Joint Likelihood-based Principal Components Regression

Karl Oskar Ekvall*

Institute of Statistics and Mathematical Methods in Economics

Vienna University of Technology, Vienna, Austria

karl.ekvall@tuwien.ac.at

April 29, 2020

Abstract

We propose a method for estimating principal components regressions by maximizing a multivariate normal joint likelihood for responses and predictors. In contrast to classical principal components regression, our method uses information in both responses and predictors to select useful linear combinations of the predictors. We show our estimators are consistent when responses and predictors have sub-Gaussian distributions and the number of observations tends to infinity faster than the number of predictors. Simulations indicate our method is substantially more accurate than classical principal components regression in estimation and prediction, and that it often compares favorably to competing methods such as partial least squares and predictor envelopes. We corroborate the simulation results and illustrate the practical usefulness of our method with a data example with cross-sectional prediction of stock returns.

*The work was supported by FWF (Austrian Science Fund, https://www.fwf.ac.at/en/) [P30690-N35]. The author thanks Efstathia Bura, Dan Eck, Erik Hjalmarsson, and Aaron Molstad for helpful comments.

1 Introduction

Principal components analysis is one of the most commonly used methods for dimension reduction (Cook, 2018a), whereby p variables are reduced to $k \leq p$ linear combinations of those variables—the principal components. The vectors of weights defining the principal components are eigenvectors of the variables' sample covariance matrix. Thus, subject to the constraint that the weight vectors be orthonormal, the k principal components are the k linear combinations with maximal sample variances. In classical principal components regression (PCR), principal components of the predictors are used to model or predict one or more response variables (e.g. Jolliffe, 2002, Chapter 8). Intuitively, one expects this to be useful when the corresponding relation holds in the population; that is, when only a few linear combinations of the predictors, with large population variances, affect the responses. This is the case in many popular models (Stock and Watson, 2002; Bair et al., 2006; Yu et al., 2006; Bai and Wang, 2016; Singer et al., 2016) and can be motivated by more general probabilistic arguments (Artemiou and Li, 2009). Accordingly, PCR has been considered in a wide range of applications, including ones in economics, chemometrics, and genomics (e.g. Næs and Martens, 1988; Chang et al., 2001; Stock and Watson, 2002; Watson, 2003; Wang and Abbott, 2008; Barbarino and Bura, 2017; Du et al., 2018).

Despite its popularity, it has also been argued that classical PCR is unreliable (Cook, 2018a) and that it is preferable to use both responses and predictors to decide which linear combinations of the predictors are relevant (Helland, 1992; Bair et al., 2006; Yu et al., 2006; Cook et al., 2010; Kelly and Pruitt, 2015). Our work shows that, in many such cases, it may be that a PCR model, which we will define in detail shortly, is still appropriate, but that classical PCR gives poor estimates of the parameters in that model. We address this by proposing a method that estimates a PCR model by maximizing a multivariate normal joint likelihood of responses and predictors. Thus, if data are multivariate normal, our estimators are maximum likelihood estimators, and otherwise they are M-estimators. As we will see, our method leads to substantially more accurate estimation and prediction than classical PCR in a wide range of settings. Other related methods typically do not give estimates that lie in the parameter space of a PCR model, and hence may be inappropriate when one is suggested by

domain knowledge or other considerations. Nevertheless, we compare our method to partial least squares (PLS) and predictor envelopes (Cook et al., 2013) since these are appropriate for similar settings and tend to perform well in practice. Using simulations, we show our method compares favorably to both when data are generated from our model. The practical relevance this model is supported by a data example where our method gives more accurate cross-sectional prediction of stock returns than classical PCR, PLS, predictor envelopes, and ordinary least squares (OLS). For a more precise discussion, we require some notation.

Let the pairs $(Y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^p$, $i = 1, \ldots, n$, be independent with $\mathbb{E}(Y_i) = 0$, $\mathbb{E}(X_i) = 0$, $\mathbb{E}(Y_i \mid X_i) = \beta^\mathsf{T} X_i$, $\beta \in \mathbb{R}^{p \times r}$, $\text{cov}(Y_i \mid X_i) = \Sigma \in \mathbb{S}^r_{++}$, and $\text{cov}(X_i) = \Sigma_X \in \mathbb{S}^p_{++}$, where $\mathbb{S}^p_{++}$ is the set of $p \times p$ symmetric and positive definite matrices. Later, we will assume (Y_i, X_i) is sub-Gaussian for some of our asymptotic results, but that is not necessary to define our model. The assumption that both responses and predictors have mean zero is not needed in practice but makes some of our theoretical arguments easier to present. To define our model, we assume there exist a semi-orthogonal $U \in \mathbb{O}^{p \times k}$, diagonal and positive semi-definite $D \in \mathbb{D}^k_+$, $\tau > 0$, and $\gamma \in \mathbb{R}^{k \times r}$ such that

$$(i) \;\; \beta = U\gamma \quad \text{and} \quad (ii) \;\; \Sigma_X = \tau(I_p + UDU^\mathsf{T}). \tag{1}$$

Model (1) implies the columns of U are leading eigenvectors of Σ_X and that $\mathbb{E}(Y_i \mid X_i) = \gamma^\mathsf{T} U^\mathsf{T} X_i$. That is, only k linear combinations of the predictors, whose weights are given by leading eigenvectors of the predictors' population covariance matrix, are relevant for the regression. Condition (ii) implies the $p - k$ smallest eigenvalues of Σ_X are all equal to τ or, equivalently, that Σ_X is spiked (Johnstone, 2001). A spiked predictor covariance matrix is not necessary to define a PCR model, but it facilitates both our theory and the implementation of our method. Spiked covariance matrices are common in the modern literature on principal components analysis and large covariance matrix estimation (Johnstone, 2001; Cai et al., 2014; Wang and Fan, 2017; Donoho et al., 2018), and they arise naturally in latent variable and factor models that motivate both our and several related methods (Tipping and Bishop, 1999; Bair et al., 2006; Yu et al., 2006; Kelly and Pruitt, 2015; Singer et al., 2016) (see also Appendix C). With $k = p$, any $\beta \in \mathbb{R}^{p \times r}$ and $\Sigma_X \in \mathbb{S}^p_{++}$ can be written as in (1). Thus, our model includes a classical multivariate linear regression model as a special case. Here, however, we are primarily interested in the cases where $k < p$ so that there is potential for dimension reduction.

We propose estimating β, Σ, and Σ_X by maximizing a multivariate normal joint log-likelihood for n observations from (1); we do not attempt to estimate U, γ, or D directly since they are unidentifiable. Nevertheless, for any β, Σ, and Σ_X in our parameter set, we can decompose them as in (1) and write the conditional log-likelihood for $Y = [Y_1, \ldots, Y_n]^\mathsf{T} \in \mathbb{R}^{n \times r}$ given $X = [X_1, \ldots, X_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$ and the marginal likelihood for X, respectively, as

$$\ell_{Y|X}(U, \gamma, \Sigma) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \operatorname{tr}\left[(Y - XU\gamma)^\mathsf{T}(Y - XU\gamma)\Sigma^{-1}\right]$$

and

$$\ell_X(U, D, \tau) = -\frac{n}{2} \log |\tau(I_p + UDU^\mathsf{T})| - \frac{1}{2\tau} \operatorname{tr}\left[X^\mathsf{T}X(I_p + UDU^\mathsf{T})^{-1}\right],$$

where $|\cdot|$ is the determinant and $\operatorname{tr}(\cdot)$ the trace when applied to square matrices. The joint likelihood our estimators maximize is then

$$\ell_n(\beta, \Sigma, \Sigma_X) = \ell_n(U, \gamma, \Sigma, D, \tau) = \ell_{Y|X}(U, \gamma, \Sigma) + \ell_X(U, D, \tau). \tag{2}$$

Since both $\ell_{Y|X}$ and ℓ_X depend on U, our estimate of Σ_X, and in particular the eigenvectors of that estimate, will in general depend on both Y and X. It is well known (Tipping and Bishop, 1999) that if $(\tilde{U}, \tilde{D}, \tilde{\tau})$ maximizes ℓ_X, then the columns of $\tilde{U}$ are leading eigenvectors of $S_X = X^\mathsf{T}X/n$. Thus, loosely speaking, our method will try to select a U whose columns are close to the leading eigenvectors of S_X and at the same time makes the (weighted) sum of squared errors of the regression of Y on XU small. By contrast, classical PCR is equivalent to a two-stage procedure where, in the first stage, ℓ_X is maximized to obtain $\tilde{U}$, effectively ignoring $\ell_{Y|X}$ and the fit of the regression of Y on $X\tilde{U}$. Then, in the second stage with $\tilde{U}$ fixed, β is estimated as $\tilde{U}\tilde{\gamma}$, where $\tilde{\gamma}$ is obtained from a least squares regression of Y on $X\tilde{U}$ (Jolliffe, 2002, Chapter 8).

The log-likelihood ℓ_n is in general unbounded if $p > n$, and hence we assume $n > p$ throughout. Our asymptotic results, showing our estimators are consistent, assume both n and p tend to infinity with $p/n \to 0$. This reflects our belief that our method will be of most interest in settings where the number of predictors is large enough for dimension reduction to be useful, but small enough in comparison to n that a likelihood-based method without penalization is applicable. For context, we note that $p/n \to c \in (0, \infty)$ is common in the random matrix literature and that $p/n \to 0$ is a special case of settings considered in sparse,

high-dimensional spiked covariance matrix estimation (Wang and Fan, 2017). Asymptotic theory for related likelihood-based methods such as predictor envelopes typically assumes p is fixed (Cook et al., 2010), but it has been noted that it may be more appropriate to let p grow with n (Cook et al., 2007).

Maximizing ℓ_n is a non-trivial, non-convex optimization problem. Helland (1992) proposed an algorithm for a somewhat similar setting but with $r = 1$, $k = 1$, and without spiked covariance matrix. The spiked covariance is essential to our development of a method that handles $r > 1$ and $k > 1$. The key step for the algorithm we propose is to re-parameterize UDU^T as LL^T for an $L \in \mathbb{R}^{p\times k}$ which we discuss in more detail in Section 3.1. Briefly, with appropriate restrictions on L, this parameterization is identifiable and the corresponding log-likelihood can be partially maximized analytically in all arguments but L. The resulting profile log-likelihood for L can then maximized numerically.

There are similarities between our method and the ones we compare to, but also important differences. Predictor envelopes assume $\beta = V\alpha$ for some $\alpha \in \mathbb{R}^{k\times r}$ and $V \in \mathbb{O}^{p\times k}$ whose columns are eigenvectors of Σ_X that, in contrast to in our model, need not be leading ones. Thus, the predictor envelope model is more flexible and less parsimonious than ours. PLS can be viewed as a moment-based estimator of a predictor envelope. With $r = 1$, PLS is equivalent to selecting linear combinations of the predictors given by the columns of the sample Krylov matrix $\tilde{K} = [S_{XY}, S_X S_{XY}, \ldots, S_X^{k-1} S_{XY}] \in \mathbb{R}^{p\times k}$, where $S_{XY} = \sum_{i=1} X_i Y_i / n$ (Helland, 1990; Cook et al., 2013). More precisely, PLS is equivalent to selecting linear combinations given by any set of basis vectors for the column space of $\tilde{K}$, and it is well known that the column space of $K \in \mathbb{R}^{p\times k}$, defined as $\tilde{K}$ but with population covariances in place of their sample versions, coincides with the predictor envelope (Cook et al., 2013). Loosely speaking, PLS and likelihood-based predictor envelope estimators attempt to both infer which eigenvectors are relevant and estimate them. Since our method assumes the leading ones are relevant, we expect it to perform better in settings where that is either true or a reasonable approximation. This intuition is confirmed in our simulations and data example.

The remainder of the paper is organized as follows. In Section 2 we give conditions that ensure consistency of our estimators and in Section 3 we outline how the to compute the estimates in practice and how to select the number of components, k. Section 4 contains

simulation results and in Section 5 we compare our method to competing ones in a data example. Concluding remarks are in Section 6.

2 Asymptotic results

2.1 Consistent estimation

This section gives conditions that ensure consistency of our estimators. The idea is to show that an appropriately scaled log-likelihood concentrates around its expectation, which is maximized by the true parameter. The majority of the work is showing that that convergence is uniform on suitably chosen subsets of the parameter set. Allowing p to grow with n and letting the true parameter depend on p (and n) makes this more complicated than in classical settings. The results in this section assume k is known; how to select k in practice is discussed in Section 3.2. Proofs not given in the main text can be found in Appendix A unless otherwise noted.

We parameterize in terms of precision matrices $\Omega = \Sigma^{-1}$ and $\Omega_X = \Sigma_X^{-1}$. Let $\bar{\Theta} = \mathbb{R}^{p\times r} \times \mathbb{S}^r_{++} \times \mathbb{S}^p_{++}$ and $\Theta \subseteq \bar{\Theta}$ be the set of $\theta = (\beta, \Omega, \Omega_X) \in \bar{\Theta}$ for which $\Sigma_X = \Omega_X^{-1}$ and β can be written as in (1). The true parameter is denoted $\theta_* = (\beta_*, \Omega_*, \Omega_{X*}) \in \Theta$. We equip $\bar{\Theta}$ with the max-norm defined by

$$\|\theta\|_M = \max\{\|\beta\|, \|\Omega\|, \|\Omega_X\|\},$$

where $\|\cdot\|$ is the spectral norm when applied to matrices. When parameterizing in terms of precision matrices, the negative log-likelihood $g_n : \bar{\Theta} \to \mathbb{R}$ corresponding to (2) is, up to scaling and an additive constant,

$$g_n(\theta) = -\log|\Omega| + \frac{1}{n}\operatorname{tr}\Big[(Y - X\beta)^\mathsf{T}(Y - X\beta)\Omega\Big] - \log|\Omega_X| + \operatorname{tr}(S_X\Omega_X). \tag{3}$$

We denote the minimizers of interest by $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} g_n(\theta)$. Clearly, if $\hat{\theta}_n = (\hat{\beta}, \hat{\Omega}, \hat{\Omega}_X)$ minimizes g_n over Θ, then $\hat{\beta}$ and $\hat{\Sigma}_X = \hat{\Omega}_X^{-1}$ can be decomposed to give maximizers of (2). The following result says a $\hat{\theta}_n$ exists under reasonable conditions. We let Q_X denote the orthogonal projection onto the orthogonal complement of the column space of X.

Theorem 2.1. *The set* $\arg\min_{\theta\in\Theta} g_n(\theta)$ *is non-empty if* $Y^\mathsf{T} Q_X Y$ *is invertible and* S_X *has rank greater than* k*, and it is empty if* S_X *has rank less than or equal to* k *or if* $Y^\mathsf{T} Q_X Y$ *is non-invertible and* $k \geq r$*.*

The conditions in Theorem 2.1 are weak in the settings we consider. In particular, if both responses and predictors have continuous distributions, $[Y, X]$ has full column rank almost surely if $n \geq p + r$, which implies the sufficient conditions in the theorem.

To state the assumptions we use for the asymptotic analysis, we use the following definitions (Vershynin, 2018). A random variable $\xi \in \mathbb{R}$, or its distribution, is sub-Gaussian if there exists a $c > 0$ such that $\mathsf{P}(|\xi| \geq t) \leq 2\exp(-ct^2)$ for all $t \geq 0$. The sub-Gaussian norm of $\xi \in \mathbb{R}$ is $\inf\{t > 0 : \mathbb{E}[\exp(\xi^2/t^2)] \leq 2\}$. A random vector $\xi \in \mathbb{R}^d$ is sub-Gaussian if $v^\mathsf{T}\xi$ is sub-Gaussian for any unit-length $v \in \mathbb{R}^d$. The sub-Gaussian norm of $\xi \in \mathbb{R}^d$ is the supremum over all unit-length $v \in \mathbb{R}^d$ of the sub-Gaussian norm of $v^\mathsf{T}\xi$. A random vector $\xi \in \mathbb{R}^d$ is sub-Gaussian if and only if its sub-Gaussian norm is finite. Common sub-Gaussian distributions include the multivariate normal and all with bounded support, such as, for example, Bernoulli and uniform distributions.

We use five assumptions to establish consistency. Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of their matrix arguments, respectively. We use c to denote a generic constant that does not depend on n or p but can otherwise change with context. The number of predictors can change as a function of n and we write $p = p(n)$.

Assumption 1. For any fixed $k \leq p < n$ and every $i = 1, \ldots, n$, $\mathbb{E}(X_i) = 0$, $\mathbb{E}(Y_i \mid X_i) = \beta_*^\mathsf{T} X_i$, $\mathrm{cov}(Y_i \mid X_i) = \Sigma_* \in \mathbb{S}^r_{++}$, and $\mathrm{cov}(X_i) = \Sigma_{X*} \in \mathbb{S}^p_{++}$; there also exist $U_* \in \mathbb{O}^{p\times k}$, $D_* \in \mathbb{D}^k_+$, $\tau_* > 0$, and $\gamma_* \in \mathbb{R}^{k\times r}$ such that $\beta_* = U_*\gamma_*$ and $\Sigma_{X*} = \tau_*(I_p + U_* D_* U_*^\mathsf{T})$.

Assumption 2. The true parameter $\theta_* = (\beta_*, \Omega_*, \Omega_{X*})$ depends on p, and can hence depend on n, but there exists a constant $c > 0$ such that, for every n and p,

1. $c^{-1} \leq \lambda_{\min}(\Omega_*) \leq \lambda_{\max}(\Omega_*) \leq c$,
2. $c^{-1} \leq \lambda_{\min}(\Omega_{X*}) \leq \lambda_{\max}(\Omega_{X*}) \leq c$, and
3. $\lambda_{\max}(\beta_*^\mathsf{T}\Omega_{X*}^{-1}\beta_*) \leq c$.

Assumption 3. The number of observations n and the number of predictors $p = p(n)$ both tend to infinity and satisfy

$$\lim_{n\to\infty} \frac{p(n)}{n} = 0;$$

the number of responses r and the number of relevant linear combinations k are fixed.

Assumption 4. The vectors $[Y_i^\mathsf{T}, X_i^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{r+p}$, $i = 1, \ldots, n$, are independent and have sub-Gaussian norms bounded by a $c < \infty$ not depending on n or p.

Assumptions 2 and 3 define the type of asymptotics we consider. Cook and Forzani (2019) discuss possible interpretations of the boundedness of $\lambda_{\max}(\beta_*^\mathsf{T}\Sigma_{X*}\beta_*)$ and $\lambda_{\max}(\Sigma_{X*})$ in high-dimensional regressions. Here, since the estimators we consider do not exist in general when $p > n$ (Theorem 2.1), we only note that $\lambda_{\max}(\beta_*^\mathsf{T}\Sigma_{X*}\beta_*) \leq c$ is required to avoid the case where $\lambda_{\max}(\mathrm{cov}(Y_i)) = \lambda_{\max}(\Sigma_* + \beta_*^\mathsf{T}\Sigma_{X*}\beta_*)$ tends to infinity with n and p. Assumption 4 is useful to us because, together with Assumption 3, it implies the sample covariance matrix $n^{-1}\sum_{i=1}^n [Y_i^\mathsf{T}, X_i^\mathsf{T}]^\mathsf{T}[Y_i^\mathsf{T}, X_i^\mathsf{T}]$ is close to its population counterpart in spectral norm, with high probability (Vershynin, 2018, Theorem 4.6.1).

The final assumption, defined below, is made so that terms of g_n that depend on $\lambda_{\max}(\Omega) = \tau = \lambda_{\min}(\Sigma_X)$ but no other arguments concentrate around their expectation. Because it can seem less intuitive than the other assumptions, we discuss this assumption in detail in Section 2.2 and show it is implied by the other assumptions in many relevant settings, including when data are multivariate normal or generated from a common latent variable model.

Assumption 5. For any $\epsilon > 0$, as $n \to \infty$,

$$\mathsf{P}\left(|\operatorname{tr}(S_X - \Sigma_{X*})| \geq \epsilon\right) \to 0.$$

We are ready to state the main result of the section.

Theorem 2.2. *If Assumptions 1 – 5 hold, then as $n \to \infty$, $\arg\min_{\theta\in\Theta} g_n(\theta)$ is non-empty with probability tending to one, and for any sequence of estimators $\{\hat{\theta}_n\}$ with $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} g_n(\theta)$, it holds for any $\epsilon > 0$ that*

$$\mathsf{P}\left(\|\hat{\theta}_n - \theta_*\|_M \geq \epsilon\right) \to 0.$$

Before proving Theorem 2.2, we discuss some intermediate results that we need in the proof. Recall, the main idea is to show that the distance between g_n and $\kappa_n(\theta) = \mathbb{E}[g_n(\theta)]$ tends to zero in probability, uniformly in θ on appropriately chosen subsets of $\bar{\Theta}$. To that end, for $0 < c_1 \leq c_2 < \infty$, we define the set $A_n = A_n(c_1, c_2)$ by

$$A_n = \{\theta \in \Theta : \|\beta\| \leq c_2, c_1 \leq \lambda_{\min}(\Omega), \lambda_{\max}(\Omega) \leq c_2, c_1 \leq \lambda_{\min}(\Omega_X), \lambda_{\max}(\Omega_X) \leq c_2\}. \quad (4)$$

The following result formalizes the necessary uniform convergence.

Lemma 2.3. *If Assumptions 1 – 5 hold, then for any $0 \leq c_1 < c_2 < \infty$ and $\epsilon > 0$, as $n \to \infty$,*

$$\mathsf{P}\left(\sup_{\theta \in A_n} |g_n(\theta) - \kappa_n(\theta)| \geq \epsilon\right) \to 0.$$

Lemma 2.3 alone is not enough to ensure consistency; because the uniformity is not over the whole of Θ, we also need to show that both $\hat{\theta}_n$ and θ_* are in A_n. Moreover, we need that κ_n is, loosely speaking, asymptotically uniquely minimized at θ_*. The following two lemmas give what we need. Results like Lemma 2.5 are common in the M-estimation literature; here, some work is needed to prove the lemma because κ_n depends on n.

Lemma 2.4. *If Assumptions 1 – 5 hold, then there exist $0 < c_1 < c_2 < \infty$ such that*

$$\mathsf{P}\left(\arg\min_{\theta \in \Theta} g_n(\theta) \subseteq A_n(c_1, c_2)\right) \to 1.$$

Lemma 2.5. *For any $0 < c_1 \leq c_2 < \infty$ such that $\theta_* \in A_n(c_1, c_2)$ and any $\theta \in A_n(c_1, c_2)$,*

$$\kappa_n(\theta) - \kappa_n(\theta_*) \geq \delta \|\theta - \theta_*\|_M^2$$

for some $\delta > 0$ that depends on c_1 and c_2 but not on n.

We are ready to prove Theorem 2.2

Proof of Theorem 2.2. To show the existence part, it suffices, by Theorem 2.1, to show that $X^\mathsf{T}X/n$ and $Y^\mathsf{T}Q_XY/n$ are invertible with probability tending to one. Lemma A.3 implies these quantities are consistent in spectral norm for Σ_{X*} and Σ_*, respectively. Thus, by Lemma A.1 and Assumption 2, $\lambda_{\min}(S_X) \geq c^{-1}/2 > 0$ and $\lambda_{\min}(Y^\mathsf{T}Q_XY/n) \geq c^{-1}/2 > 0$ with probability tending to one. It remains to prove that any existing minimizers are consistent. To that end, pick an arbitrary $\epsilon > 0$ and a c_1 and c_2 so that Lemmas 2.4 and 2.5 hold;

this is possible because $\theta_* \in A_n(c_1, c_2)$ for all small enough c_1 and large enough c_2, and making A_n larger only increases the probability in Lemma 2.4. A routine calculation shows $\{\theta_*\} = \arg\min_{\theta\in\bar{\Theta}} \kappa_n(\theta)$, and, hence, $\arg\min_{\theta\in\Theta} \kappa_n(\theta) = \arg\min_{\theta\in A_n} \kappa_n(\theta) = \{\theta_*\}$. Let $C_n = A_n \cap B_\epsilon(\theta_*)^c$, where superscript c denotes the complement and $B_\epsilon(\theta_*)$ is the open ball in $\bar{\Theta}$ with radius ϵ and center θ_*. By Lemma 2.5, $\kappa_n(\theta) \geq \kappa_n(\theta_*) + \delta\epsilon^2$ for every $\theta \in C_n$. But Lemma 2.3 says that, with probability tending to one, $|g_n(\theta) - \kappa_n(\theta)| \leq \delta\epsilon^2/4$ (say) for any $\theta \in A_n$. Thus, with probability tending to one, for any $\theta \in C_n$, $g_n(\theta) \geq \kappa_n(\theta) - \delta\epsilon^2/4 \geq \kappa_n(\theta_*) + 3\delta\epsilon^2/4 \geq g_n(\theta_*) + \delta\epsilon^2/2$. Thus, with probability tending to one, $\arg\min_{\theta\in A_n} g_n(\theta) \cap C_n = \emptyset$. Thus, since Lemma 2.4 says $\arg\min_{\theta\in\Theta} g_n(\theta) \subseteq A_n$ with probability tending to one, $\arg\min_{\theta\in\Theta} g_n(\theta) \subseteq B_\epsilon(\theta_*) \cap A_n \subseteq B_\epsilon(\theta_*)$ with probability tending to one, and that finishes the proof. □

The following corollary says the estimates of the covariance matrices are consistent when those of the precision matrices are. Its proof is straightforward by using bounds on the eigenvalues of Ω_*, Ω_{X*}, and their estimates implied by Assumption 2 and Theorem 2.2, and is hence omitted.

Corollary 2.1. *If Assumptions 1 – 5 hold, then for any sequence of estimators $\{\hat{\theta}_n\}$ with $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} g_n(\theta)$, it holds that, in probability as $n \to \infty$,*

$$\|\hat{\Omega}^{-1} - \Sigma_*\| \to 0 \quad \textit{and} \quad \|\hat{\Omega}_X^{-1} - \Sigma_{X*}\| \to 0.$$

2.2 Consistent trace

To gain some intuition for when Assumption 5 holds, we discuss a few sufficient conditions that cover several relevant settings, including multivariate normal predictors.

Let $V_* G_* V_*^\mathsf{T}$ be a spectral decomposition of Σ_{X*} and define $Z_i = V_*^\mathsf{T} X_i$, $i = 1, \ldots, n$. The Z_is have common covariance matrix G_* and are sub-Gaussian when the X_is are. Moreover, $\operatorname{tr}(S_X - \Sigma_{X*}) = \operatorname{tr}(V_*^\mathsf{T} S_X V_* - G_*) = \operatorname{tr}(S_Z - G_*) = n^{-1} \sum_{j=1}^p \sum_{i=1}^n (Z_{i,j}^2 - G_{*j,j})$. Thus, independence of the Z_is and Chebyshev's inequality gives, for any $\epsilon > 0$,

$$\mathsf{P}\left(|\operatorname{tr}(S_X - \Sigma_{X*})| \geq \epsilon\right) \leq (\epsilon n)^{-2} \sum_{i=1}^n \left[\sum_{j=1}^p \operatorname{var}(Z_{i,j}^2) + 2\sum_{j<l} \operatorname{cov}(Z_{i,j}^2, Z_{i,l}^2)\right] \tag{5}$$

If X_i has a multivariate normal distribution for every i, then the elements of Z_i are independent since they are uncorrelated, and hence their squares are independent and have covariance zero. Then, the right hand side of (5) is less than $\epsilon^2 ncp/n^2$ for any $c \geq \text{var}(Z_i^2)$, which can be chosen to not depend on n or p by Assumption 4. But $\epsilon^2 ncp/n^2 \to 0$ as $n \to \infty$ by Assumption 3, so Assumption 5 is satisfied. More generally, the same argument works if linear combinations of X_i that are uncorrelated are also independent; normality is not necessary.

If the covariance terms in (5) need not vanish, then the right hand side is less than $\epsilon^2 ncp^2/n^2$, which tends to zero if $p(n)^2/n \to 0$. Thus, Assumption 5 is redundant if Assumption 3 is appropriately strengthened.

Finally, suppose that $X_i = \Gamma_*^\mathsf{T} W_i + \sqrt{\tau_*}\varepsilon_i$, where $W_i \in \mathbb{R}^k$ and $\varepsilon_i \in \mathbb{R}^p$ are latent variables whose elements are independent with mean zero and unit variance, and $\Gamma_* \in \mathbb{R}^{k \times p}$ is a parameter. This commonly considered latent variable model (Tipping and Bishop, 1999; Singer et al., 2016; Cook, 2018b) is consistent with our assumptions since $\text{cov}(X_i) = \Gamma_*^\mathsf{T}\Gamma_* + \tau_* I_p$ is a spiked covariance matrix whose smallest $p - k$ eigenvalues are all equal to τ_*. Let $W = [W_1, \ldots, W_n]^\mathsf{T} \in \mathbb{R}^{n \times k}$ and $E = [\varepsilon_1, \ldots, \varepsilon_n]^\mathsf{T} \in \mathbb{R}^{n \times p}$. Then $X = W\Gamma_* + \sqrt{\tau_*}E$, and hence

$$\text{tr}[S_X - \Gamma_*^\mathsf{T}\Gamma_* - \tau_* I_p] = \text{tr}[\Gamma_*^\mathsf{T}(S_W - I_k)\Gamma_*] + 2\sqrt{\tau_*}\,\text{tr}(\Gamma_*^\mathsf{T} S_{WE}) + \tau_*\,\text{tr}[S_E - I_p].$$

The absolute values of the first two terms are bounded by $k\|\Gamma_*\|^2\|S_W - I_k\|$ and $2k\sqrt{\tau}\|\Gamma_*\|\|S_{WE}\|$, respectively. These tend to zero in probability since Assumption 2 implies both τ_* and $\|\Gamma_*\|$ are bounded as $n \to \infty$, and Assumptions 3 and 4 imply consistency of sample covariance matrices in spectral norm (Lemma A.3). Finally, using Chebyshev's inequality as in (5) together with independence of the elements of E, and hence their squares, gives that $\text{tr}(S_E - I_p)$ tends to zero in probability.

We collect the sufficient conditions just discussed in the following proposition. Neither of the conditions is necessary, and they are not exclusive. In particular, 2 is a special case of 3.

Proposition 2.6. *Assumption 5 is satisfied if*

1. *Assumptions 1, 2, and 4 hold and Assumption 3 is strengthened to $p(n)^2/n \to 0$,*

or, in addition to Assumptions 1 – 4, any of the following holds for every i:

2. *X_i is multivariate normal.*

3. *X_i has the property that for any $v_1, v_2 \in \mathbb{R}^p$ such that $v_1^\mathsf{T} \Sigma_{X*} v_2 = 0$, it holds that $v_1^\mathsf{T} X_i$ and $v_2^\mathsf{T} X_i$ are independent.*

4. *$X_i = \Gamma_*^\mathsf{T} W_i + \sqrt{\tau_*}\varepsilon_i$ for some $W_i \in \mathbb{R}^k$ and $\varepsilon_i \in \mathbb{R}^p$, whose entries are independent with mean zero and unit variance, and parameters $\Gamma_* \in \mathbb{R}^{k \times p}$ and $\tau_* > 0$.*

3 Implementation

3.1 Maximizing the likelihood

Recall the definition of Θ and g_n in (3). Minimization of g_n over Θ is a non-convex problem without analytical solution. However, simplifications are possible upon re-parameterizing UDU^T as LL^T, and hence $\Sigma_X = \tau(I_p + LL^\mathsf{T})$, for $L \in \mathbb{L}^{p \times k}$, the set of lower-echelon matrices with positive leading entries in every column (Cantó et al., 2015). The representation is unique, so that the parameters are identified, if D is positive definite (Cantó et al., 2015). With this parameterization, left singular vectors of L are leading eigenvectors of Σ_X. Thus, $\beta = U\gamma$ is equivalent to $\beta = L\alpha$ for some $\alpha \in \mathbb{R}^{k \times r}$. Using this, the scaled negative log-likelihood after re-parameterization is

$$\begin{aligned} h_n(\Sigma, \alpha, \tau, L) = \log|\Sigma| + \frac{1}{n}\operatorname{tr}[(Y - XL\alpha)^\mathsf{T}(Y - XL\alpha)\Sigma^{-1}] \\ + \log|\tau(I_p + LL^\mathsf{T})| + \frac{1}{n\tau}\operatorname{tr}\left[X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}\right]. \end{aligned} \tag{6}$$

Minimizing h_n over $\mathbb{S}_{++}^r \times \mathbb{R}^{k \times r} \times (0, \infty) \times \mathbb{L}^{p \times k}$ is equivalent to minimizing g_n over Θ. However, h_n is useful for practical purposes because a profile log-likelihood for L can be derived analytically. Indeed, routine calculations show the necessary partial minimizers of H_n are $\bar{\alpha} = (L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}Y$, $\bar{\Sigma} = Y^\mathsf{T}Q_{XL}Y/n$, where Q_{XL} is the orthogonal projection onto the orthogonal complement of the column space of XL; and $\bar{\tau} = \operatorname{tr}\left[X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}\right]/(np)$. Thus,

$$H_n(L) = \min_{\Sigma, \alpha, \tau} h_n(\Sigma, \alpha, \tau, L) = \log|Y^\mathsf{T}Q_{XL}Y| + p\log\operatorname{tr}\left[X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}\right].$$

If a minimizer $\hat{L}$ of H_n can be found numerically, then the parameters β_*, Σ_*, τ_*, and hence Σ_{X*}, can be estimated by plugging $\hat{L}$ in place of L in the expressions for $\bar{\alpha}$, $\bar{\Sigma}$, and $\bar{\tau}$. On

points where XL is not invertible, the inverse in the definition of $\bar{\alpha}$ can be replaced by, for example, the Moore–Penrose pseudo-inverse without affecting H_n.

We use standard first order methods to minimize H_n, which is non-convex. Software implementing our method, using a descent-based algorithm for constrained optimization (Byrd et al., 1995), is available at the author's web page. Simulations and data examples indicate the estimates found by the proposed method are useful. Calculating the gradient necessary to implement a first order method is straightforward but tedious (Appendix B).

3.2 Selecting k

Information criteria offer a principled way to select k in practice. Fix p and r, and, for any $k \in \{0, \ldots, p\}$, let $d(k)$ denote the number of parameters and $\hat{\theta}_n(k)$ a minimizer of g_n. Then, up to additive constants, many popular information criteria can be written as

$$\mathbb{I}_\rho(k) = n g_n(\hat{\theta}_n(k)) + \rho d(k),$$

where different $\rho > 0$ give different criteria. Akaike's information criterion (AIC) (Akaike, 1998) and Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978) set, respectively, $\rho = 2$ and $\rho = \log(n)$. For a given ρ, k is selected as $\hat{k} \in \arg\min_{k=0,\ldots,p} \mathbb{I}_\rho(k)$. We examine the performance of AIC and BIC in our model using simulations in Section 4. The following proposition establishes $d(k)$. Recall, we are assuming $\mathbb{E}(Y_i) = 0$ and $\mathbb{E}(X_i) = 0$; parameterizing the means requires an additional r and p parameters, respectively.

Proposition 3.1. *For a given p, r, and k, the number of parameters in our model is $d(k) = r(r+1)/2 + k[r + 1 + p - (k+1)/2] + 1$ if $k < p$, and $d(k) = r(r+1)/2 + rk + p(p+1)/2$ if $k = p$.*

4 Simulations

We compare our method to classical PCR, PLS using the SIMPLS algorithm in the `pls` package (Mevik and Wehrens, 2007), predictor envelopes using the `Renvlp` package (Lee and Su, 2019), and OLS. We use root mean squared error (RMSE) of estimating β_* and out-of-sample prediction RMSE. These are defined, respectively, for a generic estimate $\hat{\beta}$ and

independent test set $(X_{new}, Y_{new}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times r}$, as

$$\|\hat{\beta} - \beta_*\|_F/\sqrt{rp} \quad \text{and} \quad \|X_{new}\hat{\beta} - Y_{new}\|_F/\sqrt{rn},$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. We also compare the methods' bias in selecting k. For classical PCR and PLS we use leave one out cross-validation to select k. For predictor envelopes, we use AIC, BIC, and a likelihood ratio test procedure implemented in the `Renvlp` package (Lee and Su, 2019). In all simulations, both the training set $(X, Y) \in \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times r}$ and the test set are generated as n independent observations from our model with multivariate normal responses and predictors. We fix $r = 2$ and $\Sigma_* = 2I_2$ in all simulations. We consider p and n in various alignments for which it is reasonable to use likelihood-based methods (here, our and envelopes). As a baseline, we use $p = 40$ and $n = 120$. This is similar to our data example (Section 5) which has $n = 123$ and $p = 28$. We construct the predictor covariance as $\Sigma_{X*} = \tau_*(I_p + U_*D_*U_*^\mathsf{T})$ where U_* is a realization from the uniform distribution on $\mathbb{O}^{p\times k}$. We examine performance for different values of n, p, k, D_*, and β_*. In all settings, $D_* = \text{diag}(1.1d_*, \ldots, 0.9d_*)$ for some d_*, which we call the average spiked eigenvalue. The coefficient is generated as $U_*\gamma_*$, where $\gamma_* \in \mathbb{R}^{k\times r}$ is constructed by drawing its elements as independent realizations of the uniform distribution on $(-1, 1)$ and then normalizing each column to have a Euclidean norm that can change between simulations; we refer to this as the coefficient column norm or the coefficient size. Because U_* is semi-orthogonal, $\|\beta_{*j}\| = \|U_*\gamma_{*j}\| = \|\gamma_{*j}\|$ for any column $j = 1, \ldots, r$.

Before discussing the results in Figure 1 setting by setting, we note that, overall, our method has lower RMSE than classical PCR, by a substantial margin, in almost all settings we consider. In some settings, predictor envelopes with k selected using likelihood ratio tests is competitive with our method, and in some PLS is. However, neither of the other methods perform as well as ours across all the simulations. The only setting in which our method is substantially outperformed by another is when $n = 50$ and $p = 40$; that is, when n is about the same size as p and both are relatively small. In this setting, the moment-based dimension reduction methods, classical PCR and PLS, perform better than the likelihood-based, though our method is substantially more accurate than predictor envelopes and OLS. In general, our method performs best in both estimation and prediction when k is selected using BIC, and envelopes perform better with k selected using likelihood ratio tests than information criteria.

The first row in Figure 1 shows our method performs best for all considered coefficient sizes, whether k is selected using AIC or BIC, and in both estimation and prediction. The method whose RMSE is closest to that of our method is predictor envelopes with k selected using likelihood ratio tests. PLS also performs well, especially for small coefficient sizes. On the other hand, leave one out cross validation with classical PCR grossly overestimates k, and for large coefficients classical PCR is no better than OLS. Overall, the results indicate dimension reduction is particularly useful in settings with small coefficients.

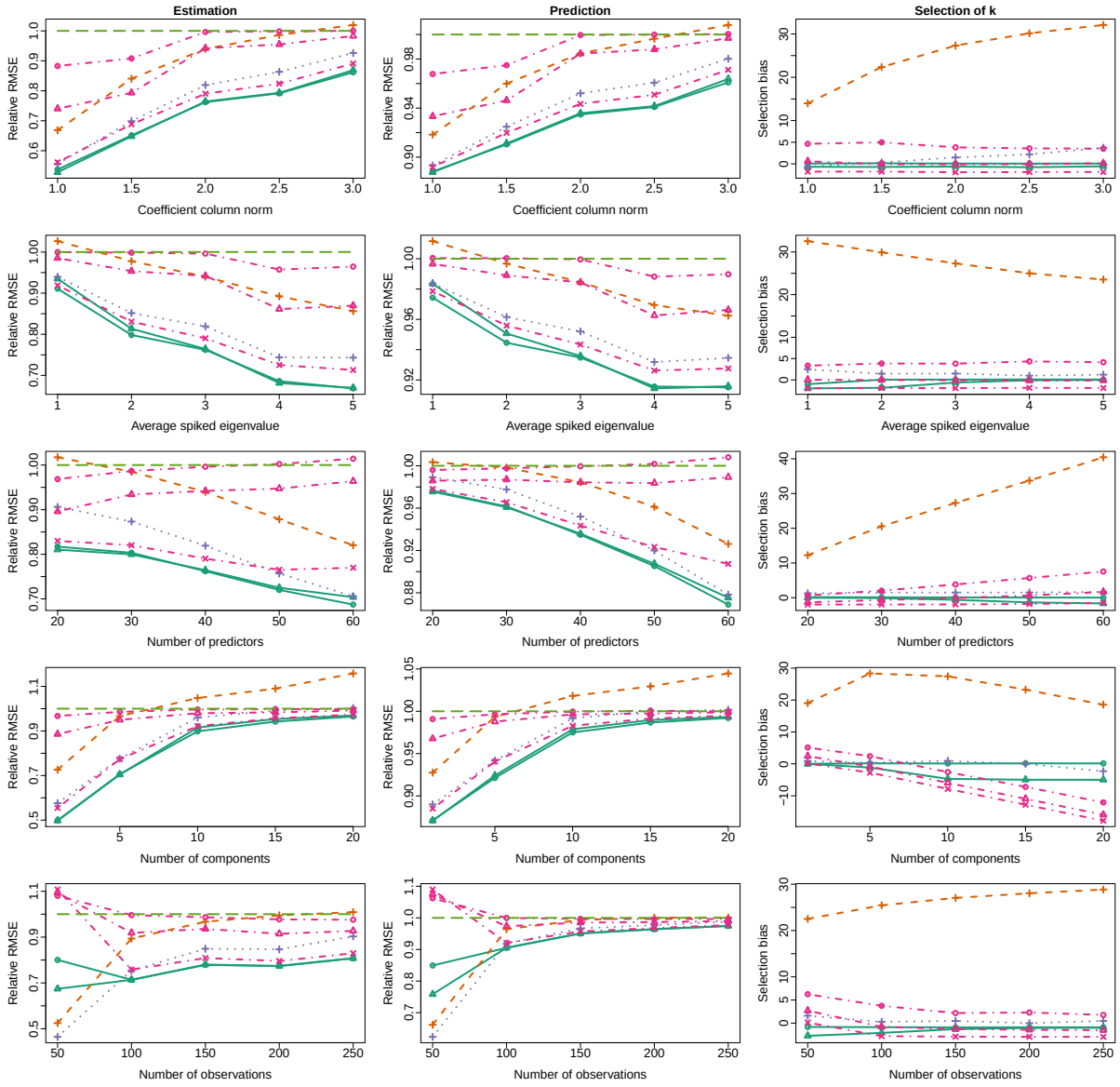
The second row in Figure 1 shows all methods perform better relative to OLS when the average spiked eigenvalue is larger, which is intuitive since the relevant eigenvectors should then be easier to identify. Our method with k selected using BIC performs best for all settings, but for the smallest d_* envelopes with k selected using likelihood ratio tests perform slightly better than our method with k selected using AIC.

The third row in Figure 1 shows our method and predictor envelopes with k selected using likleihood ratio tests are preferable to the moment-based methods when p is small, but that PLS becomes competitive as p increases. This is consistent with the fact that our method and envelopes in general require $n > p$, and our asymptotics which require $p/n \to 0$. Note, since $\|\beta_{*j}\| = 2$ for $j = 1, \ldots, r$ and all p, a large p implies a regression with relatively many predictors, but where each predictor has a small effect on the response. Conversely, a small p means each predictor has a larger effect.

In the fourth row of Figure 1, our method is shown to perform better than the other methods for all considered k. We also note that dimension reduction is decreasingly useful as the number of relevant components increases.

Finally, the last row of Figure 1 shows the moment-based methods are preferable for small n, and that in estimation, the differences between our method and the others is relatively stable as n varies. By contrast, the prediction RMSE for all methods gets closer to that of OLS as n increases. That is, dimension reduction may not be as useful when n is very large in comparison to p.

Figure 1: Monte Carlo root mean squared errors and selection biases.



NOTE: The root mean squared errors are divided by that for OLS with all predictors. The methods are our (solid lines), classical PCR (dashes), PLS (dots), predictor envelopes (dots and dashes), and OLS (long dashes). The number of components is selected by AIC (circles), BIC (triangles), leave one out cross validation (pluses), or likelihood ratio tests (crosses). When not varying as indicated on the horizontal axes, $n = 120$, $p = 40$, $k = 4$, $r = 2$, $\Sigma_* = 2I_2$, $\tau_* = 1$, $d_* = 3$, and $\|\beta_{*j}\| = 2$, $j = 1, \ldots, r$. All numbers are based on 500 Monte Carlo replications

.

5 Prediction of stock returns

We apply our method to data on $n = 123$ monthly returns, from January 2010 to March 2020, on 29 stocks in the Dow Jones Industrial Average index (one out of the usual 30 is omitted because it was introduced in 2019). Specifically, we consider cross-sectional prediction of the return on one stock using contemporaneous returns of the other $p = 28$ stocks. Code for replicating all results are available at the author's web page. Before discussing the results and how they are obtained, we give a short motivation for why stock returns are suitable for an application of our method.

First, classical models assume stock prices follow a geometric Brownian motion (Hull, 2017). This implies the (log-)returns—the first difference of the logarithm of the prices—are temporally independent, normally distributed, and have constant variance. Of course, in practice we expect neither to hold exactly, but the popularity of models based on the geometric Brownian motion suggests it is often a useful approximation. With this in mind, it seems reasonable to estimate parameters by maximizing a likelihood for independent normal observations.

Secondly, it is often hypothesized that stock returns can be decomposed into a stock-specific component, which is independent of the returns on other stocks, and a few common components, or factors (Fama and French, 2015; Bai and Wang, 2016). The latter can include, for example, a risk-free interest rate and a market return, and they can be observable or unobservable (Bai and Wang, 2016). Let $W_t = [W_{t,1}, \ldots, W_{t,k}]^\mathsf{T} \in \mathbb{R}^k$ denote a vector of such components at time $t \in \{1, \ldots, n\}$, and suppose the return on stock $j \in \{1, \ldots, p+1\}$ is

$$R_{t,j} = \mu_{*j} + \gamma_{*j}^\mathsf{T} W_t + \varepsilon_{t,j}, \tag{7}$$

where $\mu_{*j} \in \mathbb{R}$ and $\gamma_{*j} \in \mathbb{R}^k$ are parameters and $\varepsilon_{t,j}$ an unobservable error term with mean zero and variance $\tau_* > 0$, assumed to be independent and identically distributed for all t and j. Our data comprise realizations of $R_{t,j}$ but not $W_{t,j}$. Thus, the common components are latent variables that can induce dependence between contemporaneous returns of different stocks. To see how (7) relates to our model, let $Y_t = R_{t,1}$ be the return to be predicted using the vector of predictors $X_t = [R_{t,2}, \ldots, R_{t,p+1}]^\mathsf{T} \in \mathbb{R}^p$. Then with $\Gamma_* = [\gamma_{2*}, \ldots, \gamma_{p+1}]^\mathsf{T} \in \mathbb{R}^{p \times k}$ and $\Sigma_{W*} = \operatorname{cov}(W_t)$,

$$\Sigma_{X*} = \operatorname{cov}(X_t) = \Gamma_* \Sigma_{W*} \Gamma_*^\mathsf{T} + \tau_* I_p,$$

which is a spiked covariance matrix whose $p - k$ smallest eigenvalues are equal to τ_* and whose k leading eigenvectors are k leading left singular vectors of $\Gamma_* \Sigma_{W*}^{1/2}$. Moreover, (7) and normality of the returns imply

$$\mathbb{E}(Y_t \mid X_t) = \mathbb{E}(Y_t) + \Sigma_X^{-1} \operatorname{cov}(X_t, Y_t)(X_t - \mathbb{E}(X_t)) = \mathbb{E}(Y_t) + \beta_*^{\mathsf{T}}[X_t - \mathbb{E}(X_t)]$$

where $\beta_* = \Sigma_{X*}^{-1} \operatorname{cov}(X_t, Y_t) = (\Gamma_* \Sigma_{W_*} \Gamma_*^{\mathsf{T}} + \tau_* I_p)^{-1} \Gamma_* \Sigma_{W*} \gamma_{*1}$. Thus, β_* lies in the span of k leading left singular vectors of $\Gamma_* \Sigma_{W*}^{1/2}$ and, hence, (7) leads to a model consistent with our assumptions.

We split the 123 observations so that the first 70 are training data and the remaining 53 are test data. The response is centered by its training data sample mean and the predictors are centered and scaled by their training data sample means and sample standard deviations, respectively. We fit each method to the training data and compute the RMSE of prediction using the test data. For our method, we select k using AIC and BIC. For classical PCR and PLS, we select k by leave one out cross-validation, using the implementation in the `pls` package (Mevik and Wehrens, 2007). The package suggests two ways of selecting k based on cross-validation: a randomization test approach (van der Voet, 1994) or the "one standard error rule", which selects the smallest number of components that gives a cross-validation error within one standard error of the smallest cross-validation error of any number of components (Hastie et al., 2009). In our application, we found that both of these rules often selected the model with zero components, which led to poor performance and an uninteresting comparison with our method. Thus, we instead present results where the number of components are selected to minimize the cross-validation error. The predictor envelope model we fit using the `Renvlp` package (Lee and Su, 2019). This package provides three ways to select k: AIC, BIC, and likelihood ratio tests—we consider all three.

We focus on prediction of Home Depot's stock return. To highlight which results are particular to this choice of response and which hold more generally, we also present summary statistics from repeating the same analysis with the other stocks as responses. We focus on the results for Home Depot because they are relatively representative of more general patterns indicated by the summary statistics.

Table 1 shows the prediction results. The presented RMSEs are divided by the RMSE of a model without predictors, that is, the model which predicts all response realizations in the

test set are equal to the training data sample mean of the response. For Home Depot's stock, our method with k selected using BIC performs best (RMSE 0.75); PLS (0.76) and envelopes with k selected using likelihood ratio tests (0.76) also perform well. Envelopes with k selected using AIC or BIC and OLS perform worse than a prediction without predictors. In general, the results for Home Depot indicate that methods which select a relatively small k, but not $k = 0$, perform better. This finding is corroborated when looking at the summary statistics from repeating the analysis with the other stocks as responses: Our method with k selected using BIC performs best (average RMSE 0.80) and, on average, selects $k = 2$. PLS performs second best (0.82) and on average selects $k = 1.8$. Our method with k selected using AIC and classical PCR also perform relatively well. On average, all methods except OLS perform better than the model without predictors. However, our method with k selected using BIC is the only method that does no worse than the model without predictors for any choice of response.

We now focus on the Home Depot results for our method with $k = 2$, as selected by BIC. In Figure 2, the data are plotted along with fitted values and predictions. The fitted values and predictions are typically closer to the sample mean than the response, and look to be relatively close to an (scaled) average of the predictors. Accordingly, if we fit our model to the full data with $k = 2$, then the estimated leading eigenvector of Σ_{X*}, say $\hat{u}_1$, has elements which all have the same sign and absolute values between 0.10 and 0.24 (see Supplementary Material); that is, $\hat{u}_1^\mathsf{T} X_t$ is proportional to a weighted average of the predictors, which with (7) in mind could be interpreted as a market component. The second estimated eigenvector, $\hat{u}_2$, is more difficult to interpret: Its elements do not all have the same sign, and some are close to zero. In an OLS regression of Y_t on $\hat{u}_1^\mathsf{T} X_t$ and $\hat{u}_2^\mathsf{T} X_t$, the t-value for testing the null hypothesis that the first coefficient is zero is -8.25 and that for the second -0.53 (Supplementary Material). This may be an indication that only the first component has an effect on Y_t. This is still compatible with $k = 2$ as suggested by BIC since, even if only one linear combination of X_t has a direct effect on Y_t, $k < 2$ may lead to a poor model for Σ_X. Inspecting the estimates from the full data, we find $\hat{D} = \text{diag}(18.4, 3.4)$ and $\hat{\tau} = 0.56$; that is, $\lambda_{\max}(\hat{\Sigma}_X) = 10.82$, $\lambda_2(\hat{\Sigma}_X) = 2.47$, and $\lambda_{\min}(\hat{\Sigma}_X) = 0.56$, where $\lambda_j(\cdot)$ denotes the jth largest eigenvalue of its argument matrix.

Table 1: Root mean squared errors for out-of-sample predictions

Statistic	Our (A)	Our (B)	PCR	PLS	Env. (A)	Env. (B)	Env. (L)	OLS
				Home Depot				
RMSE	0.79	0.75	0.79	0.76	1.02	1.03	0.76	1.05
$\hat{k}$	4	2	3	1	5	3	1	28
				All stocks				
Ave. RMSE	0.83	0.80	0.84	0.82	0.98	0.96	0.90	1.01
Max. RMSE	1.25	0.99	1.15	1.14	1.38	1.36	1.39	1.34
Ave. $\hat{k}$	3.90	2.00	8.38	1.79	5.28	3.17	1.52	28.00

NOTE: The reported RMSEs are for the last 53 observations and are scaled by the RMSE of the training data sample mean so that a prediction assuming $\beta_* = 0$ has unit RMSE. The number of components ($\hat{k}$) is selected using Akaike's (A) or Bayesian (B) information criteria, likelihood ratio tests (L), or for (classical) PCR and PLS, leave one out cross-validation. OLS uses all 28 predictors. Numbers for "All stocks" are summary statistics of the 29 RMSEs obtained by applying the methods once with every stock return as response.

6 Discussion

We have proposed a likelihood-based method for estimating principal components regressions that outperforms classical principal components regression in many practically relevant settings. Our method does this by addressing a common criticism of classical principal components regression, namely that the responses are ignored when selecting linear combinations of the predictors.

Both our data example and simulations indicate our method performs best when the number of components are selected using BIC, rather than AIC. In general, erring on the side of using fewer components seems preferable to using too many; that is, parsimonious models are preferred. In our data example, BIC indicates that using two components is appropriate, but further analysis indicates one may be sufficient for the regression. Thus, it may be that BIC

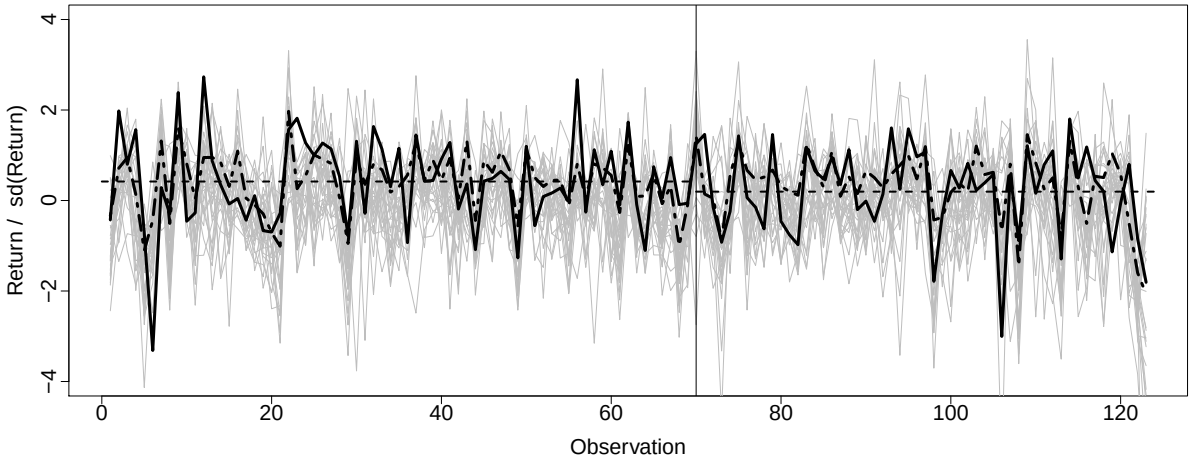


Figure 2: Logarithmic returns (log-differences)

NOTE: The returns are scaled by their sample standard deviation. Thin lines are all firms' returns except Home Depot. The thick solid line is Home Depot's returns and the thick dashed is fitted values (observations 1 – 70) and predictions (observations 71 – 123) from our model with $k = 2$. The vertical line indicates the break between training and testing data. The dashed horizontal lines indicate training and testing data sample means for Home Depot.

selects two instead of one because the model for the predictors' covariance matrix requires two components to fit well. This suggests that, more generally, it may be possible to further improve our method by implementing variable selection, or component selection. With some work, our method could be extended in that direction by, for example, encouraging sparsity in the estimate of the parameter γ in the representation $\beta = U\gamma$. Sparsity in γ would say that, while the principal components regression model holds, not all of the principal components are important for all responses. If, for example, the jth row of γ vanishes, then the jth principal components is unimportant. Moreover, if a sparsity-inducing or other penalty on γ is added to the negative log-likelihood, then the resulting objective function need not be unbounded when $n < p$. That is, it may also be possible to extend our method to high-dimensional settings by using a penalized likelihood estimator. When p is small in comparison to n, one may also be able to select which components are relevant for the regression by comparing information criteria from fitting the model with various restrictions on γ. Both of these approaches would

require a substantial amount of work as the current algorithm relies on a re-parameterization where γ and U do not appear explicitly.

In some applications it may be that only a few components are relevant, but that they do not correspond to leading eigenvectors. Our method can still perform well relative to PLS and predictor envelopes in many such cases. Intuitively, there is a price to pay for having to, as those methods aim to, infer which eigenvectors are important, so in some settings it can be preferable to fit a principal components regression model with many linear combinations, including some important and some unimportant ones.

Finally, we note that there may be room for extending our method to dependent data. Our asymptotic theory fundamentally relies on concentration of sub-Gaussian vectors; if the concentration results we use can be replaced by equivalent ones for dependent data, we expect other parts of our proofs can be adapted accordingly to show our estimators are consistent. However, an arguably more appropriate, and more ambitious, extension of our method to dependent data would use a likelihood that takes that dependence into account, for example by assuming responses and predictors follow vector autoregressive processes. This would also allow for an interesting re-analysis of the stock return data, focusing on forecasting instead of cross-sectional prediction.

A Proofs

Define

$$g_{n,1}(\theta) = -\log|\Omega| + n^{-1}\operatorname{tr}[(Y - X\beta)^{\mathsf{T}}(Y - X\beta)\Omega],$$

$$g_{n,2}(\theta) = -\log|\Omega_X| + \operatorname{tr}[S_X\Omega_X],$$

$$\kappa_{n,1}(\theta) = \mathbb{E}[g_{n,1}(\theta)] = -\log|\Omega| + \operatorname{tr}\left[(\beta - \beta_*)^{\mathsf{T}}\Omega_X^{-1}(\beta - \beta_*)\Omega\right] + \operatorname{tr}(\Omega\Omega_*^{-1}),$$

and

$$\kappa_{n,2}(\theta) = \mathbb{E}[g_{n,2}(\theta)] = -\log|\Omega_X| + \operatorname{tr}(\Omega_X\Omega_{X*}^{-1}).$$

The following is Corollary III.2.6 in Bhatia (2012), attributed to Weyl, which we state here for easy reference. We let $\lambda_j(\cdot)$ denote the jth largest eigenvalue of its argument matrix.

Lemma A.1. *If A and B are symmetric matrices for which $\|A-B\| \leq \epsilon$, then* $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \epsilon$.

The following lemma is similar to Lemma A.18 in Cook (2018b). The differences are that we focus on minimization rather than maximization, that some of our conditions are weaker, and that, consequently, our conclusion is different. The proof idea is the same, however, so we omit the proof.

Lemma A.2. *If $S \in \mathbb{R}^{p\times p}$ is symmetric and $R = \operatorname{diag}(r_1, \ldots, r_k) \in \mathbb{R}^{k\times k}$, $k \leq p$, with $r_1 \geq \cdots \geq r_k \geq 0$, then $H \mapsto \operatorname{tr}(SHRH^\mathsf{T})$ is minimized over $\mathbb{O}^{p\times k}$ by $\hat{H} = [v_p, \ldots, v_{p-k+1}]$, where v_j is a normalized eigenvector corresponding to the jth largest eigenvalue of S.*

Proof of Theorem 2.1 (non-empty argmin). Suppose $Y^\mathsf{T}Q_XY$ is invertible and S_X has rank greater than k. Routine calculations show g_n is uniquely partially minimized in Ω by $\bar{\Omega} = [(Y-X\beta)^\mathsf{T}(Y-X\beta)/n]^{-1} = [Y^\mathsf{T}Q_XY/n + (\beta-\hat{\beta}_{OLS})^\mathsf{T}S_X(\beta-\hat{\beta}_{OLS})]^{-1}$, where $\hat{\beta}_{OLS} = S_X^+S_{XY}$ and superscript $+$ denotes the Moore-Penrose pseudo-inverse. It follows that a minimizer of g_n exists if and only if the partially minimized objective $g_n^{(1)}(\beta, \Omega_X) = \min_\Omega g_n(\beta, \Omega, \Omega_X)$ has a minimizer. We have

$$g_n^{(1)}(\beta, \Omega_X) = \log|Y^\mathsf{T}Q_XY/n + (\beta - \hat{\beta}_{OLS})^\mathsf{T}S_X(\beta - \hat{\beta}_{OLS})| + \operatorname{tr}(I_r) + g_{n,2}(\Omega_X).$$

Suppose $g_n^{(1)}$ has a minimizer $(\hat{\beta}, \hat{\Omega}_X)$, then $(P_{X^\intercal}\hat{\beta}, \hat{\Omega}_X)$, where P is the orthogonal projection onto the column space of the matrix in the subscript, is also a minimizer because it gives the same objective function value. Thus, to show the existence of a minimizer it suffices to show the existence of a minimizer over the set of (β, Ω_X) which satisfies (1) and $P_{X^\intercal}\beta = \beta$; denote this set $\Theta^{(1)}$. Let $B = B(c_1, c_2)$ be the set of all $(\beta, \Omega_X) \in \Theta^{(1)}$ for which $\|\beta\| \leq c_2$, $\lambda_{\min}(\Omega_X) \geq c_1$, and $\lambda_{\max}(\Omega_X) \leq c_2$. We will show that c_1 and c_2 can be selected so that (i) $\arg\min_{(\beta,\Omega_X)\in B} g_n^{(1)}(\beta, \Omega_X)$ is non-empty and (ii) there are no global minimizers on $\Theta^{(1)} \setminus B$, which together imply the desired conclusion. We start with (ii).

By unconstrained minimization in β, spectral decomposition $\Omega_X = HRH^\mathsf{T}$, and Lemma A.2,

$$g_n^{(1)}(\beta, \Omega_X) \geq \log|Y^\mathsf{T}Q_XY/n| + \operatorname{tr}(I_r) - \log|R| + \operatorname{tr}[FR],$$

where F is diagonal with the eigenvalues of S_X in increasing order on its diagonal; that is, small diagonal elements in F correspond to large diagonal elements in R. Letting $F = \text{diag}(f_1, \ldots, f_p)$ and $R = \text{diag}(r_1, \ldots, r_p)$, and using $r_1 = \cdots = r_{p-k}$ for any Ω_X in our parameter set, we get

$$g_n^{(1)}(\beta, \Omega_X) \geq \log|Y^\mathsf{T} Q_X Y/n| + r - (p-k)\log(r_1) + r_1 \sum_{j=1}^{p-k} f_j + \sum_{j=p-k+1}^{p} \left[-\log(r_j) + f_j r_j\right]. \quad (8)$$

Since S_X has rank greater than k, f_{p-k} is strictly positive, and hence the derived lower bound in (8) tends to infinity if $\lambda_{\max}(\Omega_X) = r_1 \to \infty$. Similarly, the lower bound tends to infinity if $\lambda_{\min}(\Omega_X) = r_p \to 0$. Thus, we can pick c_1 and c_2 so that $g_n^{(1)}$ is larger than at an arbitrary, fixed, point in $\Theta^{(1)}$ whenever $\lambda_{\min}(\Omega_X) < c_1$ or $\lambda_{\max}(\Omega_X) > c_2$.

Next, using the results of Tipping and Bishop (1999) or Lemma A.2, it follows that $g_{n,2}$ is minimized by $\bar{\Omega}_X = HRH^\mathsf{T}$ with $H = [v_p, \ldots, v_1]$, v_j an eigenvector corresponding to $\lambda_j(S_X)$, and $R = \text{diag}(r_1, \ldots, r_p)$ with $r_j^{-1} = \sum_{j=k+1}^{p} \lambda_j(S_X)/(p-k)$ for $j = 1, \ldots, p-k$ and $r_j^{-1} = \lambda_{p-j+1}(S_X)$ for $j = p-k+1, \ldots, p$. Thus,

$$g_n^{(1)}(\beta, \Omega_X) \geq \log|Y^\mathsf{T} Q_X Y/n + (\beta - \hat{\beta}_{OLS})^\mathsf{T} S_X (\beta - \hat{\beta}_{OLS})| + g_{n,2}(\bar{\Omega}_X),$$

Define $\tilde{\beta} = \beta - \hat{\beta}_{OLS}$ and notice that $P_{X^\mathsf{T}}\tilde{\beta} = \tilde{\beta}$. Let VGV^T be a spectral decomposition of S_X, with $V \in \mathbb{O}^{p\times q}$, q being the rank of S_X. Then $\tilde{\beta} = V\alpha$ for some $\alpha \in \mathbb{R}^{q\times r}$. Suppose $\|\beta\| \to \infty$, then $\|\tilde{\beta}\| = \|\alpha\| \to \infty$, and thus the maximum eigenvalue of $Y^\mathsf{T} Q_X Y + \tilde{\beta}^\mathsf{T} S_X \tilde{\beta}$, which is greater than $\|\alpha\|\lambda_{\min}(G)$, also tends to infinity. Since its smallest eigenvalue is lower bounded by that of $Y^\mathsf{T} Q_X Y$, we conclude $\log|Y^\mathsf{T} Q_X Y/n + (\beta - \hat{\beta}_{OLS})^\mathsf{T} S_X (\beta - \hat{\beta}_{OLS})| \to \infty$ if $\|\beta\| \to \infty$. Thus, we have established (ii).

It remains to show (i), fow which it suffices to show $g_n^{(1)}$ is continuous on B and that B is compact. The former follows from that Ω_X and $Y^\mathsf{T} Q_X Y + \tilde{\beta}^\mathsf{T} S_X \tilde{\mathcal{B}}$ are positive definite on B (see e.g. Magnus and Neudecker, 2002), and hence we omit the proof. Since B is bounded by construction, to show it is compact it suffices to show it is closed. Consider a sequence $\{(\beta^m, \Omega_X^m)\} \in B$, $m \geq 1$, converging to $(\beta^0, \Omega_X^0) \in \mathbb{R}^{p\times r} \times \mathbb{R}^{p\times p}$ as $m \to \infty$. We must show $(\beta^0, \Omega_X^0) \in B$. That $\|\beta^0\| \leq c_2$ follows from $\|\beta^m\| \leq c_2$ and continuity of norms. That the eigenvalues of Ω_X^0 are all between c_1 and c_2 follows from Lemma A.1 and that those of Ω_X^m are. Symmetry of Ω_X^0 follows from that convergence in spectral norm implies elementwise

convergence: For any (i,j), $(\Omega_X^m)_{i,j} - (\Omega_X^m)_{j,i} = 0$ for every m and, hence, so is it for the limit as $m \to \infty$. Since $\beta^m \to \beta^0$, we have $\beta^m = P_{X^\intercal}\beta^m \to P_{X^\intercal}\beta^0 = \beta^0$. It remains only to show $(\Omega_X^0)^{-1}$ and β satisfy (1).

Let $c_1^{-1} \geq \sigma_1^m \geq \cdots \geq \sigma_p^m \geq c_2^{-1}$ denote eigenvalues of $\Sigma_X^m = (\Omega_X^m)^{-1}$ and let $v_1^m, \ldots, v_p^m$ denote a set of orthonormal eigenvectors corresponding to those eigenvalues such that $\beta^m \in \text{span}\{v_1^m, \ldots, v_k^m\}$. We can always pick such eigenvalues and eigenvectors because the sequence $\{(\beta, \Omega)^m\}$ lives in B. Now, $\|\Sigma_X^m - \Sigma_X^0\| = \|\Sigma_X^m(\Omega_X^m - \Omega_X^0)\Sigma_X^0\| \leq \|\Sigma_X^m\|\|\Sigma_X^0\|\|\Omega_X^m - \Omega_X^0\| \leq c_1^{-2}\|\Omega_X^m - \Omega_X^0\|$ which tends to zero as $m \to \infty$. Thus, by Lemma A.1, $|\sigma_j^m - \sigma_j^0| \to 0$ for every j. Consider

$$\Sigma_X^m v_j^m = \sigma_j^m v_j^m;$$

because $\|v_j^m\| = 1$, Bolzano–Weierstrass says we can, for any j, find a sub-sequence $\{v_j^{m_l}\}$ that converges to some limit v_j^0; in fact, by repeating that argument we can extract a subsequence such that this holds simultaneously for all $j = 1, \ldots, p$. Taking limits along this sub-sequence, we find

$$\Sigma_X^0 v_j^0 = \sigma_j^0 v_j^0;$$

that is, v_j^0 is an eigenvector of Σ_X^0 corresponding to σ_j^0. Since $(v_j^m)^\intercal v_{j'}^m = \mathbf{1}(j = j')$ for every $j, j' \leq k$, taking limits along the subsequence shows $\{v_1^0, \ldots, v_k^0\}$ is a set of orthonormal eigenvectors of Σ_X^0. Finally, since $\beta^{m_l} \to \beta^0$ and $v_j^{m_l} \to v_j^0$, taking limits in $\beta^{m_l} = V^{m_l}(V^{m_l})^\intercal \beta^{m_l}$, where $V^m = [v_1^m, \ldots, v_k^m] \in \mathbb{O}^{p \times k}$, shows $\beta^0 = V^0(V^0)^\intercal \beta^0$, which finishes the proof. □

Proof of Theorem 2.1 (empty argmin). Recall (8) and notice that this bound is attainable. But if S_X has rank less than or equal to k, then the coefficient on $r_1 = \lambda_{\max}(\Omega_X)$ vanishes and, hence, we can send $g_n^{(1)}(\beta, \Omega_X)$ to negative infinity by sending $r_1 \to \infty$; that is, no minimizer exists.

If $Y^\intercal Q_X Y$ is non-invertible and $k \geq r$, then we can fix an Ω_X such that the columns of $\hat{\beta}_{OLS}$ are in the span of the k leading eigenvectors of Ω_X^{-1}. Thus, we can pick $\beta = \hat{\beta}_{OLS}$, and evaluate

$$g_n(\hat{\beta}_{OLS}, \Omega, \Omega_X) = -\log|\Omega| + \text{tr}(Y^\intercal Q_X Y \Omega) + g_{n,2}(\Omega_X).$$

Now it is straightforward to see that g_n is unbounded from below: we can select Ω to have the same eigenvectors as $Y^\intercal Q_X Y$ and, say, the largest eigenvalue of Ω to correspond to the same

eigenvector as one of the vanishing eigenvalues of $Y^\mathsf{T} Q_X Y$. Thus, we can send the largest eigenvalue of Ω to infinity, sending $-\log|\Omega|$ to negative infinity, without affecting other terms of g_n, and this finishes the proof.

□

Lemma A.3 (Vershynin (2018), Theorem 4.6.1)**.** *If $Z_1, \ldots, Z_n \in \mathbb{R}^p$ are independent and identically distributed with mean zero and sub-Gaussian norm bounded by $c < \infty$; then for a $c_1 < \infty$ and any $\epsilon > 0$ and $t \geq 1$,*

$$\mathsf{P}\left(\|S_Z - \operatorname{cov}(Z_1)\| \leq \epsilon\right) \geq 1 - 2\exp(-t^2 p)$$

for every $n \geq c_1 c^4 (t/\epsilon)^2 p$

Proof of Lemma 2.3. By the triangle inequality and that the fact that the supremum of a sum is less than the sum of the suprema,

$$\mathsf{P}\left(\sup_{\theta\in A_n} |g_n(\theta) - \kappa_n(\theta)| \geq \epsilon\right) \leq \mathsf{P}\left(\sup_{\theta\in A_n} |g_{n,1}(\theta) - \kappa_{n,1}(\theta)| + \sup_{\theta\in A_n} |g_{n,2}(\theta) - \kappa_{n,2}(\theta)| \geq \epsilon\right).$$

Since the sum is greater than ϵ only if at least one of the summands is greater than $\epsilon/2$, subadditivity (union bound) says the last quantity is upper bounded by

$$\mathsf{P}\left(\sup_{\theta\in A_n} |g_{n,1}(\theta) - \kappa_{n,1}(\theta)| \geq \epsilon/2\right) + \mathsf{P}\left(\sup_{\theta\in A_n} |g_{n,2}(\theta) - \kappa_{n,2}(\theta)| \geq \epsilon/2\right).$$

We will show that both of these tend to zero and start with the first one.

We have

$$\begin{aligned} g_{n,1}(\theta) - \kappa_{n,1}(\theta) &= \frac{1}{n}\operatorname{tr}\left[(Y - X\beta)^\mathsf{T}(Y - X\beta)\Omega\right] - \frac{1}{n}\mathbb{E}\operatorname{tr}\left[(Y - X\beta)^\mathsf{T}(Y - X\beta)\Omega\right] \\ &= \operatorname{tr}\left[(S_Y - \Sigma_{Y*})\Omega\right] + 2\operatorname{tr}\left[\beta^\mathsf{T}(\Sigma_{XY*} - S_{XY})\Omega\right] \\ &\quad + \operatorname{tr}\left[\beta^\mathsf{T}[S_X - \Sigma_{X*}]\beta\Omega\right]. \end{aligned}$$

Since for any $A \in \mathbb{R}^{r\times r}$, $|\operatorname{tr}(A)| \leq r\|A\|$, the absolute value of the last right hand side is upper bounded by

$$r\|S_Y - \Sigma_{Y*}\|\|\Omega\| + 2r\|\Omega\beta^\mathsf{T}\|\|\Sigma_{XY*} - S_{XY}\| + r\|\beta\Omega\beta^\mathsf{T}\|\|S_X - \Sigma_{X*}\|.$$

Thus, since $\|\beta\| \leq c_2$ and $\lambda_{\max}(\Omega) \leq c_2$ on A_n, $\sup_{\theta \in A_n} |g_{n,1}(\theta) - \kappa_{n,1}(\theta)|$ is upper bounded by

$$r\|S_Y - \Sigma_{Y*}\|c_2 + 2rc2^2\|\Sigma_{XY*} - S_{XY}\| + rc_2^3\|S_X - \Sigma_{X*}\|,$$

which tends to zero in probability by Lemma A.3.

For the second term we have by picking $\tau > 0$, $D \in \mathbb{D}_+^k$, and $U \in \mathbb{O}^{p \times k}$ such that $\Sigma_X = \tau(I_p + UDU^\mathsf{T})$,

$$|g_{n,2}(\theta) - \kappa_{n,2}(\theta)| = \tau^{-1}|\operatorname{tr}[(S_X - \Sigma_{X*})(I_p + UDU^\mathsf{T})^{-1}],$$

which, by using $(I_p + UDU^\mathsf{T})^{-1} = I_p - U(D + I_k)^{-1}DU^\mathsf{T}$, can be written

$$\tau^{-1}|\operatorname{tr}[S_X - \Sigma_{X*}] - \operatorname{tr}[U^\mathsf{T}(S_X - \Sigma_{X*})U(D + I_k)^{-1}D]|.$$

Using the triangle inequality and relating the trace to the spectral norm, the last line is upper bounded by

$$\tau^{-1}|\operatorname{tr}[S_X - \Sigma_{X*}]| + \tau^{-1}k\|S_X - \Sigma_{X*}\|,$$

where the scaling k on the second term comes from that $U \in \mathbb{O}^{p \times k}$ and $\|(I_p + D)^{-1}D\| \leq 1$. Now we are done upon observing that $\tau^{-1} \leq c_2$ on A_n, the trace term tends to zero in probability by Assumption 5, and $\|S_X - \Sigma_{X*}\|$ tends to zero in probability by Lemma A.3. □

Proof of Lemma 2.4. Because $\lambda_{\min}(\Sigma_{X*}) = \tau_* \geq c^{-1} > 0$, where c is that given by Assumption 2, Lemmas A.1 and A.3 imply S_X is invertible with probability tending to one. Thus, it suffices to consider outcomes with invertible S_X. Pick an arbitrary $\bar{\theta} = (\bar{\beta}, \bar{\Omega}, \bar{\Omega}_X) \in \arg\min_{\theta \in \Theta} g_n(\theta)$; if none exists we are done trivially, so assume one does. We show that, with probability tending to one, $\bar{\theta} \in A_n(c_1, c_2)$ for some small enough $c_1 > 0$ and large enough $c_2 < \infty$ not depending on n or p.

Because $\bar{\theta} \in \Theta$, there exist $\bar{U} \in \mathbb{O}^{p \times k}$, $\bar{D} \in \mathbb{D}_+^k$, $\bar{\tau} > 0$, and $\bar{\gamma} \in \mathbb{R}^{p \times k}$ such that $\bar{\Sigma}_X = \bar{\Omega}_X^{-1} = \bar{\tau}(I_p + \bar{U}\bar{D}\bar{U}^\mathsf{T})$ and $\bar{\beta} = \bar{U}\bar{\gamma}$. Since $\bar{\theta}$ is a minimizer, $\bar{\gamma}$ is a partial minimizer and, hence, minimizes

$$\gamma \mapsto \operatorname{tr}[(Y - X\bar{U}\gamma)^\mathsf{T}(Y - X\bar{U}\gamma)\bar{\Omega}];$$

that is,

$$\bar{\gamma} = (\bar{U}^\mathsf{T}S_X\bar{U})^{-1}\bar{U}^\mathsf{T}S_{XY}.$$

Thus, using that the spectral norm is submultiplicative,

$$\|\bar{\beta}\| = \|\bar{\gamma}\| \leq \|(\bar{U}^\mathsf{T} S_X \bar{U})^{-1}\| \|\bar{U}^\mathsf{T} S_{XY}\| \leq \lambda_{\min}(S_X)^{-1} \|S_X\|^{1/2} \|S_Y\|^{1/2},$$

which by Lemmas A.1 and A.3 is less than, say, $\lambda_{\max}(\Sigma_{X*})^{1/2} \lambda_{\max}(\Sigma_{Y*})^{1/2} / (2\lambda_{\min}(\Sigma_{X*})) \leq c^{3/2}(c+c^3)^{1/2}/2$ with probability tending to one. A similar argument shows $\bar{\Sigma} = \bar{\Omega}^{-1}$ must satisfy

$$\bar{\Sigma} = Y^\mathsf{T} Q_{X\bar{U}} Y / n.$$

Since the column space of $X\bar{U}$ is a subset of that of X, it follows that

$$\lambda_{\min}(Y^\mathsf{T} Q_X Y/n) \leq \lambda_{\min}(\bar{\Sigma}) \leq \lambda_{\max}(\bar{\Sigma}) \leq \lambda_{\max}(Y^\mathsf{T} Y/n).$$

Thus, by Lemmas A.1 and A.3, with probability tending to one,

$$c^{-1}/2 \leq \lambda_{\min}(\Sigma_*)/2 \leq \lambda_{\min}(\bar{\Sigma}) \leq \lambda_{\max}(\bar{\Sigma}) \leq 2\lambda_{\min}(\Sigma_*) \leq 2c.$$

Finally, let VGV^T be a spectral decomposition of Ω_X, where $G_{1,1} = G_{2,2} = \cdots = G_{p-k,p-k} \geq G_{p-k+1,p-k+1} \geq \cdots \geq G_{p,p}$. The terms of g_n that depend on G are

$$-\log|VGV^\mathsf{T}| + \operatorname{tr}(VGV^\mathsf{T} S_X) = \sum_{j=1}^{p} \left[-\log(G_{j,j}) + (V^\mathsf{T} S_X V)_{j,j} G_{j,j} \right].$$

Let $a_j = (V^\mathsf{T} S_X V)_{j,j}$. We will show that

$$2c \geq 2\lambda_{\max}(\Omega_{X*}) \geq \bar{G}_{1,1} \geq \bar{G}_{p,p} \geq \lambda_{\min}(\Omega_{X*})/2 \geq c^{-1}/2$$

with probability tending to one by using the facts that (i) $\lambda_{\min}(S_X) \leq a_j \leq \lambda_{\max}(S_X)$ for every j since V is orthogonal, and (ii) $x \mapsto -\log(x) + x a_j$ is strictly convex on $(0,\infty)$ and minimized at $1/a_j$. First, it must be that $\bar{G}_{p,p} \leq 1/\lambda_{\min}(S_X)$ since, by observations (i) and (ii), setting $G_{j,j} = 1/\lambda_{\min}(S_X)$ for all j gives a lower objective than any point for which $G_{j,j} > 1/\lambda_{\min}(S_X)$ for every j. Now pick an arbitrary $j < p$ and suppose $\bar{G}_{j,j} \leq 1/\lambda_{\min}(S_X)$. Then by observations (i) and (ii), it must be that $\bar{G}_{j+1,j+1} \leq 1/\lambda_{\min}(S_X)$ since, if it were not, the objective could be decreased by decreasing $\bar{G}_{j+1,j+1}$ without affecting the ordering. It follows by induction that $\bar{G}_{j,j} \leq 1/\lambda_{\min}(S_X)$ for all j. A similar argument shows $\bar{G}_{j,j} \geq 1/\lambda_{\max}(S_X)$ for all j, and the desired conclusion follows from Lemmas A.1 and A.3. □

Proof of Lemma 2.5. Pick an arbitrary $\theta \in A_n$ and let $\|\theta - \theta_*\|_M = \epsilon$. By definition of $\|\cdot\|_M$, it must hold that (i) $\|\beta - \beta_*\| = \epsilon$, (ii) $\|\Omega - \Omega_*\| = \epsilon$, or (iii) $\|\Omega_X - \Omega_{X*}\| = \epsilon$. We also have, because both $\kappa_{n,1}$ and $\kappa_{n,2}$ are minimized by θ_*, that

$$\kappa_n(\theta) - \kappa_n(\theta_*) \geq \kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*)$$

and

$$\kappa_n(\theta) - \kappa_n(\theta_*) \geq \kappa_{n,2}(\Omega_X) - \kappa_{n,2}(\Omega_{X*}).$$

Thus, it suffices to show that if either of (i), (ii), or (iii) hold, then at least one of the last two right hand sides are greater than $\epsilon^2\delta$ for some $\delta > 0$. We first suppose that (iii) holds and consider the second right hand side. More precisely, consider the function defined by $\mathrm{vec}(\Omega_X) \mapsto \kappa_{n,2}(\Omega_X) - \kappa_{n,2}(\Omega_{X*})$ for $\Omega_X \in \mathbb{S}^p_{++}$. Its gradient vanishes at $\mathrm{vec}(\Omega_{X*})$ and its Hessian is $H(\Omega_X) = \Omega_X^{-1} \otimes \Omega_X^{-1}$. Thus, $\kappa_{n,2}(\Omega_X) - \kappa_{n,2}(\Omega_{X*}) = \mathrm{vec}(\Omega_X - \Omega_{X*})^\mathsf{T} H(\tilde{\Omega}_X)\,\mathrm{vec}(\Omega_X - \Omega_{X*})/2$ for some $\tilde{\Omega}_X$ on the line segment connecting Ω_X and Ω_{X*}. The greatest eigenvalues of Ω_X and Ω_{X*} are both smaller than c_2, and hence that of $\tilde{\Omega}_X$ is too, and hence the smallest eigenvalue of $H(\tilde{\Omega}_X)$ is greater than c_2^{-2}. Thus,

$$\kappa_{n,2}(\Omega_X) - \kappa_{n,2}(\Omega_{X*}) \geq c_2^{-2}\|\Omega_X - \Omega_{X*}\|_F^2/2 \geq c_2^{-2}\epsilon^2/2,$$

where we have used that the spectral norm is upper bounded by the Frobenius norm.

Suppose next that (i) holds and consider $\kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*)$. Then,

$$\begin{aligned}
\kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*) &\geq \min_{\Omega \in \mathbb{S}^r_{++}} \{\kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*)\} \\
&= \log|(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*) + \Sigma_*| + \mathrm{tr}(I_r) - \log|\Sigma_*| - \mathrm{tr}(I_r) \\
&= \mathrm{tr}\left[\tilde{\Omega}^\mathsf{T}(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*)\right],
\end{aligned}$$

where $\tilde{\Omega}$ is the inverse of some matrix on the line segment between Σ_* and $(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*) + \Sigma_*$. The greatest eigenvalue of the latter is smaller than $c_2 + 4c_2^3$ on A_n, and hence the smallest eigenvalue of $\tilde{\Omega}$ is greater than $1/(c_2 + 4c_2^3)$. Thus,

$$\kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*) \geq \frac{1}{c_2 + 4c_2^3}\|\beta - \beta_*\|_F^2 \geq \frac{1}{c_2 + 4c_2^3}\epsilon^2.$$

Finally, by partially minimizing in β we get

$$\begin{aligned}
\kappa_{n,1}(\beta, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*) &\geq \kappa_{n,1}(\beta_*, \Omega) - \kappa_{n,1}(\beta_*, \Omega_*) \\
&= -\log|\Omega| + \mathrm{tr}(\Omega\Omega_*^{-1}) + \log|\Omega_*| - \mathrm{tr}(I_r),
\end{aligned}$$

which, by the arguments used to lower bound $\kappa_{n,2}$, is lower bounded by $c_2^{-2}\epsilon^2/2$. Thus, the conclusion of the lemma holds with $\delta = \min\left\{c_2^{-2}/2, 1/(c_2 + 4c_2^3)\right\} = 1/(c_2 + 4c_2^3) > 0$. □

Proof of Proposition 3.1. Suppose $k < p$ and consider (1). It takes $r(r+1)/2$ parameters to specify Σ, k to specify D, and one to specify τ. The first column in U is an unrestricted vector in the unit sphere in $\mathbb{R}^p$, and hence takes $p-1$ parameters. More generally, the jth column of U is in the unit sphere and orthogonal to the $j-1$ previous columns, and hence takes $p-j$ parameters to specify. Thus, U takes $\sum_{j=1}^{k}(p-j) = pk - k(k+1)/2$ parameters. With U fixed, γ is unrestricted and hence takes rk parameters. Thus, in total our model requires

$$d(k) = r(r+1)/2 + rk + k + 1 + pk - k(k+1)/2 = r(r+1)/2 + k[r + 1 + p - (k+1)/2] + 1.$$

If $k = p$, then Σ_X is an unrestricted covariance matrix and $\beta \in \mathbb{R}^{p\times r}$ also unrestricted. Thus, $d(k) = r(r+1)/2 + pr + p(p+1)/2$. □

B Computing details

We derive the gradient of H_n assuming its argument L is an unconstrained matrix; the gradient under the restriction that $L_{i,j} = 0$ for $j > i$ is obtained by setting the corresponding elements of the unconstrained gradient to zero. The differential of $Q_{XL} = I_n - XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}$ is

$$\begin{aligned} dQ_{XL} = &-X(dL)(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T} \\ &+ XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}[(dL)^\mathsf{T}X^\mathsf{T}XL + L^\mathsf{T}X^\mathsf{T}XdL](L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T} \\ &- XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}(dL)^\mathsf{T}X^\mathsf{T}. \end{aligned}$$

Thus, the differential of $\log|Y^\mathsf{T}Q_{XL}Y|$ is, with $S = Y^\mathsf{T}Q_{XL}Y$,

$$\begin{aligned}
d\log|Y^\mathsf{T}Q_{XL}Y| &= -\operatorname{tr}\left[S^{-1}Y^\mathsf{T}X(dL)(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}Y\right] \\
&\quad + \operatorname{tr}\left[S^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}[(dL)^\mathsf{T}X^\mathsf{T}XL + L^\mathsf{T}X^\mathsf{T}XdL](L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}Y\right] \\
&\quad - \operatorname{tr}\left[S^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}(dL)^\mathsf{T}X^\mathsf{T}Y\right] \\
&= -\operatorname{tr}\left[(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}YS^{-1}Y^\mathsf{T}XdL\right] \\
&\quad + \operatorname{tr}\left[(dL)^\mathsf{T}X^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}YS^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}\right] \\
&\quad + \operatorname{tr}\left[(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}YS^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}XdL\right] \\
&\quad - \operatorname{tr}\left[(dL)^\mathsf{T}X^\mathsf{T}YS^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}\right].
\end{aligned}$$

Thus, $\nabla\log|Y^\mathsf{T}Q_{XL}Y|$ is

$$\begin{aligned}
\nabla\log|Y^\mathsf{T}Q_{XL}Y| &= -2X^\mathsf{T}YS^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1} \\
&\quad + 2X^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}L^\mathsf{T}X^\mathsf{T}YS^{-1}Y^\mathsf{T}XL(L^\mathsf{T}X^\mathsf{T}XL)^{-1}
\end{aligned}$$

Finally,

$$\nabla\log|I_p + LL^\mathsf{T}| = 2(I_p + LL^\mathsf{T})^{-1}L,$$

and

$$\nabla\log\operatorname{tr}\left[X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}\right] = -\frac{2}{\operatorname{tr}\left[X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}\right]}(I_p + LL^\mathsf{T})^{-1}X^\mathsf{T}X(I_p + LL^\mathsf{T})^{-1}L.$$

C Latent variables

We here elaborate on the connection between out model and a latent variable model that has been used to motivate supervised probabilistic principal components (Tipping and Bishop, 1999) and PLS (Singer et al., 2016; Cook, 2018b). It can also be analyzed in terms of (predictor) envelopes (Cook, 2018b, Section 4.1.2). Let $E \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{n\times p}$, and $W \in \mathbb{R}^{n\times k}$ be latent random variables whose elements are all independent with mean zero and unit variance. Let also $\Gamma_Y \in \mathbb{R}^{k\times r}$, $\Gamma_X \in \mathbb{R}^{k\times p}$, $\tau_E > 0$, and $\tau_V > 0$. Suppose that

$$Y = W\Gamma_Y + \tau_E E \quad \text{and} \quad X = W\Gamma_X + \tau_V V. \tag{9}$$

The latent matrix W connects X to Y, and E and V are noise variables. By construction, rows of Y are independent, as are the rows of X. The following proposition gives moments of the joint distribution of (Y, X).

Proposition C.1. *Under the model in* (9)*, for* $i = 1, \ldots, n$*,*

1. $\mathrm{cov}(Y_i) = \Gamma_Y^\mathsf{T}\Gamma_Y + \tau_E^2 I_r$,
2. $\mathrm{cov}(X_i) = \Gamma_X^\mathsf{T}\Gamma_X + \tau_V^2 I_p$,
3. $\mathrm{cov}(Y_i, X_i) = \Gamma_Y^\mathsf{T}\Gamma_X$,

and, hence, if Y_i *and* X_i *are multivariate normal,*

4. $\mathbb{E}(Y_i \mid X_i) = \Gamma_Y^\mathsf{T}\Gamma_X(\Gamma_X^\mathsf{T}\Gamma_X + \tau_V^2 I_p)^{-1}X_i$*, and*
5. $\mathrm{cov}(Y_i \mid X_i) = I_r\tau_E^2 + \Gamma_Y^\mathsf{T}\left[I_k - \Gamma_X(\Gamma_X^\mathsf{T}\Gamma_X + \tau_V^2 I_p)^{-1}\Gamma_X^\mathsf{T}\right]\Gamma_Y$.

Proposition C.1 shows the column space of Γ_X is spanned by the leading k eigenvectors of $\mathrm{cov}(X_i)$, and it holds for $\beta = \mathbb{E}(X_i X_i^\mathsf{T})^{-1}\mathbb{E}(X_i^\mathsf{T} Y_i)$ that $\beta = \Gamma_X\alpha$ for some $\alpha \in \mathbb{R}^{k\times r}$. However, there are some important differences to our model. For example, the parameters in the latent variable model are not identified: the distributions of the observable variables are the same if Γ_Y and Γ_X are left-multiplied by the same orthogonal matrix. Additionally, Proposition C.1 shows the covariance matrix for Y_i is spiked when $r > k$ but ours is unstructured.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Artemiou, A. and Li, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon. *Statistica Sinica*, 19(4):1557–1565.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Bair, E., Hastie, T. J., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Barbarino, A. and Bura, E. (2017). A unified framework for dimension reduction in forecasting. *Finance and Economics Discussion Series*, 2017(004).

Bhatia, R. (2012). *Matrix Analysis*. Springer New York.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cai, T., Ma, Z., and Wu, Y. (2014). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3-4):781–815.

Cantó, R., Peláez, M. J., and Urbano, A. M. (2015). Full rank Cholesky factorization for rank deficient matrices. *Applied Mathematics Letters*, 40:17–22.

Chang, C.-W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 65(2):480.

Cook, R. D. (2018a). Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application*, 5(1):533–559.

Cook, R. D. (2018b). *An Introduction to Envelopes*. Wiley-Blackwell, Hoboken, NJ, USA.

Cook, R. D. and Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics*, 47(2):884–908.

Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(5):851–877.

Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20(3):927–960.

Donoho, D., Gavish, M., and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742–1778.

Du, C., Wei, J., Wang, S., and Jia, Z. (2018). Genomic selection using principal component regression. *Heredity*, 121(1):12–23.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Hastie, T. J., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning*. Springer New York.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17(2):97–114.

Helland, I. S. (1992). Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(2):637–647.

Hull, J. C. (2017). *Options, Futures, and Other Derivatives*. Pearson, New York, NY, 10 edition.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York.

Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316.

Lee, M. and Su, Z. (2019). *Renvlp: Computing Envelope Estimators*. R package version 2.7.

Magnus, J. R. and Neudecker, H. (2002). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley John & Sons.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2).

Næs, T. and Martens, H. (1988). Principal component regression in NIR analysis: Viewpoints, background details and selection of components. *Journal of Chemometrics*, 2(2):155–167.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Singer, M., Krivobokova, T., Munk, A., and de Groot, B. (2016). Partial least squares for dependent data. *Biometrika*, 103(2):351–362.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, 25(2):313–323.

Vershynin, R. (2018). *High-Dimensional Probability*. Cambridge University Press, Cambridge.

Wang, K. and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology*, 32(2):108–118.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374.

Watson, M. W. (2003). Macroeconomic forecasting using many predictors. *Econometric Society Monographs*, 37:87–114.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–473. ACM.