{\delta}, \mathbf{e}\rangle - C\|\mathbf{Z}_S\boldsymbol{\delta}\| \\ &\times \left(\sqrt{t + |S \cap \{j : \boldsymbol{\delta}_j \neq 0\}|\log p + r(K + |S \cap \{j : \boldsymbol{\delta}_j \neq 0\}| - r)}\right) \geq 0. \end{aligned} \tag{A3}$$

Note in the above we have put t inside the squared root so the probability is Ce^{-Ct} instead of Ce^{-Ct^2}. First, obviously $\mathbf{Z}_S\boldsymbol{\delta} = \mathbf{Z}_{S\cap\{j:\boldsymbol{\delta}_j\neq 0\}}\boldsymbol{\delta}$ and thus in the above we can assume without loss of generality that $\{j : \boldsymbol{\delta}_j \neq 0\} = S$ and replace $|S \cap \{j : \boldsymbol{\delta}_j \neq 0\}|$ simply by $|S|$. For any fixed S with $|S| = s$, (A2) means with probability at most $Ce^{-C(t+s\log p)}$,

$$\inf_{\boldsymbol{\delta}:\boldsymbol{\delta}=\mathrm{vec}(\boldsymbol{\Delta}),\mathrm{rank}(\boldsymbol{\Delta})\leq r} \langle \mathbf{Z}_S\boldsymbol{\delta}, \mathbf{e}\rangle - C\|\mathbf{Z}_S\boldsymbol{\delta}\|(\sqrt{t + |S|\log p + r(K + |S| - r)}) \geq 0.$$

Taking union over all S with $|S| = s$, we get with probability at most $\binom{p}{s}Ce^{-C(t+s\log p)} \leq Ce^{-C(t+s\log p)}$,

$$\inf_{\substack{S,\boldsymbol{\delta}:|S|=s \\ \boldsymbol{\delta}=\mathrm{vec}(\boldsymbol{\Delta}),\mathrm{rank}(\boldsymbol{\Delta})\leq r}} \langle \mathbf{Z}_S\boldsymbol{\delta}, \mathbf{e}\rangle - C\|\mathbf{Z}_S\boldsymbol{\delta}\|(\sqrt{t + |S|\log p + r(K + |S| - r)}) \geq 0.$$

Then taking union over $s \le 2s_{max}$, we have with probability at most $\sum_{s=1}^{2s_{max}} Ce^{-C(t+s\log p)} \le Ce^{-Ct}$,

$$\inf_{\substack{S,\boldsymbol{\delta}:|S|\le 2s_{max}\\ \boldsymbol{\delta}:\boldsymbol{\delta}=\mathrm{vec}(\boldsymbol{\Delta}),\mathrm{rank}(\boldsymbol{\Delta})\le r}} \langle \mathbf{Z}_S\boldsymbol{\delta},\mathbf{e}\rangle - C\|\mathbf{Z}_S\boldsymbol{\delta}\|(\sqrt{t+|S|\log p + r(K+|S|-r)}) \ge 0.$$

As argued before, this is equivalent to (A3).

Thus, with probability at least $1-Ce^{-Ct}$,

$$\begin{aligned}\langle \mathbf{Z}_S\boldsymbol{\delta},\mathbf{e}\rangle &\le C\|\mathbf{Z}_S\boldsymbol{\delta}\|\left(\sqrt{t+|S\cap\{j:\boldsymbol{\delta}_j\neq 0\}|\log p + r(K+|S\cap\{j:\boldsymbol{\delta}_j\neq 0\}|-r)}\right)\\ &\le \frac{1}{4}\|\mathbf{Z}_S\boldsymbol{\delta}\|^2 + C\left(t+|S\cap\{j:\boldsymbol{\delta}_j\neq 0\}|\log p + r(K+|S\cap\{j:\boldsymbol{\delta}_j\neq 0\}|-r)\right). \qquad (A4)\end{aligned}$$

Let $p_{H,\lambda}(x) = (-x^2/2+\lambda|x|)I(|x|\le\lambda) + (\lambda^2/2)I(|x|>\lambda)$ be the hard-thresholding penalty. We will show that uniformly over $\{S:|S|\le 2s_{\max}\}$ and $\{\boldsymbol{\delta}=\mathrm{vec}(\boldsymbol{\Delta}):\mathrm{rank}(\boldsymbol{\Delta})\le r\}$, with probability at least $1-Ce^{-Ct}$,

$$\langle \mathbf{Z}_S\boldsymbol{\delta},\mathbf{e}\rangle \le \frac{1}{4}\|\mathbf{Z}_S\boldsymbol{\delta}\|^2 + C\left(t+\sum_{\boldsymbol{\delta}_j\neq 0} p_{H,\lambda_0}(\sqrt{n\kappa}\|\boldsymbol{\delta}_j\|) + r(K-r)\right), \qquad (A5)$$

where $\lambda_0 \asymp \sqrt{r+\log p}$ and κ is the uniform upper bound for the eigenvalue as stated in (A2') (κ is assumed to be a bounded constant).

Define

$$\widetilde{\boldsymbol{\delta}} \in \arg\min_{\boldsymbol{\delta}} \frac{1}{4}\|\mathbf{Z}_S\boldsymbol{\delta}\|^2 - \langle \mathbf{Z}_S\boldsymbol{\delta},\mathbf{e}\rangle + C\sum_{\boldsymbol{\delta}_j\neq\ 0} p_{H,\lambda_0}(\sqrt{n\kappa}\|\boldsymbol{\delta}_j\|). \qquad (A6)$$

Obviously we only need to show that for all S there exists such a (possibly non-unique) minimizer $\widetilde{\boldsymbol{\delta}}$ which also satisfies

$$\frac{1}{4}\|\mathbf{Z}\widetilde{\boldsymbol{\delta}}\|^2 - \langle \mathbf{Z}_S\widetilde{\boldsymbol{\delta}},\mathbf{e}\rangle + C\left(t+\sum_{\boldsymbol{\delta}_j\neq\ 0} p_{H,\lambda_0}(\sqrt{n\kappa}\|\widetilde{\boldsymbol{\delta}}_j\|) + r(K-r)\right) > 0.$$

By (A4) and the definition of λ_0, it is sufficient to show that

$$\sum_{j\in S,\widetilde{\boldsymbol{\delta}}_j\neq\ 0} p_{H,\lambda_0}(\sqrt{n\kappa}\|\widetilde{\boldsymbol{\delta}}_j\|) = (\lambda_0^2/2)|S\cap\{j:\widetilde{\boldsymbol{\delta}}_j\neq\ 0\}|. \qquad (A7)$$

The above can be shown as follows. By the rank constraint, as in (7), we can write $\boldsymbol{\Delta}=\mathbf{E}\mathbf{D}^{\mathrm{T}}$ with $\mathbf{E}^{\mathrm{T}}\mathbf{E}=\mathbf{I}$. Denote by $\mathbf{W}$ the matrix with rows $(\mathbf{B}^{\mathrm{T}}(T_i)\mathbf{E})\otimes\mathbf{X}_i^{\mathrm{T}}$, the minimization of (A6) over $\mathbf{d}=(\mathbf{d}_1^{\mathrm{T}},\ldots,\mathbf{d}_p^{\mathrm{T}})^{\mathrm{T}}=\mathrm{vec}(\mathbf{D}^{\mathrm{T}})$ for any given $\mathbf{E}$ is

$$\frac{1}{4}\|2\mathbf{e}-\mathbf{W}\mathbf{d}\|^2 + C\sum_j p_{H,\lambda_0}(\sqrt{n\kappa}\|\mathbf{d}_j\|).$$

Thus Lemma A.2 shows there is a minimizer with either $\mathbf{d}_j=0$ or $\sqrt{n\kappa}\|\mathbf{d}_j\|>\lambda_0$ for all j. This shows (A7).

Since $\widehat{\boldsymbol{\theta}}$ is the minimizer of (5), we have

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \le \|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + 2\langle\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}),\mathbf{e}+\mathbf{R}\rangle + n\sum_j p_\lambda(\|\boldsymbol{\theta}_j\|) - n\sum_j p_\lambda(\|\widehat{\boldsymbol{\theta}}_j\|), \qquad (A8)$$

Using (A5) in (A8), we get

$$\begin{aligned}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2 \le{}& \|\mathbf{Z}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2 + \frac{1}{2}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|^2 + 2\|\mathbf{Z}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|\|\mathbf{R}\| + Ct + Cr(K-r)\\ &+ n\sum_j p_\lambda(\|\boldsymbol{\theta}_j\|) - n\sum_j p_\lambda(\|\widehat{\boldsymbol{\theta}}_j\|) + C\sum_j p_{H,\lambda_0}(\sqrt{n\kappa}\|\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j\|).\end{aligned}$$

Now using the elementary property that, when $\lambda \geq C\lambda_0/\sqrt{n}$ for sufficiently large C, $Cp_{H,\lambda_0}(\sqrt{n\kappa}\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|) \leq Cp_{H,\lambda_0}(\sqrt{n\kappa}\|\widehat{\boldsymbol{\theta}}_j\|) + Cp_{H,\lambda_0}(\sqrt{n\kappa}\|\boldsymbol{\theta}_j\|) \leq np_\lambda(\|\boldsymbol{\theta}_j\|) + np_\lambda(\|\widehat{\boldsymbol{\theta}}_j\|)$, and thus

$$\begin{aligned}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 &\leq \|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + \frac{1}{2}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 + 2\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|\|\mathbf{R}\| + Ct + Cr(K - r) \\ &\quad + Cn\sum_j p_\lambda(\|\boldsymbol{\theta}_j\|) \\ &\leq \|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + \frac{1}{2}\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 + 2\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|\|\mathbf{R}\| + Ct + Cr(K - r) + Cn\lambda^2 s(\boldsymbol{\theta}).\end{aligned}$$

Using Cauchy-Schwarz inequality and that $\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leq \|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\| + \|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|$, we get

$$\|\mathbf{Z}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 \leq C(\|\mathbf{Z}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 + \|\mathbf{R}\|^2 + Ct + Cr(K + s(\boldsymbol{\theta}) - r) + Cs(\boldsymbol{\theta})\log p,$$

with probability $1 - Ce^{-Ct}$, which immediately implies the theorem. ■

Lemma A.2: *For any* $\mathbf{W}$ *with its operator norm bounded by* $\sqrt{n\kappa}$, *any vector* $\mathbf{a}$, *and* $C > 2$, *there exists some global minimizer* $\widetilde{\mathbf{d}}$ *of*

$$g(\mathbf{d}) := \frac{1}{2}\|\mathbf{a} - \mathbf{W}\mathbf{d}\|^2 + C\sum_j p_{H,\lambda}(\sqrt{n\kappa}\|\mathbf{d}_j\|)$$

such that for all j, either $\widetilde{\mathbf{d}}_j = 0$ *or* $\sqrt{n\kappa}\|\widetilde{\mathbf{d}}_j\| > \lambda$.

Proof of Lemma A.2: Define the function

$$F(\mathbf{d}, \boldsymbol{\gamma}) = \frac{1}{2}\|\mathbf{a} - \mathbf{W}\mathbf{d}\|^2 + C\sum_j p_{H,\lambda}\left(\sqrt{n\kappa}\|\mathbf{d}_j\|\right) + \frac{1}{2}\langle(n\kappa\mathbf{I} - \mathbf{W}^{\mathrm{T}}\mathbf{W})(\mathbf{d} - \boldsymbol{\gamma}), \mathbf{d} - \boldsymbol{\gamma}\rangle.$$

Given $\boldsymbol{\gamma}$, minimizing F over $\mathbf{d}$ is equivalent to

$$\min_{\mathbf{d}} \frac{n\kappa}{2}\left\|\mathbf{d} - \left(\mathbf{I} - \frac{\mathbf{W}^{\mathrm{T}}\mathbf{W}}{n\kappa}\right)\boldsymbol{\gamma} - \frac{\mathbf{W}^{\mathrm{T}}\mathbf{a}}{n\kappa}\right\|^2 + C\sum_j p_{H,\lambda}(\sqrt{n\kappa}\|\mathbf{d}_j\|).$$

Denote $\mathbf{w} = (\mathbf{I} - \frac{\mathbf{W}^{\mathrm{T}}\mathbf{W}}{n\kappa})\boldsymbol{\gamma} + \frac{\mathbf{W}^{\mathrm{T}}\mathbf{a}}{n\kappa}$. By direct calculation, the above has an explicit solution $\widetilde{\mathbf{d}}_j = \mathbf{w}_j I(\sqrt{n\kappa}\|\mathbf{w}_j\| > \lambda)$. On the other hand, given $\mathbf{d}$, minimizing F over $\boldsymbol{\gamma}$ gives $\boldsymbol{\gamma} = \mathbf{d}$. Thus given any minimizer $\mathbf{d}$ of g, we have

$$g(\mathbf{d}) = F(\mathbf{d}, \mathbf{d}) \geq F(\widetilde{\mathbf{d}}, \widetilde{\mathbf{d}}) = g(\widetilde{\mathbf{d}}).$$

Thus $\widetilde{\mathbf{d}}$ is also a global minimizer of g with the desired property. ■