FACTOR MODELS AND VARIABLE SELECTION IN HIGH-DIMENSIONAL REGRESSION ANALYSIS

By Alois Kneip and Pascal Sarda

Universität Bonn and Université Paul Sabatier

The paper considers linear regression problems where the number of predictor variables is possibly larger than the sample size. The basic motivation of the study is to combine the points of view of model selection and functional regression by using a factor approach: it is assumed that the predictor vector can be decomposed into a sum of two uncorrelated random components reflecting common factors and specific variabilities of the explanatory variables. It is shown that the traditional assumption of a sparse vector of parameters is restrictive in this context. Common factors may possess a significant influence on the response variable which cannot be captured by the specific effects of a small number of individual variables. We therefore propose to include principal components as additional explanatory variables in an augmented regression model. We give finite sample inequalities for estimates of these components. It is then shown that model selection procedures can be used to estimate the parameters of the augmented model, and we derive theoretical properties of the estimators. Finite sample performance is illustrated by a simulation study.

1. Introduction. The starting point of our analysis is a high-dimensional linear regression model of the form

$$(1.1) \qquad Y_i = \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are i.i.d. random pairs with $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$. We will assume without loss of generality that $\mathbb{E}(X_{ij}) = 0$ for all $j = 1, \ldots, p$. Furthermore, $\boldsymbol{\beta}$ is a vector of parameters in $\mathbb{R}^p$ and $(\varepsilon_i)_{i=1,\ldots,n}$ are centered i.i.d. real random variables independent with $\mathbf{X}_i$ with $\mathrm{Var}(\varepsilon_i) = \sigma^2$. The dimension p of the vector of parameters is assumed to be typically larger than the sample size n.

Roughly speaking, model (1.1) comprises two main situations which have been considered independently in two separate branches of statistical literature. On one side, there is the situation where $\mathbf{X}_i$ represents a (high-dimensional) vector of different predictor variables. Another situation arises when the regressors are p discretizations (e.g., at different observations times) of a same curve. In this case model (1.1) represents a discrete version of an underlying continuous *functional linear model*. In the two setups, very different strategies for estimating $\boldsymbol{\beta}$ have been adopted, and underlying structural assumptions seem to be largely incompatible. In this paper we will study similarities and differences of these methodologies, and we will show that a combination of ideas developed in the two settings leads to new estimation procedures which may be useful in a number of important applications.

The first situation is studied in a large literature on model selection in high-dimensional regression. The basic structural assumptions can be described as follows:

- There is only a relatively small number of predictor variables with $|\beta_j| > 0$ which have a significant influence on the outcome Y. In other words, the set of nonzero coefficients is sparse, $S := \#\{j|\beta_j \neq 0\} \ll p$.
- The correlations between different explanatory variables X_{ij} and X_{il}, $j \neq l$, are "sufficiently" weak.

The most popular procedures to identify and estimate nonzero coefficients β_j are Lasso and the Dantzig selector. Some important references are Tibshirani (1996), Meinshausen and Bühlmann (2006), Zhao and Yu (2006), van de Geer (2008), Bickel, Ritov and Tsybakov (2009), Candes and Tao (2007) and Koltchinskii (2009). Much work in this domain is based on the assumption that the columns $(X_{1j}, \ldots, X_{nj})^T$, $j = 1, \ldots, p$, of the design matrix are almost orthogonal. For example, Candes and Tao (2007) require that "every set of columns with cardinality less than S approximately behaves like an orthonormal system." More general conditions have been introduced by Bickel, Ritov and Tsybakov (2009) or Zhou, van de Geer and Bülhmann (2009). The theoretical framework developed in these papers also allows one to study model selection for regressors with substantial amount of correlation, and it provides a basis for the approach presented in our paper.

In sharp contrast, the setup considered in the literature on functional regression rests upon a very different type of structural assumptions. We will consider the simplest case that $X_{ij} = X_i(t_j)$ for random functions $X_i \in L^2([0,1])$ observed at an equidistant grid $t_j = \frac{j}{p}$. Structural assumptions on coefficients and correlations between variables can then be subsumed as follows:

- $\beta_j := \frac{\beta(t_j)}{p}$, where $\beta(t) \in L^2([0,1])$ is a continuous slope function, and as $p \to \infty$, $\sum_j \beta_j X_{ij} = \sum_j \frac{\beta(t_j)}{p} X_i(t_j) \to \int_0^1 \beta(t) X_i(t)\, dt$.

- There are very high correlations between explanatory variables $X_{ij} = X_i(t_j)$ and $X_{il} = X_i(t_l)$, $j \neq l$. As $p \to \infty$, $\mathrm{corr}(X_i(t_j), X_i(t_{j+m})) \to 1$ for any fixed m.

Some important applications as well as theoretical results on functional linear regression are, for example, presented in Ramsay and Dalzell (1991), Cardot, Ferraty and Sarda (1999), Cuevas, Febrero and Fraiman (2002), Yao, Müller and Wang (2005), Cai and Hall (2006), Hall and Horowitz (2007), Cardot, Mas and Sarda (2007) and Crambes, Kneip and Sarda (2009). Obviously, in this setup no variable $X_{ij} = X_i(t_j)$ corresponding to a specific observation at grid point t_j will possess a particulary high influence on Y_i, and there will exist a large number of small, but nonzero coefficients β_j of size proportional to $1/p$. One may argue that dimensionality reduction and therefore some underlying concept of "sparseness" is always necessary when dealing with high-dimensional problems. However, in functional regression sparseness is usually not assumed with respect to the coefficients β_j, but the model is rewritten using a "sparse" expansion of the predictor functions X_i.

The basic idea relies on the so-called Karhunen–Loève decomposition which provides a decomposition of random functions in terms of functional principal components of the covariance operator of X_i. In the discretized case analyzed in this paper this amounts to consider an approximation of X_i by the principal components of the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^T)$. In practice, often a small number k of principal components will suffice to achieve a small L^2-error. An important points is now that even if $p > n$ the eigenvectors corresponding to the leading eigenvalues $\mu_1, \ldots, \mu_k$ of $\boldsymbol{\Sigma}$ can be well estimated by the eigenvectors (estimated principal components) $\widehat{\boldsymbol{\psi}}_r$ of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$. This is due to the fact that if the predictors X_{ij} represent discretized values of a continuous functional variable, then for sufficiently small k the eigenvalues $\mu_1, \ldots, \mu_k$ will necessarily be of an order larger than $\frac{p}{\sqrt{n}}$ and will thus exceed the magnitude of purely random components. From a more general point of view the underlying theory will be explained in detail in Section 4.

Based on this insight, the most frequently used approach in functional regression is to approximate $\mathbf{X}_i \approx \sum_{r=1}^{k} (\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i) \widehat{\boldsymbol{\psi}}_r$ in terms of the first k estimated principal components $\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_k$, and to rely on the approximate model $Y_i \approx \sum_{r=1}^{k} \alpha_r \widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i + \varepsilon_i$. Here, k serves as smoothing parameter. The new coefficients $\boldsymbol{\alpha}$ are estimated by least squares, and $\widehat{\beta}_j = \sum_{r=1}^{k} \widehat{\alpha}_r \widehat{\psi}_{rj}$. Resulting rates of convergence are given in Hall and Horowitz (2007).

The above arguments show that a suitable regression analysis will have to take into account the underlying structure of the explanatory variables X_{ij}. The basic motivation of this paper now is to combine the points of view of the above branches of literature in order to develop a new approach for model

adjustment and variable selection in the practically important situation of strongly correlated regressors. More precisely, we will concentrate on *factor models* by assuming that the $\mathbf{X}_i \in \mathbb{R}^p$ can be decomposed in the form

$$(1.2) \qquad \mathbf{X}_i = \mathbf{W}_i + \mathbf{Z}_i, \qquad i = 1, \ldots, n,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated random vectors in $\mathbb{R}^p$. The random vector $\mathbf{W}_i$ is intended to describe high correlations of the X_{ij} while the components Z_{ij}, $j = 1, \ldots, p$, of $\mathbf{Z}_i$ are uncorrelated. This implies that the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}_i$ adopts the decomposition

$$(1.3) \qquad \boldsymbol{\Sigma} = \boldsymbol{\Gamma} + \boldsymbol{\Psi},$$

where $\boldsymbol{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$, while $\boldsymbol{\Psi}$ is a diagonal matrix with diagonal entries $\mathrm{var}(Z_{ij})$, $j = 1, \ldots, p$.

Note that factor models can be found in any textbook on multivariate analysis and must be seen as one of the major tools in order to analyze samples of high-dimensional vectors. Also recall that a standard factor model is additionally based on the assumption that a finite number k of factors suffices to approximate $\mathbf{W}_i$ precisely. This means that the matrix $\boldsymbol{\Gamma}$ only possesses k nonzero eigenvalues. In the following we will more generally assume that a small number of eigenvectors of $\boldsymbol{\Gamma}$ suffices to approximate $\mathbf{W}_i$ with high accuracy.

We want to emphasize that the typical structural assumptions to be found in the literature on high-dimensional regression are special cases of (1.2). If $\mathbf{W}_i = 0$ and thus $\mathbf{X}_i = \mathbf{Z}_i$, we are in the situation of uncorrelated regressors which has been widely studied in the context of model selection. On the other hand, $\mathbf{Z}_i = 0$ and thus $\mathbf{X}_i = \mathbf{W}_i$ reflect the structural assumption of functional regression.

In this paper we assume that W_{ij} as well as Z_{ij} represent nonnegligible parts of the variance of X_{ij}. We believe that this approach may well describe the situation encountered in many relevant applications. Although standard factor models are usually considered in the case $p \ll n$, (1.2) for large values of p may be of particular interest in time series or spatial analysis. Indeed, factor models for large p with a finite number k of nonzero eigenvalues of $\boldsymbol{\Gamma}$ play an important role in the econometric study of multiple time series and panel data. Some references are Forni and Lippi (1997), Forni et al. (2000), Stock and Watson (2002), Bernanke and Boivin (2003) and Bai (2003, 2009).

Our objective now is to study linear regression (1.1) with respect to explanatory variables which adopt decomposition (1.2). Each single variable X_{ij}, $j = 1, \ldots, p$, then possesses a *specific* variability induced by Z_{ij} and may thus explain some part of the outcome Y_i. One will, of course, assume that only few variables have a significant influence on Y_i which enforces the use of model selection procedures.

On the other hand, the term W_{ij} represents a *common* variability. Corresponding principal components quantify a simultaneous variation of many individual regressors. As a consequence, such principal components may possess some additional power for predicting Y_i which may go beyond the effects of individual variables. A rigorous discussion will be given in Section 3. We want to note that the concept of "latent variables," embracing the common influence of a large group of individual variables, plays a prominent role in applied, parametric multivariate analysis.

These arguments motivate the main results of this paper. We propose to use an "augmented" regression model which includes principal components as additional explanatory variables. Established model selection procedures like the Dantzig selector or the Lasso can then be applied to estimate the nonzero coefficients of the augmented model. We then derive theoretical results providing bounds for the accuracy of the resulting estimators.

The paper is organized as follows: in Section 2 we formalize our setup. We show in Section 3 that the traditional sparseness assumption is restrictive and that a valid model may have to include principal components. The augmented model is thus introduced with an estimation procedure. Section 4 deals with the problem how accurately true principal components can be estimated from the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Finite sample inequalities are given, and we show that it is possible to obtain sensible estimates of those components which explain a considerable percentage of the total variance of all X_{ij}, $j = 1, \ldots, p$. Section 5 focuses on theoretical properties of the augmented model, while in Section 6 we present simulation results illustrating the finite sample performance of our estimators.

2. The setup. We study regression of a response variable Y_i on a set of i.i.d. predictors $\mathbf{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, which adopt decomposition (1.2) with $\mathbb{E}(X_{ij}) = \mathbb{E}(W_{ij}) = \mathbb{E}(Z_{ij}) = 0$, $\mathbb{E}(Z_{ij}Z_{ik}) = 0$, $\mathbb{E}(W_{ij}Z_{il}) = 0$, $\mathbb{E}(Z_{ij}Z_{ik}Z_{il}Z_{im}) = 0$ for all $j, k, l, m \in \{1, \ldots, p\}$, $j \notin \{k, l, m\}$. Throughout the following sections we additionally assume that there exist constants $D_0, D_3 < \infty$ and $0 < D_1 \le D_2 < \infty$ such that with $\sigma_j^2 := \text{Var}(Z_{ij})$ the following assumption (A.1) is satisfied for all p:

(A.1) $0 < D_1 \le \sigma_j^2 \le D_2, \mathbb{E}(X_{ij}^2) \le D_0, \mathbb{E}(Z_{ij}^4) \le D_3$ for all $j = 1, \ldots, p$.

Recall that $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^T)$ is the covariance matrix of $\mathbf{X}_i$ with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} + \boldsymbol{\Psi}$, where $\boldsymbol{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ and $\boldsymbol{\Psi}$ is a diagonal matrix with diagonal entries σ_j^2, $j = 1, \ldots, p$. We denote as $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T$ the empirical covariance matrix based on the sample $\mathbf{X}_i$, $i = 1, \ldots, n$.

Eigenvalues and eigenvectors of the standardized matrices $\frac{1}{p}\boldsymbol{\Gamma}$ and $\frac{1}{p}\boldsymbol{\Sigma}$ will play a central role. We will use $\lambda_1 \ge \lambda_2 \ge \cdots$ and $\mu_1 \ge \mu_2 \ge \cdots$ to denote the eigenvalues of $\frac{1}{p}\boldsymbol{\Gamma}$ and $\frac{1}{p}\boldsymbol{\Sigma}$, respectively, while $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ and $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots$

denote corresponding orthonormal eigenvectors. Note that all eigenvectors of $\frac{1}{p}\Sigma$ and Σ (or $\frac{1}{p}\Gamma$ and Γ) are identical, while eigenvalues differ by the factor $1/p$. Standardization is important to establish convergence results for large p, since the largest eigenvalues of Σ tend to infinity as $p \to \infty$.

From a conceptional point of view we will concentrate on the case that p is large compared to n. Another crucial, *qualitative* assumption characterizing our approach is the dimensionality reduction of $\mathbf{W}_i$ using a small number $k \ll p$ of eigenvectors (principal components) of $\frac{1}{p}\Gamma$ such that (in a good approximation) $\mathbf{W}_i \approx \sum_{r=1}^{k} \xi_{ir}\psi_r$. We also assume that $\mathcal{D}_X = \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}(X_{ij}^2) > \mathcal{D}_W = \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}(W_{ij}^2) \gg \frac{1}{p}$. *Then all leading principal components of $\frac{1}{p}\Gamma$ corresponding to the k largest eigenvalues explain a considerable percentage of the total variance of $\mathbf{W}_i$ and $\mathbf{X}_i$.*

Indeed, if $\mathbf{W}_i = \sum_{r=1}^{k}\xi_{ir}\psi_r$, we necessarily have $\lambda_1 \geq \frac{\mathcal{D}_W}{k} \gg \frac{1}{p}$ and $\mu_1 \geq \lambda_1 \geq \frac{\mathcal{D}_W}{k} \gg \frac{1}{p}$. Then $\mathrm{tr}(\frac{1}{p}\Gamma) = \sum_{r=1}^{p}\lambda_r = \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}(W_{ij}^2)$, and the first principal component of $\frac{1}{p}\Gamma$ explains a considerable proportion $\frac{\lambda_1}{(1/p)\sum_{j=1}^{p}\mathbb{E}(W_{ij}^2)} \geq \frac{1}{k} \gg \frac{1}{p}$ of the total variance of $\mathbf{W}_i$.

We want to emphasize that this situation is very different from the setup which is usually considered in the literature on the analysis of high-dimensional covariance matrices; see, for example, Bickel and Levina ([2008]). It is then assumed that the variables of interest are only weakly correlated and that the largest eigenvalue μ_1 of the corresponding scaled covariance matrix $\frac{1}{p}\Sigma$ is of order $\frac{1}{p}$. This means that for large p the first principal component only explains a negligible percentage of the total variance of $\mathbf{X}_i$, $\frac{\mu_1}{(1/p)\sum_{j=1}^{p}\mathbb{E}(X_{ij}^2)} = O(\frac{1}{p})$. It is well known that in this case no consistent estimates of eigenvalues and principal components can be obtained from an eigen-decomposition of $\frac{1}{p}\widehat{\Sigma}$.

However, we will show in Section 4 that principal components which are able to explain a considerable proportion of total variance can be estimated consistently. These components will be an intrinsic part of the augmented model presented in Section 3.

We will need a further assumption which ensures that all covariances between the different variables are well approximated by their empirical counterparts:

(A.2) There exists a $C_0 < \infty$ such that

$$(2.1) \qquad \sup_{1 \leq j,l \leq p}\left|\frac{1}{n}\sum_{i=1}^{n}W_{ij}W_{il} - \mathrm{cov}(W_{ij}, W_{il})\right| \leq C_0\sqrt{\frac{\log p}{n}},$$

$$(2.2) \qquad \sup_{1 \leq j,l \leq p}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}Z_{il} - \mathrm{cov}(Z_{ij}, Z_{il})\right| \leq C_0\sqrt{\frac{\log p}{n}},$$

$$(2.3) \qquad \sup_{1 \le j, l \le p} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ij} W_{il} \right| \le C_0 \sqrt{\frac{\log p}{n}},$$

$$(2.4) \qquad \sup_{1 \le j, l \le p} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} X_{il} - \mathrm{cov}(X_{ij}, X_{il}) \right| \le C_0 \sqrt{\frac{\log p}{n}}$$

hold simultaneously with probability $A(n, p) > 0$, where $A(n, p) \to 1$ as $n, p \to \infty$, $\frac{\log p}{n} \to 0$.

The following proposition provides a general sufficient condition on random vectors for which (A.2) is satisfied provided that the rate of convergence of $\frac{\log p}{n}$ to 0 is sufficiently fast.

PROPOSITION 1. *Consider independent and identically distributed random vectors* $\mathbf{V}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, *such that for* $j = 1, \ldots, p$, $\mathbb{E}(V_{ij}) = 0$ *and*

$$(2.5) \qquad \mathbb{E}(e^{a|V_{ij}|}) \le C_1$$

for positive constants a *and* C_1 *with moreover* $\mathbb{E}(V_{ij}^4) \le C_1$. *Then, for any positive constant* C_0 *such that* $C_1^{1/2} \le \frac{1}{2} \sqrt{\frac{C_0 n}{\log p}}$ *and* $C_1 \le \frac{1}{8} C_0 e^{a \sqrt{C_0 n/(\log p)}} \sqrt{\frac{\log p}{n}}$

$$(2.6) \qquad P \left(\sup_{1 \le j, l \le p} \left| \frac{1}{n} \sum_{i=1}^{n} V_{ij} V_{il} - \mathrm{cov}(V_{ij}, V_{il}) \right| \le C_0 \sqrt{\frac{\log p}{n}} \right)$$
$$\ge 1 - p^{2 - C_0^2/(8(C_1 + C_0^{3/2}/3))} + 2p^2 n C_1 e^{-(a/\sqrt{2})(C_0 n/\log p)^{1/4}}.$$

Note that as $n, p \to \infty$ the right-hand side of (2.6) converges to 1 provided that C_0 is chosen sufficiently large and that $p/e^{n^{1-\tau}} = O(1)$ for some $4/5 < \tau < 1$. Therefore, assumption (A.2) is satisfied if the components of the random variables X_{ij} possess some exponential moments. For the specific case of centered normally distributed random variables, a more precise bound in (2.6) may be obtained using Lemma 2.5 in Zhou, van de Geer and Bülhmann (2009) and large deviations inequalities obtained by Zhou, Lafferty and Wassermn (2008). In this case it may also be shown that for sufficiently large C_0 events (2.1)–(2.4) hold with probability tending to 1 as $p \to \infty$ without any restriction on the quotient $\log p/n$. Of course, the rate $\sqrt{\frac{\log p}{n}}$ in (2.1)–(2.4) depends on the tails of the distributions: it would be possible to replace this rate with a slower one in case of heavier tails than in Proposition 1. Our theoretical results could be modified accordingly.

3. The augmented model. Let us now consider the structural model (1.2) more closely. It implies that the vector $\mathbf{X}_i$ of predictors can be decomposed

into two uncorrelated random vectors $\mathbf{W}_i$ and $\mathbf{Z}_i$. Each of these two components separately may possess a significant influence on the response variable Y_i. Indeed, if $\mathbf{W}_i$ and $\mathbf{Z}_i$ were known, a possibly substantial improvement of model (1.1) would consist in a regression of Y_i on the $2p$ variables $\mathbf{W}_i$ and $\mathbf{Z}_i$

$$(3.1) \qquad Y_i = \sum_{j=1}^{p} \beta_j^* W_{ij} + \sum_{j=1}^{p} \beta_j Z_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with different sets of parameters β_j^* and β_j, $j = 1, \ldots, p$, for each contributor. We here again assume that ε_i, $i = 1, \ldots, n$, are centered i.i.d. random variables with $\mathrm{Var}(\varepsilon_i) = \sigma^2$ which are independent of W_{ij} and Z_{ij}.

By definition, W_{ij} and Z_{ij} possess substantially different interpretations. Z_{ij} describes the part of X_{ij} which is *uncorrelated with all other variables*. A nonzero coefficient $\beta_j \neq 0$ then means that the variation of X_{ij} has a *specific* effect on Y_i. We will of course assume that such nonzero coefficients are *sparse*, $\sharp\{j | \beta_j \neq 0\} \leq S$ for some $S \ll p$. The true variables Z_{ij} are unknown, but with $\beta_j^{**} = \beta_j^* - \beta_j$ model (3.1) can obviously be rewritten in the form

$$(3.2) \qquad Y_i = \sum_{j=1}^{p} \beta_j^{**} W_{ij} + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The variables W_{ij} are heavily correlated. It therefore does not make any sense to assume that for some $j \in \{1, \ldots, p\}$ any particular variable W_{ij} possesses a specific influence on the predictor variable. However, the term $\sum_{j=1}^{p} \beta_j^{**} W_{ij}$ may represent an important, *common* effect of all predictor variables. The vectors $\mathbf{W}_i$ can obviously be rewritten in terms of principal components. Let us recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ denote the eigenvalues of the standardized covariance matrix of $\mathbf{W}_i$, $\frac{1}{p}\boldsymbol{\Gamma} = \frac{1}{p}\mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ and $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ corresponding orthonormal eigenvectors. We have

$$\mathbf{W}_i = \sum_{r=1}^{p} (\boldsymbol{\psi}_r^T \mathbf{W}_i)\boldsymbol{\psi}_r \quad \text{and} \quad \sum_{j=1}^{p} \beta_j^{**} W_{ij} = \sum_{r=1}^{p} \alpha_r^* (\boldsymbol{\psi}_r^T \mathbf{W}_i),$$

where $\alpha_r^* = \sum_{j=1}^{p} \beta_j^{**} \psi_{rj}$. As outlined in the previous sections we now assume that the use of principal components allows for a considerable reduction of dimensionality, and that a small number of leading principal components will suffice to describe the effects of the variable $\mathbf{W}_i$. This may be seen as an analogue of the sparseness assumption made for the Z_{ij}. More precisely, subsequent analysis will be based on the assumption that the following *augmented model* holds for some suitable $k \geq 1$:

$$(3.3) \qquad Y_i = \sum_{r=1}^{k} \alpha_r \xi_{ir} + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i,$$

where $\xi_{ir} = \boldsymbol{\psi}_r^T \mathbf{W}_i / \sqrt{p\lambda_r}$ and $\alpha_r = \sqrt{p\lambda_r}\alpha_r^*$. The use of ξ_{ir} instead of $\boldsymbol{\psi}_r^T \mathbf{W}_i$ is motivated by the fact that $\mathrm{Var}(\boldsymbol{\psi}_r^T \mathbf{W}_i) = p\lambda_r$, $r = 1, \ldots, k$. Therefore the ξ_{ir} are standardized variables with $\mathrm{Var}(\xi_{i1}) = \cdots = \mathrm{Var}(\xi_{i1}) = 1$. Fitting an augmented model requires us to select an appropriate k as well as to determine sensible estimates of $\xi_{i1}, \ldots, \xi_{ik}$. Furthermore, model selection procedures like Lasso or Dantzig have to be applied in order to retrieve the nonzero coefficients α_r, $r = 1, \ldots, k$, and β_j, $j = 1, \ldots, p$. These issues will be addressed in subsequent sections.

Obviously, the augmented model may be considered as a synthesis of the standard type of models proposed in the literature on functional regression and model selection. It generalizes the classical multivariate linear regression model (1.1). If a k-factor model holds exactly, that is, $\mathrm{rank}(\boldsymbol{\Gamma}) = k$, then the only substantial restriction of (3.1)–(3.3) consists in the assumption that Y_i depends *linearly* on W_i and Z_i.

We want to emphasize, however, that our analysis does not require the validity of a k-factor model. It is only assumed that there exists "some" Z_i and W_i satisfying our assumptions which lead to (3.3) for a sparse set of coefficients β_j.

3.1. *Identifiability.* Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$. Since $\boldsymbol{\psi}_r$, $r = 1, \ldots, k$, are eigenvectors of $\boldsymbol{\Gamma}$ we have $\mathbb{E}(\boldsymbol{\psi}_r^T \mathbf{W}_i \boldsymbol{\psi}_s^T \mathbf{W}_i) = 0$ for all $r, s = 1, \ldots, p$, $r \neq s$. By assumption the random vectors $\mathbf{W}_i$ and $\mathbf{Z}_i$ are uncorrelated, and hence $\mathbb{E}(\boldsymbol{\psi}_r^T \mathbf{W}_i Z_{ij}) = 0$ for all $r, j = 1, \ldots, p$. Furthermore, $\mathbb{E}(Z_{il} Z_{ij}) = 0$ for all $l \neq j$. If the augmented model (3.3) holds, some straightforward computations then show that under (A.1) for any alternative set of coefficients $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$, $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_k^*)^T$,

$$\mathbb{E}\left(\left[\sum_{r=1}^k (\alpha_r - \alpha_r^*)\xi_{ir} + \sum_{j=1}^p (\beta_j - \beta_j^*)X_{ij}\right]^2\right)$$

(3.4)
$$\geq \sum_{r=1}^k (\alpha_r - \alpha_r^* + \sqrt{p\lambda_r}\boldsymbol{\psi}_r(\boldsymbol{\beta} - \boldsymbol{\beta}^*))^2$$
$$+ D_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

We can conclude that the coefficients α_r, $r = 1, \ldots, k$ and β_j, $j = 1, \ldots, p$, in (3.3) are uniquely determined.

Of course, an inherent difficulty of (3.3) consists of the fact that it contains the unobserved, "latent" variables $\xi_{ir} = \boldsymbol{\psi}_r^T \mathbf{W}_i / \sqrt{p\lambda_r}$. To study this problem, first recall that our setup imposes the decomposition (1.3) of the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}_i$. If a factor model with k factors holds exactly, then the $\boldsymbol{\Gamma}$ possesses rank k. It then follows from well-established results in multivariate analysis that if $k < p/2$ the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ are uniquely

identified. If $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$, then also $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ are uniquely determined (up to sign) from the structure of $\boldsymbol{\Gamma}$.

However, for large p, identification is possible under even more general conditions. It is not necessary that a k-factor model holds exactly. We only need an additional assumption on the magnitude of the eigenvalues of $\frac{1}{p}\boldsymbol{\Gamma}$ defining the k principal components of $\mathbf{W}_i$ to be considered.

(A.3) The eigenvalues of $\frac{1}{p}\boldsymbol{\Gamma}$ are such that

$$\min_{j,l \leq k, j \neq l} |\lambda_j - \lambda_l| \geq v(k), \qquad \min_{j \leq k} \lambda_j \geq v(k)$$

for some $1 \geq v(k) > 0$ with $pv(k) > 6D_2$.

In the following we will qualitatively assume that $k \ll p$ as well as $v(k) \gg 1/p$. More specific assumptions will be made in the sequel. Note that eigenvectors are only unique up to sign changes. In the following we will always assume that the right "versions" are used. This will go without saying.

THEOREM 1. Let $\xi_{ir}^* := \frac{\boldsymbol{\delta}_r^T \mathbf{X}_i}{\sqrt{p\mu_r}}$ and $\mathbf{P}_k = \mathbf{I}_p - \sum_{j=1}^k \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T$. Under assumptions (A.1) and (A.2) we have for all $r = 1, \ldots, k$, $j = 1, \ldots, p$ and all k, p satisfying (A.3):

$$(3.5) \qquad |\mu_r - \lambda_r| \leq \frac{D_2}{p},$$

$$\|\boldsymbol{\psi}_r - \boldsymbol{\delta}_r\|_2 \leq \frac{2D_2}{pv(k)},$$
$$(3.6)$$
$$\delta_{rj}^2 \leq \frac{D_0}{pv(k)},$$

$$\mathbb{E}([\xi_{ir} - \xi_{ir}^*]^2) \leq \frac{D_2}{p\mu_r} + \frac{(8\lambda_1 + 1)D_2^2}{p^2 v(k)^2 \mu_r},$$
$$(3.7)$$
$$\mathbb{E}\left(\left[\xi_{ir} - \frac{\boldsymbol{\psi}_r^T \mathbf{X}_i}{\sqrt{p\lambda_r}}\right]^2\right) \leq \frac{D_2}{p\lambda_r},$$

$$(3.8) \quad \mathbb{E}\left(\left[\sum_{r=1}^k \alpha_r \xi_{ir} - \sum_{r=1}^k \alpha_r \xi_{ir}^*\right]^2\right) \leq k \sum_{r=1}^k \alpha_r^2 \left(\frac{D_2}{p\mu_r} + \frac{(8\lambda_1 + 1)D_2^2}{p^2 v(k)^2 \mu_r}\right).$$

For small p, standard factor analysis uses special algorithms in order to identify $\boldsymbol{\psi}_r$. The theorem tells us that for large p this is unnecessary since then the eigenvectors $\boldsymbol{\delta}_r$ of $\frac{1}{p}\boldsymbol{\Sigma}$ provide a good approximation. The predictor ξ_{ir}^* of ξ_{ir} possesses an error of order $1/\sqrt{p}$. The error decreases as p increases, and ξ_{ir}^* thus yields a good approximation of ξ_{ir} *if p is large*. Indeed, if $p \to \infty$

[for fixed μ_r, $v(k)$, D_1 and D_2] then by (3.7) we have $\mathbb{E}([\xi_{ir} - \xi_{ir}^*]^2) \to 0$. Furthermore, by (3.8) the error in predicting $\sum_{r=1}^{k} \alpha_r \xi_{ir} + \sum_{j=1}^{p} \beta_j X_{ij}$ by $\sum_{r=1}^{k} \alpha_r \xi_{ir}^* + \sum_{j=1}^{p} \beta_j X_{ij}$ converges to zero as $p \to \infty$.

A crucial prerequisite for a reasonable analysis of the model is sparseness of the coefficients β_j. Note that if p is large compared to n, then by (3.8) the error in replacing ξ_{ir} by ξ_{ir}^* is negligible compared to the estimation error induced by the existence of the error terms ε_i. If $k \ll p$ and $\sharp\{j|\beta_j \neq 0\} \ll p$, then the true coefficients α_r and β_j provide a *sparse* solution of the regression problem.

Established theoretical results [see Bickel, Ritov and Tsybakov (2009)] show that under some regularity conditions (validity of the "restricted eigen-value conditions") model selection procedures *allow to identify such sparse solutions* even if there are multiple vectors of coefficients satisfying the normal equations. The latter is of course always the case if $p > n$. Indeed, we will show in the following sections that factors can be consistently estimated from the data, and that a suitable application of Lasso or the Dantzig-selector leads to consistent estimators $\widehat{\alpha}_r$, $\widehat{\beta}_j$ satisfying $\sup_r |\alpha_r - \widehat{\alpha}_r| \to_P 0$, $\sup_j |\beta_j - \widehat{\beta}_j| \to_P 0$, as $n, p \to \infty$.

When replacing ξ_{ir} by ξ_{ir}^*, there are alternative sets of coefficients leading to the same prediction error as in (3.8). This is due to the fact that $\xi_{ir}^* = \sum_{j=1}^{p} \frac{\delta_{rj} X_{ij}}{\sqrt{p\mu_r}}$. However, all these alternative solutions are *nonsparse* and cannot be identified by Lasso or other procedures. In particular, it is easily seen that

$$\sum_{r=1}^{k} \alpha_r \xi_{ir}^* + \sum_{j=1}^{p} \beta_j X_{ij} = \sum_{j=1}^{p} \beta_j^{LR} X_{ij}$$

(3.9)

$$\text{with } \beta_j^{LR} := \beta_j + \sum_{r=1}^{k} \alpha_r \frac{\delta_{rj}}{\sqrt{p\mu_r}}.$$

By (3.6) all values δ_{rj}^2 are of order $1/(pv(k))$. Since $\sum_j \delta_{rj}^2 = 1$, this implies that many δ_{rj}^2 are nonzero. Therefore, if $\alpha_r \neq 0$ for some $r \in \{1, \ldots, k\}$, then $\{j|\beta_j^{LR} \neq 0\}$ contains a large number of small, nonzero coefficients and is not at all sparse. If p is large compared to n no known estimation procedure will be able to provide consistent estimates of these coefficients.

Summarizing the above discussion we can conclude:

(1) If the variables X_{ij} are heavily correlated and follow an approximate factor model, then one may reasonably expect substantial effects of the common, joint variation of all variables and, consequently, nonzero coefficients β_j^* and α_r in (3.1) and (3.3). But then a "bet on sparsity" is unjustifiable when dealing with the standard regression model (1.1). It follows from (3.9)

that for large p model (1.1) holds approximately for a nonsparse set of co-efficients β_j^{LR}, since many small, nonzero coefficients are necessary in order to capture the effects of the common joint variation.

(2) The augmented model offers a remedy to this problem by pooling possible effects of the joint variation using a small number of additional variables. Together with the familiar assumption of a small number of variables possessing a specific influence, this leads to a sparse model with at most $k + S$ nonzero coefficients which can be recovered from model selection procedures like Lasso or the Dantzig-selector.

(3) In practice, even if (3.3) only holds approximately, since a too-small value of k has been selected, it may be able to quantify at least some important part of the effects discussed above. Compared to an analysis based on a standard model (1.1), this may lead to a substantial improvement of model fit as well as to more reliable interpretations of significant variables.

3.2. *Estimation.* For a pre-specified $k \geq 1$ we now define a procedure for estimating the components of the corresponding augmented model (3.3) from given data. This obviously specifies suitable procedures for approximating the unknown values ξ_{ir} as well as to apply subsequent model selection procedures in order to retrieve nonzero coefficients α_r and β_j, $r = 1, \ldots, k$, $j = 1, \ldots, p$. A discussion of the choice of k can be found in the next section.

Recall from Theorem 1 that for large p the eigenvectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ of $\frac{1}{p}\boldsymbol{\Gamma}$ are well approximated by the eigenvectors of the standardized covariance matrix $\frac{1}{p}\boldsymbol{\Sigma}$. This motivates us to use the empirical principal components of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in order to determine estimates of $\boldsymbol{\psi}_r$ and ξ_{ir}. Theoretical support will be given in the next section. Define $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots$ as the eigenvalues of the standardized empirical covariance matrix $\frac{1}{p}\widehat{\boldsymbol{\Sigma}} = \frac{1}{np}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$, while $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ are associated orthonormal eigenvectors. We then estimate ξ_{ir} by

$$\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i / \sqrt{p\widehat{\lambda}_r}, \qquad r = 1, \ldots, k, i = 1, \ldots, n.$$

When replacing ξ_{ir} by $\widehat{\xi}_{ir}$ in (3.3), a direct application of model selection procedures does not seem to be adequate, since $\widehat{\xi}_{ir}$ and the predictor variables X_{ij} are heavily correlated. We therefore rely on a projected model. Consider the projection matrix on the orthogonal space of the space spanned by the eigenvectors corresponding to the k largest eigenvalues of $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$

$$\widehat{\mathbf{P}}_k = \mathbf{I}_p - \sum_{r=1}^{k} \widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T.$$

Then model (3.3) can be rewritten for $i = 1, \ldots, n$,

$$(3.10) \quad Y_i = \sum_{r=1}^{k} \widetilde{\alpha}_r \widehat{\xi}_{ir} + \sum_{j=1}^{p} \widetilde{\beta}_j \frac{(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j}{((1/n)\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2)^{1/2}} + \widetilde{\varepsilon}_i + \varepsilon_i,$$

where $\widetilde{\alpha}_r = \alpha_r + \sqrt{p\widehat{\lambda}_r} \sum_{j=1}^p \widehat{\psi}_{rj}\beta_j$, $\widetilde{\beta}_j = \beta_j(\frac{1}{n}\sum_{i=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2)^{1/2}$ and $\widetilde{\varepsilon}_i = \sum_{r=1}^k \alpha_r(\xi_{ir} - \widehat{\xi}_{ir})$. It will be shown in the next section that for large n and p the additional error term $\widetilde{\varepsilon}$ can be assumed to be reasonably small.

In the following we will use $\widetilde{\mathbf{X}}_i$ to denote the vectors with entries $\widetilde{X}_{ij} := \frac{(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j}{((1/n)\sum_{l=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_l)_j^2)^{1/2}}$. Furthermore, consider the $(k+p)$-dimensional vector of predictors $\mathbf{\Phi}_i := (\widehat{\xi}_{i1}, \ldots, \widehat{\xi}_{ik}, \widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip})^T$. The Gram matrix in model (3.10) is a block matrix defined as

$$\frac{1}{n}\sum_{i=1}^n \mathbf{\Phi}_i \mathbf{\Phi}_i^T = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{X}}_i \widetilde{\mathbf{X}}_i^T \end{pmatrix},$$

where $\mathbf{I}_k$ is the identity matrix of size k. Note that the normalization of the predictors in (3.10) implies that the diagonal elements of the Gram matrix above are equal to 1.

Arguing now that the vector of parameters $\boldsymbol{\theta} := (\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_k, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^T$ in model (3.10) is $(k+S)$-sparse, we may use a selection procedure to recover/estimate the nonnull parameters. In the following we will concentrate on the Lasso estimator introduced in Tibshirani (1996). For a pre-specified parameter $\rho > 0$, an estimator $\boldsymbol{\theta}$ is then obtained as

(3.11) $$\widehat{\boldsymbol{\theta}} = \arg\min_{\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{k+p}} \frac{1}{n}\|\mathbf{Y} - \mathbf{\Phi}\widetilde{\boldsymbol{\theta}}\|_2 + 2\rho\|\widetilde{\boldsymbol{\theta}}\|_1,$$

$\mathbf{\Phi}$ being the $n \times (k+p)$-dimensional matrix with rows $\mathbf{\Phi}_i$. We can alternatively use the Dantzig selector introduced in Candes and Tao (2007).

Finally, from $\widehat{\boldsymbol{\theta}}$, we define corresponding estimators for α_r, $r = 1, \ldots, k$, and β_j, $j = 1, \ldots, p$, in the unprojected model (3.3).

$$\widehat{\beta}_j = \frac{\widehat{\widetilde{\beta}}_j}{((1/n)\sum_{i=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2)^{1/2}}, \qquad j = 1, \ldots, p,$$

and

$$\widehat{\alpha}_r = \widehat{\widetilde{\alpha}}_r - \sqrt{p\widehat{\lambda}_r} \sum_{j=1}^p \widehat{\psi}_{rj}\widehat{\beta}_j, \qquad r = 1, \ldots, k.$$

4. High-dimensional factor analysis: Theoretical results. The following theorem shows that principal components which are able to explain a considerable proportion of total variance can be estimated consistently.

For simplicity, we will concentrate on the case that n as well as $p > \sqrt{n}$ are large enough such that

(A.4) $C_0(\log p/n)^{1/2} \geq \frac{D_0}{p}$ and $v(k) \geq 6(D_2/p + C_0(\log p/n)^{1/2})$.

THEOREM 2. *Under assumptions* (A.1)–(A.4) *and under events* (2.1)–(2.4) *we have for all* $r = 1, \ldots, k$ *and all* $j = 1, \ldots, p$,

$$(4.1) \qquad |\lambda_r - \widehat{\lambda}_r| \leq \frac{D_2}{p} + C_0(\log p/n)^{1/2}$$

$$(4.2) \qquad \|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 2\frac{D_2/p + C_0(\log p/n)^{1/2}}{v(k)},$$

$$(4.3) \qquad \psi_{rj}^2 \leq \frac{D_0 - D_1}{p\lambda_r} \leq \frac{D_0 - D_1}{pv(k)},$$

$$(4.4) \qquad \widehat{\psi}_{rj}^2 \leq \frac{D_0 + C_0(\log p/n)^{1/2}}{p\widehat{\lambda}_r}$$

$$\leq \frac{6}{5}\frac{D_0 + C_0(\log p/n)^{1/2}}{pv(k)}.$$

Theorem 2 shows that for sufficiently large p ($p > \sqrt{n}$) the eigenvalues and eigenvectors of $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$ provide reasonable estimates of λ_r and $\boldsymbol{\psi}_r$ for $r = 1, \ldots, k$. Quite obviously it is not possible to determine sensible estimates of *all* p principal components of $\frac{1}{p}\boldsymbol{\Sigma}$. Following the proposition it is required that λ_r as well as μ_r be of order at least $\sqrt{\frac{\log p}{n}}$. Any smaller component cannot be distinguished from pure "noise" components. Up to the $\log p$-term this corresponds to the results of Hall and Hosseini-Nasab (2006) who study the problem of the number of principal components that can be consistently estimated in a functional principal component analysis.

The above insights are helpful for selecting an appropriate k in a real data application. In tendency, a suitable factor model will incorporate k components which explain a large percentage of the total variance of $\mathbf{X}_i$, while λ_{k+1} is very small. If for a sample of high-dimensional vectors $\mathbf{X}_i$ a principal component analysis leads to the conclusion that the first (or second, third, . . .) principal components explains a *large* percentage of the total (empirical) variance of the observations, then such a component cannot be generated by "noise" but reflects an underlying structure. In particular, such a component may play a crucial role in modeling a response variable Y_i according to an augmented regression model of the form (3.3).

Bai and Ng (2002) develop criteria of selecting the dimension k in a high-dimensional factor model. They rely on an adaptation of the well-known AIC and BIC procedures in model selection. One possible approach is as follows: Select a maximal possible dimension $k_{\max}$ and estimate $\bar{\sigma}^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_j^2$ by

$\widehat{\sigma}^2 = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \sum_{r=1}^{k_{\max}} (\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i) \widehat{\psi}_{rj})^2$. Then determine an estimate $\widehat{k}$ by minimizing

$$(4.5) \quad \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(X_{ij} - \sum_{r=1}^{\kappa} (\widehat{\boldsymbol{\psi}}_r \mathbf{X}_i) \widehat{\psi}_{rj} \right)^2 + \kappa \widehat{\sigma}^2 \left(\frac{n+p}{np} \right) \log \min\{n,p\}$$

over $\kappa = 1, \ldots, k_{\max}$. Bai and Ng (2002) show that under some regularity conditions this criterium (as well as a number of alternative versions) provides asymptotically consistent estimates of the true factor dimension k as $n, p \to \infty$. In our context these regularity conditions are satisfied if (A.1)–(A.4) hold for all n and p, $\sup_{j,p} \mathbb{E}(Z_{ij}^8) < \infty$ and if there exists some $B_0 > 0$ such that $\lambda_k \geq B_0 > 0$, for all n, p.

Now recall the modified version (3.10) of the augmented model used in our estimation procedure. The following theorem establishes bounds for the projections $(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j$ as well as for the additional error terms $\widetilde{\varepsilon}_i$. Let $\mathbf{P}_k = \mathbf{I}_p - \sum_{j=1}^{k} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T$ denote the population version of $\widehat{\mathbf{P}}_k$.

THEOREM 3. *Assume* (A.1) *and* (A.2). *There then exist constants* M_1, M_2, $M_3 < \infty$, *such that for all* n, p, k *satisfying* (A.3) *and* (A.4), *all* $j, l \in \{1, \ldots, p\}$, $j \neq l$,

$$(4.6) \quad \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2 \geq \sigma_j^2 - M_1 \frac{k n^{-1/2} \sqrt{\log p}}{v(k)^{1/2}},$$

$$(4.7) \quad \left| \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2 - \sigma_j^2 \right| \leq \mathbb{E}((\mathbf{P}_k \mathbf{W}_i)_j^2) + M_2 \frac{k n^{-1/2} \sqrt{\log p}}{v(k)^{3/2}}$$

hold with probability $A(n, p)$, *while*

$$(4.8) \quad \begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \widetilde{\varepsilon}_i^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{r=1}^{k} (\widehat{\xi}_{ir} - \xi_{ir}) \alpha_r \right)^2 \\ &\leq \frac{k \alpha_{\mathrm{sum}}^2 M_3}{v(k)^3} \left(\frac{\log p}{n} + \frac{v(k)^2}{p} \right) \end{aligned}$$

holds with probability at least $A(n, p) - \frac{k}{n}$. *Here,* $\alpha_{\mathrm{sum}}^2 = \sum_{r=1}^{k} \alpha_r^2$.

Note that if $\mathbf{X}_i$ satisfies a k-dimensional factor model, that is, if the rank of $\frac{1}{p} \boldsymbol{\Gamma}$ is equal to k, then $\mathbf{P}_k \mathbf{W}_i = 0$. The theorem then states that for large n and p the projected variables $(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j$, $j = 1, \ldots, k$, "in average" behave similarly to the specific variables Z_{ij}. Variances will be close to $\sigma_j^2 = \mathrm{Var}(Z_{ij})$.

5. Theoretical properties of the augmented model. We come back to model (3.3). As shown in Section 3.2, the Lasso or the Dantzig selector may be used to determine estimators of the parameters of the model. Identification of sparse solutions as well as consistency of estimators require structural assumptions on the explanatory variables. The weakest assumption on the correlations between different variables seems to be the so-called *restricted eigenvalue* condition introduced by Bickel, Ritov and Tsybakov (2009); see also Zhou, van de Geer and Bülhmann (2009).

We first provide a theoretical result which shows that for large n, p the design matrix of the projected model (3.10) satisfies the restricted eigenvalue conditions given in Bickel, Ritov and Tsybakov (2009) with high probability. We will additionally assume that n, p are large enough such that

(A.5) $D_1/2 > M_1 \frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{1/2}}$,

where M_1 is defined as in Theorem 3.

Let J_0 denote an arbitrary subset of indices, $J_0 \subset \{1, \ldots, p\}$ with $|J_0| \leq k + S$. For a vector $\mathbf{a} \in \mathbb{R}^{k+p}$, let $\mathbf{a}_{J_0}$ be the vector in $\mathbb{R}^{k+p}$ which has the same coordinates as $\mathbf{a}$ on J_0 and zero coordinates on the complement J_0^c of J_0. We define in the same way $\mathbf{a}_{J_0^c}$. Now for $k + S \leq (k+p)/2$ and for an integer $m \geq k + S$, $S + m \leq p$, denote by J_m the subset of $\{1, \ldots, k+p\}$ corresponding to m largest in absolute value coordinates of $\mathbf{a}$ outside of J_0, and define $J_{0,m} := J_0 \cup J_m$. Furthermore, let $(x)_+ = \max\{x, 0\}$.

PROPOSITION 2.   *Assume* (A.1) *and* (A.2). *There then exists a constant* $M_4 < \infty$, *such that for all* n, p, k, S, $k + S \leq (k+p)/2$, *satisfying* (A.3)–(A.5), *and* $c_0 = 1, 3$

$$\kappa(k+S, k+S, c_0)$$

$$:= \min_{J_0 \subset \{1,\ldots,k+p\}\,:\,|J_0|\leq k+S} \min_{\boldsymbol{\Delta} \neq 0\,:\,\|\boldsymbol{\Delta}_{J_0^c}\|_1 \leq c_0 \|\boldsymbol{\Delta}_{J_0}\|_1} \frac{[\boldsymbol{\Delta}^T (1/n) \sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T \boldsymbol{\Delta}]^{1/2}}{\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_2}$$

(5.1)

$$\geq \left(\frac{D_1}{D_0 + C_0 n^{-1/2}\sqrt{\log p}} - \frac{8(k+S)c_0 M_4 k^2 n^{-1/2}\sqrt{\log p}}{v(k)(D_1 - M_1 k v(k)^{1/2} n^{-1/2}\sqrt{\log p})} \right)^{1/2}_+$$

$$=: \mathbf{K}_{n,p}(k, S, c_0)$$

holds with probability $A(n, p)$.

Asymptotically, if n and p are large, then $\mathbf{K}_{n,p}(k, S, c_0) > 0$, $c_0 = 1, 3$, provided that k, S and $1/v(k)$ are sufficiently small compared to n, p. In this case the proposition implies that with high probability the restricted eigenvalue condition $\mathrm{RE}(k + S, k + S, c_0)$ of Bickel, Ritov and Tsybakov (2009) [i.e., $\kappa(k + S, k + S, c_0) > 0$] is satisfied. The same holds for the conditions $\mathrm{RE}(k + S, c_0)$ which require $\kappa(k + S, c_0) > 0$, where

$$\kappa(k + S, c_0)$$

$$:= \min_{J_0 \subset \{1,\ldots,k+p\}\,:\,|J_0| \le k+S} \min_{\boldsymbol{\Delta} \ne 0\,:\,\|\boldsymbol{\Delta}_{J_0^c}\|_1 \le c_0 \|\boldsymbol{\Delta}_{J_0}\|_1} \frac{[\boldsymbol{\Delta}^T (1/n) \sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T \boldsymbol{\Delta}]^{1/2}}{\|\boldsymbol{\Delta}_{J_0}\|_2}$$

$$\ge \kappa(k + S, k + S, c_0).$$

The following theorem now provides bounds for the L^1 estimation error and the L^2 prediction loss for the Lasso estimator of the coefficients of the augmented model. It generalizes the results of Theorem 7.2 of Bickel, Ritov and Tsybakov (2009) obtained under the standard linear regression model. In our analysis merely the values of $\kappa(k + S, c_0)$ for $c_0 = 3$ are of interest. However, only slight adaptations of the proofs are necessary in order to derive generalizations of the bounds provided by Bickel, Ritov and Tsybakov (2009) for the Dantzig selector ($c_0 = 1$) and for the L^q loss, $1 < q \le 2$. In the latter case, $\kappa(k + S, c_0)$ has to be replaced by $\kappa(k + S, k + S, c_0)$. In the following, let M_1 and M_3 be defined as in Theorem 3.

THEOREM 4. *Assume* (A.1), (A.2) *and suppose that the error terms* ε_i *in model (3.3) are independent* $\mathcal{N}(0, \sigma^2)$ *random variables with* $\sigma^2 > 0$. *Now consider the Lasso estimator* $\widehat{\boldsymbol{\theta}}$ *defined by (3.11) with*

$$\rho = A\sigma \sqrt{\frac{\log(k+p)}{n}} + \frac{M_5 \alpha_{\mathrm{sum}}}{v(k)^{3/2}} \sqrt{\frac{\log p}{n}},$$

where $A > 2\sqrt{2}$, M_5 *is a positive constant and* $\alpha_{\mathrm{sum}} = \sum_{r=1}^k |\alpha_r|$.

If $M_5 < \infty$ *is sufficiently large, then for all* $n, p, k,\ k + S \le (k + p)/2$, *satisfying* (A.3)–(A.5) *as well as* $\mathbf{K}_{n,p}(k, S, 3) > 0$, *the following inequalities hold with probability at least* $A(n, p) - (p + k)^{-A^2/2}$:

(5.2)
$$\sum_{r=1}^k |\widehat{\alpha}_r - \alpha_r| \le \frac{16(k + S)}{\kappa^2}$$

$$\times \rho \left(1 + \frac{k(D_0 + C_0 n^{-1/2}\sqrt{\log p})^{1/2}}{(D_1 - M_1(kn^{-1/2}\sqrt{\log p}/v(k)^{1/2}))^{1/2}}\right),$$

(5.3)
$$\sum_{j=1}^p |\widehat{\beta}_j - \beta_j| \le \frac{16(k + S)}{\kappa^2(D_1 - M_1(kn^{-1/2}\sqrt{\log p}/v(k)^{1/2}))^{1/2}} \rho,$$

where $\kappa = \kappa(k + S, 3)$. *Moreover,*

(5.4)
$$\frac{1}{n} \sum_{i=1}^n \left(\sum_{r=1}^k \widehat{\xi}_{ir} \widehat{\alpha}_r + \sum_{j=1}^p X_{ij} \widehat{\beta}_j - \left(\sum_{r=1}^k \xi_{ir} \alpha_r + \sum_{j=1}^p X_{ij} \beta_j\right)\right)^2$$

$$\le \frac{32(k + S)}{\kappa^2} \rho^2 + \frac{2k\alpha_{\mathrm{sum}}^2 M_3}{v(k)^3} \left(\frac{\log p}{n} + \frac{v(k)^2}{p}\right)$$

holds with probability at least $A(n, p) - (p + k)^{-A^2/2} - \frac{k}{n}$.

Of course, the main message of the theorem is asymptotic in nature. If n, p tend to infinity for fixed values of k and S, then the L_1 estimation error and the L^2 prediction error converge at rates $\sqrt{\log p/n}$ and $\log p/n + 1/p$, respectively. For values of k and S tending to infinity as the sample size tends to infinity, the rates are more complicated. In particular, they depend on how fast $v(k)$ converges to zero as $k \to \infty$. Similar results hold for the estimators based on the Dantzig selector.

REMARK 1. Note that Proposition 2 as well as the results of Theorem 4 heavily depend on the validity of assumption (A.1) and the corresponding value $0 < D_1 \leq \inf_j \sigma_j^2$, where $\sigma_j^2 = \mathrm{var}(Z_{ij})$. It is immediately seen that the smaller the D_1 the smaller the value of $\kappa(k + S, k + S, c_0)$ in (5.1). This means that *all* variables $X_{ij} = W_{ij} + Z_{ij}$, $j = 1, \ldots, p$ have to possess a sufficiently large specific variation which is not shared by other variables. For large p this may be seen as a restrictive assumption. In such a situation one may consider a restricted version of model (3.3), where variables with extremely small values of σ_j^2 are eliminated. But for large n, p we can infer from Theorem 3 that a small value of $\frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2$ indicates that also σ_j^2 is small. Hence, an extension of our method consists of introducing some threshold $D_{\mathrm{thresh}} > 0$ and discarding all those variables X_{ij}, $j \in \{1, \ldots, p\}$, with $\frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2 < D_{\mathrm{thresh}}$. A precise analysis is not in the scope of the present paper.

REMARK 2. If $\alpha_1 = \cdots = \alpha_k = 0$ the augmented model reduces to the standard linear regression model (1.1) with a sparse set of coefficients, $\sharp\{j|\beta_j \neq 0\} \leq S$ for some $S \leq p$. An application of our estimation procedure is then unnecessary, and coefficients may be estimated by traditional model selection procedures. Bounds on estimation errors can therefore be directly obtained from the results of Bickel, Ritov and Tsybakov (2009), provided that the restricted eigenvalue conditions are satisfied. But in this situation a slight adaptation of the proof of Proposition 2 allows us to establish a result similar to (5.1) for the standardized variables $X_{ij}^* := X_{ij}/(\frac{1}{n} \sum_{i=1}^n X_{ij}^2)^{1/2}$. Define $\mathbf{X}^*$ as the $n \times p$-matrix with generic elements X_{ij}^*. When assuming (A.1), (A.2) as well as $D_1 - 3C_0 n^{-1/2}\sqrt{\log p} > 0$, then for $S \leq p/2$ the following inequality holds with probability $A(n, p)$:

$$\kappa(S, S, c_0)$$

$$(5.5) \quad := \min_{J_0 \subset \{1, \ldots, p\} : |J_0| \leq S} \min_{\boldsymbol{\Delta} \neq 0 : \|\boldsymbol{\Delta}_{J_0^c}\|_1 \leq c_0 \|\boldsymbol{\Delta}_{J_0}\|_1} \frac{[\boldsymbol{\Delta}^T (1/n) \sum_{i=1}^n \mathbf{X}_i^* \mathbf{X}_i^{*T} \boldsymbol{\Delta}]^{1/2}}{\|\boldsymbol{\Delta}_{J_0, S}\|_2}$$

$$\geq \left(\frac{D_1}{D_0 + C_0 n^{-1/2}\sqrt{\log p}} - \frac{8Sc_0C_0 n^{-1/2}\sqrt{\log p}}{D_1 - 3C_0 n^{-1/2}\sqrt{\log p}} \right)_+^{1/2},$$

where $c_0 = 1, 3$. Recall, however, from the discussion in Section 3.1 that $\alpha_1 = \cdots = \alpha_k = 0$ is a restrictive condition in the context of highly correlated regressors.

6. Simulation study. In this section we study the finite sample performance of the estimators discussed in the proceeding sections. We consider a factor model with $k = 2$ factors. The first factor is $\psi_{1j} = 1/\sqrt{p}$, $j = 1, \ldots, p$, while the second factor is given by $\psi_{2j} = 1/\sqrt{p}$, $j = 1, \ldots, p/2$, and $\psi_{2j} = -1/\sqrt{p}$, $j = p/2 + 1, \ldots, p$. For different values of n, p, α_1, α_2 and $0 < \lambda_1 < 1, 0 < \lambda_2 < 1$ observations $(\mathbf{X}_i, Y_i)$ with $\text{var}(X_{ij}) = 1$, $j = 1, \ldots, p$, are generated according to the model

$$(6.1) \qquad X_{ij} = \sqrt{p\lambda_1}\xi_{i1}\psi_{1j} + \sqrt{p\lambda_2}\xi_{i2}\psi_{2j} + Z_{ij},$$

$$(6.2) \qquad Y_i = \alpha_1\xi_{i1} + \alpha_2\xi_{i1} + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i,$$

where $\xi_{ir} \sim N(0, 1)$, $r = 1, 2$, $Z_{ij} \sim N(0, 1 - \lambda_1 - \lambda_2)$, and $\varepsilon_i \sim N(0, \sigma^2)$ are independent variables. Our study is based on $S = \sharp\{j|\beta_j \neq 0\} = 4$ nonzero β-coefficients whose values are $\beta_{10} = 1$, $\beta_{20} = 0.3$, $\beta_{21} = -0.3$ and $\beta_{40} = -1$, while the error variance is set to $\sigma^2 = 0.1$.

The parameters of the augmented model with $k = 2$ are estimated by using the Lasso-based estimation procedure described in Section 3.2. The behavior of the estimates is compared to the Lasso estimates of the coefficients of a standard regression model (1.1). All results reported in this section are obtained by applying the LARS-package by Hastie and Efron implemented in R. All tables are based on 1,000 repetitions of the simulation experiments. The corresponding R-code can be obtained from the authors upon request.

Figure 1 and Table 1 refer to the situation with $\lambda_1 = 0.4$, $\lambda_2 = 0.2$, $\alpha_1 = 1$ and $\alpha_2 = -0.5$. We then have $\text{var}(X_{ij}) = 1$, $j = 1, \ldots, p$, while the first and second factor explain 40% and 20% of the total variance of X_{ij}, respectively.

Figure 1 shows estimation results of one typical simulation with $n = p = 100$. The left panel contains the parameter estimates for the augmented model. The paths of estimated coefficients $\widehat{\beta}_j$ for the 4 significant variables (black lines), the 96 variables with $\beta_j = 0$ (red lines), as well as of the untransformed estimates $\widehat{\widetilde{\alpha}}_r$ (blue lines) of α_r, $r = 1, 2$, are plotted as a function of ρ. The four significant coefficients as well as α_1 and α_2 can immediately been identified in the figure. The right panel shows a corresponding plot of estimated coefficients when Lasso is directly applied to the standard regression model (1.1). As has to be expected by (3.9) the necessity of compensating the effects of α_1, α_2 by a large number of small, nonzero coefficients generates a general "noise level" which makes it difficult to identify the four significant variables in (6.2). The penalties ρ in the figure as well as in subsequent tables have to be interpreted in terms of the scaling used by the
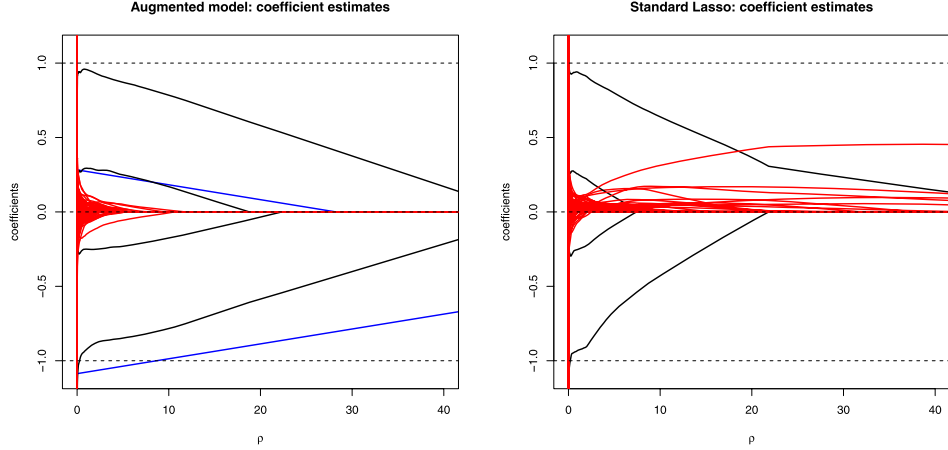
FIG. 1.   *Paths of Lasso estimates for the augmented model (left panel) and the standard linear model (right panel) in dependence of ρ; black—estimates of nonzero β_j; red—estimates of coefficients with $\beta_j = 0$; blue—$\widehat{\widehat{\alpha}}_r$.*

TABLE 1
Estimation errors for different sample sizes $(\lambda_1 = 0.4, \lambda_2 = 0.2, \alpha_1 = 1, \alpha_2 = -0.5)$

Sample sizes		Parameter estimates			Prediction			
n	p	$\sum|\widehat{\alpha}_r - \alpha_r|$	$\sum|\widehat{\beta}_r - \beta_r|$	Opt. ρ	Sample	Exact	Opt. ρ	C_P
			Lasso applied to augmented model:					
50	50	0.3334	0.8389	4.53	0.0498	0.1004	1.76	1.55
100	100	0.2500	0.5774	6.84	0.0328	0.0480	3.50	3.29
250	250	0.1602	0.3752	12.27	0.0167	0.0199	7.55	7.22
500	500	0.1150	0.2752	18.99	0.0096	0.0106	12.66	12.21
5,000	100	0.0378	0.0733	48.33	0.0152	0.0154	27.48	26.74
100	2,000	0.2741	0.8664	10.58	0.0420	0.0651	5.42	5.24
n	p	$\sum|\widehat{\beta}_r - \beta_r^{LR}|$	$\sum|\widehat{\beta}_r - \beta_r|$		Sample	Exact	Opt. ρ	
			Lasso applied to standard linear regression model:					
50	50	2.2597	1.8403		0.0521	0.1370	0.92	
100	100	2.2898	1.9090		0.0415	0.0725	1.90	
250	250	2.3653	1.7661		0.0257	0.0345	4.12	
500	500	2.4716	1.7104		0.0174	0.0207	6.87	
5,000	100	0.5376	1.5492		0.0161	0.0168	10.14	
100	2,000	3.7571	2.2954		0.0523	0.1038	3.17	

LARS-algorithm and have to be multiplied with $2/n$ in order to correspond to the standardization used in the proceeding sections.

The upper part of Table 1 provides simulation results with respect to the augmented model for different sample sizes n and p. In order to access

the quality of parameter estimates we evaluate $\sum_{r=1}^{2} |\widehat{\alpha}_r - \alpha_r|$ as well as $\sum_{j=1}^{p} |\widehat{\beta}_r - \beta_r|$ at the optimal value of ρ, where the minimum of $\sum_{r=1}^{2} |\widehat{\alpha}_r - \alpha_r| + \sum_{j=1}^{p} |\widehat{\beta}_r - \beta_r|$ is obtained. Moreover, we record the value of ρ where the minimal sample prediction error

$$(6.3) \quad \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{r=1}^{2} \alpha_r \xi_{ir} + \sum_{j=1}^{p} \beta_j X_{ij} - \left(\sum_{r=1}^{2} \widehat{\alpha}_r \widehat{\xi}_{ir} + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij} \right) \right)^2$$

ia attained. For the same value ρ we also determine the exact prediction error

$$(6.4) \quad \mathbb{E} \left(\sum_{r=1}^{2} \alpha_r \xi_{n+1,r} + \sum_{j=1}^{p} \beta_j X_{n+1,j} - \left(\sum_{r=1}^{2} \widehat{\alpha}_r \widehat{\xi}_{n+1,r} + \sum_{j=1}^{p} \widehat{\beta}_j X_{n+1,j} \right) \right)^2$$

for a new observation $\mathbf{X}_{n+1}$ independent of $X_1, \ldots, X_n$. The columns of Table 1 report the average values of the corresponding quantities over the 1,000 replications. To get some insight into a practical choice of the penalty, the last column additionally yields the average value of the parameters ρ minimizing the C_P-statistics. C_P is computed by using the R-routine "summary.lars" and plugging in the true error variance $\sigma^2 = 0.1$. We see that in all situations the average value of ρ minimizing C_P is very close to the average ρ providing the smallest prediction error. The penalties for optimal parameter estimation are, of course, larger.

It is immediately seen that the quality of estimates considerably increases when going from $n = p = 50$ to $n = p = 500$. An interesting result consists of the fact that the prediction error is smaller for $n = p = 500$ than for $n = 5,000$, $p = 100$. This may be interpreted as a consequence of (3.8).

The lower part of Table 1 provides corresponding simulation results with respect to Lasso estimates based on the standard regression model. In addition to the minimal error $\sum_{r=1}^{p} |\widehat{\beta}_r - \beta_r|$ in estimating the parameters β_j of (6.2) we present the minimal L_1-distance $\sum_{r=1}^{p} |\widehat{\beta}_r - \beta_r^{LR}|$, where $\beta_1^{LR}, \ldots, \beta_p^{LR}$ is the (nonsparse) set of parameters minimizing the population prediction error. Sample and exact prediction errors are obtained by straightforward modifications of (6.3) and (6.4). Quite obviously, no reasonable parameter estimates are obtained in the cases with $p \geq n$. Only for $n = 5,000$, $p = 100$, the table indicates a comparably small error $\sum_{r=1}^{p} |\widehat{\beta}_r - \beta_r^{LR}|$. The prediction error shows a somewhat better behavior. It is, however, always larger than the prediction error of the augmented model. The relative difference increases with p.

It was mentioned in Section 4 that a suitable criterion to estimate the dimension k of an approximate factor model consists in minimizing (4.5). This criterion proved to work well in our simulation study. Recall that the

TABLE 2
Estimation errors under different setups ($n = 100, p = 250$)

λ_1	λ_2	α_1	α_2	Parameter estimates			Prediction		
				$\sum |\widehat{\alpha}_r - \alpha_r|$	$\sum |\widehat{\beta}_r - \beta_r|$	Opt. ρ	Sample	Exact	Opt. ρ
				Lasso applied to augmented model:					
0.06	0.03	1	-0.5	0.3191	0.6104	10.37	0.0670	0.2259	2.46
0.2	0.1	1	-0.5	0.2529	0.5335	8.19	0.0414	0.0727	3.92
0.4	0.2	1	-0.5	0.2500	0.6498	7.86	0.0319	0.0454	4.35
0.6	0.3	1	-0.5	0.2866	1.1683	8.46	0.0273	0.0350	4.56
0.06	0.03	0	0	0.0908	0.4238	7.42	0.0257	0.0311	4.62
0.2	0.1	0	0	0.1044	0.4788	7.64	0.0257	0.0316	4.69
0.4	0.2	0	0	0.1192	0.6400	7.84	0.0250	0.0314	4.74
0.6	0.3	0	0	0.1825	1.1745	8.78	0.0221	0.0276	5.13
λ_1	λ_2	α_1	α_2	$\sum |\widehat{\beta}_r - \beta_r^{LR}|$	$\sum |\widehat{\beta}_r - \beta_r|$		Sample	Exact	Opt. ρ
				Lasso applied to standard linear regression model:					
0.06	0.03	1	-0.5	5.0599	1.9673		0.0777	0.3758	1.95
0.2	0.1	1	-0.5	3.4465	2.3662		0.0583	0.1403	2.63
0.4	0.2	1	-0.5	2.9215	2.0191		0.0425	0.0721	2.45
0.6	0.3	1	-0.5	3.2014	2.2246		0.0277	0.0387	1.47
0.06	0.03	0	0	0.4259	0.4259		0.0216	0.0285	4.80
0.2	0.1	0	0	0.4955	0.4955		0.0222	0.0295	4.24
0.4	0.2	0	0	0.6580	0.6580		0.0228	0.0303	3.17
0.6	0.3	0	0	1.1990	1.1990		0.0215	0.0283	1.66

true factor dimension is $k = 2$. For $n = p = 50$ the average value of the estimate $\widehat{k}$ determined by (4.5) is 2.64. In all other situations reported in Table 1 an estimate $\widehat{k} = 2$ is obtained in each of the 1,000 replications.

Finally, Table 2 contains simulations results for $n = 100$, $p = 250$, and different values of $\lambda_1, \lambda_2, \alpha_1, \alpha_2$. All columns have to be interpreted similar to those of Table 1. For $\alpha_1 = 1, \alpha_2 = -0.5$ suitable parameter estimates can obviously only been determined by applying the augmented model. For $\alpha_1 = \alpha_2 = 0$ model (6.2) reduces to a sparse, standard linear regression model. It is then clearly unnecessary to apply the augmented model. Both methods then lead to roughly equivalent parameter estimates.

We want to emphasize that $\lambda_1 = 0.06$, $\lambda_2 = 0.03$ constitutes a particularly difficult situation. Then the first and second factor only explain 6% and 3% of the variance of X_{ij}. Consequently, $v(k)$ is very small and one will expect a fairly large error in estimating ξ_{ir}. Somewhat surprisingly the augmented model still provides reasonable parameter estimates, the only problem in this case seems to be a fairly large prediction error.

Another difficult situation in an opposite direction is $\lambda_1 = 0.6$, $\lambda_2 = 0.3$. Then both factors together explain 90% of the variability of X_{ij}, while Z_{ij} only explains the remaining 10%. Consequently, D_1 is very small and one

may expect problems in the context of the restricted eigenvalue condition. The table shows that this case yields the smallest prediction error, but the quality of parameter estimates deteriorates.

APPENDIX

PROOF OF PROPOSITION 1. Define $Q_{ijl} = V_{ij}V_{il} - \mathbb{E}(V_{ij}V_{il})$, $i = 1, \ldots, n$, $1 \leq j, l \leq p$. For any $C > 0$ and $\varepsilon > 0$, noting that $\mathbb{E}(Q_{ijl}) = 0$, we have

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - \mathbb{E}(V_{ij}V_{il})\right| > \varepsilon\right)
$$

$$
= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}\right| > \varepsilon\right)
$$

$$
= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}I(|Q_{ijl}| \leq C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| \leq C))\right.\right.
$$

$$
\left.\left. + Q_{ijl}I(|Q_{ijl}| > C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| > C))\right| > \varepsilon\right)
$$

$$
\leq P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}I(|Q_{ijl}| \leq C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| \leq C))\right| > \varepsilon/2\right)
$$

$$
+ P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}I(|Q_{ijl}| > C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| > C))\right| > \varepsilon/2\right),
$$

where $I(\cdot)$ is the indicator function. We have

$$
|Q_{ijl}I(|Q_{ijl}| \leq C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| \leq C))| \leq 2C
$$

and

$$
\mathbb{E}((Q_{ijl}I(|Q_{ijl}| \leq C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| \leq C)))^2) \leq \mathrm{Var}(V_{ij}V_{il})
$$

$$
\leq (\mathbb{E}(V_{ij}^4)\mathbb{E}(V_{il}^4))^{1/2}
$$

$$
\leq C_1.
$$

Applying the Bernstein inequality for bounded centered random variables [see Hoeffding (1963)] we get

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}I(|Q_{ijl}| \leq C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| \leq C))\right| > \frac{\varepsilon}{2}\right)
$$

(A.1)

$$
\leq \exp\left\{\frac{-\varepsilon^2 n}{8(C_1 + C\varepsilon/3)}\right\}.
$$

We have now

$$(A.2) \quad P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ijl}I(|Q_{ijl}| > C) - \mathbb{E}(Q_{ijl}I(|Q_{ijl}| > C))\right| > \varepsilon/2\right)$$

$$\leq \sum_{i=1}^{n} P(|Q_{ijl}| > C) + P(\mathbb{E}(|Q_{ijl}|I(|Q_{ijl}| > C)) > \varepsilon/4).$$

Using Markov's inequality and (2.5) we obtain

$$(A.3) \quad P(|Q_{ijl}| > C) \leq P(|V_{ij}| > \sqrt{C/2}) + P(|V_{il}| > \sqrt{C/2})$$
$$+ P((\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2} > C/2)$$
$$\leq \frac{2C_1}{e^{a\sqrt{C/2}}} + P((\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2} > C/2).$$

Choose $\varepsilon = C_0\sqrt{\log p/n}$ and $C = \sqrt{C_0 n/\log p}$, where C_0 is a positive constant such that $C_1^{1/2} \leq \frac{1}{2}\sqrt{C_0 n \log p}$ and $C_1 \leq \frac{1}{8}C_0 e^{a\sqrt{C_0 n/\log p}}\sqrt{\log p/n}$. Note now that

$$P((\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2} > C/2) = 0,$$

while

$$\mathbb{E}(|Q_{ijl}|I(|Q_{ijl}| > C)) \leq (\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2}P(|Q_{ijl}| > C) \leq \frac{2C_1^{3/2}}{e^{a\sqrt{C_0 n/\log p}}},$$

which implies

$$P(\mathbb{E}(|Q_{ijl}|I(|Q_{ijl}| > C)) > \varepsilon/4) = 0.$$

Inequalities (A.1), (A.2) and (A.3) lead finally to

$$(A.4) \quad P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - \mathbb{E}(V_{ij}V_{il})\right| > C_0\sqrt{\frac{\log p}{n}}\right)$$
$$\leq p^{-C_0^2/(8(C_1 + C_0^{3/2}/3))} + 2nC_1 e^{-(a/2)(n/\log p)^{1/4}}.$$

The result (2.6) is now a consequence of (A.4) since

$$P\left(\sup_{1 \leq j,l \leq p}\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - \mathbb{E}(V_{ij}V_{il})\right| > C_0\sqrt{\frac{\log p}{n}}\right)$$
$$\leq \sum_{j=1}^{p}\sum_{l=1}^{p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - \mathbb{E}(V_{ij}V_{il})\right| > C_0\sqrt{\frac{\log p}{n}}\right). \qquad \square$$

PROOF OF THEOREM 1. For any symmetric matrix $\mathbf{A}$, we denote by $\lambda_1(\mathbf{A}) > \lambda_2(\mathbf{A}) > \cdots$ its eigenvalues. Weyl's perturbation theorem [see, e.g., Bhatia (1997), page 63] implies that for any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ and all $r = 1, 2, \ldots$

$$(A.5) \qquad |\lambda_r(\mathbf{A} + \mathbf{B}) - \lambda_r(\mathbf{A})| \leq \|\mathbf{B}\|,$$

where $\|\mathbf{B}\|$ is the usual matrix norm defined as

$$\|\mathbf{B}\| = \sup_{\|\mathbf{u}\|_2 = 1} (\mathbf{u}^T \mathbf{B} \mathbf{B}^T \mathbf{u})^{1/2}.$$

Since $\frac{1}{p}\boldsymbol{\Sigma} = \frac{1}{p}\boldsymbol{\Gamma} + \frac{1}{p}\boldsymbol{\Psi}$, (A.5) leads to $|\mu_r - \lambda_r| \leq \|\frac{1}{p}\boldsymbol{\Psi}\|$. By assumption, $\frac{1}{p}\boldsymbol{\Psi}$ is a diagonal matrix with diagonal entries $\frac{D_1}{p} \leq \frac{\sigma_i^2}{p} \leq \frac{D_2}{p}$, $j = 1, \ldots, p$. Therefore $\|\frac{1}{p}\boldsymbol{\Psi}\| \leq \frac{D_2}{p}$ and (3.5) is an immediate consequence.

In order to verify (3.6) first note that Lemma A.1 of Kneip and Utikal (2001) implies that for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$

$$(A.6) \qquad \begin{aligned} \|\boldsymbol{\psi}_r(\mathbf{A} + \mathbf{B}) - \boldsymbol{\psi}_r(\mathbf{A})\|_2 &\leq \frac{\|\mathbf{B}\|}{\min_{j \neq r} |\lambda_j(\mathbf{A}) - \lambda_r(\mathbf{A})|} \\ &\quad + \frac{6\|\mathbf{B}\|^2}{\min_{j \neq r} |\lambda_j(\mathbf{A}) - \lambda_r(\mathbf{A})|^2}, \end{aligned}$$

where $\boldsymbol{\psi}_1(\mathbf{A}), \boldsymbol{\psi}_2(\mathbf{A}), \ldots$ are the eigenvectors corresponding to the eigenvalues $\lambda_1(\mathbf{A}) > \lambda_2(\mathbf{A}) > \cdots$. By assumption (A.3) this implies

$$\|\mu_r - \psi_r\|_2 \leq \frac{\|(1/p)\boldsymbol{\Psi}\|}{v(k)} + \frac{6\|(1/p)\boldsymbol{\Psi}\|^2}{v(k)^2} \leq \frac{2D_2}{pv(k)}$$

for all $r = 1, \ldots, k$. Since $D_0 \geq \mathbb{E}(X_{ij}^2) = \sum_{r=1}^{p} \delta_{rj}^2 p\mu_r$, the second part of (3.6) follows from $D_0 \geq \delta_{rj}^2 p\mu_r$, $j = 1, \ldots, p$.

By (A.3) we necessarily have $\delta_r \geq \lambda_r \geq v(k)$ for all $r = 1, \ldots, k$. Consequently, $\sqrt{\mu_r} - \sqrt{\lambda_r} = \frac{\mu_r - \lambda_r}{\sqrt{\mu_r} + \sqrt{\lambda_r}} \leq \frac{\mu_r - \lambda_r}{2\sqrt{v(k)}}$. Furthermore, note that $\xi_{ir}^* = \xi_{ir} + \frac{\boldsymbol{\delta}_r^T \mathbf{Z}_i}{\sqrt{p\mu_r}} + (\frac{\boldsymbol{\delta}_r}{\sqrt{p\mu_r}} - \frac{\boldsymbol{\psi}_r}{\sqrt{p\lambda_r}})^T \mathbf{W}_i$. Since W_i and Z_i are uncorrelated, (3.5) and (3.6) lead to

$$\begin{aligned} \mathbb{E}([\xi_{ir} - \xi_{ir}^*]^2) &= \frac{1}{\mu_r}\boldsymbol{\delta}_r^T \frac{1}{p}\boldsymbol{\Psi}\boldsymbol{\delta}_r \\ &\quad + \left(\frac{\boldsymbol{\delta}_r - \boldsymbol{\psi}_r}{\sqrt{\mu_r}} + \frac{\sqrt{\mu_r} - \sqrt{\lambda_r}}{\sqrt{\mu_r}\sqrt{\lambda_r}}\boldsymbol{\psi}_r\right)^T \frac{1}{p}\boldsymbol{\Gamma}\left(\frac{\boldsymbol{\delta}_r - \boldsymbol{\psi}_r}{\sqrt{\mu_r}} + \frac{\sqrt{\mu_r} - \sqrt{\lambda_r}}{\sqrt{\mu_r}\sqrt{\lambda_r}}\boldsymbol{\psi}_r\right) \\ &\leq \frac{D_2}{p\mu_r} + 2\frac{\lambda_1}{\mu_r}\|\boldsymbol{\delta}_r - \boldsymbol{\psi}_r\|^2 + 2\frac{1}{\mu_r}(\sqrt{\mu_r} - \sqrt{\lambda_r})^2 \end{aligned}$$

$$\leq \frac{D_2}{p\mu_r} + \frac{8\lambda_1 D_2^2}{\mu_r p^2 v(k)^2} + \frac{D_2^2}{2\mu_r p^2 v(k)}.$$

Since $\frac{\boldsymbol{\psi}_r}{\sqrt{p\lambda_r}}^T \mathbf{X}_i = \xi_{ir} + \frac{\boldsymbol{\psi}_r}{\sqrt{p\lambda_r}}^T \mathbf{Z}_i$ the second part of (3.7) follows from similar arguments. Finally, using the Cauchy–Schwarz inequality (3.8) is a straightforward consequence of (3.7). $\quad\square$

PROOF OF THEOREM 2. With $\mathbf{A} = \boldsymbol{\Gamma}$ and $\mathbf{B} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Gamma} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} + \boldsymbol{\Psi}$, inequality (A.5) implies that for all $r \in \{1, \ldots, k\}$

$$(A.7) \qquad |\widehat{\lambda}_r - \lambda_r| \leq \left\| \frac{1}{p}\boldsymbol{\Psi} + \frac{1}{p}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \right\| \leq \left\| \frac{1}{p}\boldsymbol{\Psi} \right\| + \left\| \frac{1}{p}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \right\|.$$

But under events (2.1)–(2.4) we have

$$\left\| \frac{1}{p}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \right\|$$

$$= \sup_{\|\mathbf{u}\|_2 = 1} \left[\mathbf{u}^T \frac{1}{p^2}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^2 \mathbf{u} \right]^{1/2}$$

$$= \sup_{\|\mathbf{u}\|_2 = 1} \left[\frac{1}{p^2} \sum_{j=1}^{p} \left(\sum_{l=1}^{p} \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}X_{i,l} - \mathrm{Cov}(X_{i,j}, X_{i,l}) \right) u_l \right)^2 \right]^{1/2}$$

$$\leq \sup_{\|\mathbf{u}\|_2 = 1} \left[\frac{1}{p^2} \sum_{j=1}^{p} \|\mathbf{u}\|_2^2 \sum_{l=1}^{p} \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}X_{i,l} - \mathrm{Cov}(X_{i,j}, X_{i,l}) \right)^2 \right]^{1/2}$$

$$\leq C_0 \sqrt{\frac{\log p}{n}}.$$

On the other hand, $\|\frac{1}{p}\boldsymbol{\Psi}\| \leq \frac{D_2}{p}$, and we can conclude that

$$(A.8) \qquad\qquad \left\| \frac{1}{p}\boldsymbol{\Psi} \right\| + \left\| \frac{1}{p}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \right\| \leq \frac{D_2}{p} + C_0 \sqrt{\frac{\log p}{n}}.$$

Relation (4.1) now is an immediate consequence of (A.7) and (A.8).

Relations (A.6) and (A.8) together with (A.3), (A.4) and (2.1)–(2.4) lead to

$$\|\widehat{\boldsymbol{\psi}}_r - \boldsymbol{\psi}_r\|_2 \leq \frac{\|(1/p)\boldsymbol{\Psi} + (1/p)(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|}{\min_{j \neq l} |\lambda_j - \lambda_l|}$$

$$+ \frac{6\|(1/p)\boldsymbol{\Psi} + (1/p)(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_2^2}{\min_{j \neq l} |\lambda_j - \lambda_l|^2}$$

$$\leq \frac{D_2/p + C_0(\log p/n)^{1/2}}{v(k)}$$

$$+ \frac{6(D_2/p + C_0(\log p/n)^{1/2})^2}{v(k)^2}$$

$$\leq 2\frac{D_2/p + C_0(\log p/n)^{1/2}}{v(k)},$$

which gives (4.2). It remains to show (4.3) and (4.4). Note that the spectral decompositions of $\boldsymbol{\Psi}$ and $\widehat{\boldsymbol{\Sigma}}$ imply that for all $j = 1, \ldots, p$

$$\mathbb{E}(W_{ij}^2) = \sum_{r=1}^{p} \psi_{rj}^2 p\lambda_r, \qquad \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = \sum_{r=1}^{p} \widehat{\psi}_{rj}^2 p\widehat{\lambda}_r.$$

Under events (2.1)–(2.4) we therefore obtain for all $r \leq k$

$$\text{(A.9)} \qquad \psi_{rj}^2 \leq \frac{\mathbb{E}(W_{ij}^2)}{p\lambda_r} \leq \frac{D_0 - D_1}{p\lambda_r} \leq \frac{D_0 - D_1}{pv(k)},$$

$$\text{(A.10)} \qquad \widehat{\psi}_{rj}^2 \leq \frac{(1/n)\sum_{i=1}^{n} X_{ij}^2}{p\widehat{\lambda}_r} \leq \frac{D_0 + C_0(\log p/n)^{1/2}}{p\widehat{\lambda}_r}.$$

But by assumptions (A.3) and (A.4), relation (4.1) leads to $\widehat{\lambda}_r \geq \frac{5v(k)}{6}$. Equations (4.3) and (4.4) then are immediate consequences of (A.9) and (A.10)
$\square$

PROOF OF THEOREM 3. Choose an arbitrary $j \in \{1, \ldots, p\}$. Note that $(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j = X_{ij} - \sum_{r=1}^{k} \widehat{\psi}_{rj}\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i$. Since $X_{ij} = W_{ij} + Z_{ij}$ we obtain the decomposition

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(W_{ij} - \sum_{r=1}^{k}\widehat{\psi}_{rj}\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i\right)^2$$

$$\text{(A.11)} \qquad\qquad + 2\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\left(W_{ij} - \sum_{r=1}^{k}\widehat{\psi}_{rj}\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i\right)$$

$$+ \frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2.$$

Under events (2.1)–(2.4), we have $|\sigma_j^2 - \frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2| \leq C_0 n^{-1/2}\sqrt{\log p}$ as well as $|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}W_{ij}| \leq C_0 n^{-1/2}\sqrt{\log p}$. Furthermore, $\mathbb{E}(Z_{ij}X_{ij}) = \sigma_j^2$ and $\mathbb{E}(Z_{ij}X_{il}) = 0$ for $j \neq l$. Therefore,

$$\left|\sum_{r=1}^{k}\widehat{\psi}_{rj}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i\right)\right| \leq \sum_{r=1}^{k}\widehat{\psi}_{rj}^2\sigma_j^2 + 2C_0\sqrt{\frac{\log p}{n}}\sum_{r=1}^{k}|\widehat{\psi}_{rj}|\left(\sum_{l=1}^{p}|\widehat{\psi}_{rl}|\right).$$

Obviously, $\sum_{l=1}^{p} |\widehat{\psi}_{rl}| \leq \sqrt{p \sum_{l=1}^{p} \widehat{\psi}_{rl}^2} = \sqrt{p}$. It now follows from Theorem 2 that there exists a constant $M_1 < \infty$, which can be chosen independently of all values n, p, k, S satisfying assumptions (A.3) and (A.4), such that

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 - \sigma_j^2 \geq \frac{1}{n}\sum_{i=1}^{n}\left(W_{ij} - \sum_{r=1}^{k}\widehat{\psi}_{rj}\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i\right)^2$$

(A.12)
$$- M_1\frac{k}{v(k)^{1/2}}\sqrt{\frac{\log p}{n}}$$

$$\geq -M_1\frac{k}{v(k)^{1/2}}\sqrt{\frac{\log p}{n}}.$$

Since events (2.1)–(2.4) have probability $A(n, p)$, assertion (4.6) is an immediate consequence.

In order to show (4.7) first recall that the eigenvectors of $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$ possess the well-known "best basis" property, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{P}}_k\mathbf{X}_i\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{X}_i - \sum_{r=1}^{k}\widehat{\boldsymbol{\psi}}_r(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)\right\|_2^2$$

$$= \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\min_{\theta_1,\ldots,\theta_k\in\mathbb{R}}\left\|X_i - \sum_{r=1}^{k}\theta_r\mathbf{w}_r\right\|_2^2.$$

For $j = 1, \ldots, p$ and $r = 1, \ldots, k$ define $\widetilde{\boldsymbol{\psi}}_r^{(j)} \in \mathbb{R}^p$ by $\widetilde{\psi}_{rj}^{(j)} = \psi_{rj}$ and $\widetilde{\psi}_{rl}^{(j)} = \widehat{\psi}_{rl}$, $l \neq j$. The above property then implies that for any j

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{P}}_k\mathbf{X}_i\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left\|X_i - \sum_{r=1}^{k}\widetilde{\boldsymbol{\psi}}_r^{(j)}(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)\right\|_2^2.$$

Since the vectors $\widehat{\mathbf{P}}_k\mathbf{X}_i$ and $\mathbf{X}_i - \sum_{r=1}^{p}\widetilde{\boldsymbol{\psi}}_r^{(j)}(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)$ only differ in the jth element, one can conclude that for any $j = 1, \ldots, p$

(A.13)
$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \sum_{r=1}^{k}\psi_{rj}(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)\right)^2.$$

The spectral decomposition of $\widehat{\boldsymbol{\Sigma}}$ implies that $\widehat{\boldsymbol{\Sigma}} = \sum_{r=1}^{p}p\widehat{\lambda}_r\widehat{\boldsymbol{\psi}}_r\widehat{\boldsymbol{\psi}}_r^T$ with

$$p\widehat{\lambda}_r = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)^2, \qquad \frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i)(\widehat{\boldsymbol{\psi}}_s^T\mathbf{X}_i) = 0, \qquad s \neq r.$$

It therefore follows from (A.13) that

(A.14)
$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 \leq \frac{1}{n}\sum_{i=1}^{n}X_{ij}^2 - 2\sum_{r=1}^{k}\psi_{rj}\widehat{\boldsymbol{\psi}}_r^T\left(\frac{1}{n}\sum_{i=1}^{n}X_{ij}\mathbf{X}_i\right)$$
$$+\sum_{r=1}^{k}\psi_{rj}^2 p\widehat{\lambda}_r.$$

We obtain $\mathbb{E}(X_{ij}^2) = \mathbb{E}(W_{ij}^2) + \sigma_j^2$ as well as $\mathbb{E}(X_{ij}X_{il}) = \mathbb{E}(W_{ij}W_{il})$ for $j \neq l$. At the same time under events (2.1)–(2.4),

$$\left|\sum_{r=1}^{k}\psi_{rj}\widehat{\boldsymbol{\psi}}_r^T\left(\frac{1}{n}\sum_{i=1}^{n}X_{ij}\mathbf{X}_i\right) - \sum_{r=1}^{k}\psi_{rj}\boldsymbol{\psi}_r^T\mathbb{E}(W_{ij}\mathbf{W}_i)\right|$$
$$\leq C_0 n^{-1/2}\sqrt{\log p}\sum_{r=1}^{k}|\psi_{rj}|\left(\sum_{l=1}^{p}|\widehat{\psi}_{rl}|\right)$$
$$+\sum_{r=1}^{k}|\psi_{rj}||(\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r)^T\mathbb{E}(W_{ij}\mathbf{W}_i)|$$
$$+\sum_{r=1}^{k}|\psi_{rj}||\widehat{\psi}_{rj}|\sigma_j^2.$$

Note that $\mathbb{E}(W_{ij}W_{ik}) \leq D_0 - D_1$ for all $j, k = 1, \ldots, p$. By the Cauchy–Schwarz inequality, Theorem 2 and Assumption (A.4), we have

$$|(\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r)^T\mathbb{E}(W_{ij}\mathbf{W}_i)| \leq \|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|_2\|\mathbb{E}(W_{ij}\mathbf{W}_i)\|_2$$
$$\leq \frac{10C_0 n^{-1/2}\sqrt{\log p}}{v(k)}\sqrt{p(D_0 - D_1)}$$

as well as $\sum_{l=1}^{p}|\widehat{\psi}_{rl}| \leq \sqrt{p\sum_{l=1}^{p}\widehat{\psi}_{rl}^2} = \sqrt{p}$. The bounds for ψ_{rj} and $\widehat{\psi}_{rl}$ derived in Theorem 2 then imply that under events (2.1)–(2.4) there exists a constant $\widetilde{M}_2 < \infty$, which can be chosen independently of all values n, p, k, S satisfying assumptions (A.3) and (A.4), such that

(A.15)
$$\left|\sum_{r=1}^{k}\psi_{rj}\widehat{\boldsymbol{\psi}}_r^T\left(\frac{1}{n}\sum_{i=1}^{n}X_{ij}\mathbf{X}_i\right) - \sum_{r=1}^{k}\psi_{rj}\boldsymbol{\psi}_r^T\mathbb{E}(W_{ij}\mathbf{W}_i)\right|$$
$$\leq \widetilde{M}_2\frac{k}{v(k)^{3/2}}n^{-1/2}\sqrt{\log p}.$$

At the same time, by Theorem 2 it follows that there exist constants $\widetilde{M}_2^*, \widetilde{M}_2^{**} < \infty$ such that and $r = 1, \ldots, k$,

(A.16)
$$|p\widehat{\lambda}_r - p\lambda_r| \leq \widetilde{M}_2^* p n^{-1/2} \sqrt{\log p},$$

$$|\psi_{rj}^2 p\widehat{\lambda}_r - \psi_{rj}^2 p\lambda_r| \leq \frac{\widetilde{M}_2^{**}}{v(k)} n^{-1/2} \sqrt{\log p}.$$

Note that $\mathbb{E}((\boldsymbol{\psi}_r^T \mathbf{W}_i)^2) = p\lambda_r$. Under events (2.1)–(2.4), we can now conclude from (A.14)–(A.16) that there exists a constant $\widetilde{M}_2^{***} < \infty$, which can be chosen independently of all values n, p, k, S satisfying Assumptions (A.3) and (A.4), such that

$$\frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2$$

(A.17)
$$\leq \left| \sigma_j^2 + \mathbb{E}(W_{ij}^2) - 2 \sum_{r=1}^k \psi_{rj} \boldsymbol{\psi}_r^T \mathbb{E}(W_{ij} \mathbf{W}_i) + \sum_{r=1}^k \psi_{rj}^2 p\lambda_r \right|$$
$$+ \widetilde{M}_2^{***} \frac{k}{v(k)} n^{-1/2} (\sqrt{\log p})^2$$
$$= \sigma_j^2 + \mathbb{E}((\mathbf{P}_k \mathbf{W}_i)_j^2) + \widetilde{M}_2^{***} \frac{k}{v(k)} n^{-1/2} \sqrt{\log p}.$$

Relations (A.12) and (A.17) imply that under (2.1)–(2.4)

(A.18)
$$\frac{1}{n} \sum_{i=1}^n \left(W_{ij} - \sum_{r=1}^k \widehat{\psi}_{rj} \widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i \right)^2$$
$$\leq \mathbb{E}((\mathbf{P}_k \mathbf{W}_i)_j^2) + M_2^* \frac{k}{v(k)^{3/2}} n^{-1/2} \sqrt{\log p}$$

holds with $M_2^* \leq M_1 + \widetilde{M}_2^{***}$. Since events (2.1)–(2.4) have probability $A(n, p)$, assertion (4.7) of Theorem 3 now is an immediate consequence of (A.12), (A.17) and (A.18).

It remains to show (4.8). We have

$$\frac{1}{n} \sum_{i=1}^n \left(\sum_{r=1}^k (\widehat{\xi}_{ir} - \xi_{ir}) \alpha_r \right)^2$$

(A.19) $$\leq \alpha_{\text{sum}}^2 \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^k (\widehat{\xi}_{ir} - \xi_{ir})^2$$

$$\leq \alpha_{\text{sum}}^2 \left(2 \sum_{r=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i}{\sqrt{p\widehat{\lambda}_r}} - \frac{\boldsymbol{\psi}_r^T \mathbf{X}_i}{\sqrt{p\lambda_r}} \right)^2 + 2 \sum_{r=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\boldsymbol{\psi}_r^T \mathbf{Z}_i}{\sqrt{p\lambda_r}} \right)^2 \right).$$

But for all $r = 1, \ldots, k$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i}{\sqrt{p\widehat{\lambda}_r}} - \frac{\boldsymbol{\psi}_r^T \mathbf{X}_i}{\sqrt{p\lambda_r}} \right)^2 \leq \frac{2}{n} \sum_{i=1}^{n} \frac{(\sqrt{\widehat{\lambda}_r} - \sqrt{\lambda_r})^2 (\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i)^2}{p\lambda_r \widehat{\lambda}_r}$$

(A.20)

$$+ \frac{2}{n} \sum_{i=1}^{n} \frac{((\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r)^T \mathbf{X}_i)^2}{p\lambda_r},$$

and Theorem 2 and assumptions (A.1)–(A.4) imply that under events (2.1)–(2.4) there exist some constants $M_3^*, M_3^{**} < \infty$, which can be chosen independently of all values n, p, k, S satisfying assumptions (A.3) and (A.4), such that

(A.21) $\quad \dfrac{1}{n} \sum_{i=1}^{n} \dfrac{(\sqrt{\widehat{\lambda}_r} - \sqrt{\lambda_r})^2 (\widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i)^2}{p\lambda_r \widehat{\lambda}_r} = \dfrac{(\widehat{\lambda}_r - \lambda_r)^2}{(\sqrt{\widehat{\lambda}_r} + \sqrt{\lambda_r})^2 \lambda_s} \leq \dfrac{M_3^*}{v(k)^2} \dfrac{\log p}{n}$

and

(A.22) $\quad \dfrac{1}{n} \sum_{i=1}^{n} \dfrac{((\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r)^T \mathbf{X}_i)^2}{p\lambda_r} \leq \dfrac{\|\|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|\|_2^2}{p\lambda_r} \dfrac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i\|_2^2 \leq \dfrac{M_3^{**}}{v(k)^3} \dfrac{\log p}{n}$

hold for all $r = 1, \ldots, k$.

Now note that our setup implies that $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}((\frac{\boldsymbol{\psi}_r^T \mathbf{Z}_i}{\sqrt{p\lambda_r}})^2) \leq \frac{D_2}{pv(k)}$ and $\text{Var}(\frac{1}{n} \sum_{i=1}^{n} (\frac{\boldsymbol{\psi}_r^T \mathbf{Z}_i}{\sqrt{p\lambda_r}})^2) \leq \frac{D_3 + 2D_2^2}{nv(k)^2 p^2}$ hold for all $r = 1, \ldots, k$. The Chebyshev inequality thus implies that the event

(A.23) $\quad \dfrac{1}{n} \sum_{i=1}^{n} \left(\dfrac{\boldsymbol{\psi}_r^T \mathbf{Z}_i}{\sqrt{p\lambda_r}} \right)^2 \leq \dfrac{D_2 + \sqrt{D_3 + 2D_2^2}}{pv(k)} \qquad \text{for all } r = 1, \ldots, k$

holds with probability at least $1 - \frac{k}{n}$. We can thus infer from (A.26)–(A.22) that there exists some positive constant $M_3 < \infty$, which can be chosen independently of the values n, p, k, S satisfying (A.3)–(A.5), such that under events (2.1)–(2.4) and (A.23)

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{r=1}^{k} (\widehat{\xi}_{ir} - \xi_{ir}) \alpha_r \right)^2$$

$$\leq \alpha_{\text{sum}}^2 \left(\frac{4k(M_3^* + M_3^{**})}{v(k)^3} \frac{\log p}{n} + \frac{2k(D_2 + \sqrt{D_3 + 2D_2^2})}{pv(k)} \right).$$

Recall that events (2.1)–(2.4) and (A.24) simultaneously hold with probability at least $A(n,p) - (p+k)^{-A^2/8}$, while (A.23) is satisfied with probability at least $1 - \frac{k}{n}$. This proves assertion (4.8) with $M_3 = 4(M_3^* + M_3^{**}) + 2(D_2 + \sqrt{D_3 + 2D_2^2})$. $\square$

PROOF OF PROPOSITION 2. Let $\widehat{\mathbf{Q}}_k$ denote the $p \times p$ diagonal matrix with diagonal entries $1/\sqrt{\frac{1}{n}\sum_{l=1}^n (\widehat{\mathbf{P}}_k\mathbf{X}_l)_1^2}, \ldots, 1/\sqrt{\frac{1}{n}\sum_{l=1}^n (\widehat{\mathbf{P}}_k\mathbf{X}_l)_p^2}$ and split the $(k+p)$-dimensional vector $\boldsymbol{\Delta}$ in two vectors $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$, where $\boldsymbol{\Delta}_1$ is the k-dimensional vector with the k upper components of $\boldsymbol{\Delta}$, and $\boldsymbol{\Delta}_2$ is the p-dimensional vector with the p lower components of $\boldsymbol{\Delta}$. Then

$$\boldsymbol{\Delta}^T \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T \boldsymbol{\Delta} = \boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2^T \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{Q}}_k \widehat{\mathbf{P}}_k \mathbf{X}_i \mathbf{X}_i^T \widehat{\mathbf{P}}_k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$\geq \boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \widehat{\mathbf{P}}_k \boldsymbol{\Psi} \widehat{\mathbf{P}}_k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$+ \boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \widehat{\mathbf{P}}_k (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \widehat{\mathbf{P}}_k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2.$$

The matrix $\boldsymbol{\Psi}$ is a diagonal matrix with entries $\sigma_1^2, \ldots, \sigma_p^2$, and $\widehat{\boldsymbol{\psi}}_r^T \boldsymbol{\Psi} \widehat{\boldsymbol{\psi}}_s \leq D_2$ for all r, s. Together with the bounds for $\widehat{\psi}_{rj}$ derived in Theorem 2 we can conclude that under (2.1)–(2.4) there exists a constant $M_4^* < \infty$, which can be chosen independently of all values n, p, k, S satisfying assumptions (A.3) and (A.4), such that

$$\boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \widehat{\mathbf{P}}_k \boldsymbol{\Psi} \widehat{\mathbf{P}}_k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$= \boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \boldsymbol{\Psi} \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2 - 2\sum_{r=1}^k \boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T \boldsymbol{\Psi} \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$+ \sum_{r=1}^k \sum_{s=1}^k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2^T \widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T \boldsymbol{\Psi} \widehat{\boldsymbol{\psi}}_s \widehat{\boldsymbol{\psi}}_s^T \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$\geq \left(\frac{D_1}{\max_j ((1/n)\sum_{l=1}^n (\widehat{\mathbf{P}}_k\mathbf{X}_l)_j^2)} - \frac{2kM_4^* + k^2 M_4^*}{pv(k)\min_j((1/n)\sum_{l=1}^n(\widehat{\mathbf{P}}_k\mathbf{X}_l)_j^2)} \right) \|\boldsymbol{\Delta}_2\|_2^2.$$

We have

$$\max_j \frac{1}{n}\sum_{i=1}^n (\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 \leq \max_j \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \leq D_0 + \max_j \left| \frac{1}{n}\sum_{i=1}^n X_{ij}^2 - \mathbb{E}(X_{ij}^2) \right|,$$

and since $D_1 \leq D_0$, this leads under (2.1)–(2.4) to

$$\boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2^T \widehat{\mathbf{Q}}_k \widehat{\mathbf{P}}_k \boldsymbol{\Psi} \widehat{\mathbf{P}}_k \widehat{\mathbf{Q}}_k \boldsymbol{\Delta}_2$$

$$\geq \left(\frac{D_1}{D_0 + C_0 n^{1/2}\sqrt{\log p}} - \frac{2kM_4^* + k^2 M_4^*}{pv(k)\min_j((1/n)\sum_{l=1}^n(\widehat{\mathbf{P}}_k\mathbf{X}_l)_j^2)} \right) \|\boldsymbol{\Delta}\|_2^2.$$

On the other hand

$$\boldsymbol{\Delta}_2^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_2$$

$$= (\boldsymbol{\Delta}_{2,J_{0,k+S}} + \boldsymbol{\Delta}_{2,J_{0,k+S}^C})^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k(\boldsymbol{\Delta}_{2,J_{0,k+S}} + \boldsymbol{\Delta}_{2,J_{0,k+S}^C})$$

$$= \boldsymbol{\Delta}_{2,J_{0,k+S}}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_{2,J_{0,k+S}}$$

$$+ \boldsymbol{\Delta}_{2,J_{0,k+S}^C}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_{2,J_{0,k+S}^C}$$

$$+ 2\boldsymbol{\Delta}_{2,J_{0,k+S}}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_{2,J_{0,k+S}^C},$$

where $\boldsymbol{\Delta}_{2,J_{0,k+S}}$, respectively, $\boldsymbol{\Delta}_{2,J_{0,k+S}^C}$, is the p-dimensional vector with the last p coordinates of $\boldsymbol{\Delta}_{J_{0,k+S}}$, respectively, $\boldsymbol{\Delta}_{J_{0,k+S}^C}$. The Cauchy–Schwarz inequality leads to $\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_1 \le (2(k+S))^{1/2}\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_2$. Since $\|\boldsymbol{\Delta}_{J_{0,k+S}^C}\|_1 \le c_0\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_1$ we have

$$|\boldsymbol{\Delta}_{2,J_{0,k+S}}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_{2,J_{0,k+S}^C}|$$

$$\le \max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|\|\boldsymbol{\Delta}_{2,J_{0,k+S}}\|_1\|\boldsymbol{\Delta}_{2,J_{0,k+S}^C}\|_1$$

$$\le \max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_1\|\boldsymbol{\Delta}_{J_{0,k+S}^C}\|_1$$

$$\le c_0\max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_1^2$$

$$\le 2(k+S)c_0\max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_2^2,$$

and the same upper bound holds for the terms $\boldsymbol{\Delta}_{2,J_{0,k+S}}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k \times \boldsymbol{\Delta}_{2,J_{0,k+S}}$ and $\boldsymbol{\Delta}_{2,J_{0,k+S}^C}^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_{2,J_{0,k+S}^C}$ so that

$$\boldsymbol{\Delta}_2^T\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k\boldsymbol{\Delta}_2$$

$$\le 8(k+S)c_0\max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|\|\boldsymbol{\Delta}_{J_{0,k+S}}\|_2^2.$$

Obviously, $\widehat{\boldsymbol{\psi}}_r^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\boldsymbol{\psi}}_s \le p\max_{j,l}|\frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,l} - \mathrm{Cov}(X_{i,j}, X_{i,l})|$ for all r, s. Using Theorem 2, one can infer that under (2.1)–(2.4),

$$\max_{j,l}|(\widehat{\mathbf{Q}}_k\widehat{\mathbf{P}}_k(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\widehat{\mathbf{P}}_k\widehat{\mathbf{Q}}_k)_{j,l}|$$

$$\le \frac{\max_{j,l}|(1/n)\sum_{i=1}^n X_{i,j}X_{i,l} - \mathrm{Cov}(X_{i,j}, X_{i,l})|}{\min_j((1/n)\sum_{l=1}^n(\widehat{\mathbf{P}}_k\mathbf{X}_l)_j^2)}$$

$$+ 2\sum_{r=1}^k\max_{j,l}|(\widehat{\boldsymbol{\psi}}_r\widehat{\boldsymbol{\psi}}_r^T(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}))_{j,l}|$$

$$+ \sum_{r=1}^{k} \sum_{s=1}^{k} \max_{j,l} |(\widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \widehat{\boldsymbol{\psi}}_s \widehat{\boldsymbol{\psi}}_s^T)_{j,l}|$$

$$\leq \frac{M_4^{**}(k^2/v(k))n^{-1/2}\sqrt{\log p}}{\min_j((1/n)\sum_{l=1}^{n}(\widehat{\mathbf{P}}_k \mathbf{X}_l)_j^2)},$$

where the constant $M_4^{**} < \infty$ can be chosen independently of all values n, p, k, S satisfying assumptions (A.3) and (A.4). When combining the above inequalities, the desired result follows from (A.5) and the bound on $\min_j \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2$ to be obtained from (4.6)   □

PROOF OF THEOREM 4.   The first step of the proof consists of showing that under events (2.1)–(2.4) the following inequality holds with probability at least $1 - (p+k)^{1-A^2/8}$

$$(A.24) \qquad\qquad 2\left\| \frac{1}{n}\boldsymbol{\Phi}^T(\mathbf{Y} - \boldsymbol{\Phi}\boldsymbol{\theta}) \right\|_{\infty} \leq \rho,$$

where $\rho = A\sigma\sqrt{\frac{\log(k+p)}{n}} + \frac{M_5\boldsymbol{\alpha}_{\text{sum}}}{v(k)^{3/2}}\sqrt{\frac{\log p}{n}}$, $A > 2\sqrt{2}$ and M_5 is a sufficiently large positive constant.

Since W_{ij}, Z_{ij} and, hence, $\widehat{\xi}_{ir}$ and $\widetilde{X}_{ij}$ are independent of the i.i.d. error terms $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, it follows from standard arguments that

$$(A.25) \qquad \sup_{1 \leq r \leq k, 1 \leq j \leq p} \left\{ \frac{2}{n}\left|\sum_{i=1}^{n}\widehat{\xi}_{ir}\varepsilon_i\right|, \frac{2}{n}\left|\sum_{i=1}^{n}\widetilde{X}_{ij}\varepsilon_i\right| \right\} \leq A\sigma\sqrt{\frac{\log(k+p)}{n}}$$

holds with probability at least $1 - (p+k)^{1-A^2/8}$. Therefore, in order to prove (A.24) it only remains to show that under events (2.1)–(2.4) there exists a positive constant $M_5 < \infty$, which can be chosen independently of the values n, p, k, S satisfying (A.3)–(A.5), such that

$$(A.26) \qquad \sup_{1 \leq r \leq k, 1 \leq j \leq p} \left\{ \frac{2}{n}\left|\sum_{i=1}^{n}\widehat{\xi}_{ir}\widetilde{\varepsilon}_i\right|, \frac{2}{n}\left|\sum_{i=1}^{n}\widetilde{X}_{ij}\widetilde{\varepsilon}_i\right| \right\} \leq \frac{M_5\boldsymbol{\alpha}_{\text{sum}}}{v(k)^{3/2}}\sqrt{\frac{\log p}{n}}.$$

We will now prove (A.26). For all $r = 1, \ldots, k$ we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_{ir}\widetilde{\varepsilon}_i\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_{ir}\sum_{s=1}^{k}\alpha_s(\widehat{\xi}_{is} - \xi_{is})\right|$$

$$= \left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_{ir}\sum_{s=1}^{k}\alpha_s\left(\frac{\boldsymbol{\psi}_s^T\mathbf{W}_i}{\sqrt{p\lambda_s}} - \frac{\widehat{\boldsymbol{\psi}}_s^T\mathbf{X}_i}{\sqrt{p\widehat{\lambda}_s}}\right)\right|$$

$$= \left| \sum_{s=1}^{k} \alpha_s \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \left(\frac{\boldsymbol{\psi}_s^T \mathbf{X}_i}{\sqrt{p\lambda_s}} - \frac{\widehat{\boldsymbol{\psi}}_s^T \mathbf{X}_i}{\sqrt{p\widehat{\lambda}_s}} - \frac{\boldsymbol{\psi}_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right) \right|$$

$$\leq \alpha_{\text{sum}} \sup_s \left(\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{(\sqrt{\widehat{\lambda}_s} - \sqrt{\lambda_s}) \widehat{\boldsymbol{\psi}}_s^T \mathbf{X}_i}{\sqrt{p\lambda_s \widehat{\lambda}_s}} \right| \right.$$

$$\left. + \left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{(\boldsymbol{\psi}_s - \widehat{\boldsymbol{\psi}}_s)^T \mathbf{X}_i}{\sqrt{p\lambda_s}} \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{\boldsymbol{\psi}_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| \right).$$

Using the Cauchy–Schwarz inequality and the fact that $\frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir}^2 = 1$, inequalities (A.21) and (A.22) imply that under events (2.1)–(2.4), one obtains

$$\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \widetilde{\varepsilon}_i \right|$$

$$\leq \alpha_{\text{sum}} \sup_s \left(\left(\frac{1}{n} \sum_{i=1}^{n} \frac{(\sqrt{\widehat{\lambda}_s} - \sqrt{\lambda_s})^2 (\widehat{\boldsymbol{\psi}}_s^T \mathbf{X}_i)^2}{p\lambda_s \widehat{\lambda}_s} \right)^{1/2} \right.$$

(A.27)

$$\left. + \left(\frac{1}{n} \sum_{i=1}^{n} \frac{((\boldsymbol{\psi}_s - \widehat{\boldsymbol{\psi}}_s)^T \mathbf{X}_i)^2}{p\lambda_s} \right)^{1/2} + \left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{\boldsymbol{\psi}_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| \right)$$

$$\leq \alpha_{\text{sum}} \left(\frac{\sqrt{M_3^*}}{v(k)} \sqrt{\frac{\log p}{n}} + \frac{\sqrt{M_3^{**}}}{v(k)^{3/2}} \sqrt{\frac{\log p}{n}} + \sup_s \left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{\boldsymbol{\psi}_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| \right).$$

Since also $\frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_{ij}^2 = 1$, similar arguments show that under (2.1)–(2.4)

$$\left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_{ij} \widetilde{\varepsilon}_i \right| \leq \alpha_{\text{sum}} \left(\frac{\sqrt{M_3^*}}{v(k)} \sqrt{\frac{\log p}{n}} + \frac{\sqrt{M_3^{**}}}{v(k)^{3/2}} \sqrt{\frac{\log p}{n}} \right.$$

(A.28)

$$\left. + \sup_s \left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_{ij} \frac{\boldsymbol{\psi}_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| \right)$$

for all $j = 1, \ldots, p$. The Cauchy–Schwarz inequality yields $\sum_{l=1}^{p} |\widehat{\psi}_{rl}| \leq \sqrt{p}$, $\sum_{l=1}^{p} |\psi_{rl}| \leq \sqrt{p}$, as well as

$$\sum_{l=1}^{p} \left| \sum_{r=1}^{k} \widehat{\psi}_{rj} \widehat{\psi}_{rl} \right| \leq \sqrt{p} \sqrt{\sum_{l=1}^{p} \left(\sum_{r=1}^{k} \widehat{\psi}_{rj} \widehat{\psi}_{rl} \right)^2} = \sqrt{p} \sqrt{\sum_{r=1}^{k} \sum_{s=1}^{k} \widehat{\psi}_{rj} \widehat{\psi}_{sj} \sum_{l=1}^{p} \widehat{\psi}_{rl} \widehat{\psi}_{sl}}$$

$$\leq \sqrt{kp} \sup_r |\widehat{\psi}_{rj}|.$$

Necessarily, $v(k) \leq D_0/k$ and hence $k \leq D_0/v(k)$. It therefore follows from (4.3), (4.4), (4.6) and (A.5) that under events (2.1)–(2.4) there are some constants M_5^{***}, $\widetilde{M}_5^{***}$ such that for all $r, s = 1, \ldots, k$ and $j = 1, \ldots, p$,

$$
\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{\xi}_{ir} \frac{\psi_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{p\sqrt{\widehat{\lambda}_r \lambda_s}} \sum_{j=1}^{p} \sum_{j'=1}^{p} \widehat{\psi}_{rj} \psi_{sj'} X_{ij} Z_{ij'} \right|
$$

(A.29)
$$
\leq \frac{M_5^{***}}{v(k)} \sqrt{\frac{\log p}{n}}
$$

and

$$
\left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_{ij} \frac{\psi_s^T \mathbf{Z}_i}{\sqrt{p\lambda_s}} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j'=1}^{p} \frac{(X_{ij} - \sum_{r=1}^{k} \widehat{\psi}_{rj} \widehat{\psi}_r^T \mathbf{X}_i) \psi_{sj'} Z_{ij'}}{((1/n) \sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2)^{1/2} \sqrt{p\lambda_s}} \right|
$$

(A.30)
$$
\leq \frac{\widetilde{M}_5^{***}}{v(k)^{3/2}} \sqrt{\frac{\log p}{n}}.
$$

Result (A.26) is now a direct consequence of (A.27)–(A.30). Note that all constants in (A.27)–(A.30) and thus also the constant $M_5 < \infty$ can be chosen independently of the values n, p, k, S satisfying (A.3)–(A.5).

Under event (A.24) as well as $\mathbf{K}_{n,p}(k, S, 3) > 0$, inequalities (B.1), (4.1), (B.27) and (B.30) of Bickel, Ritov and Tsybakov (2009) may be transferred in our context which yields

(A.31) $\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{J_0}\|_2 \leq 4\rho\sqrt{k+S}/\kappa^2, \qquad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 \leq 4\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{J_0}\|_1,$

where J_0 is the set of nonnull coefficients of $\boldsymbol{\theta}$. This implies that

(A.32)
$$
\sum_{r=1}^{k} |\widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r| + \sum_{j=1}^{p} |\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j| \leq 16 \frac{k+S}{\kappa^2} \rho.
$$

Events (2.1)–(2.4) hold with probability $A(n,p)$, and therefore the probability of event (A.24) is at least $A(n,p) - (p+k)^{1-A^2/8}$. When combining (4.4), (4.6) and (A.32), inequalities (5.2) and (5.3) follow from the definitions of $\widehat{\beta}_j$ and $\widetilde{\alpha}_r$, since under (2.1)–(2.4)

$$
\sum_{j=1}^{p} |\widehat{\beta}_j - \beta_j| = \sum_{j=1}^{p} \frac{|\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j|}{((1/n) \sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2)^{1/2}}
$$

$$
\leq \frac{\sum_{j=1}^{p} |\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j|}{(D_1 - M_1(kn^{-1/2}\sqrt{\log p}/v(k)^{1/2}))^{1/2}}
$$

and

$$\sum_{r=1}^{k} |\widehat{\alpha}_r - \alpha_r| = \sum_{r=1}^{k} \left| \widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r - \sqrt{p\widehat{\lambda}_r} \sum_{j=1}^{p} \widehat{\psi}_{rj}(\widehat{\beta}_j - \beta_j) \right|$$

$$\leq \sum_{r=1}^{k} |\widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r| + k(D_0 + C_0 n^{-1/2}\sqrt{\log p})^{1/2} \sum_{j=1}^{p} |\widehat{\beta}_j - \beta_j|.$$

It remains to prove assertion (5.4) on the prediction error. We have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{r=1}^{k}\widehat{\xi}_{ir}\widehat{\alpha}_r - \xi_{ir}\alpha_r + \sum_{j=1}^{p}X_{ij}(\widehat{\beta}_j - \beta_j)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{r=1}^{k}\widehat{\xi}_{ir}(\widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r) + \sum_{j=1}^{p}\widetilde{X}_{ij}(\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j) + \sum_{r=1}^{k}(\widehat{\xi}_{ir} - \xi_{ir})\alpha_r\right)^2$$

$$(A.33) \qquad \leq \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{r=1}^{k}\widehat{\xi}_{ir}(\widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r) + \sum_{j=1}^{p}\widetilde{X}_{ij}(\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j)\right)^2$$

$$+ \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{r=1}^{k}(\widehat{\xi}_{ir} - \xi_{ir})\alpha_r\right)^2.$$

Under event (A.24) as well as $\mathbf{K}_{n,p}(k,S,3) > 0$, the first part of inequalities (B.31) in the proof of Theorem 7.2 of Bickel, Ritov and Tsybakov (2009) leads to

$$(A.34) \quad \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{r=1}^{k}\widehat{\xi}_{ir}(\widehat{\widetilde{\alpha}}_r - \widetilde{\alpha}_r) + \sum_{j=1}^{p}\widetilde{X}_{ij}(\widehat{\widetilde{\beta}}_j - \widetilde{\beta}_j)\right)^2 \leq \frac{32(k+S)}{\kappa^2}\rho^2.$$

Under events (2.1)–(2.4), (A.24) as well as (A.23), inequality (5.4) now follows from (A.33), (A.34) and (4.8). The assertion then is a consequence of the fact that (2.1)–(2.4) are satisfied with probability $A(n,p)$, while (A.24) and (A.23) hold with probabilities at least $1 - (p+k)^{-A^2/8}$ and $1 - \frac{k}{n}$, respectively. $\square$

REFERENCES

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

BAI, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279. MR2547073

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BERNANKE, B. S. and BOIVIN, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics* **50** 525–546.

BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

CARDOT, H., MAS, A. and SARDA, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138** 325–361. MR2299711

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canad. J. Statist.* **30** 285–300. MR1926066

FORNI, M. and LIPPI, M. (1997). *Aggregation and the Microfoundations of Dynamic Macroeconomics.* Oxford Univ. Press, Oxford.

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* **82** 540–554.

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. MR2212577

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

KNEIP, A. and UTIKAL, K. J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.* **96** 519–542. With comments and a rejoinder by the authors. MR1946423

KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

RAMSAY, J. O. and DALZELL, C. J. (1991). Some tools for functional data analysis (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 539–572. MR1125714

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Zhou, S., Lafferty, J. and Wassermn, L. (2008). Time varying undirected graphs. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT'08)*. Available at arXiv:0903.2515.

Zhou, S., van de Geer, S. and Bülhmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. Preprint.

Statistische Abteilung
Department of Economics
  and Hausdorff
Center for Mathematics
Universität Bonn
Adenauerallee 24-26
53113 Bonn
Germany
E-mail: akneip@uni-bonn.de

Institut de Mathématiques
Laboratoire de Statistique
  et Probabilités
Université Paul Sabatier
UMR 5219
118, Route de Narbonne
31062 Toulouse Cedex
France
E-mail: Pascal.Sarda@math.ups-tlse.fr