ARTICLE IN PRESS

Journal of Statistical Planning and Inference xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Principal envelope model

Jia Zhang [a], Xin Chen [b,*]

[a] School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China
[b] Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 25 February 2019
Received in revised form 2 September 2019
Accepted 2 October 2019
Available online xxxx

Keywords:
Envelope model
Grassmann manifolds
Principal components
Sufficient dimension reduction

ABSTRACT

Principal component analysis (PCA) is widely used in various fields to reduce high dimensional data sets to lower dimensions. Traditionally, the first a few principal components that capture most of the variance in the data are thought to be important. Tipping and Bishop (1999) introduced probabilistic principal component analysis (PPCA) in which they assumed an isotropic error in a latent variable model. Motivated by a general error structure and incorporating the novel idea of "envelope" proposed by Cook et al. (2010), we construct principal envelope models (PEM) which demonstrate the possibility that any subset of the principal components could retain most of the sample's information. The useful principal components can be found through maximum likelihood approaches. We also embed the PEM to a factor model setting to illustrate its reasonableness and validity. Numerical results indicate the potentials of the proposed method.

1. Introduction

Principal component analysis (PCA) is a popular data processing and dimension reduction technique. First introduced by Pearson (1901), PCA has a long history and is now widely used in various areas, including agriculture, ecology, genetics and economics. PCA seeks uncorrelated linear combinations of the original variables that capture maximal variance. Suppose we have n observations on p features $x_1, \ldots, x_p$. Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ denote the ith observation, $i = 1, \ldots, n$, and $\mathbf{x} = (x_1, \ldots, x_p)^T$ be the vector variable. Let $\tilde{\mathbf{x}}^{(i)}$ denote the centered observation vectors, $i = 1, \ldots, n$, and $\mathbb{X}$ be the $n \times p$ centered data matrix with row $\tilde{\mathbf{x}}^{(i)}$ and rank $r \leq \min(n, p)$. Since there is no response involved, this article is mainly about unsupervised multivariate dimension reduction method.

Let $\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_r$ be the eigenvectors of the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$ corresponding to its non-zero eigenvalues. Without loss of generality, $\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_r$ are ordered by descending eigenvalues. Let $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_r$ be the non-zero eigenvalues with descending order. The principal component directions $\hat{\mathbf{g}}_k$, $k = 1, \ldots, r$, can also be obtained by maximizing $\boldsymbol{\alpha}_k^T(\mathbb{X}^T\mathbb{X})\boldsymbol{\alpha}_k$ successively subject to $\boldsymbol{\alpha}_k^T\boldsymbol{\alpha}_k = 1$ and $\boldsymbol{\alpha}_h^T\boldsymbol{\alpha}_k = 0$, $\forall h < k$. This demonstrates that PCA pursues the linear combinations of the original variables such that the derived variables capture maximal variance. The sample variance of the ith principal component (PC) equals $\hat{\lambda}_i$ (Anderson, 1963). There are many methods for selecting the number of principal components, depending on specific requirements for different applications, see Jolliffe (2002).

PCA enjoys high popularity, but it is not based on a probability model. Tipping and Bishop (1999) introduced probabilistic principal component analysis (PPCA) in which the first few principal component directions can be obtained through maximum likelihood estimation. However, the assumption of an isotropic error in the PPCA model is quite limited.

* Corresponding author.
E-mail addresses: zhangjia@swufe.edu.cn (J. Zhang), chenx8@sustech.edu.cn (X. Chen).

https://doi.org/10.1016/j.jspi.2019.10.001
0378-3758/

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

By assuming a general error structure and incorporating the novel "envelope" idea of Cook et al. (2010), we establish principal envelope models that encompass PPCA as a special case and demonstrate the possibility that any subset of principal components could retain most of the sample's information. Since the introduction of "envelope" into statistical literature by Cook et al. (2010), various envelope models have been developed, including partial envelopes (Su and Cook, 2011), inner envelopes (Su and Cook, 2012) and simultaneous envelopes (Cook and Zhang, 2015a). Cook and Zhang (2015b) provides a comprehensive overview of "envelope" development.

We revisit PPCA in Section 2.1 to investigate the link between PPCA and principal envelope models. In Section 2.2 we describe the concept of an envelope and demonstrate the possibility that any subset of principal components could retain most of the sample's information. We build some intermediate models in Section 2.3. The log-likelihood function of one specific principal envelope model has the same form as probabilistic extreme components analysis (PXCA) (Welling et al., 2003) if the dimension of the envelope is the same as the minimum dimension reduction subspace. However, the concepts and statistical meanings of these two approaches are quite different. In Section 2.4, we employ the likelihood ratio test to determine the dimension of the envelop. Results of simulation studies are presented in Section 3. An extension to factor model is given in Section 4. Real data analysis is presented in Section 5. A brief discussion about the proposed methods can be found in Section 6. Technical details are given in Appendix.

2. Principal envelope model

2.1. Probabilistic principal component analysis revisited

Tipping and Bishop (1999) proposed a probabilistic principal component model as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\nu} + \sigma\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\mu}$ permits $\mathbf{x}$ to have non-zero mean and the $p \times d$ matrix $\boldsymbol{\beta}$ relates the observable variable $\mathbf{x}$ and the latent variable $\boldsymbol{\nu}$, which is assumed to be normally distributed with mean 0 and identity covariance matrix. The error $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_p)$ is assumed to be independent of $\boldsymbol{\nu}$ and d is assumed to be known. The parameter $\boldsymbol{\beta}$ is not identified since $\boldsymbol{\beta}\boldsymbol{\nu} = (\boldsymbol{\beta}\mathbf{O})(\mathbf{O}^T\boldsymbol{\nu})$ for any orthogonal matrix $\mathbf{O}$, resulting in an equivalent model. However, the subspace $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ is identified and estimable. Tipping and Bishop showed that the maximum likelihood estimator of $\mathcal{B}$ is the span of the first d eigenvectors of $\hat{\boldsymbol{\Sigma}}$. A Grassmann manifold, which is defined as the set of all d-dimensional subspaces in $\mathbb{R}^p$, is the natural parameter space for $\mathcal{B}$. For more background on Grassmann manifold optimization, see Edelman et al. (1998).

We reformulate model (1) as

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\delta}\boldsymbol{\nu} + \sigma\boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\Gamma}$ is a $p \times d$ semi-orthogonal matrix ($\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$), $\boldsymbol{\delta}$ is a full rank $d \times d$ coordinate matrix, $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$ are defined previously. Let $\mathcal{S}_{\boldsymbol{\Gamma}}$ denote the column space of $\boldsymbol{\Gamma}$. The population covariance matrix of $\mathbf{x}$ is

$$\boldsymbol{\Gamma}\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_p = \boldsymbol{\Gamma}(\boldsymbol{\delta}\boldsymbol{\delta}^T + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T = \boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$$

where $\mathbf{V} = \boldsymbol{\delta}\boldsymbol{\delta}^T + \sigma^2\mathbf{I}_d$. The full log-likelihood function $L_0(\mathcal{S}_{\boldsymbol{\Gamma}}, \mathbf{V}, \sigma^2)$ can be calculated straightforwardly:

$$\begin{aligned} &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)^{-1}\tilde{x}_i \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\mathbf{V}^{-1}) - \frac{n}{2}(p-d)\log(\sigma^2) - \frac{n}{2\sigma^2}\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0). \end{aligned}$$

In the above expression, $|A|$ denotes the determinant of the matrix A. In later sections, we will ignore the common constant $-(np/2)\log(2\pi)$ in the log-likelihood functions.

If we maximize over $\mathbf{V}$ and σ^2 separately, we arrive at the same partially maximized likelihood function as Tipping and Bishop (1999). However, the parameters $\mathbf{V}$ and σ^2 are not in proper product spaces because the eigenvalues of $\mathbf{V}$ are bounded below by σ^2. Thus it seems inappropriate to maximize over $\mathbf{V}$ and σ^2 separately. The result of Proposition 1 is the same as Tipping and Bishop (1999), but we present a totally different proof in the appendix.

Proposition 1. *The maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ in $L_0(\mathcal{S}_{\boldsymbol{\Gamma}}, \mathbf{V}, \sigma^2)$ is the subspace spanned by the first d principal component directions and can be obtained by maximizing* $\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma})$ *subject to* $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$.

2.2. Motivation: general error structure

Instead of assuming an isotropic error structure, which is very limiting, we assume

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\Phi}$ is a general positive definite matrix. The latent variable $\boldsymbol{\nu}$ represents extrinsic variation in $\mathbf{x}$, while the error $\boldsymbol{\epsilon}$ represents intrinsic variation. Traditional PCA reduces dimensionality while keeping most of its total variation. Our goal

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

is different and we reduce the dimension of $\mathbf{x}$ accounting for its extrinsic variation. Under this case, we can show that $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$ (Cook, 1998, 2007). Thus $R = \boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$ is the reduction we would like to estimate because $\mathbf{x}$ contains no further information about $\boldsymbol{\nu}$ given $\boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$. Let $\mathcal{T} = \text{span}(\boldsymbol{\Phi}^{-1}\boldsymbol{\beta})$. Since any full rank linear transformation $\mathbf{A}$ of $\mathbf{R}$ results in an equivalent reduction; $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{R}$ if and only if $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{AR}$, it is sufficient to estimate $\mathcal{T}$. Additionally, if $\mathcal{T}$ is minimal and if $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{B}^T\mathbf{x}$, then $\mathcal{T} \subseteq \text{span}(\mathbf{B})$.

Remark 1. Under Model (3) with an isotropic error structure, reducing dimensionality to keep most of $\mathbf{x}$'s total variation is equivalent to reducing dimensionality to keep most of its extrinsic variation. Here, the extrinsic variation echoes the exogenous variables in linear regression, and the intrinsic variation comes from the noise. The reason is that the isotropic error does not disrupt the order of the eigenvalues. When $\mathbf{x}$ is contaminated with non-negligible noise, it is more reasonable to reduce dimensionality with respect to extrinsic variation in comparison with total variation, which can be demonstrated to some extent by the factor model in Section 4. Intuitively, extrinsic variation represents the true sample information. Thus, dimension reduction in terms of the total variation may be misleading.

Under model (3), we see that $\mathbf{x}$ is normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\beta}\boldsymbol{\beta}^T$. The maximum likelihood estimator of $\boldsymbol{\mu}$ is simply the sample mean of $\mathbf{x}$, however $\boldsymbol{\Phi}$ and $\boldsymbol{\beta}$ are confounded, thus $\mathcal{T}$ cannot be estimated without assuming additional structure. The principal envelope idea is to estimate an upper bound on $\mathcal{T}$. By doing so, we do not lose any information on its extrinsic variation. Before we explain the concept of an envelope, we review the concept of reducing subspace.

Definition 1 (*Conway, 1990*)**.** A subspace $\mathscr{R}$ is a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathbf{M}\mathscr{R} \subseteq \mathscr{R}$ and $\mathbf{M}\mathscr{R}^{\perp} \subseteq \mathscr{R}^{\perp}$ where $\mathscr{R}^{\perp}$ stands for the complement of $\mathscr{R}$ in the usual inner product.

Definition 2 (*Cook et al., 2007, 2010*)**.** Suppose that the symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ and let the subspace $\mathcal{K} \subseteq \text{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{K}$, to be written as $\mathcal{E}_{\mathbf{M}}(\mathcal{K})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{K}$.

with a little abuse of notation, let $\boldsymbol{\Gamma}$ be an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$ and $\boldsymbol{\Gamma}_0$ be the orthogonal complement of $\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ and $u \geq d$. By definition, we have $\mathcal{T} \subseteq \mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$ and $\boldsymbol{\Gamma}^T\mathbf{x} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{x}|\boldsymbol{\nu}$. Then model (3) can be re-written as

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \\ \boldsymbol{\Phi} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta}$ is a $u \times d$ matrix with rank d, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are positive definite matrices. This model is referred as principal envelope model (PEM). Of note is that this model has been mentioned in Section 8.3 of Cook et al. (2010) but without any detail discussion. We also note that $\mathcal{S}_{\boldsymbol{\Gamma}}$ is the same as $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$. In the likelihood function, we prefer to use $\mathcal{S}_{\boldsymbol{\Gamma}}$ so that it is clear that what parameter we are going to estimate. The estimate of $\mathcal{S}_{\boldsymbol{\Gamma}}$ provides an upper bound on the estimate of $\mathcal{T}$. The parameter d is not estimable under this model. When $u = d$, $\boldsymbol{\Omega} = \sigma^2\mathbf{I}_d$ and $\boldsymbol{\Omega}_0 = \sigma^2\mathbf{I}_{p-d}$, model (4) reduces to PPCA.

The population covariance of $\mathbf{x}$ can be calculated straightforwardly: $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\Omega}+\boldsymbol{\eta}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. Let $\boldsymbol{\Psi} = \boldsymbol{\Omega}+\boldsymbol{\eta}\boldsymbol{\eta}^T$. The parameter $\boldsymbol{\eta}$ is confounded with $\boldsymbol{\Omega}$ and cannot be estimated here and in later sections, but $\boldsymbol{\Psi}$ can be estimated. After some algebra, we have the log-likelihood function

$$\begin{aligned} &-\frac{n}{2}\log|\boldsymbol{\Gamma}\boldsymbol{\Psi}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Psi}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}\tilde{x}_i \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^T)\tilde{x}_i - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\tilde{x}_i \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}) - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}). \end{aligned}$$

Maximizing over $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}_0$, we have the partially maximized log-likelihood function

$$\begin{aligned} L_1(\mathcal{S}_{\boldsymbol{\Gamma}}) &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0| - \frac{n}{2}p \\ &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}p - \frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}| \end{aligned}$$

The function $L_1(\mathcal{S}_{\boldsymbol{\Gamma}})$ requires $n > p$ as $\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}$ must not be singular.

Proposition 2.

(i) We have $L_1(\mathcal{S}_{\boldsymbol{\Gamma}}) \leq -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$ *for all* $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$.

(ii) Let $\mathcal{J}$ *be a subset with u elements of the index set* $\{1, \ldots, p\}$. *Define* $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}} = \text{span}(\mathbf{g}_{\mathcal{J}_1}, \ldots, \mathbf{g}_{\mathcal{J}_u})$ *where* $\mathbf{g}_{\mathcal{J}_1}, \ldots, \mathbf{g}_{\mathcal{J}_u}$ *denotes any u principal component directions, then* $L_1(\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}) = -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$.

From Proposition 2, we see that the span of any u principal component directions is the maximum likelihood estimator of $\mathcal{S}_{\Gamma}$. In other words, any subset with cardinality u of principal component directions is equally supported by the likelihood function. It also tells us that we need extra information to tell which subset is useful.

Remark 2. Proposition 2 seems a bit strange, but it makes sense. Note that

$$\begin{aligned}\Sigma &= \Gamma(\Omega + \eta\eta^T)\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T \\ &= \Gamma\Psi\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T \\ &= \Gamma\mathbf{G}\Lambda\mathbf{G}^T\Gamma^T + \Gamma_0\mathbf{G}_0\Lambda_0\mathbf{G}_0^T\Gamma_0^T,\end{aligned}$$

where $\mathbf{G}\Lambda\mathbf{G}^T$ and $\mathbf{G}_0\Lambda_0\mathbf{G}_0^T$ are the spectral decompositions of Ψ and Ω_0 respectively. The principal component directions under Model (4) can be written unordered as $\Gamma\mathbf{G}$ and $\Gamma_0\mathbf{G}_0$ with eigenvalues given by the corresponding elements of the diagonal Λ and Λ_0. $\mathcal{S}_{\Gamma}$ is just the column span of $\Gamma\mathbf{G}$. If the smallest eigenvalue in Λ is bigger than the largest eigenvalue in Λ_0, $\hat{\mathcal{S}}_{\Gamma}$ is the subspace spanned by the first u principal component directions. Whereas, if nothing is known about the structures of Ψ and Ω_0, we can certainly not discern which subsets of the principal component directions would represent most of the extrinsic variation, thus all the subsets yield the maximum likelihood.

2.3. Specific principal envelope models

Assuming that we can model $\Psi = \sigma^2\mathbf{I}_u$ and $\Omega_0 = \sigma_0^2\mathbf{I}_{p-u}$, we have the log-likelihood function:

$$\begin{aligned}&-\frac{n}{2}\log|\sigma^2\Gamma\Gamma^T + \sigma_0^2\Gamma_0\Gamma_0^T| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\sigma^2\Gamma\Gamma^T + \sigma_0^2\Gamma_0\Gamma_0^T)^{-1}\tilde{x}_i \\ =\ &-\frac{n}{2}u\log(\sigma^2) - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\tilde{x}_i^T(\Gamma\Gamma^T)\tilde{x}_i - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\tilde{x}_i^T(\Gamma_0\Gamma_0^T)\tilde{x}_i \\ =\ &-\frac{n}{2}u\log(\sigma^2) - \frac{n}{2\sigma^2}\mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma) - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{n}{2\sigma_0^2}\mathrm{trace}(\Gamma_0^T\hat{\Sigma}\Gamma_0).\end{aligned}$$

Maximizing over σ^2 and σ_0^2, we have the partially maximized log-likelihood function

$$\begin{aligned}L_2(\mathcal{S}_{\Gamma}) =\ &-\frac{n}{2}u\log(\mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma)) - \frac{n}{2}(p-u)\log(\mathrm{trace}(\Gamma_0^T\hat{\Sigma}\Gamma_0)) \\ &-\frac{n}{2}p + \frac{n}{2}u\log(u) + \frac{n}{2}(p-u)\log(p-u) \\ =\ &-\frac{n}{2}u\log(\mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma)) - \frac{n}{2}(p-u)\log(\mathrm{trace}(\hat{\Sigma}) - \mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma)) \\ &-\frac{n}{2}p + \frac{n}{2}u\log(u) + \frac{n}{2}(p-u)\log(p-u)\end{aligned}$$

The function $L_2(\mathcal{S}_{\Gamma})$ requires $n > p - u + 1$ to ensure $\mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma) > 0$.

Proposition 3. *When $\Psi = \sigma^2\mathbf{I}_u$ and $\Omega_0 = \sigma_0^2\mathbf{I}_{p-u}$, the maximum likelihood estimator $\hat{\mathcal{S}}_{\Gamma}$ is the span of either the first u principal component directions ($\sigma^2 > \sigma_0^2$) or the last u principal component directions ($\sigma^2 < \sigma_0^2$).*

It is quite restrictive that Ψ is modeled as isotropic (Cook, 2007), however we demonstrate a situation where the last a few principal component directions can retain most of the sample's information.

Assuming that only $\Omega_0 = \sigma_0^2\mathbf{I}_{p-u}$, we have the model

$$\begin{aligned}\mathbf{x} &= \mu + \Gamma\eta\nu + \Phi^{1/2}\epsilon \\ \Phi &= \Gamma\Omega\Gamma^T + \sigma_0^2\Gamma_0\Gamma_0^T.\end{aligned} \tag{5}$$

Then we have the log-likelihood function of model (5):

$$\begin{aligned}&-\frac{n}{2}\log|\Gamma\Psi\Gamma^T + \sigma_0^2\Gamma_0\Gamma_0^T| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\Gamma\Psi\Gamma^T + \sigma_0^2\Gamma_0\Gamma_0^T)^{-1}\tilde{x}_i \\ =\ &-\frac{n}{2}\log|\Psi| - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\Gamma\Psi^{-1}\Gamma^T)\tilde{x}_i - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\tilde{x}_i^T(\Gamma_0\Gamma_0^T)\tilde{x}_i \\ =\ &-\frac{n}{2}\log|\Psi| - \frac{n}{2}\mathrm{trace}(\Gamma^T\hat{\Sigma}\Gamma\Psi^{-1}) - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{n}{2\sigma_0^2}\mathrm{trace}(\Gamma_0^T\hat{\Sigma}\Gamma_0).\end{aligned}$$

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

Maximizing over $\boldsymbol{\Psi}$ and σ_0^2, we have the partially log-likelihood function

$$\begin{aligned} L_3(\mathcal{S}_{\boldsymbol{\Gamma}}) &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}(p-u)\log(\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0)) \\ &\quad -\frac{n}{2}p + \frac{n}{2}(p-u)\log(p-u) \\ &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}(p-u)\log(\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma})) \\ &\quad -\frac{n}{2}p + \frac{n}{2}(p-u)\log(p-u). \end{aligned}$$

The function $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$ requires $n > p$ as $\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}$ must not be singular.

Proposition 4. *Under model (5), the maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ is the span of the first k and last $u-k$ principal component directions that maximizes $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$ subject to $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$ where k needs to be determined.*

Let λ_i, $i = 1, 2, \ldots, u$ be the population eigenvalues of $\boldsymbol{\Psi}$, $\lambda_1 < \lambda_2 < \cdots < \lambda_u$. The setting of model (5) basically says that the signals can have different scales but the noises have the same magnitude. If $\lambda_i > \sigma_0^2$ for all $i = 1, 2, \ldots, u$, then we have the same solution as the usual principal component analysis. It is equivalent to say that if the signal is strong enough, the usual principal component analysis is doing a sensible thing. If $\sigma_0^2 > \lambda_i$ for all $i = 1, 2, \ldots, u$, then we have the last u principal component directions as the solution. If σ_0^2 lies among λ_i for $i = 1, 2, \ldots, u$, then the solution is the span of the first k principal component directions and $u-k$ last principal component directions where k ranges from 1 to u. This provides a fast algorithm to search the maximizer of $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$ which was also addressed by Welling et al. (2003). Welling et al. (2003) proposed a probabilistic model for "extreme components analysis" (PXCA) which extracts an optimal combination of principal and minor components at the maximum likelihood solution. Instead of adding isotropic Gaussian noise in all directions in the probabilistic PCA model, PXCA adds the noise only in the directions orthogonal to the column space of $\boldsymbol{\beta}$:

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon} \\ \boldsymbol{\Phi} &= \mathbf{I}_p - \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T, \end{aligned}$$

where $\boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_0^2\mathbf{I}_p)$. Notice that the covariance matrix of $\mathbf{x}$ in this model is

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma_0^2\{\mathbf{I}_p - \boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T\},$$

which is equivalent to the covariance matrix in Model (5):

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\eta}\boldsymbol{\eta}^T + \boldsymbol{\Omega})\boldsymbol{\Gamma}^T + \sigma_0^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T,$$

when $u = d$, where $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ is the orthogonal complement of $\boldsymbol{\Gamma}$.

Therefore, the log-likelihood function of Model (5) has the same form as that of PXCA when $u = d$, which means that Model (5) reduces to the PXCA model (Model (6) or (8) in Welling et al. (2003)) when $u = d$. The main difference between these two approaches lies in the fact that model (5) aims to estimate an upper bound on the minimum dimension reduction subspaces.

If we assume $\boldsymbol{\Omega}_0$ in model (4) to be a diagonal matrix, it is equivalent to model (4) because $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}\boldsymbol{\Gamma}_0^T$ can always be re-parameterized as $\boldsymbol{\Gamma}_0'\Lambda\boldsymbol{\Gamma}_0'^T$. The positive definite matrix $\boldsymbol{\Omega}_0$ in model (4) can be considered as the covariance matrix for $\boldsymbol{\Gamma}_0^T\mathbf{x}$ and it may not be always a diagonal matrix given $\boldsymbol{\Gamma}_0$. Suppose we can model $\boldsymbol{\Omega}_0$ as

$$\sigma_0^2\begin{pmatrix} 1 & c & c & \ldots & c \\ c & 1 & c & \ldots & c \\ c & c & 1 & \ldots & c \\ \vdots & & & \ddots & \vdots \\ c & c & c & \ldots & 1 \end{pmatrix}. \tag{6}$$

This means that the correlation coefficients for $\boldsymbol{\Gamma}_0^T\mathbf{x}$ are modeled as constant c where $-1/(p-u-1) < c < 1$. We can represent $\boldsymbol{\Omega}_0$ as $\sigma_0^2\{(1-c)\mathbf{Q}_1 + (1+(p-u-1)c)\mathbf{P}_1\}$ where $\mathbf{P}_1$ is the projection matrix onto the $(p-u)\times 1$ vector of ones and $\mathbf{Q}_1 = \mathbf{I}_{p-u} - \mathbf{P}_1$. The log-likelihood function can be calculated as

$$\begin{aligned} &-\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}) - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}) \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}) - \frac{n}{2}(p-u)\log\sigma_0^2 - \frac{n}{2}(p-u-1)\log(1-c) \\ &\quad -\frac{n}{2}\log(1+(p-u-1)c) - \frac{n}{2\sigma_0^2}\text{trace}\{\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\left(\frac{\mathbf{Q}_1}{1-c} + \frac{\mathbf{P}_1}{1+(p-u-1)c}\right)\}. \end{aligned}$$

Table 1
PEM models.

Model setting	Structure specification under model (4)	$\hat{\mathcal{S}}_{\Gamma}$ (Span of)
Model (3)	Φ is a general positive definite matrix	Cannot be estimated
Model (4)	None	Any u principal component directions
None	$\Omega_0 = (6)$	A subset of the first k' and last $u+1-k'$ principal component directions
Model (5)	$\Omega_0 = \sigma_0^2 \mathbf{I}_{p-u}$	The first k and last $u-k$ principal component directions
	$\Psi = \sigma^2 \mathbf{I}_u$ and $\Omega_0 = \sigma_0^2 \mathbf{I}_{p-u}$	The first or the last u principal component directions
Model (1) or (2) (PPCA)	$\Omega = \sigma^2 \mathbf{I}_d$ and $\Omega_0 = \sigma_0^2 \mathbf{I}_{p-d}$	The first d principal component directions

Maximizing over $\boldsymbol{\Psi}$ and σ_0^2 first, we have

$$-\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}(p-u-1)\log(1-c) - \frac{n}{2}\log(1+(p-u-1)c)$$
$$-\frac{n}{2}\log\{\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\left(\frac{\mathbf{Q}_1}{1-c}+\frac{\mathbf{P}_1}{1+(p-u-1)c}\right))\} - \frac{n}{2}p + \frac{n(p-u)}{2}\log(p-u).$$

Then maximizing above over c, we have the partially maximized log-likelihood

$$L_5(\mathcal{S}_{\Gamma}) = -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n(p-u-1)}{2}\log\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{Q}_1)$$
$$-\frac{n}{2}\log\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{P}_1) - \frac{n}{2}p + \frac{n(p-u-1)}{2}\log(p-u-1),$$

where the maximum likelihood estimators $\hat{\sigma}_0^2 = \mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0)/(p-u)$ and

$$\hat{c} = 1 - \frac{(p-u)\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{Q}_1)}{(p-u-1)\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0)}.$$

Proposition 5. *When $\boldsymbol{\Psi} \succ 0$ and $\boldsymbol{\Omega}_0 = (6)$, the maximum likelihood estimator $\hat{\mathcal{S}}_{\Gamma}$ is the span of one subset of u principal component directions that maximizes $L_5(\mathcal{S}_{\Gamma})$ subject to $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$.*

In this setting, the coordinate matrix $\boldsymbol{\Omega}_0$ has two different eigenvalues, one with $p-u-1$ replicates. From Proposition 5, we have a simple algorithm to find the maximum likelihood estimator $\hat{\mathcal{S}}_{\Gamma}$. Among the first k and last $u+1-k$ principal component directions where k ranges from 0 to $u+1$, we record any subset with dimension u as a possible candidate. The total number of candidates is less than $(u+2)(u+1)$. Then we search among all candidates and find the one that maximizes L_5.

We summarize all the results of the above models in Table 1.

2.4. Selection of the dimension u

We use the likelihood ratio test to determine the dimensionality of the envelope denoted by u. For example, let us consider model (5). The hypothesis $u = u_0$ can be tested by using the likelihood ratio statistic $\Lambda(u_0) = 2(\hat{L}_{fm} - \hat{L}^{(u_0)})$, where $\hat{L}_{fm}$ denotes the maximum value of the log likelihood for the full model ($u = p$), and $\hat{L}^{(u_0)}$ the maximum value of the log likelihood when $u = u_0$. In fact, $\hat{L}_{fm} = -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$. The total number of parameters needed to estimate model (5) is

$$df(u) = p + \frac{u(u+1)}{2} + u(p-u) + 1.$$

The first term on the right hand side corresponds to the estimation of the grand mean $\boldsymbol{\mu}$. The second term corresponds to the estimation of the unconstrained symmetric matrix $\boldsymbol{\Psi}$. The third term corresponds the number of parameters needed to describe the subspace $\mathcal{S}_{\Gamma}$ (Edelman et al., 1998). The last term corresponds to σ_0^2. Following standard likelihood theory, under the null hypothesis, $\Lambda(u_0)$ is distributed asymptotically as a chi-squared random variable with $(p-u_0+2)(p-u_0-1)/2$ degrees of freedom.

3. Simulation studies

Two small simulation studies are conducted in this section. We generate data from model (5). $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ are generated randomly by R function "randortho" in Package "pracma" with dimensions $p \times u$ and $p \times (p-u)$ respectively. $\boldsymbol{\eta}$ is a $u \times d$ matrix with stand normal distributed elements and rank d. $\boldsymbol{\Omega} = \mathbf{I}_u$, $\boldsymbol{\mu} = (0, \ldots, 0)^T$ and $\sigma_0 = 2$. In this setting, $\boldsymbol{\Psi} = \mathbf{I}_u + \boldsymbol{\eta}\boldsymbol{\eta}^T$. We set dimensions (p, u, d) to be (20, 2, 1), (50, 5, 3) and (80, 8, 5).

There is no maximum likelihood estimator for $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, however we can estimate $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$, an upper bound of $\mathcal{T}$. By Proposition 4, the population maximum likelihood estimator, denoted as $\mathcal{S}_{\Gamma}$, is the span of the first k and the last $u-k$

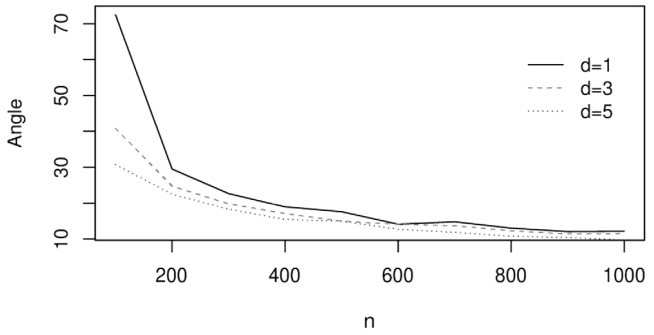

Fig. 1. Maximum angles versus n.

eigenvectors of $\boldsymbol{\Sigma}$ where k is need to be determined. It is clear to see that the usual principal component analysis fails. On each of 100 replications with fixed n we compute the maximum angle between $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ and span($\boldsymbol{\beta}$). Fig. 1 summarizes the maximum angles for $n = (100, 1000; 100)$. It can be seen clearly that the span of the principal envelope solution is very efficient at estimating an upper bound of span($\boldsymbol{\beta}$). Moreover, the maximum angle tends to become smaller when d gets larger.

In the second simulation study, we generate data from

$$\mathbf{x} = \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \quad \boldsymbol{\Phi} = 0.1\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + 10\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T,$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & -1 & \dots & 1 & -1 \end{pmatrix}^T / \sqrt{20} \in \mathbb{R}^{20\times 2}$$

and $\boldsymbol{\nu}$ is generated as points on a 2×2 square instead of a normal distribution to visualize the effectiveness of principal envelope model (5). We have $p = 20$, $u = 2$ and $d = 2$. Fig. 2 shows the scatter plot of $\mathbf{PE}_2 = \mathbb{X}\hat{\mathbf{e}}_2$ versus $\mathbf{PE}_1 = \mathbb{X}\hat{\mathbf{e}}_1$, and the scatter plot of $\mathbf{PC}_2 = \mathbb{X}\hat{\mathbf{g}}_2$ versus $\mathbf{PC}_1 = \mathbb{X}\hat{\mathbf{g}}_1$ for two different sample size $n = 200$ and $n = 400$ with one replication where $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ represent the two directions of $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$. From Fig. 2, we can see that the solution of principal envelope model (5) can recover the square well while the solution of principal component analysis fails. We notice that in this setting $\boldsymbol{\nu}$ does not follow a normal distribution which is required by model (5), however the solution is still reasonable. This tells us that principal envelope model can be robust.

A friend suggested that we should exhibit some results for higher dimensions. We extend 2-dimensional square to 3-dimensional sphere, where the latent vector is generated, to illustrate the reasonableness of PEM. In the 3-dimensional setting, we set $u = d = 3$, $p = 20$ and $n = 1000$. Data is generated from

$$\mathbf{x} = \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \quad \boldsymbol{\Phi} = 0.01\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + 10\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T,$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & -1 & \dots & 1 & -1 \\ 1 & 1 & \dots & -1 & -1 \end{pmatrix}^T / \sqrt{p} \in \mathbb{R}^{p\times 3},$$

and $\boldsymbol{\nu}$ is uniformly distributed on a sphere. Fig. 3 shows the scatter plot of $\mathbf{PE}_1$, $\mathbf{PE}_2$ and $\mathbf{PE}_3$ and the scatter plot of $\mathbf{PC}_1$, $\mathbf{PC}_2$ and $\mathbf{PC}_3$. It can be seen clearly that the solution of principal envelope model (5) can recover the sphere well while the solution of principal component analysis fails. In this setting, the distribution of the latent vector ν is far away from the normal distribution, and that is why we reduce the coefficient 0.1 to 0.01 when we formulate Φ.

4. An extension to factor model

Consider a factor model with a general error structure base on Model (4), called envelope factor model (EFM):

$$\begin{aligned} y &= c + \boldsymbol{\alpha}^T\boldsymbol{\nu} + \zeta, \\ \mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \\ \boldsymbol{\Phi} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \end{aligned} \tag{7}$$

where y is the target variable, c is some constant, $\boldsymbol{\alpha}$ is a $d \times 1$ vector, ζ is the stochastic error independent of $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$, and the other notations stay the same as those in Model (4). Here, $\boldsymbol{\nu}$ can be seen as common factors driving both the response and the predictors $\mathbf{x}$ and $\boldsymbol{\Gamma}\boldsymbol{\eta}$ as factor loadings. Obviously, it is the extrinsic variation with respect to $\boldsymbol{\nu}$ that is

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

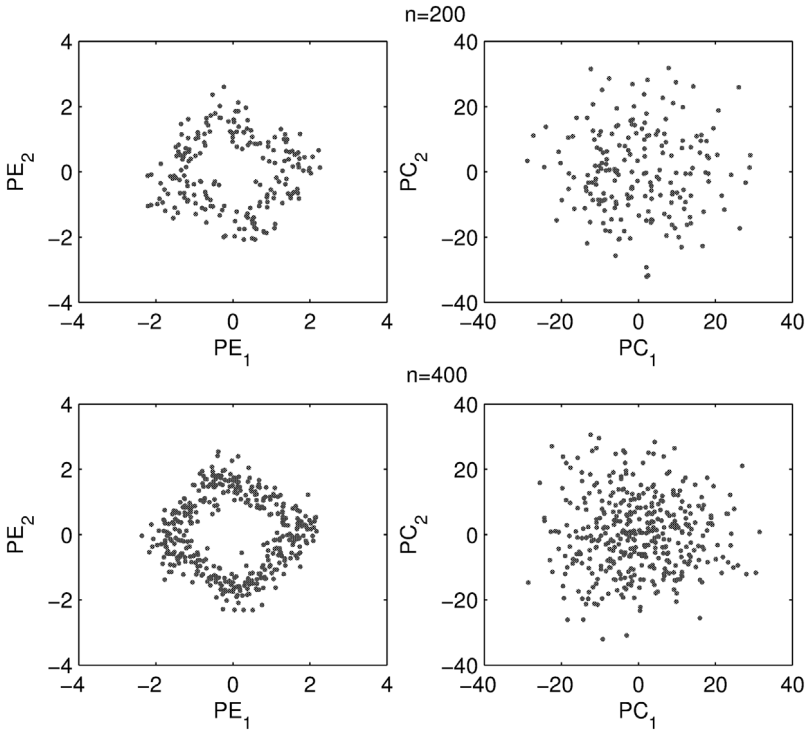

Fig. 2. Recover the square using PEM and PCA methods.

directly related to the target variable. Hence, the PEM model can be a more effective technique to extract the latent factors from the predictors. When an isotropic error structure is imposed on Model (4), EFM model is reduced to the traditional factor models where the factor loadings are the eigenvectors corresponding to the largest d eigenvalues of $\boldsymbol{\Sigma}$, see among others (Stock and Watson, 2002), Bai and Ng (2002), Fan et al. (2013) and Bai and Ng (2013) for reference. In the EFM model setting, we give a simple simulation to demonstrate the effectiveness of the PEM model framework.

Consider Model (4) with $\boldsymbol{\Psi} = \sigma^2\mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = \sigma_0^2\mathbf{I}_{p-u}$ aforementioned. Let $\sigma = 1$ and $\sigma_0 = 2$ such that the maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ is the span of the last u principal component directions. $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ are generated randomly by R function "randortho" in Package "pracma" with dimensions $p \times u$ and $p \times (p-u)$ respectively. $\boldsymbol{\eta}$ is a $u \times d$ matrix with stand normal distributed elements and rank d and is generated to ensure that $\boldsymbol{\Omega}$ is a positive definite matrix. Let $n = 400$, $p = 50$, $u = 5$, $d = 3$, and $\boldsymbol{\mu} = (0, \ldots, 0)^T$. Under model (7), let $c = 0$, $\boldsymbol{\alpha} = (1, -1, 1)^T$ and ζ follow the stand normal distribution. We employ the first u (PPCA) and the last u (PEM) principal component directions to formulate the latent factors respectively which are then used to predict the target variable y. The adjusted R^2 of y being regressed on the real factors $\boldsymbol{\nu}$ is about 0.93. Under PEM model, the adjusted R^2 reduces to 0.65. By contrast, using factors generated by the first u principal component directions causes the adjusted R^2 to drop sharply to -0.01. Clearly, the first u principal component directions do not contain any information with respect to the target variable y. PEM framework makes sense and performs well.

5. Data analysis

We first applied our method to a data set about agricultural economics studies. This data contains 17 cases, one response and 8 explanatory variables. The response variable y is retail food price index adjusted by the CPI. Principal envelope models do not utilize any information on the response and the response is only used to verify that PEM solutions can be efficient. The 8 explanatory variables are price of beef (cents/lb) (x_1); consumption of beef per capita (lbs) (x_2); price of pork (cents/lb) (x_3); consumption of pork per capita (lbs) (x_4); retail food price index (x_5); disposable income per capita index (x_6); food consumption per capita index (x_7) and index of real disposable income per capital ((x_8). According to the multivariate normality test (Royston, 1982), this data seems to follow a multivariate normal distribution which is the premise of principal envelope models. The data can be downloaded from the web page http://lib.stat.cmu.edu/DASL/Datafiles/agecondat.html.

Let $\mathbf{m}$ be the regression coefficient vector of y on $x_1, \ldots, x_8$. We use Model (5) to fit the data and the likelihood ratio tests suggest $u = 4$ using significance level $\alpha = 0.01$ and $u = 5$ using $\alpha = 0.05$. When $u = 4$, the corresponding principal envelope solution consists the last 4 eigenvectors of the covariance matrix and the largest principal angle

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

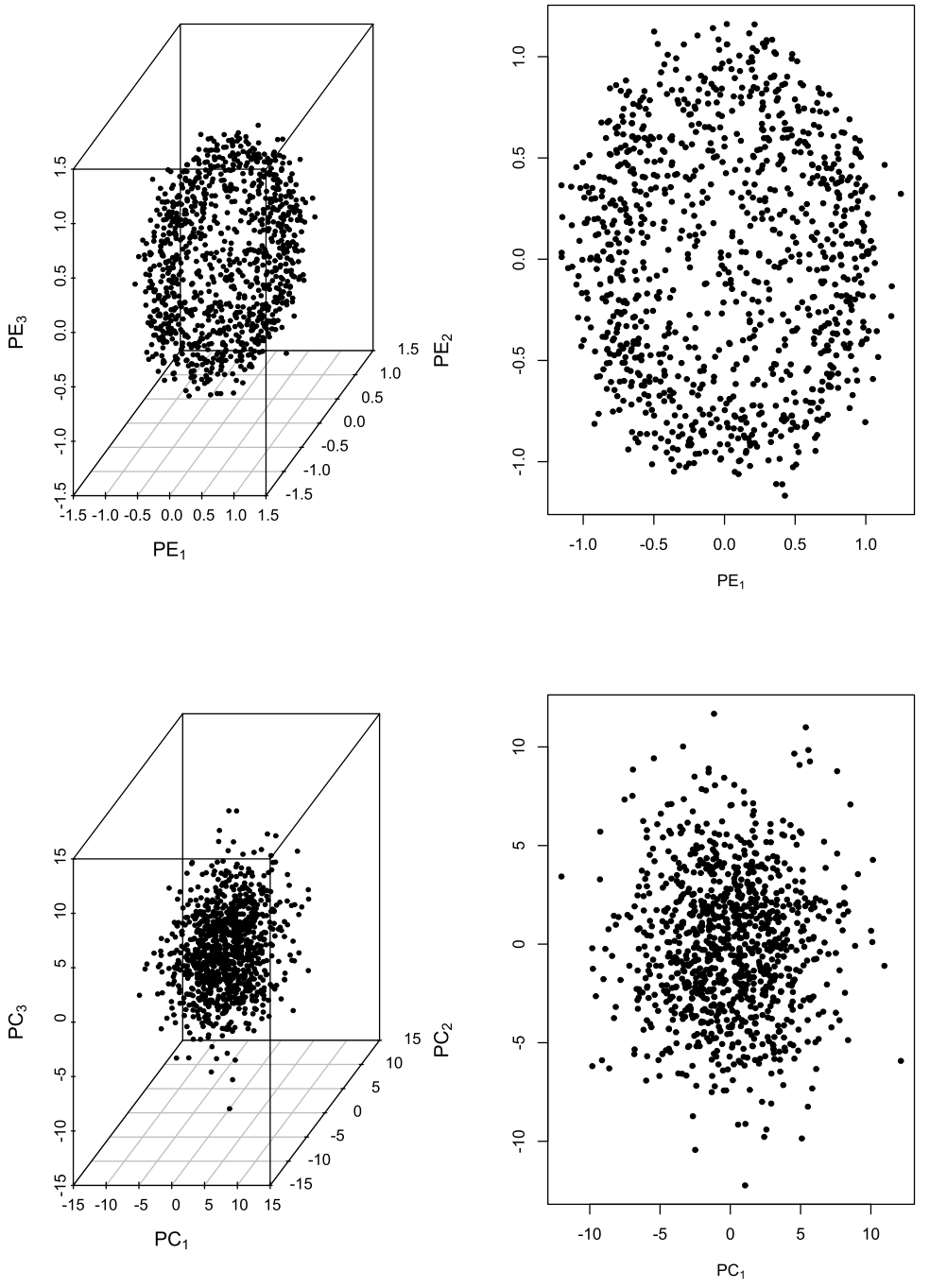

Fig. 3. Recover the sphere using PEM and PCA methods.

(Knyazev and Argentati, 2002) between **m** and the subspace spanned by the envelope solution is 7.359 degrees. When $u = 5$, the corresponding principal envelope solution consists the last 5 eigenvectors and the largest principal angle between **m** and the subspace spanned by the envelope solution is 7.356 degrees. The first 5 principal component explain 99.9% total variance while the largest principal angle between **m** and the subspace spanned by the first 5 principal component directions is about 74.47 degrees.

Then we applied our method to the man hours data, a good example of multicollinearity data structure. This data contains 25 observations, one response y and 7 explanatory variables $x_1, \ldots, x_7$. The response variable y is monthly man hours needed to operate an establishment of U.S. Navy bachelor officers' quarters. The 7 explanatory variables are average

daily occupancy (x_1); monthly average number of check-ins (x_2); weekly hours of service desk operation (x_3); common use area (in square feet) (x_4); number of building wings (x_5); operational berthing capacity (x_6); number of rooms (x_7). The correlation coefficients between x_6 and x_7, x_2 and x_7, and x_2 and x_6 are 0.98, 0.86, 0.85 respectively. The data can be downloaded from the web page http://www.stat.umn.edu/~xchen/manhours.txt.

According to the Royston normality test (with almost zero p-value), this data does not seem to follow a multivariate normal distribution. This means that model (5) might not be valid for this data set anymore. Usually if the data set follows a normal distribution, we do not standardize the data set as the standardization will destroy the marginal normality. However in this non-normality case, we standardize $x^{(1)}, \ldots, x^{(7)}$ to $z^{(1)}, \ldots, z^{(7)}$ with mean 0 and standard deviation 1 and apply model (5) to fit the data anyway.

The likelihood ratio tests suggest $u = 2$ in both $\alpha = 0.01$ and $\alpha = 0.05$. The corresponding principal envelope solution consists the first and the seventh eigenvector. Let $\mathbf{m}$ be the regression coefficient vector of y on $z^{(1)}, \ldots, z^{(7)}$. The principal envelope solution, the first and the seventh principal components explain 67% total variance while the largest principal angle between $\mathbf{m}$ and the subspace spanned by principal envelope solution is about 12 degrees. The first 5 principal component explain 98% total variance while the largest principal angle between $\mathbf{m}$ and the subspace spanned by the first 5 principal component directions is about 75 degrees.

Both data sets demonstrate that the principal envelope model can be much more efficient than PCA in some situations.

6. Discussion

We have seen that if the error structure deviates from the isotropic error in the model (1), the usual PCA may not work anymore. Motivated by a general error structure, we establish probabilistic models named as principal envelope models that show any combination of principal component directions could contain most of the sample's information. Under more specific principal envelope models, we are able to discern which combination is useful via maximum likelihood estimators. Hence we provide an alternative to PCA in multivariate analysis when PCA fails. We also studied several different structure of $\boldsymbol{\Omega}_0$ defined in model (4).

Acknowledgments

The authors would like to thank R. Dennis Cook for his helps that lead to substantial improvement for the paper. The authors are also grateful to the Editor, an Associate Editor and two anonymous referees for constructive comments. This research was supported in part by the Fundamental Research Funds for the Central Universities, China (Grant No. JBK171121, JBK170161, JBK150501) and the Joint Lab of Data Science and Business Intelligence at SWUFE, China. This research was also supported by SUSTech start-up fund, China (Y01286224).

Appendix

A few lemmas

Lemma 1. *Let $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d]$ where $\boldsymbol{\gamma}_i$ stands for the ith column of $\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\Gamma}} = [\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_d]$ where $\hat{\mathbf{g}}_i$ denotes the ith principal component direction. Then $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ maximizes the objective function* $\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$ *subject to* $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$.

Proof. Let $\boldsymbol{\gamma}_i = \sum_{j=1}^{p} c_{ij} \hat{\mathbf{g}}_j$. Then

$$\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}) = \sum_{i=1}^{d} \boldsymbol{\gamma}_i^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\gamma}_i = \sum_{i=1}^{d} \sum_{j=1}^{p} \lambda_j c_{ij}^2 = \sum_{j=1}^{p} \lambda_j \left(\sum_{i=1}^{d} c_{ij}^2 \right).$$

where $\hat{\lambda}_j$ denotes the jth eigenvalue.

Since $\sum_{i=1}^{d} c_{ij}^2 \le 1$ and $\sum_{j=1}^{p} \sum_{i=1}^{d} c_{ij}^2 = d$, to maximize $\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$, the optimum situation arrives when $\sum_{i=1}^{d} c_{i1}^2 = \sum_{i=1}^{d} c_{i2}^2 = \cdots = \sum_{i=1}^{d} c_{id}^2 = 1$ and $\sum_{i=1}^{d} c_{i(d+1)}^2 = \cdots = \sum_{i=1}^{d} c_{ip}^2 = 0$. It is clear to see that $\hat{\boldsymbol{\gamma}}_i = \hat{\mathbf{g}}_i$ for $i = 1, \ldots, d$ satisfying the optimum condition. The global maximum value of $\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$ equals $\sum_{i=1}^{d} \hat{\lambda}_i$.

Lemma 2. *Let $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d]$ where $\boldsymbol{\gamma}_i$ stands for the ith column of $\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\Gamma}} = [\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_d]$ where $\hat{\mathbf{g}}_i$ denotes the ith principal component direction. Then $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ maximizes the objective function*

$$\text{trace}\left(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} \text{diag}(k_1, k_2, \ldots, k_d) \right)$$

subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$ where $k_1 \ge k_2 \ge \cdots \ge k_d$ are positive real numbers.

Proof. Let $\boldsymbol{\Gamma}_i = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{d-i}]$ for $i = 1, \ldots, d-1$. It is clear to see that

$$\begin{aligned}&\text{trace}\left(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} \text{diag}(k_1, k_2, \ldots, k_d)\right)\\ &= k_d \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}) + (k_{d-1} - k_d)\text{trace}(\boldsymbol{\Gamma}_1^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}_1) + \cdots + (k_1 - k_2)\text{trace}(\boldsymbol{\Gamma}_{d-1}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}_{d-1}).\end{aligned}$$

From Lemma 1, we know $\hat{\boldsymbol{\Gamma}}$ maximizes every term on the right side of the formula above. The global maximum value of

$$\text{trace}\left(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} \text{diag}(k_1, k_2, \ldots, k_d)\right)$$

equals $\sum_{i=1}^{d} \lambda_i k_i$. If $k_1, k_2, \ldots, k_d$ are not in descending order, then we need permutate columns of $\hat{\boldsymbol{\Gamma}}$ such that it corresponds the order of k_i. However the subspace spanned by $\hat{\boldsymbol{\Gamma}}$ does not change.

Lemma 3. *Let $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d]$ where $\boldsymbol{\gamma}_i$ stands for the ith column of $\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\Gamma}} = [\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_d]$ where $\hat{\mathbf{g}}_i$ denotes the ith principal component direction. Then $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ maximizes the objective function* $\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} \boldsymbol{\Phi})$ *subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$ where $\boldsymbol{\Phi}$ is a $d \times d$ positive definite matrix.*

Proof. Let $A\text{diag}(m_1, \ldots, m_d)A^T$ be the spectral decomposition of $\boldsymbol{\Phi}$ where $m_1 \geq m_2 \geq \cdots \geq m_d$ are positive real numbers. Then

$$\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} \boldsymbol{\Phi}) = \text{trace}\left(A^T \boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} A \text{diag}(m_1, \ldots, m_d)\right).$$

Since $(\boldsymbol{\Gamma}A)^T \boldsymbol{\Gamma}A = \mathbf{I}_d$, from Lemma 2, $\text{span}(\hat{\boldsymbol{\Gamma}})$ is the maximum likelihood estimator of $\text{span}(\boldsymbol{\Gamma}A)$ in the objective function above while $\text{span}(\boldsymbol{\Gamma}A) = \text{span}(\boldsymbol{\Gamma})$.

Lemma 4. *Let a real function $f(x) = \log(x) + C\log(K - x)$ defined on the interval $[a, b]$, $0 < a < K/(1 + C) < b < K$, then $f(x)$ reaches its maximum at $K/(1 + C)$ and reaches its minimum at either a or b.*

It is easy to calculate the first derivative of $f(x)$

$$f'(x) = \frac{K - (1 + C)x}{x(K - x)},$$

and the second derivative

$$f''(x) = -\frac{1}{x^2} - \frac{C}{(K - x)^2} < 0.$$

We see that $f(x)$ is concave with the only stationary point $K/(1+C)$. So we can conclude that $f(x)$ reaches its maximum at $K/(1 + C)$ and reaches its minimum at the boundary point, either a or b.

Proof of Proposition 1

We can rewrite $L_0(\mathcal{S}_{\boldsymbol{\Gamma}}, \mathbf{V}, \sigma^2)$ as

$$-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}(p - d)\log(\sigma^2) - \frac{n}{2\sigma^2}\text{trace}(\hat{\boldsymbol{\Sigma}}) + \frac{n}{2}\text{trace}[\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}\{(1/\sigma^2)\mathbf{I}_d - \mathbf{V}^{-1}\}].$$

It is clear to see that

$$\frac{1}{\sigma^2}\mathbf{I}_d - \mathbf{V}^{-1}$$

is a positive definite matrix by the definition of $\mathbf{V}$, from Lemma 3, we can conclude that the subspace spanned by the first d principal component directions is the maximum likelihood estimator of $\text{span}(\boldsymbol{\Gamma})$ in this case.

Proof of Proposition 2

Since

$$\log|\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}_0| \geq \log|\hat{\boldsymbol{\Sigma}}|,$$

we have $L_1(\boldsymbol{\Gamma}) \leq -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$ for all $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_u$. Also

$$\begin{aligned}&\log|\hat{\boldsymbol{\Gamma}}_{\mathcal{J}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Gamma}}_{\mathcal{J}}| + \log|\hat{\boldsymbol{\Gamma}}_{\mathcal{J}_0}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Gamma}}_{\mathcal{J}_0}|\\ &= \log|\hat{\boldsymbol{\Gamma}}_{\mathcal{J}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Gamma}}_{\mathcal{J}}| + \log|\hat{\boldsymbol{\Gamma}}_{\mathcal{J}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Gamma}}_{\mathcal{J}}| + \log|\hat{\boldsymbol{\Sigma}}|\end{aligned}$$

$$= \log(\hat{\lambda}_{\mathcal{J}_1} \ldots \hat{\lambda}_{\mathcal{J}_u}) + \log(\hat{\lambda}_{\mathcal{J}_1}^{-1} \ldots \hat{\lambda}_{\mathcal{J}_u}^{-1}) + \log|\hat{\boldsymbol{\Sigma}}|$$
$$= \log|\hat{\boldsymbol{\Sigma}}|.$$

Then we know $L_1(\hat{\boldsymbol{\Gamma}}_{\mathcal{J}}) = -(np)/2 - n/2 \log|\hat{\boldsymbol{\Sigma}}|$.

Proof of Proposition 3

Maximizing $L_2(\mathcal{S}_{\boldsymbol{\Gamma}})$ is equivalent to minimizing

$$\log(\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})) + \frac{p-u}{u} \log(\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})).$$

Let $x = \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$, $C = (p-u)/u$ and $K = \text{trace}(\hat{\boldsymbol{\Sigma}})$. With probability one, we have

$$\min\{\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})\} < \frac{K}{1+C} = \frac{u}{p}\text{trace}(\hat{\boldsymbol{\Sigma}}) < \max\{\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})\}$$

subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$.

Following Lemma 4, we know that the maximum value of $L_2(\boldsymbol{\Gamma})$ is reached at either $\max\{\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})\}$ or $\min\{\text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})\}$. That is to say, the maximum likelihood estimator of $\boldsymbol{\Gamma}$ is either the first u principal component directions or the last u principal component directions.

Proof of Proposition 4

Maximizing $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$ is equivalent to minimize

$$\log|\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}| + (p-u)\log(\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})) \tag{8}$$

subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_u$. Using the Lagrange multiplier rule, the solution $\hat{\boldsymbol{\Gamma}}$ can be gotten by finding the stationary points of the following unconstrained function:

$$\log|\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}| + (p-u)\log(\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})) + \text{trace}(\mathbf{U}(\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} - \mathbf{I}_u))$$

where $\mathbf{U}$ is a $u \times u$ matrix of the Lagrange multipliers. Taking derivatives with respect to $\boldsymbol{\Gamma}$ and $\mathbf{U}$, we have the condition that the stationary points must satisfy

$$2\hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})^{-1} - \frac{2(p-u)\hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}}{\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})} + \boldsymbol{\Gamma}(\mathbf{U} + \mathbf{U}^T) = 0$$

subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} - \mathbf{I}_u = 0$. Let $w = (\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}))/(p-u)$. Then $\mathbf{U} + \mathbf{U}^T = (2/w)\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma} - 2\mathbf{I}_u$. Substituting $\mathbf{U} + \mathbf{U}^T$ into the condition above, we have

$$\hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}\{(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})^{-1} - \frac{1}{w}\mathbf{I}_u\} = \boldsymbol{\Gamma}\{\mathbf{I}_u - \frac{1}{w}(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})\} \tag{9}$$

subject to $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} - \mathbf{I}_u = 0$.

If w is not equal to any eigenvalue of the $u \times u$ matrix $\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}$, the matrices $(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})^{-1} - (1/w)\mathbf{I}_u$ and $\mathbf{I}_u - (1/w)(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$ are of full rank. Then $\text{span}(\hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$ must equal $\text{span}(\boldsymbol{\Gamma})$, implying that $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ has to be the span of one subset of u principal component directions.

If w equals an eigenvalue of $\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}$, we will show that $\hat{\boldsymbol{\Gamma}}$ cannot be the maximizer of $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$. Then the matrices $(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})^{-1} - (1/w)\mathbf{I}_u$ and $\mathbf{I}_u - (1/w)(\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma})$ are singular with rank $u-1$. Let the spectral decomposition of $\boldsymbol{\Gamma}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}$ be $\boldsymbol{\theta}\text{diag}(\kappa_1, \ldots, \kappa_u)\boldsymbol{\theta}^T$ where $\boldsymbol{\theta}$ is a $u \times u$ orthogonal matrix. With probability one, $\kappa_1, \ldots, \kappa_u$ are positive and distinct. Let $\boldsymbol{\Gamma}\boldsymbol{\theta} = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_u)$. Then

$$(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_u)^T \hat{\boldsymbol{\Sigma}} (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_u) = \text{diag}(\kappa_1, \ldots, \kappa_u).$$

Without loss of generality, assume $w = \kappa_u$ as κ_i are not ordered here.

The condition (9) is equivalent to

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_u)\text{diag}(\frac{1}{\kappa_1} - \frac{1}{\kappa_u}, \ldots, \frac{1}{\kappa_{u-1}} - \frac{1}{\kappa_u}, 0) = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_u)\text{diag}(1 - \frac{\kappa_1}{\kappa_u}, \ldots, 1 - \frac{\kappa_{u-1}}{\kappa_u}, 0),$$

subject to $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} - \mathbf{I}_u = 0$. We have $\hat{\boldsymbol{\Sigma}}\boldsymbol{\tau}_i = \kappa_i\boldsymbol{\tau}_i$ for $i = 1, \ldots, u-1$. In another word, $\boldsymbol{\tau}_i$ are eigenvectors of $\hat{\boldsymbol{\Sigma}}$ for $i = 1, \ldots, u-1$. The formula (8) equals

$$\sum_1^u \log(\kappa_i) + (p-u)\log\{(\text{trace}(\hat{\boldsymbol{\Sigma}}) - (\kappa_1 + \cdots + \kappa_u))\} \tag{10}$$

$$= \sum_1^{u-1} \log(\kappa_i) + \log(\kappa_u) + (p-u)\log\{(\text{trace}(\hat{\boldsymbol{\Sigma}}) - (\kappa_1 + \cdots + \kappa_{u-1})) - \kappa_u\}. \tag{11}$$

Since

$$\begin{aligned} w &= \kappa_u = (\text{trace}(\hat{\boldsymbol{\Sigma}}) - \text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}))/(p-u) \\ &= \text{trace}(\hat{\boldsymbol{\Sigma}}) - (\kappa_1 + \cdots + \kappa_u)/(p-u), \end{aligned}$$

we have $\kappa_u = (\text{trace}(\hat{\boldsymbol{\Sigma}}) - (\kappa_1 + \cdots + \kappa_{u-1}))/(p-u+1)$.

Fixing $\kappa_1, \ldots, \kappa_{u-1}$, by Lemma 4, κ_u reaches the maximum value for (11). Replacing κ_u with any other eigenvalues of $\hat{\boldsymbol{\Sigma}}$ that is different to $\kappa_1, \ldots, \kappa_{u-1}$ would make (11) smaller. This is to say, if w equals to one eigenvalue of $\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}$, $\hat{\boldsymbol{\Gamma}}$ cannot reach the minimum of (8), i.e. the maximum of $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$. From the discussion, we can conclude that the maximum likelihood estimator is one subset of u principal component directions.

Now assume $\kappa_1, \ldots, \kappa_u$ is one subset of u eigenvalues of $\hat{\boldsymbol{\Sigma}}$ that minimizes (10). Let $\kappa_{u+1}, \ldots, \kappa_p$ denote the complement of $\kappa_1, \ldots, \kappa_u$. Suppose there exists $\kappa_i < \kappa_l < \kappa_j$ where $1 \le l \le u$ and $u+1 \le i, j \le p$. Fixing $\kappa_1, \kappa_{l-1}, \kappa_{l+1} \ldots, \kappa_u$, by Lemma 4, the formula (10) can be reduced by replacing κ_l with either κ_i or κ_j. This tells us that $\kappa_{u+1}, \ldots, \kappa_p$ must form a "continuum block" of the eigenvalues. In other words, the maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ is the span of the first k and last $u-k$ principal component directions where k needs to be determined by the maximization of $L_3(\mathcal{S}_{\boldsymbol{\Gamma}})$ subject to $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$.

Proof of Proposition 5

Using the Lagrange multiplier rule, the solution $\hat{\boldsymbol{\Gamma}}$ can be gotten by finding the stationary points of the following unconstrained function:

$$-\frac{2}{n}L_5(\mathcal{S}_{\boldsymbol{\Gamma}}) + \text{trace}(\mathbf{U}(\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma}_0 - \mathbf{I}_u))$$

where $\mathbf{U}$ is a $(p-u)\times(p-u)$ matrix that stands for the Lagrange multipliers. Taking derivatives with respect to $\boldsymbol{\Gamma}_0$ and $\mathbf{U}$, we have the condition that the stationary points must satisfy

$$2\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Gamma}_0)^{-1} + \frac{2(p-u-1)\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{Q}_1}{\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{Q}_1)} + \frac{2\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{P}_1}{\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\mathbf{P}_1)} + \boldsymbol{\Gamma}_0(\mathbf{U}+\mathbf{U}^T) = 0$$

subject to $\boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma}_0 - \mathbf{I}_{p-u} = 0$. It is straightforward to get the expression of $\mathbf{U}+\mathbf{U}^T$. Substituting the expression $\mathbf{U}+\mathbf{U}^T$ into the condition above and after simplification, we will finally find out that

$$\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Gamma}_0 = 0.$$

Then we can conclude that the maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ in $L_5(\mathcal{S}_{\boldsymbol{\Gamma}})$ is the span of one subset of u principal component directions.

References

Anderson, T.W., 1963. Asymptotic theory for principal component analysis. Ann. Math. Stat. 34, 122–148.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bai, J., Ng, S., 2013. Principal components estimation and identification of static factors. J. Econometrics 176, 18–29.
Conway, J.B., 1990. A Course in Functional Analysis. Springer, New York.
Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regressions Through Graphics. Wiley, New York.
Cook, R.D., 2007. Fisher lecture: dimension reduction in regression (with discussion). Statist. Sci. 22, 1–26.
Cook, R.D., Li, B., Chiaromonte, F., 2007. Dimension reduction in regression without matrix inversion. Biometrika 94, 569–584.
Cook, R.D., Li, B., Chiaromonte, F., 2010. Envelope models for parsimonious and efficient multivariate linear regression. Statist. Sinica 20, 927–1010.
Cook, R.D., Zhang, X., 2015a. Simultaneous envelopes for multivariate linear regression. Technometrics 57, 11–25.
Cook, R.D., Zhang, X., 2015b. Foundations for envelope models and methods. J. Amer. Statist. Assoc. 110, 599–611.
Edelman, A., Arias, T.A., Smith, S.T., 1998. The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20, 303–353.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 603–680.
Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Sringer-Verlag.
Knyazev, A.V., Argentati, M.E., 2002. Principal angles between subspaces in an a-based scalar product: Algorithms and perturbation estimates. SIAM J. Sci. Comput. 23 (6), 2009–2041.
Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag. 2, 559–572.

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.

Royston, J.P., 1982. An extension of shapiro and wilk's w test for normality to large samples. Appl. Statist. 31, 115–124.
Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. J. Amer. Stat. Assoc. 97, 1167–1179.
Su, Z., Cook, R.D., 2011. Partial envelopes for efficient estimation in multivariate linear regression. Biometrika 98, 133–146.
Su, Z., Cook, R.D., 2012. Inner envelopes: efficient estimation in multivariate linear models. Biometrika 99, 687–702.
Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B 61, 611–622.
Welling, M., Agakov, F., Williams, C.K.I., 2003. Extreme Components Analysis. In: Neural Information Processing Systems, vol. 16, Vancouver, Canada.

Please cite this article as: J. Zhang and X. Chen, Principal envelope model. Journal of Statistical Planning and Inference (2019), https://doi.org/10.1016/j.jspi.2019.10.001.