OPERATIONS RESEARCH
Vol. 62, No. 4, July–August 2014, pp. 957–972
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2014.1290

Directed Principal Component Analysis

Yi-Hao Kao, Benjamin Van Roy
Stanford University, Stanford, California 94305
{yhkao@alumni.stanford.edu, bvr@stanford.edu}

We consider a problem involving estimation of a high-dimensional covariance matrix that is the sum of a diagonal matrix and a low-rank matrix, and making a decision based on the resulting estimate. Such problems arise, for example, in portfolio management, where a common approach employs principal component analysis (PCA) to estimate factors used in constructing the low-rank term of the covariance matrix. The decision problem is typically treated separately, with the estimated covariance matrix taken to be an input to an optimization problem. We propose *directed PCA*, an efficient algorithm that takes the decision objective into account when estimating the covariance matrix. Directed PCA effectively adjusts factors that would be produced by PCA so that they better guide the specific decision at hand. We demonstrate through computational studies that directed PCA yields significant benefit, and we prove theoretical results establishing that the degree of improvement over conventional PCA can be arbitrarily large.

Subject classifications: principal component analysis; covariance matrix estimation; high-dimensional data; portfolio management; convex optimization; decision analysis; stochastic programming.
Area of review: Stochastic Models.
History: Received May 2012; revision received September 2013; accepted April 2014. Published online in *Articles in Advance* June 20, 2014.

1. Introduction

We consider a problem that involves estimating a covariance matrix from independent identically distributed sample vectors and then making a decision based on this estimate. The payoff depends on the decision and an additional independent sample vector, observed after the decision is made. We focus on the case where the dimension of sample vectors is large relative to the number of observed samples.

Our formulation is relevant, for example, to the area of portfolio management, where it is common to estimate asset return covariances and to use these estimates to guide investment decisions. The prototypical decision problem here is to select a portfolio that maximizes risk-adjusted return, with risk measured in terms of return variance. This optimization problem is commonly formulated as a quadratic program in which return expectations and covariances serve as problem data.

To produce a meaningful estimate of a high-dimensional covariance matrix from a limited number of samples, we must assume that the matrix obeys some simplifying structure. In this paper we consider a scenario where the covariance matrix can be well-approximated by the sum of a diagonal matrix and a low-rank symmetric matrix. Such a covariance matrix can be viewed as representing a factor model, in which each observed variable is a noise-corrupted linear combination of latent common factors. A common approach to estimating such a covariance matrix is through principal component analysis (PCA). This approach focuses on explaining the observed data without regard to the objective of the subsequent decision. In other words, covariance matrix estimation and decision optimization are carried out independently. This separation leaves room for improvement, as we shall demonstrate in this paper.

We propose a new approach—*directed PCA*—which estimates the covariance matrix in a way that is tailored to the objective of the subsequent decision problem. As with PCA, directed PCA produces an estimate that is the sum of a diagonal matrix and a low-rank symmetric matrix. To understand directed PCA, it is useful to first consider an approach that we will refer to as *empirical optimization*. Let us assume for the purpose of this discussion that we are restricted to select a covariance matrix from a set S, consisting of positive semidefinite matrices each of which is the sum of a diagonal matrix and a symmetric matrix with rank that does not exceed some prespecified value K. In empirical optimization, one assumes that the future data sample is drawn uniformly from the set of previously observed samples, and the covariance matrix is selected from S to maximize expected payoff of the resulting decision. Equivalently, one can view empirical optimization as selecting from S a covariance matrix that would lead to a decision strategy that optimizes "in-sample performance." As such, empirical optimization does not explicitly aim to select a covariance matrix that explains historical data; the focus is on historical performance of hypothetical decisions.

Unfortunately, empirical optimization does not generally lead to effective future decisions. The problem is overfitting; the selected covariance matrix is too specialized to decisions that would have been effective in the face of previously observed data samples, and the performance of the resulting decision strategy does not generalize well to future

samples. Directed PCA aims to rectify this shortcoming by combining the merits of empirical optimization and PCA-based methods for covariance estimation. Directed PCA essentially carries out empirical optimization, but subject to a constraint that the resulting covariance matrix explains the data well. In particular, the point estimate is selected from a confidence region around a maximum *a posteriori* (MAP) estimate.

A new formulation for estimating covariance matrices from high-dimensional data is not practically useful without an efficient estimation algorithm. An important contribution of this paper is an efficient algorithm for directed PCA, with computational requirements comparable to those of conventional PCA. To assess merits of directed PCA, we apply the algorithm to study two data sets: one is a synthetic data set designed for this specific purpose, whereas the other is an empirical time series of S&P 500 stock returns. We find that directed PCA yields significant improvements over conventional approaches. Specifically, applying directed PCA to a portfolio management example based on empirical data increases certain-equivalent payoff by 7%. With our synthetic data set, we identify plausible examples where the gain reaches 34%.

Aside from our formulation, algorithm, and empirical study, a significant contribution of this paper is in a pair of theoretical results that elucidate the benefits of directed PCA. These results assume that the covariance matrix of the generating distribution is the sum of a diagonal matrix and a low-rank symmetric matrix. As such, each data sample can be viewed as a noise-corrupted linear combination of factor loading vectors. The first of our two theoretical results establishes that when the decision objective is aligned with one of the factor loading vectors, the absolute performance increase from using directed PCA instead of MAP estimation can grow linearly in the dimension of the data. The second result establishes that when the decision objective is orthogonal to the span of factor loading vectors, the percentage performance increase from using directed PCA instead of MAP estimation can grow linearly in the dimension of the data. These two results offer striking indications of the importance of accounting for the decision objective in the estimation process, especially when dealing with high-dimensional data.

A conceptual thrust in the development of directed PCA is in the use of a decision objective to guide model-fitting. In order to put our work in perspective relative to prior literature, we will discuss here three threads of research that relate to this spirit: robust optimization, operational statistics, and statistical decision theory.

In robust optimization, one makes a decision assuming that uncertain parameters are chosen adversarially, though constrained to a prescribed confidence region (see, e.g., Ben-Tal et al. 2009). The work from this area that most closely relates to ours is that of Goldfarb and Iyengar (2003), which treated a robust portfolio selection problem in which confidence regions for expected returns and covariances are produced through a regression analysis of historical data. This approach selects a portfolio that maximizes risk-adjusted expected return, assuming worst case point estimates within the confidence regions. Also related is the subsequent work of Delage and Ye (2008), which treated a more general formulation that accommodates uncertainty in statistics of return distributions beyond expectations and covariances; this formulation synthesizes those of Goldfarb and Iyengar (2003) and Popescu (2007).

Similarly with robust optimization, our approach constrains the choice of point estimate to a confidence region. However, an important difference is that, instead of selecting a worst case point estimate, our approach selects one that optimizes in-sample performance. Relative to using the MAP estimate, our approach is in a sense more aggressive whereas robust optimization is more conservative. Being aggressive helps when the MAP estimate exhibits significant bias and reasonably low variance. Bias can stem from model misspecification or MAP estimation itself. For example, in our context, empirical data may not be generated by a factor model with a small number of factors, and even if they were, the MAP estimate can be different from the conditional expectation. On the other hand, assuming a factor model structure controls the variance of the resulting estimate so that it is sufficiently robust without imposing any additional conservatism. Indeed, in the course of research that led to this paper, we tried robust optimization using our confidence regions but found resulting decisions to be overly conservative and to perform worse than MAP estimates.

Operational statistics is another line of work that emphasizes the relevance of decision objectives in estimating model parameters. This terminology first appeared in Liyanage and Shanthikumar (2005), where the authors describe a method that factors objectives into how demand distributions are estimated when they are to be used in a newsvendor inventory control problem. In particular, instead of estimating the parameters of the distribution, the method estimates an optimal order quantity directly from the data. This approach is subsequently elaborated by Chu et al. (2008) using Bayesian analysis with a non-informative prior. In a related vein, Besbes et al. (2010) develop a statistical test that incorporates decision performance into a measure of statistical validity and illustrate their approach in the context of a revenue management problem. Our work can be seen as contributing to this line of research by developing a method that is similar in spirit but designed for a different class of problems involving estimation of a factor model.

Our work also relates to the broad and well-studied area of statistical decision theory (see, e.g., Berger 1985). In this area, it is common to apply a Bayesian approach, which begins with a prior distribution over all possible models and then evaluates a posterior distribution conditioned on observed data. The Bayes optimal decision is then taken to be the one that maximizes expected performance with respect to this posterior distribution. Estimation and decision-making are decoupled, but in a coherent way that does not leave

room for improvement. However, this approach is often computationally intractable for relevant decision objectives and model classes. As a consequence, practitioners tend to appeal to approximate solutions. In our context, MAP estimation provides an approximate solution, and the nature of this approximation does leave room for improvement from coupling estimation and decision-making, as we establish in this paper.

This paper is organized as follows. A mathematical formulation of our problem is presented in §2. In §3 we present two PCA-based estimation methods and introduce directed PCA. To simplify the exposition, the discussion of §3 assumes that the diagonal component of the covariance matrix estimate is constrained to be a nonnegative multiple of the identity matrix. We provide generalizations in §4 that accommodate arbitrary positive semidefinite diagonal matrices. We present computational studies in §5 and theoretical results in §6. Section 7 concludes the paper with a discussion of some possible extensions.

2. Problem Formulation

Consider a stochastic optimization problem that involves selecting a decision $\mathbf{u} \in \mathbb{R}^M$ in order to maximize the expected value of an objective function

$$g(\mathbf{u}, \mathbf{x}) = \mathbf{c}^{\mathrm{T}}\mathbf{u} - (\mathbf{u}^{\mathrm{T}}\mathbf{x})^2, \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^M$ is a given objective vector, and $\mathbf{x} \in \mathbb{R}^M$ is a random vector drawn from a zero-mean Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma}_*)$, where $\mathbf{\Sigma}_*$ is unknown to us. This problem could be easily solved if we had access to the covariance matrix $\mathbf{\Sigma}_*$. Indeed, since

$$\mathrm{E}[g(\mathbf{u}, \mathbf{x})] = \mathbf{c}^{\mathrm{T}}\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_*\mathbf{u},$$

we can rewrite this decision problem as

$$\max_{\mathbf{u}\in\mathbb{R}^M} \mathbf{c}^{\mathrm{T}}\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}_*\mathbf{u}, \tag{2}$$

which has an analytical solution $\mathbf{u}_* = \frac{1}{2}\mathbf{\Sigma}_*^{-1}\mathbf{c}$. Now suppose we have observed N sample vectors drawn i.i.d. from $\mathcal{N}(0, \mathbf{\Sigma}_*)$, denoted by a set $\mathscr{X} = \{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}\}$. Our goal is to compute a point estimate $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}_*$ based on the dataset $\mathscr{X}$, and use that estimate to guide our decision making. One natural approach of doing so is to use $\hat{\mathbf{\Sigma}}$ as a surrogate for $\mathbf{\Sigma}_*$ in the decision problem (2). This leads to a decision

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}\in\mathbb{R}^M} \mathbf{c}^{\mathrm{T}}\mathbf{u} - \mathbf{u}^{\mathrm{T}}\hat{\mathbf{\Sigma}}\mathbf{u} = \tfrac{1}{2}\hat{\mathbf{\Sigma}}^{-1}\mathbf{c}. \tag{3}$$

The out-of-sample performance of this resulting decision, given by $\mathrm{E}[g(\hat{\mathbf{u}}, \mathbf{x}) \mid \mathbf{\Sigma}_*] = \mathbf{c}^{\mathrm{T}}\hat{\mathbf{u}} - \hat{\mathbf{u}}^{\mathrm{T}}\mathbf{\Sigma}_*\hat{\mathbf{u}}$, is therefore the evaluation criterion for an estimate $\hat{\mathbf{\Sigma}}$.

A typical Bayesian treatment of this problem goes as follows. We first choose a prior distribution $p(\mathbf{\Sigma})$ that reflects our belief about the properties of $\mathbf{\Sigma}_*$. We then evaluate the posterior probability $p(\mathbf{\Sigma} \mid \mathscr{X})$ conditioned on observation $\mathscr{X}$, and solve for a Bayes optimal decision by optimizing

$$\max_{\mathbf{u}} \mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathscr{X}]. \tag{4}$$

Recall that

$$\mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathscr{X}] = \int_{\mathbf{\Sigma}} \mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathbf{\Sigma}] p(\mathbf{\Sigma} \mid \mathscr{X})\, d\mathbf{\Sigma}. \tag{5}$$

This integral is generally hard to evaluate, and one common relaxation is to assume $p(\mathbf{\Sigma} \mid \mathscr{X})$ peaks sharply at its mode. Specifically, let $\mathbf{\Sigma}_{\mathrm{MAP}}$ denote the *maximum a posteriori* (MAP) estimate, defined as

$$\mathbf{\Sigma}_{\mathrm{MAP}} = \arg\max_{\mathbf{\Sigma}} p(\mathbf{\Sigma} \mid \mathscr{X}). \tag{6}$$

We can approximate (5) by

$$\begin{aligned}\mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathscr{X}] &\simeq \int_{\mathbf{\Sigma}} \mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathbf{\Sigma}] \delta(\mathbf{\Sigma} - \mathbf{\Sigma}_{\mathrm{MAP}})\, d\mathbf{\Sigma} \\ &= \mathrm{E}[g(\mathbf{u}, \mathbf{x}) \mid \mathbf{\Sigma}_* = \mathbf{\Sigma}_{\mathrm{MAP}}],\end{aligned} \tag{7}$$

which, together with (4) and (3), suggest a decision $\mathbf{u} = \frac{1}{2}\mathbf{\Sigma}_{\mathrm{MAP}}^{-1}\mathbf{c}$. Thus, instead of solving a convoluted decision problem, practitioners often adopt the above rationale and compute the MAP estimate as a solution.

To produce a meaningful MAP estimate from a limited number of samples in high-dimensional space, the prior $p(\mathbf{\Sigma})$ is usually designed to put stronger weights on those covariance matrices that obey some simplifying structure. In this paper, we focus on a scenario in which the covariance matrix $\mathbf{\Sigma}_*$ is believed to be dominated by a few components. Mathematically speaking, we will assume that $\mathbf{\Sigma}_*$ can be well-approximated by the sum of a diagonal matrix $\mathbf{R}_* \succeq 0$ and a symmetric matrix $\mathbf{F}_* \succeq 0$ whose rank is much smaller than the dimension M. This assumption effectively corresponds to a factor model that generates each sample vector by

$$\mathbf{x}_{(n)} = \mathbf{F}_*^{1/2}\mathbf{z}_{(n)} + \mathbf{w}_{(n)},$$

where $\mathbf{F}_*^{1/2}$ is a thin matrix that represents the factor loadings, $\mathbf{z}_{(n)} \sim \mathcal{N}(0, \mathbf{I})$ represents a parsimonious set of independent factors, and $\mathbf{w}_{(n)} \sim \mathcal{N}(0, \mathbf{R}_*)$ represents the idiosyncratic residual noise not captured by these factors. Such a model has been widely applied in economics, finance, medicine, psychology, and various other natural and social sciences (Harman 1976). We will discuss two popular choices of prior that express this factor model assumption in next section.

Although using MAP estimate to guide decision making is a common approach widely adopted by practitioners due to its simplicity, it generally suffers from two disadvantages. First, the prior $p(\mathbf{\Sigma})$ is usually chosen in a way that not only reflects our assumption but also enables efficient computation of the MAP estimate. Such mathematical convenience is often attained at the price of introducing model bias, or equivalently, prior mis-specification. Second, MAP estimation does not take into account the decision objective when computing an estimate. As we will see in the following sections, these disadvantages together leave room for improvement.

3. Estimation Algorithms: Uniform Residual Case

We will begin our discussion with a simplified scenario in which the residual variances are further assumed to be identical. In other words, we will assume $\mathbf{R}_*$ is a multiple of identity matrix, denoted by $\sigma_*^2\mathbf{I}$, and therefore $\mathbf{\Sigma}_* = \mathbf{F}_* + \sigma_*^2\mathbf{I}$. This assumption helps us better illustrate our main idea, and will be relaxed in the next section.

3.1. Regularized Maximum-Likelihood Estimates

Instead of solving (6) directly for an MAP estimate, in practice one might convert it into a *regularized* maximum-likelihood estimation problem, where the regularization reflects our prior belief that $\mathbf{\Sigma}_*$ obeys the factor model assumption. We now consider two types of regularization for this purpose, both of which are popular partly due to the fact that they can be efficiently computed via PCA.

3.1.1. Constraining the Rank. Since we assume the data is generated by a factor model with a parsimonious set of factors, a natural choice of regularization is to constrain the number of factors, or equivalently, the rank of matrix $\mathbf{F}_*$. Such regularized maximum-likelihood estimation can be formulated as an optimization problem

$$\begin{aligned} \max_{\mathbf{F}\in\mathbb{S}_+^M,\,\sigma^2\geqslant 0} \quad & \log p(\mathscr{X} \mid \mathbf{\Sigma}) \\ \text{s.t.} \quad & \mathbf{\Sigma} = \mathbf{F} + \sigma^2\mathbf{I}, \\ & \operatorname{rank}(\mathbf{F}) \leqslant K, \end{aligned} \tag{8}$$

where $\mathbb{S}_+^M$ denotes the set of all $M \times M$ positive semidefinite symmetric matrices, and K is the number of factors exogenously specified by the user. Recall that the log likelihood of the observation $\mathscr{X}$ can be written as

$$\begin{aligned} &\log p(\mathscr{X} \mid \mathbf{\Sigma}) \\ &\quad = -\frac{N}{2}\big(M\log(2\pi) + \log\det(\mathbf{\Sigma}) + \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{SAM}})\big), \end{aligned} \tag{9}$$

where $\mathbf{\Sigma}_{\text{SAM}} = (1/N)\sum_{n=1}^N \mathbf{x}_{(n)}\mathbf{x}_{(n)}^{\mathrm{T}}$ denotes the sample covariance matrix. Although $\log p(\mathscr{X} \mid \mathbf{\Sigma})$ is not concave in $\mathbf{\Sigma}$ and therefore (8) is not a convex program, Tipping and Bishop (1999) has shown that its solution can be efficiently computed via PCA. This involves first computing an eigendecomposition of the sample covariance matrix $\mathbf{\Sigma}_{\text{SAM}} = \mathbf{BSB}^{\mathrm{T}}$, where $\mathbf{B} \in \mathbb{R}^{M\times M}$ is an orthonormal matrix and $\mathbf{S}$ is a diagonal matrix whose diagonal elements are sorted in decreasing order. Our estimate for the residual variance is then given by

$$\hat{\sigma}^2 = \frac{1}{M-K}\sum_{k=K+1}^{M} \mathbf{S}_{k,k}.$$

Furthermore, letting $\mathbf{H}$ be a $K \times K$ diagonal matrix with diagonal entries $\mathbf{H}_{k,k} = \mathbf{S}_{k,k} - \sigma^2$ and $\mathbf{B}_{1:K}$ be the $M \times K$ matrix made up of the first K columns of $\mathbf{B}$, the estimate for the factor loadings is given by $\hat{\mathbf{F}} = \mathbf{B}_{1:K}\mathbf{H}\mathbf{B}_{1:K}^{\mathrm{T}}$. We will refer to this method as *uniform-residual rank-constrained maximum-likelihood*, and use $\mathbf{\Sigma}_{\text{URM}}^K = \hat{\mathbf{F}} + \hat{\sigma}^2\mathbf{I}$ to denote the covariance matrix resulting from this procedure.

In our implementation, we employ a version of cross-validation for the selection of K. Details of the procedure can be found in the appendix. Through selection of K, this procedure arrives at a covariance matrix which we will denote by $\mathbf{\Sigma}_{\text{URM}}$.

3.1.2. Penalizing the Trace. Instead of constraining the rank of factor loadings matrix $\mathbf{F}$, Kao and Van Roy (2013) regularized the maximum-likelihood estimation by introducing a trace penalty term. Specifically, they formulated a convex program by using three facts: first, the log-likelihood function (9) is concave in the *inverse* covariance matrix $\mathbf{\Sigma}^{-1}$; second, if $\mathbf{\Sigma} = \mathbf{F} + \sigma^2\mathbf{I}$ with $\mathbf{F} \in \mathbb{S}_+^M$, then the matrix defined by $\mathbf{G} = \sigma^{-2}\mathbf{I} - \mathbf{\Sigma}^{-1}$ is in $\mathbb{S}_+^M$ with $\operatorname{rank}(\mathbf{G}) = \operatorname{rank}(\mathbf{F})$; third, for any $\mathbf{G} \in \mathbb{S}_+^M$, it is a common technique to use $\operatorname{tr}(\mathbf{G})$ as a convex surrogate for $\operatorname{rank}(\mathbf{G})$. These facts together suggest working with inverse covariance matrix $\mathbf{\Sigma}^{-1}$ and penalizing $\operatorname{tr}(\mathbf{G})$ when computing maximum-likelihood estimate, as formally described by a convex program

$$\begin{aligned} \max_{\mathbf{G}\in\mathbb{S}_+^M,\,v\geqslant 0} \quad & \log p(\mathscr{X} \mid \mathbf{\Sigma}) - \lambda\operatorname{tr}(\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{\Sigma}^{-1} = v\mathbf{I} - \mathbf{G}. \end{aligned} \tag{10}$$

Here, the variable v represents the reciprocal of residual variance. We will refer to this method as *uniform-residual trace-penalized maximum-likelihood*, and use $\mathbf{\Sigma}_{\text{UTM}}^{\lambda}$ to denote the covariance matrix derived from the optimal solution to this convex program. Similarly to URM, λ here can be selected by cross-validation.

Kao and Van Roy (2013) also provided an analytical solution to (10) using PCA. Their solution is based on soft-thresholding eigenvalues, as defined below.

Definition 1. For all symmetric $M \times M$ matrices, denoted by a set $\mathbb{S}^M$, we define an operator $\mathscr{F}_\lambda\colon \mathbb{S}^M \to \mathbb{S}^M$ such that $\mathscr{F}_\lambda(\mathbf{A}) = \mathbf{B}$ if $\mathbf{A}$ and $\mathbf{B}$ share the same eigenvectors, and their corresponding eigenvalues, denoted by $a_1, a_2, \ldots, a_M$ and $b_1, b_2, \ldots, b_M$, sum to the same trace and satisfy

$$b_i = \max\left\{a_i - \frac{2\lambda}{N}, \frac{1}{v}\right\}, \quad i = 1, \ldots, M,$$

for some scalar v.

They have shown that

$$\mathbf{\Sigma}_{\text{UTM}}^{\lambda} = \mathscr{F}_\lambda(\mathbf{\Sigma}_{\text{SAM}}). \tag{11}$$

Therefore, to compute $\mathbf{\Sigma}_{\text{UTM}}^{\lambda}$, we first compute the eigendecomposition of $\mathbf{\Sigma}_{\text{SAM}}$, and then determine the value of v such that the eigenvalues given by the above formula sum to the desired trace. Note that for reasonably large values of λ, the operator $\mathscr{F}_\lambda$ will typically convert most eigenvalues to a constant and subtract another constant from the other,

larger, eigenvalues. This produces a matrix that is the sum of a low-rank one and a multiple of the identity matrix, as desired.

One of the main differences between URM and UTM is the way they deal with the large eigenvalues of the sample covariance matrix. While URM preserves those values in its estimate, UTM subtracts a constant $2\lambda/N$ from them. Such subtraction has been shown to correct the bias induced by sample eigenvalues, and generally yields more accurate estimates (Kao and Van Roy 2013).

3.2. Posterior-Constrained Empirical Optimization

Both of the methods discussed above focus on the goodness of fit and ignore the subsequent decision guided by the estimate. We now propose a method that takes into account the decision objective by maximizing the in-sample performance of the resulting decision. Recall from (3) that, given an estimate $\mathbf{\Sigma}$, the resulting decision $\mathbf{u}$ can be viewed as a function of it. Let us denote this relation by $\mathbf{u}(\mathbf{\Sigma}) = \frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{c}$. The *in-sample performance* of $\mathbf{\Sigma}$ is therefore defined as the payoff we receive as if we apply the resulting decision $\mathbf{u}(\mathbf{\Sigma})$ over the observed data, i.e.,

$$\tilde{g}(\mathbf{\Sigma}) = \sum_{n=1}^{N} g(\mathbf{u}(\mathbf{\Sigma}), \mathbf{x}_{(n)}).$$

A simple *empirical optimization* approach would seek a $\mathbf{\Sigma}$ that maximizes the in-sample performance $\tilde{g}(\mathbf{\Sigma})$. Although this approach explicitly takes into account the decision objective, it generally suffers from over-fitting since the selected estimate is too specialized for in-sample data, and the resulting decision does not generalize well to future samples.

One remedy to this problem is to require $\mathbf{\Sigma}$ to be selected from a set of covariance matrices that are coherent with our model assumption and well explain the historical data. Since posterior probability $p(\mathbf{\Sigma} \mid \mathscr{X})$ effectively reflects these two criteria, we propose selecting a $\mathbf{\Sigma}$ that maximizes in-sample performance $\tilde{g}(\mathbf{\Sigma})$ subject to a constraint that its posterior probability $p(\mathbf{\Sigma} \mid \mathscr{X})$ is sufficiently high, as formerly described by

$$\begin{aligned} \max_{\mathbf{\Sigma}} \quad & \tilde{g}(\mathbf{\Sigma}) \\ \text{s.t.} \quad & p(\mathbf{\Sigma} \mid \mathscr{X}) \geqslant p_0. \end{aligned}$$

To select a prior for this purpose, note that the trace penalty introduced in §3.1.2 is particularly suitable. Specifically, if we view $\exp(-\lambda \operatorname{tr}(\mathbf{G}))$ as a prior for $\mathbf{\Sigma}$, then $\log p(\mathbf{\Sigma} \mid \mathscr{X})$ equals to $\log p(\mathscr{X} \mid \mathbf{\Sigma}) - \lambda \operatorname{tr}(\mathbf{G}) +$ a constant, which is concave in $\mathbf{\Sigma}^{-1}$ and leads to a desirable convex level-set. This observation together with the fact that $\tilde{g}$ is a concave function of $\mathbf{\Sigma}^{-1}$ suggest a convex formulation:

$$\begin{aligned} \max_{\mathbf{G}\in\mathbb{S}_+^M,\, v\geqslant 0} \quad & \tilde{g}(\mathbf{\Sigma}) \\ \text{s.t.} \quad & \log p(\mathscr{X} \mid \mathbf{\Sigma}) - \lambda \operatorname{tr}(\mathbf{G}) \geqslant \bar{p} - \epsilon, \\ & \mathbf{\Sigma}^{-1} = v\mathbf{I} - \mathbf{G}, \end{aligned} \tag{12}$$

where $\bar{p}$ is the optimum of (10), and ϵ specifies the "radius" of the confidence set. We will refer to this method as *posterior-constrained empirical optimization*, and denote the resulting estimate by $\mathbf{\Sigma}_{\text{PEO}}^{\lambda,\epsilon}$.

To make this formulation practically useful, we also need an efficient solution method. One typical approach for solving (12) is to price out the inequality constraint to produce the following related problem

$$\begin{aligned} \max_{\mathbf{G}\in\mathbb{S}_+^M,\, v\geqslant 0} \quad & \gamma\tilde{g}(\mathbf{\Sigma}) + \log p(\mathscr{X} \mid \mathbf{\Sigma}) - \lambda \operatorname{tr}(\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{\Sigma}^{-1} = v\mathbf{I} - \mathbf{G}, \end{aligned} \tag{13}$$

where $\gamma \geqslant 0$ is a scalar that adjusts the weights between the in-sample performance and the posterior probability. It is easy to see that γ increases with ϵ, and $\gamma = 0$ when $\epsilon = 0$. Therefore, to solve (12), we can solve (13) for a sequence of γ, and pick the solution that corresponds to the largest γ while remaining feasible with respect to (12). The problem then boils down to how to solve (13) efficiently.

Since (13) involves a semidefinite constraint and a trace penalty, one might consider solving the problem through application of an alternating direction method of multipliers (ADMM) (see, e.g., Boyd et al. 2011). This approach, however, gives rise to onerous computational demands when the data dimension M is large. One important contribution of this paper is an efficient algorithm that solves (13) using a variation of PCA. This method is justified by a theoretical result that relates the decision objective to the principal components of the PEO estimate. Specifically, letting $(\mathbf{G}_\gamma, v_\gamma)$ be the solution to (13) and $\mathbf{\Sigma}_\gamma = (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}$ be the resulting estimate, we have:

THEOREM 1. *For any $\mathbf{\Sigma} \in \mathbb{S}_{++}^M$, $\mathbf{\Sigma} = \mathbf{\Sigma}_\gamma$ if and only if it is a fixed point that satisfies*

$$\mathbf{\Sigma} = \mathscr{F}_\lambda(\mathbf{\Sigma}_{\text{SAM}} + \gamma\mathbf{C} \otimes \mathbf{D}), \tag{14}$$

where $\mathbf{C} = \mathbf{c}\mathbf{c}^{\mathrm{T}}$, $\mathbf{D} = \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{SAM}} - \mathbf{I}$, *and* $\mathbf{C} \otimes \mathbf{D} \triangleq \frac{1}{2}(\mathbf{C}\mathbf{D} + \mathbf{D}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}})$.

To interpret this result, note that $\mathbf{C}$ essentially contains the information about the decision objective, $\mathbf{D}$ represents the discrepancy between the estimate and the sample covariance matrix (as $\mathbf{D}$ vanishes when $\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{SAM}}$), and the operation $\otimes$ captures the alignment between the objective and the discrepancy. Furthermore, the soft-thresholding operator $\mathscr{F}_\lambda$ produces a covariance matrix that is the sum of a low-rank matrix and a multiple of identity matrix, as desired.

Based on Theorem 1, we can design an iterative algorithm for solving (13). This algorithm evaluates the right-hand side (RHS) of (14) at each iteration, and use that result as the search direction for parameters update. In the rare and degenerate case where (14) outputs a non-positive definite matrix, or the case where no point along that search direction can increase the objective value, we simply update the parameters according to a projected gradient ascent method. Algorithm 1 describes this procedure. For the sake of clarity

and simplicity, we put off some implementation details into the appendix. These parts are marked by asterisks in Algorithm 1.

Since Algorithm 1 uses (14) as its termination criterion, it returns the solution to (13) whenever it terminates. Although we do not provide formal convergence analysis here, we have applied Algorithm 1 to a total of 2800 randomly generated problem instances, and for all of them Algorithm 1 terminated successfully. More details on these experiments will be reported in §5.

We will refer to this algorithm as *directed PCA*, since the selected components are tailored for the decision objective. To measure the performance of directed PCA, we benchmark it against a common implementation of ADMM based on L-BFGS algorithm (Liu and Nocedal 1989), as suggested in Boyd et al. (2011). We found in our experiments that directed PCA generally requires three orders of magnitude less compute time than the ADMM to attain the same level of accuracy. Table 1 gives the average compute time of directed PCA versus ADMM on a typical workstation for ten randomly generated problem instances over dimension $M = 100$, 200, and 500.

Algorithm 1 (Directed PCA)
Input: $\mathscr{X}$, $\mathbf{c}$, λ, γ, initial point $\mathbf{\Sigma}_0$
Output: $\mathbf{\Sigma}_\gamma$
(For steps marked with asterisks, details are provided in the appendix.)
$(v\mathbf{I}-\mathbf{G})^{-1} \leftarrow \mathbf{\Sigma}_0$ // initialize $(\mathbf{G}, v)$
repeat
$\hat{\mathbf{D}} \leftarrow (v\mathbf{I}-\mathbf{G})\mathbf{\Sigma}_{\mathrm{SAM}} - \mathbf{I}$
// evaluate the discrepancy between estimate and sample
$\hat{\mathbf{\Sigma}} \leftarrow \mathscr{F}_\lambda(\mathbf{\Sigma}_{\mathrm{SAM}} + \gamma\mathbf{C} \otimes \hat{\mathbf{D}})$
// evaluate the RHS of (14) using PCA
if $\hat{\mathbf{\Sigma}} \in \mathbb{S}^M_{++}$ **then**
$(\hat{v}\mathbf{I}-\hat{\mathbf{G}})^{-1} \leftarrow \hat{\mathbf{\Sigma}}$
$(\Delta\mathbf{G}, \Delta v) \leftarrow (\hat{\mathbf{G}}, \hat{v}) - (\mathbf{G}, v)$
// generate search direction
Use backtracking line search to select an appropriate step size $\alpha \in [0, 1]^*$
$(\mathbf{G}, v) \leftarrow (\mathbf{G}, v) + \alpha(\Delta\mathbf{G}, \Delta v)$
end if
if $\hat{\mathbf{\Sigma}} \notin \mathbb{S}^M_{++}$ or $\alpha = 0$ **then**
// this means $\hat{\mathbf{\Sigma}}$ fails to give a legitimate search direction
Update $(\mathbf{G}, v)$ by projected gradient ascent*
end if
until $(v\mathbf{I}-\mathbf{G})^{-1}$ satisfies (14)*
return $(v\mathbf{I}-\mathbf{G})^{-1}$.

4. Estimation Algorithms: Nonuniform Residual Case

All of the algorithms discussed in §3 assume the residual variances are identical for each variable, which is not always the case in practice. We will relax such an assumption in this section and discuss three algorithms that can be viewed as extensions of URM, UTM, and PEO.

Table 1. The average compute time of directed PCA versus ADMM over different problem dimensions.

Dimension	$M = 100$	$M = 200$	$M = 500$
ADMM	44.5 sec	4.31 min	3.01 hr
Directed PCA	77.1 ms	0.319 sec	3.47 sec

4.1. Expectation Maximization

Without the assumption of uniform residual variances, the rank-constrained maximum-likelihood method can be written as

$$\begin{aligned} \max_{\mathbf{F}\in\mathbb{S}^M_+,\,\mathbf{R}\in\mathbb{D}^M_+} \quad & \log p(\mathscr{X} \mid \mathbf{\Sigma}) \\ \text{s.t.} \quad & \mathbf{\Sigma} = \mathbf{F} + \mathbf{R}, \\ & \operatorname{rank}(\mathbf{F}) \leqslant K, \end{aligned} \tag{15}$$

where $\mathbb{D}^M_+$ denotes the set of all $M \times M$ diagonal matrices whose entries are non-negative. Unlike (8), this formulation does not have an analytical solution. One common approach to solving it approximately is through the expectation-maximization algorithm (EM). We now give a sketch of this method, while more details can be found in Rubin and Thayer (1982).

The algorithm generates a sequence of iterates $\mathbf{F}^{1/2} \in \mathbb{R}^{M\times K}$ and $\mathbf{R} \in \mathbb{D}^M_+$, such that the covariance matrix $\mathbf{\Sigma} = \mathbf{F}^{1/2}\mathbf{F}^{\mathrm{T}/2} + \mathbf{R}$ increases the log-likelihood of $\mathscr{X}$ with each iteration. Each iteration involves an estimation step in which we compute expectations $\mathrm{E}[\mathbf{z}_{(n)} \mid \mathbf{x}_{(n)}]$ and $\mathrm{E}[\mathbf{z}_{(n)}\mathbf{z}^{\mathrm{T}}_{(n)} \mid \mathbf{x}_{(n)}]$, for $n = 1, \ldots, N$, assuming the data are generated according to the covariance matrix $\mathbf{\Sigma} = \mathbf{F}^{1/2}\mathbf{F}^{\mathrm{T}/2} + \mathbf{R}$. A maximization step then updates $\mathbf{F}$ and $\mathbf{R}$ according to these expectations. This algorithm is guaranteed to converge, though not necessarily to a global optimum.

4.2. Scaled Trace-Penalized Maximum-Likelihood

To relax the uniform residual assumption for trace-penalized maximum-likelihood, Kao and Van Roy (2013) propose a method based on componentwise scaling of the data. Specifically, let $\mathbf{T} \in \mathbb{D}^M_+$ be a scaling matrix, and let $\mathbf{T}\mathscr{X} = \{\mathbf{T}\mathbf{x}_{(1)}, \ldots, \mathbf{T}\mathbf{x}_{(N)}\}$. It is easy to show that $p(\mathscr{X} \mid \mathbf{\Sigma}) = p(\mathbf{T}\mathscr{X} \mid \mathbf{T}\mathbf{\Sigma}\mathbf{T})$ if $\mathbf{T}$ has unit determinant. This observation motivates an approach that simultaneously seeks an appropriate scaling matrix $\mathbf{T}$ and a factor model that best explains the scaled data, as formerly described by the following optimization problem:

$$\begin{aligned} \max_{\mathbf{G}\in\mathbb{S}^M_+,\,v\in\mathbb{R}_+,\,\mathbf{T}\in\mathbb{D}^M_+} \quad & \log p(\mathbf{T}\mathscr{X} \mid \mathbf{\Sigma}) - \lambda \operatorname{tr}(\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{\Sigma}^{-1} = v\mathbf{I} - \mathbf{G}, \\ & \log\det\mathbf{T} \geqslant 0. \end{aligned} \tag{16}$$

The solution to this problem identifies a linear transformation that allows the data to be best-explained by a factor model with uniform residual variances. Given an optimal solution, $1/\mathbf{T}_{i,i}^2$ should be approximately proportional to the variance of the ith residual, so that normalizing by $\mathbf{T}_{i,i}$ makes residual variances uniform. Note that the optimization problem constrains $\log\det\mathbf{T}$ to be nonnegative rather than zero. This makes the feasible region convex, and this constraint is binding at the optimal solution. Denote the optimal solution to (16) by $(\mathbf{G}_*, v_*, \mathbf{T}_*)$. The estimate is thus given by $\mathbf{T}_*^{-1}(v_*\mathbf{I}-\mathbf{G}_*)^{-1}\mathbf{T}_*^{-\mathrm{T}}$.

The objective function of (16) is not concave in $(\mathbf{G}, v, \mathbf{T})$, but is biconcave in $(\mathbf{G}, v)$ and $\mathbf{T}$. The authors solve it by coordinate ascent, alternating between optimizing $(\mathbf{G}, v)$ and $\mathbf{T}$. This procedure is guaranteed to converge, though still possibly to a local optimum. They have tested different heuristics for choosing initial points and found their results insensitive to this choice. In this paper, we will choose the initial point via UTM in our implementation and refer to this method as *scaled trace-penalized maximum-likelihood* (STM).

4.3. Scaled Posterior-Constrained Empirical Optimization

We now extend PEO to deal with nonuniform residuals by a scaling technique similar to that presented in §4.2. Note that for any scaling matrix $\mathbf{T}$, we have

$$\mathbf{c}^{\mathrm{T}}\mathbf{u}-(\mathbf{u}^{\mathrm{T}}\mathbf{x})^2=(\mathbf{T}\mathbf{c})^{\mathrm{T}}(\mathbf{T}^{-1}\mathbf{u})-((\mathbf{T}^{-1}\mathbf{u})^{\mathrm{T}}\mathbf{T}\mathbf{x})^2.$$

This implies if we scale the data, objective vector, and decision by the same scaling matrix, then the resulting performance does not change. Based on this equivalence, we propose *scaled posterior-constrained empirical optimization* (SPEO), formally described by the following procedure:

1. Use STM to produce a scaling matrix $\mathbf{T}$.
2. Scale the data and objective vector by $\mathbf{T}$.
3. Apply PEO to the scaled data and objective to produce a scaled covariance estimate $\tilde{\mathbf{\Sigma}}$.
4. Compute a prescaled decision $\tilde{\mathbf{u}}=\frac{1}{2}\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{T}\mathbf{c})$.
5. Output a decision $\hat{\mathbf{u}}=\mathbf{T}\tilde{\mathbf{u}}$.

We will refer to SPEO and PEO collectively as *directed PCA*.

5. Computational Experiments

To compare the performance of aforementioned algorithms, we conducted two sets of experiments. The first one uses synthetic data, whereas the second one is based on the real data from S&P 500 stocks returns.

5.1. Synthetic Data

For synthetic data experiment, we further divided it into two cases: one with uniform residual variances, and the other with nonuniform residual variances. The former used the following procedure to generate data:

1. Sample M orthonormal vectors $\phi_1, \phi_2, \ldots, \phi_M \in \mathbb{R}^M$ isotropically.
2. Sample $f_1, f_2, \ldots, f_M$ iid from $\mathcal{N}(-1, 2)$.
3. Let $\mathbf{F}_*^{1/2}=[e^{f_1}\phi_1\ e^{f_2}\phi_2\ \ldots\ e^{f_M}\phi_M]$.
4. Let $\mathbf{\Sigma}_*=\mathbf{F}_*^{1/2}\mathbf{F}_*^{\mathrm{T}/2}+\mathbf{I}$.
5. Sample $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$ iid from $\mathcal{N}(0, \mathbf{\Sigma}_*)$.

Note that we sampled the magnitudes of factors from a log-normal distribution, and therefore only a small fraction of them would be significant while the others close to zero. This is intended to simulate the scenario where the covariance matrix can be well approximated by but not exactly equal to the sum of a low-rank matrix and a diagonal one, as in many real-world cases.

We repeated this procedure 100 times for each $N \in \{25, 50, 100, 200\}$ with $M = 100$, and tested URM, UTM, and PEO on these uniform-residual data sets. For URM and UTM, the parameters of the prior distribution K and λ were selected via cross-validation, where about 70% of each dataset was used for training and 30% for validation. For PEO, we chose their λ to be the same value as UTM's, and set their ϵ to be $8, 6, 4, 2$, for $N = 25, 50, 100, 200$, respectively.

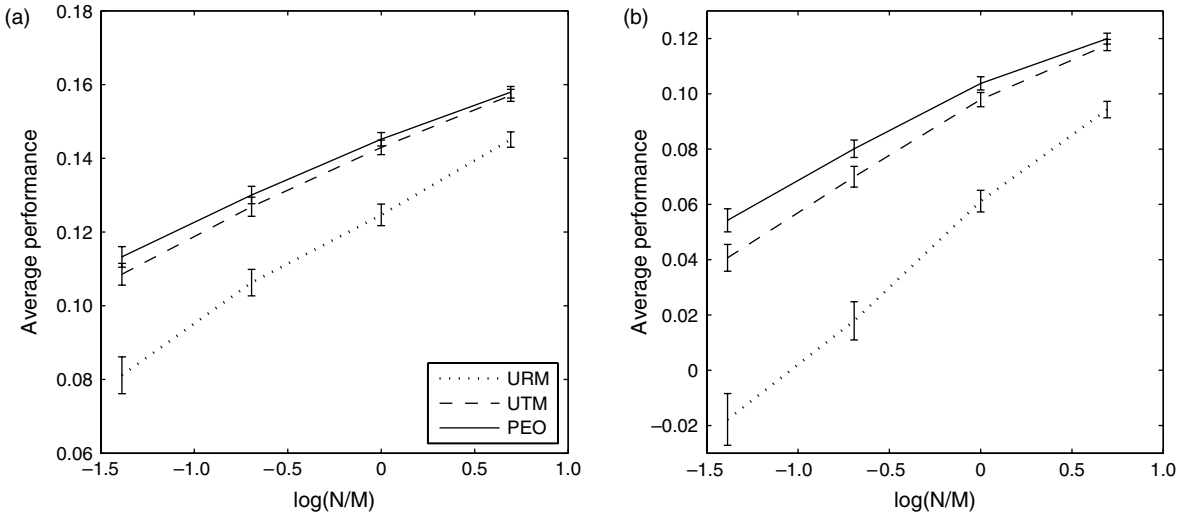
To generate an objective vector $\mathbf{c}$, we considered two cases: independent objective and aligned objective. In the first case, we simply sampled a $\mathbf{c}$ from the unit sphere in $\mathbb{R}^M$ isotropically. In the second case, we generated a $\mathbf{c}$ that was relatively more aligned with the top twenty primary factor loading vectors. Specifically, let us assume without loss of generality that $f_1 > f_2 > \cdots > f_M$. We sampled $p_1, p_2, \ldots, p_{20}$ iid from $\mathcal{N}(0, 4)$, and set $\check{\mathbf{c}}=\sum_{i=1}^{20} p_i\phi_i+\bar{\mathbf{c}}$, where $\bar{\mathbf{c}}$ was drawn from $\mathcal{N}(0, \mathbf{I}_M)$. This $\check{\mathbf{c}}$ was then normalized to produce a unit-length $\mathbf{c}$. We can see that such $\mathbf{c}$ vector tends to have larger projections on $\phi_1, \ldots, \phi_{20}$ than other directions. As we shall see in the results, directed PCA has more prominent advantage in such scenario. It is also worth mentioning that such scenario is not unusual in practice. For example, in portfolio management, $\mathbf{c}$ represents the expected return of assets and can often be weakly aligned with some market factors.

To compare the performance, we consider out-of-sample objective value delivered by the decision generated by each algorithm. Specifically, once each algorithm produces a covariance matrix estimate $\hat{\mathbf{\Sigma}}$, we compute a decision $\hat{\mathbf{u}}=\frac{1}{2}\hat{\mathbf{\Sigma}}^{-1}\mathbf{c}$, and then evaluate the out-of-sample objective value by $\mathrm{E}[g(\mathbf{x}, \hat{\mathbf{u}})]=\mathbf{c}^{\mathrm{T}}\hat{\mathbf{u}}-\hat{\mathbf{u}}^{\mathrm{T}}\mathbf{\Sigma}_*\hat{\mathbf{u}}$. Figure 1 plots the average out-of-sample objective value delivered by each algorithm for the independent and aligned objective cases. Here the x-axis is the log-ratio of the number of samples to the number of variables. This measure represents the availability of data relative to the number of variables, and is expected to drive performance differences. It is easy to see that PEO has an advantage over UTM, and the advantage is particularly large when the objective vector is weakly aligned with primary factors. Indeed, the gain of PEO over UTM can be as high as 34% when the size of dataset is small.

Our second type of synthetic data was generated using a similar procedure except Step 4 was replaced by

$$\mathbf{\Sigma}_*=\mathbf{F}_*^{1/2}\mathbf{F}_*^{\mathrm{T}/2}+\mathrm{diag}(e^{r_1}, \ldots, e^{r_M}),$$

Figure 1. The average out-of-sample objective value delivered by URM, UTM, and PEO, for (a) independent objective, and (b) aligned objective.



Note. The error bars denote one standard deviation.

where $r_1, \ldots, r_M$ were drawn iid from $\mathcal{N}(0, 0.6^2)$. This way we effectively introduced moderate variation into residual variances. EM, STM, and SPEO were tested on this nonuniform residual data set, and Figure 2 plots the average out-of-sample objective value delivered by these algorithms for both independent and aligned objective vectors. These results has similar trend as in Figure 1, except a mild increase in the advantage of SPEO.

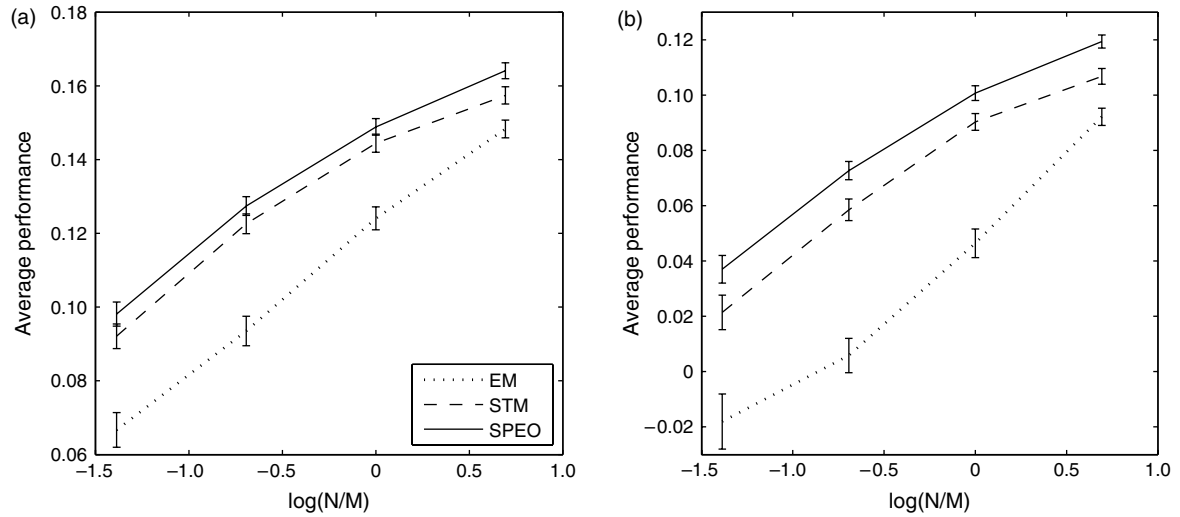
5.2. Real Data

Since portfolio optimization is an important application of directed PCA, in this subsection we present our experiment results for such setting using real stock return data. A typical portfolio optimization problem aims to select a portfolio that maximizes *certain-equivalent payoff*, defined as the expected payoff of the portfolio minus the variance of the payoff weighted by a risk-aversion coefficient. Let us denote the future returns of M assets by a vector $\mathbf{y} \in \mathbb{R}^M$, and denote the dollar amounts we would like to invest on these assets by a vector $\mathbf{u} \in \mathbb{R}^M$. We can write the objective as

$$\max_{\mathbf{u}} \mathrm{E}[\mathbf{u}^{\mathrm{T}}\mathbf{y}] - \xi \,\mathrm{Var}[\mathbf{u}^{\mathrm{T}}\mathbf{y}],$$

where the risk-aversion coefficient ξ is usually set to some small value in the order of 10^{-6}. Now suppose we are given a good estimate of the expected future returns $\mathrm{E}[\mathbf{y}]$, denoted by a vector $\hat{\mu} \in \mathbb{R}^M$. In general, the magnitude of $\hat{\mu}$ is much smaller than that of $\mathbf{y}$, and therefore $\mathrm{Var}[\mathbf{u}^{\mathrm{T}}\mathbf{y}] \simeq \mathrm{E}[(\mathbf{u}^{\mathrm{T}}\mathbf{y})^2]$.

Figure 2. The average out-of-sample objective value delivered by EM, STM, and SPEO, for (a) independent objective, and (b) aligned objective.



Based on this approximation, we can recast the objective by $\mathrm{E}[g(\mathbf{u}, \mathbf{y})]$, where g is defined in (1) with objective vector $\mathbf{c} = \xi^{-1}\hat{\mu}$. We now demonstrate how directed PCA can help generate better portfolio decision $\mathbf{u}$ in this scenario. It is worth mentioning that this scenario is indeed a common industry practice where multiple groups of researchers collaboratively work on different aspects of portfolio management. Specifically, one group of researchers would use various methods to produce an estimate for asset returns $\hat{\mu}$, while another group take this estimate as input and produce investment decisions based on a risk metric. Here we focus on the latter part of this process.

Our experiments involve estimation of covariance matrices from historical daily returns of stocks represented in the S&P 500 index as of March, 2011. We use price data collected from the period starting November 2, 2001, and ending August 9, 2007. This period was chosen to avoid the erratic market behavior observed during the bursting of the dot-com bubble in 2000 and the financial crisis that began in 2008. Daily returns were computed from closing prices while outliers were clipped. Over this duration, there were 1,450 trading days, indexed by $1, \ldots, 1450$, and 453 stocks under consideration. Let us denote the daily return of stock i on day t by $y_{i,t}$, and use $\mathbf{y}_{(t)} = [y_{1,t} \ \cdots \ y_{M,t}]^{\mathrm{T}}$, $M = 453$, to represent the return vector on day t. More details on this preprocessing procedure can be found in the appendix.

We generated estimates corresponding to each among a subset of the 1,450 days. As would be done in real-time application, for each such day t we used N data points $\{\mathbf{y}_{(t-N+1)}, \ldots, \mathbf{y}_{(t)}\}$ that would have been available on that day to compute the estimate and subsequent data to assess performance. In particular, we generated estimates every 20 days beginning on day 1,350 and ending on day 1,430. For each of these days, we evaluated the certain-equivalent payoff over the next 20 days. Figure 3 illustrates this sliding-window procedure.

However, those $\mathbf{y}_{(t)}$ vectors can hardly be regarded as stationary, mainly due to the fluctuation of volatility over time. Thus, instead of naively applying covariance learning algorithms on $\mathbf{y}_{(t)}$, we considered a more sophisticated approach. We first computed the 50-day-average volatility at each time point $\tau > 50$ for every asset i, formally defined as

$$\eta_{i,\tau} \triangleq \sqrt{\frac{1}{50}\sum_{j=1}^{50} y_{i,\tau-j}^2}.$$

We then used these quantities to normalize daily returns, resulting in normalized daily return vectors

$$\mathbf{x}_{(\tau)} \triangleq \left[\frac{y_{1,\tau}}{\eta_{1,\tau}} \ \cdots \ \frac{y_{M,\tau}}{\eta_{M,\tau}}\right]^{\mathrm{T}}.$$

The distribution of these $\mathbf{x}_{(\tau)}$ vectors is closer to being stationary than that of $\mathbf{y}_{(\tau)}$, and hence it makes more sense to apply covariance learning algorithms on $\mathbf{x}_{(\tau)}$.

Because of this normalization process, the objective vector $\mathbf{c}$ and decision $\mathbf{u}$ need special handling, too, as suggested in §4.3. Recall that our objective is to construct a decision $\mathbf{u}$ from $\{\mathbf{y}_{(t-N+1)}, \ldots, \mathbf{y}_{(t)}\}$ that aims to optimize the certain-equivalent payoff over the subsequent 20 days, which can be written as

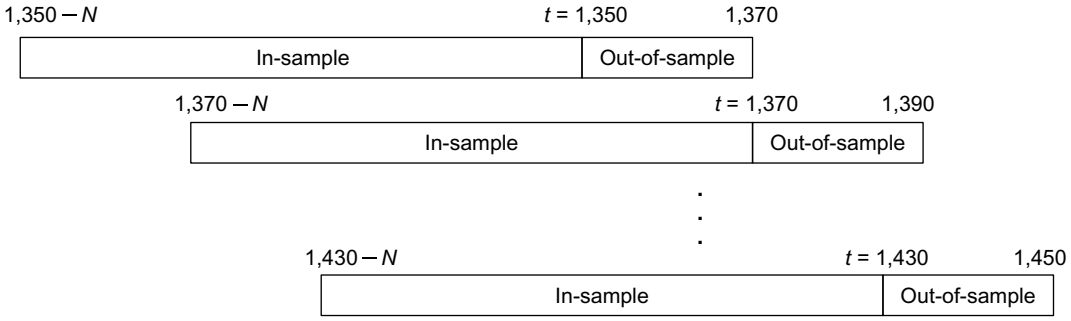
$$\mathbf{c}^{\mathrm{T}}\mathbf{u} - \tfrac{1}{20}\sum_{j=1}^{20}(\mathbf{u}^{\mathrm{T}}\mathbf{y}_{(t+j)})^2 = \mathbf{c}^{\mathrm{T}}\mathbf{u} - \tfrac{1}{20}\sum_{j=1}^{20}(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Upsilon}_{(t+j)}\mathbf{x}_{(t+j)})^2$$

where $\boldsymbol{\Upsilon}_{(\tau)} = \mathrm{diag}(\eta_{1,\tau}, \eta_{2,\tau}, \ldots, \eta_{M,\tau})$. Since $\eta_{i,\tau}$ is a 50-day moving-average, we have $\eta_{i,t} \simeq \eta_{i,t+j}$, for $j = 1, 2, \ldots, 20$, and therefore

$$\begin{aligned}
&\mathbf{c}^{\mathrm{T}}\mathbf{u} - \tfrac{1}{20}\sum_{j=1}^{20}(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Upsilon}_{(t+j)}\mathbf{x}_{(t+j)})^2 \\
&\quad\simeq \mathbf{c}^{\mathrm{T}}\mathbf{u} - \tfrac{1}{20}\sum_{j=1}^{20}(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Upsilon}_{(t)}\mathbf{x}_{(t+j)})^2 \\
&\quad= (\boldsymbol{\Upsilon}_{(t)}^{-1}\mathbf{c})^{\mathrm{T}}(\boldsymbol{\Upsilon}_{(t)}\mathbf{u}) - \tfrac{1}{20}\sum_{j=1}^{20}((\boldsymbol{\Upsilon}_{(t)}\mathbf{u})^{\mathrm{T}}\mathbf{x}_{(t+j)})^2.
\end{aligned}$$

Letting $\mathbf{c}' = \boldsymbol{\Upsilon}_{(t)}^{-1}\mathbf{c}$ and $\mathbf{u}' = \boldsymbol{\Upsilon}_{(t)}\mathbf{u}$, we can rewrite the last equation in the canonical form $\mathbf{c}'^{\mathrm{T}}\mathbf{u}' - \frac{1}{20}\sum_{j=1}^{20}(\mathbf{u}'^{\mathrm{T}}\mathbf{x}_{(t+j)})^2$, which suggests we apply various learning algorithms to normalized returns $\{\mathbf{x}_{(t-N+1)}, \ldots, \mathbf{x}_{(t)}\}$ with scaled objective

Figure 3. The sliding-window procedure for testing.



vectors $\mathbf{c}'$, and scale the resulting decisions $\mathbf{u}'$ to arrive at final decisions $\mathbf{u} = \boldsymbol{\Upsilon}_{(t)}^{-1}\mathbf{u}'$. Algorithm 2 formalizes this procedure. For each learning algorithm $\mathscr{A}$, we took the average of these five 20-day averages to be its out-of-sample performance, defined as

$$\frac{1}{5}\sum_{j=0}^{4}\mathscr{T}(\hat{\mu}, \mathscr{A}, N, 1350+20j).$$

In our implementation, we set $\xi = 10^{-6}$, $\epsilon = 40$, and the regularization parameters K and λ were selected by a cross-validation procedure, whose details can be found in the appendix.

Algorithm 2 (Testing Procedure $\mathscr{T}$)

Input: expected returns $\hat{\mu}$, learning algorithm $\mathscr{A}$, window size N, time point t

Output: average certain-equivalent payoff over test period $t+1, \ldots, t+20$

$$\mathscr{X} \leftarrow \left\{ \mathbf{x}_{(\tau)} \,\middle|\, \mathbf{x}_{(\tau)} = \left[\frac{y_{1,\tau}}{\eta_{1,\tau}} \cdots \frac{y_{M,\tau}}{\eta_{M,\tau}} \right]^{\mathrm{T}}, \tau = t-N+1, \ldots, t \right\}$$

// normalized returns

$$\mathbf{c}' \leftarrow \xi^{-1}\left[\frac{\hat{\mu}_1}{\eta_{1,t}} \cdots \frac{\hat{\mu}_M}{\eta_{M,t}} \right]^{\mathrm{T}}$$

// scale objective vector by the latest volatility

$$\hat{\boldsymbol{\Sigma}} \leftarrow \mathscr{A}(\mathscr{X}, \mathbf{c})$$

$$\mathbf{u}' \leftarrow \tfrac{1}{2}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{c}$$

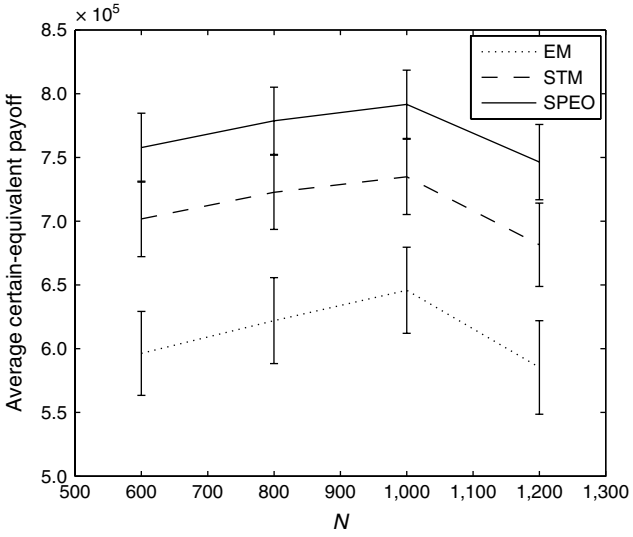
$$\mathbf{u} \leftarrow \left[\frac{\mathbf{u}'_1}{\eta_{1,t}} \cdots \frac{\mathbf{u}'_M}{\eta_{M,t}} \right]^{\mathrm{T}}$$

// scale the decision by the latest volatility

return $\hat{\mu}^{\mathrm{T}}\mathbf{u} - \xi \cdot \frac{1}{20}\sum_{\tau=t+1}^{t+20}(\mathbf{u}^{\mathrm{T}}\mathbf{y}_{(\tau)})^2$.

To evaluate the efficacy of our method for a wide range of possible scenarios, we repeated the above procedure one hundred times, each with a different expected returns vector $\hat{\mu}$ randomly sampled from $\mathscr{N}(0, 10^{-6}\mathbf{I})$. Figure 4 plots the average certain-equivalent payoff delivered by EM, STM, and SPEO for each $N \in \{600, 800, 1{,}000, 1{,}200\}$. SPEO is the dominant solution, generally outperforming the runner-up STM by 7%. It is worth noting that the performance of each algorithm peaks at $N = 1{,}000$. If the time series were stationary, one would expect performance to monotonically improve with N, as we have observed in the synthetic data experiment. However, this is a real time series and might not be perfectly stationary. We believe that its distribution changes enough over about a thousand trading days so that using historical data collected further back degrades the estimates. This observation indeed suggests that in real applications SPEO is likely to deliver significantly superior performance than STM or EM even when a large amount of data is provided. This is in contrast with the synthetic data experiment, which may have led to an impression that the performance difference could be made arbitrarily small by using more data.

Figure 4. The average certain-equivalent payoff delivered by EM, STM, and SPEO, for 100 randomly generated objectives.



6. Analysis

Through computational studies, we have demonstrated that directed PCA leads to better decisions than conventional PCA. In this section we provide an analysis that helps to explain the sources of improvement. We will focus on analysis of the uniform-residual case, though extension to nonuniform-residual case is straightforward. Our analysis will focus on a comparison between PEO and UTM, since PEO can be viewed as a variation of UTM that takes the decision objective into account and both PEO and UTM outperform URM. For further discussion comparing UTM against URM, we refer the reader to Kao and Van Roy (2013).

Let us start by developing some intuition for what directed PCA does. Recall that PEO aims to maximize $\tilde{g}(\boldsymbol{\Sigma})$ while maintaining high $p(\boldsymbol{\Sigma} \mid \mathscr{X})$. Since $\boldsymbol{\Sigma}_{\mathrm{SAM}}$ maximizes $\tilde{g}(\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_{\mathrm{UTM}}^{\lambda}$ maximizes $p(\boldsymbol{\Sigma} \mid \mathscr{X})$, we can think of $\boldsymbol{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon}$ as an estimate that deviates from $\boldsymbol{\Sigma}_{\mathrm{UTM}}^{\lambda}$ toward $\boldsymbol{\Sigma}_{\mathrm{SAM}}$ in order to increase $\tilde{g}(\boldsymbol{\Sigma})$. Furthermore, because the value of $\tilde{g}(\boldsymbol{\Sigma})$ is more sensitive to changes along the direction of $\mathbf{c}\mathbf{c}^{\mathrm{T}}$, such deviations tend to be larger for components that are aligned with $\mathbf{c}$. The following example illustrates this property.

Suppose

$$\boldsymbol{\Sigma}_{\mathrm{SAM}} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

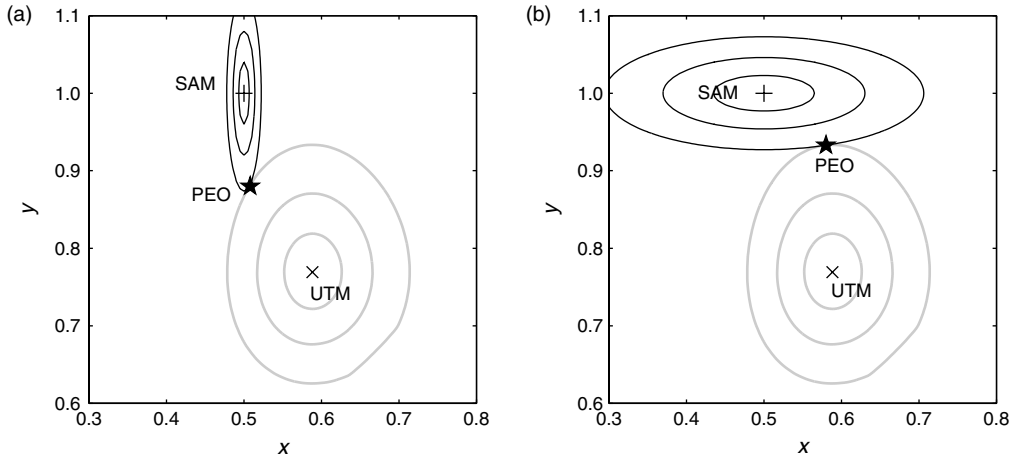
and $2\lambda/N = 0.3$. By (11), we have

$$\boldsymbol{\Sigma}_{\mathrm{UTM}} = \begin{bmatrix} 1.7 & 0 \\ 0 & 1.3 \end{bmatrix}.$$

To simplify illustration, let us restrict attention to covariance matrices that take the form

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}.$$

Figure 5. The level sets of $p(\mathbf{\Sigma} \mid \mathscr{X})$ is given by the gray curves, whereas the black curves plot the level sets of $\tilde{g}(\mathbf{\Sigma})$ for (a) $\mathbf{c} = [0.8\ 0.2]^{\mathrm{T}}$ and (b) $\mathbf{c} = [0.2\ 0.8]^{\mathrm{T}}$.



Note. The cross, plus sign, and star represent $\mathbf{\Sigma}_{\mathrm{UTM}}$, $\mathbf{\Sigma}_{\mathrm{SAM}}$, and $\mathbf{\Sigma}_{\mathrm{PEO}}$, respectively.

Figure 5(a) plots the level sets of $\tilde{g}(\mathbf{\Sigma})$ and $p(\mathbf{\Sigma} \mid \mathscr{X})$ in the x-y plane, for an objective vector $\mathbf{c} = [0.8\ 0.2]^{\mathrm{T}}$ that is closely aligned with $[1\ 0]^{\mathrm{T}}$. We label $\mathbf{\Sigma}_{\mathrm{UTM}}$, $\mathbf{\Sigma}_{\mathrm{SAM}}$, and $\mathbf{\Sigma}_{\mathrm{PEO}}$ in the same plot. It is easy to see that the displacement between PEO and SAM is much smaller in the x-direction than in the y-direction. On the other hand, Figure 5(b) plots the same contours for an objective vector $\mathbf{c} = [0.2\ 0.8]^{\mathrm{T}}$, and in this case PEO deviates from UTM mostly in y-direction.

Let us now move on to establish more formal results. We will consider an idealized, analytically tractable scenario in which the sample covariance matrix $\mathbf{\Sigma}_{\mathrm{SAM}}$ turns out to be identical to $\mathbf{\Sigma}_*$. As we shall see, such simplifying assumption can largely facilitate our analysis and lead to directly interpretable results.

Let an eigendecomposition of $\mathbf{\Sigma}_*$ be $\mathbf{ALA}^{\mathrm{T}}$, where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M]$ is orthonormal and $\mathbf{L} = \mathrm{diag}(l_1, l_2, \ldots, l_M)$, with $l_1 > l_2 > \cdots > l_K > l_{K+1} = l_{K+2} = \cdots = l_M = \sigma_*^2$. Recall that we evaluate the quality of an estimate $\hat{\mathbf{\Sigma}}$ by the out-of-sample performance of the resulting decision $\frac{1}{2}\hat{\mathbf{\Sigma}}^{-1}\mathbf{c}$. Let us denote such performance measure by a function

$$\mathscr{G}(\hat{\mathbf{\Sigma}}) = \mathrm{E}\big[g\big(\tfrac{1}{2}\hat{\mathbf{\Sigma}}^{-1}\mathbf{c}, \mathbf{x}\big)\big].$$

Under the simplifying assumption that $\mathbf{\Sigma}_{\mathrm{SAM}} = \mathbf{\Sigma}_*$, we can demonstrate the advantage of PEO over UTM by the following result.

Proposition 1. *If* $\mathbf{c} \in \{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$, *for any* $\lambda > 0$ *such that* $\mathscr{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}) \neq \sup_{\mathbf{\Sigma}} \mathscr{G}(\mathbf{\Sigma})$, *we have*

$$\mathscr{G}(\mathbf{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon}) > \mathscr{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}), \quad \forall\, \epsilon > 0.$$

Furthermore, the gap of this inequality monotonically increases with ϵ.

This proposition indicates that, under the simplifying assumption, if the objective vector is an eigenvector of $\mathbf{\Sigma}_*$, and the estimate produced by UTM is not already optimal, then PEO generally outperforms UTM. To further quantify the degree of this improvement, we now give a pair of results that illustrate how the improvement can be arbitrarily large as the data dimension grows. These results will be based on an additional condition on the λ parameter. Specifically, we will assume λ is fixed to $M\sigma_*^2$ from now on. Such setting is recommended by Kao and Van Roy (2013), who has shown through random matrix theory that this parameter choice optimizes the prediction accuracy of UTM estimate in terms of the expected log-likelihood of out-of-sample data.

Using this additional condition, our next result identifies an asymptotic regime where the improvement of PEO over UTM grows with the data dimension.

Proposition 2. *Fixing* N, K, *and* σ_*^2, *consider a sequence of covariance matrices* $\mathbf{\Sigma}_*^{(M)}$ *and objective vectors* $\mathbf{c}^{(M)}$, *indexed by the dimension* M, *that satisfy* $l_i^{(M)} \in [2\lambda/N - \delta_i, 2\lambda/N + \delta_i]$ *for* $i = 1, 2, \ldots, K$ *and constants* $\delta_1, \ldots, \delta_K$. *If* $\mathbf{c}^{(M)} \in \{\mathbf{a}_1^{(M)}, \mathbf{a}_2^{(M)}, \ldots, \mathbf{a}_K^{(M)}\}$, *we have*

$$\lim_{\epsilon \to \infty} \mathscr{G}(\mathbf{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon}) - \mathscr{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}) = \Omega(M).$$

The above result implies the performance gap between PEO and UTM is particularly large when the objective vector is aligned with the factor loading vectors, as we have observed in our experiment results. For the opposite case where the objective vector is orthogonal to the factor loading vectors, we can still illustrate the advantage of PEO over UTM in terms of performance ratio, as formerly described below.

Proposition 3. *Fixing* N *and* σ_*^2, *consider a sequence of covariance matrices* $\mathbf{\Sigma}_*^{(M)}$ *and objective vectors* $\mathbf{c}^{(M)}$, *indexed by the dimension* M, *that satisfy* $K^{(M)} = \lceil \alpha M \rceil$, *where* $\alpha \in (0, 1)$ *is a constant, and* $l_i^{(M)} > 2\lambda/N + \sigma_*^2$

for $i = 1, 2, \ldots, K^{(M)}$. *If* $\mathbf{c}^{(M)} \perp \text{span}\{\mathbf{a}_1^{(M)}, \mathbf{a}_2^{(M)}, \ldots, \mathbf{a}_{K^{(M)}}^{(M)}\}$, *we have*

$$\lim_{\epsilon \to \infty} \frac{\mathcal{G}(\mathbf{\Sigma}_{\text{PEO}}^{\lambda,\epsilon})}{\mathcal{G}(\mathbf{\Sigma}_{\text{UTM}}^{\lambda})} = \Omega(M).$$

This pair of results together suggest that directed PCA generally offers substantial improvement over conventional PCA methods for high-dimensional problems, and indeed manifest the importance of accounting for the decision objective in the estimation process.

7. Conclusion

We have proposed a new approach to covariance matrix estimation, which we refer to as *directed PCA*. The idea is to produce a covariance matrix estimate that corresponds to a factor model with factor loadings and residual variances estimated in a way that optimizes in-sample performance of the resulting decision strategy subject to a constraint that the model explains the data well. Such a method effectively incorporates the decision objective into the model fitting procedure. Computational and theoretical analyses demonstrate that our approach indeed outperforms conventional methods.

There is a growing body of research on variations of PCA, including methods that generate sparse factor loadings (Jolliffe et al. 2003, Zou et al. 2006, D'Aspremont et al. 2004, Johnstone and Lu 2009, Amini and Wainwright 2009) and methods that are resistant to corrupted data (Pison et al. 2003, Candès et al. 2009, Xu et al. 2010). It would be interesting to explore how to incorporate decision objectives into those settings. Furthermore, the portfolio management application we consider in this paper has a relatively simple setting, whereas more sophisticated variations have been proposed (see Steinbach 2001 for a detailed survey). Extending our method to deal with those settings is also a potential direction for future research.

Acknowledgments

The authors would like to thank Stephen Boyd for his pointers and discussion on ADMM. This work was supported in part by the National Science Foundation Award CMMI-0968707.

Appendix A. Proof

THEOREM 1. *For any* $\mathbf{\Sigma} \in \mathbb{S}_{++}^M$, $\mathbf{\Sigma} = \mathbf{\Sigma}_\gamma$ *if and only if it is a fixed point that satisfies*

$$\mathbf{\Sigma} = \mathcal{F}_\lambda(\mathbf{\Sigma}_{\text{SAM}} + \gamma \mathbf{C} \otimes \mathbf{D}),$$

where $\mathbf{C} = \mathbf{c}\mathbf{c}^{\mathrm{T}}$, $\mathbf{D} = \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\text{SAM}} - \mathbf{I}$, *and* $\mathbf{C} \otimes \mathbf{D} \triangleq \frac{1}{2}(\mathbf{C}\mathbf{D} + \mathbf{D}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}})$.

PROOF. We first show that $\mathbf{\Sigma}_\gamma$ satisfies (14). Expanding $\tilde{g}(\mathbf{\Sigma})$ and $\log p(\mathscr{X} \mid \mathbf{\Sigma})$, we can re-write (13) as

$$\begin{aligned} \min_{\mathbf{G}, v} \quad & \gamma\big(\tfrac{1}{2}\mathbf{c}^{\mathrm{T}}(v\mathbf{I} - \mathbf{G})\mathbf{\Sigma}_{\text{SAM}}(v\mathbf{I} - \mathbf{G})\mathbf{c} - \mathbf{c}^{\mathrm{T}}(v\mathbf{I} - \mathbf{G})\mathbf{c}\big) \\ & - \log\det(v\mathbf{I} - \mathbf{G}) + \text{tr}\big((v\mathbf{I} - \mathbf{G})\mathbf{\Sigma}_{\text{SAM}}\big) + \lambda' \text{tr}(\mathbf{G}) \\ \text{s.t.} \quad & \mathbf{G} \in \mathbb{S}_+^M, \end{aligned}$$

where $\lambda' = 2\lambda/N$. Note that the constraint $v \geqslant 0$ is implied in the domain of the objective function. We associate a Lagrange multiplier $\mathbf{\Omega} \in \mathbb{S}_+^M$ with the $\mathbf{G} \in \mathbb{S}_+^M$ constraint and write down the Lagrangian as

$$\begin{aligned} \mathscr{L}(\mathbf{G}, v, \mathbf{\Omega}) = {} & \gamma\big(\tfrac{1}{2}\mathbf{c}^{\mathrm{T}}(v\mathbf{I} - \mathbf{G})\mathbf{\Sigma}_{\text{SAM}}(v\mathbf{I} - \mathbf{G})\mathbf{c} - \mathbf{c}^{\mathrm{T}}(v\mathbf{I} - \mathbf{G})\mathbf{c}\big) \\ & - \log\det(v\mathbf{I} - \mathbf{G}) + \text{tr}\big((v\mathbf{I} - \mathbf{G})\mathbf{\Sigma}_{\text{SAM}}\big) \\ & + \lambda' \text{tr}(\mathbf{G}) - \text{tr}(\mathbf{\Omega}\mathbf{G}). \end{aligned}$$

Let $(\mathbf{G}_\gamma, v_\gamma)$ be the optimal solution, and let $\mathbf{\Omega}_\gamma$ be the corresponding Lagrangian multiplier. By KKT conditions we have:

$$\begin{aligned} & \nabla_{\mathbf{G}} \mathscr{L}\big|_{\mathbf{G}_\gamma, v_\gamma, \mathbf{\Omega}_\gamma} \\ & \quad = \gamma\big(\mathbf{c}\mathbf{c}^{\mathrm{T}} - \tfrac{1}{2}\mathbf{c}\mathbf{c}^{\mathrm{T}}(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)\mathbf{\Sigma}_{\text{SAM}} - \tfrac{1}{2}\mathbf{\Sigma}_{\text{SAM}}(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)\mathbf{c}\mathbf{c}^{\mathrm{T}}\big) \\ & \qquad + (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} - \mathbf{\Sigma}_{\text{SAM}} + \lambda'\mathbf{I} - \mathbf{\Omega}_\gamma = 0, \end{aligned} \tag{A1}$$

$$\begin{aligned} & \frac{\partial \mathscr{L}}{\partial v}\bigg|_{\mathbf{G}_\gamma, v_\gamma, \mathbf{\Omega}_\gamma} \\ & \quad = -\gamma \,\text{tr}\big(\mathbf{c}\mathbf{c}^{\mathrm{T}} - \tfrac{1}{2}\mathbf{c}\mathbf{c}^{\mathrm{T}}(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)\mathbf{\Sigma}_{\text{SAM}} - \tfrac{1}{2}\mathbf{\Sigma}_{\text{SAM}}(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)\mathbf{c}\mathbf{c}^{\mathrm{T}}\big) \\ & \qquad + \text{tr}(\mathbf{\Sigma}_{\text{SAM}}) - \text{tr}((v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}) = 0, \end{aligned} \tag{A2}$$

$$\mathbf{\Omega}_\gamma, \mathbf{G}_\gamma \in \mathbb{S}_+^M, \tag{A3}$$

$$\text{tr}(\mathbf{\Omega}_\gamma\mathbf{G}_\gamma) = 0. \tag{A4}$$

Using the fact that $\mathbf{\Sigma}_\gamma = (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}$ and $\mathbf{D} = \mathbf{\Sigma}_\gamma^{-1}\mathbf{\Sigma}_{\text{SAM}} - \mathbf{I}$, we can rewrite (A1) as

$$(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} - \gamma\mathbf{C} \otimes \mathbf{D} - \mathbf{\Sigma}_{\text{SAM}} + \lambda'\mathbf{I} - \mathbf{\Omega}_\gamma = 0$$

and rewrite (A2) as

$$\text{tr}\big(\gamma\mathbf{C} \otimes \mathbf{D} + \mathbf{\Sigma}_{\text{SAM}} - (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}\big) = 0.$$

To simplify notation, let us define $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}_{\text{SAM}} + \gamma\mathbf{C} \otimes \mathbf{D}$, and further rewrite (A1) and (A2) as

$$\hat{\mathbf{\Sigma}} = (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} + \lambda'\mathbf{I} - \mathbf{\Omega}_\gamma, \tag{A5}$$

$$\text{tr}(\hat{\mathbf{\Sigma}}) = \text{tr}\big((v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}\big) \tag{A6}$$

Let an eigendecomposition of $\mathbf{G}_\gamma$ be $\mathbf{A}\mathbf{Q}\mathbf{A}^{\mathrm{T}}$ for which $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ is orthonormal. Plugging this into (A4) we get

$$0 = \text{tr}(\mathbf{\Omega}_\gamma\mathbf{G}_\gamma) = \text{tr}(\mathbf{\Omega}_\gamma\mathbf{A}\mathbf{Q}\mathbf{A}^{\mathrm{T}}) = \text{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{\Omega}_\gamma\mathbf{A}\mathbf{Q}) = \sum_{i=1}^M \mathbf{Q}_{i,i}\mathbf{a}_i^{\mathrm{T}}\mathbf{\Omega}_\gamma\mathbf{a}_i.$$

By (A3), $\mathbf{Q}_{i,i} \geqslant 0$ and $\mathbf{a}_i^{\mathrm{T}}\mathbf{\Omega}_\gamma\mathbf{a}_i \geqslant 0$, $\forall\, i = 1, \ldots, M$, which implies

$$\mathbf{a}_i^{\mathrm{T}}\mathbf{\Omega}_\gamma\mathbf{a}_i = 0 \quad \text{if } \mathbf{Q}_{i,i} > 0, \quad \forall\, i = 1, \ldots, M.$$

Let $\mathscr{I}_+ = \{i\colon \mathbf{Q}_{i,i} > 0\}$. Since $\mathbf{\Omega}_\gamma$ is positive semidefinite, for all $i_0 \in \mathscr{I}_+$ we also have $\mathbf{\Omega}_\gamma\mathbf{a}_{i_0} = 0$. Furthermore, since

$$(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} = \mathbf{A}\,\text{diag}\bigg(\frac{1}{v_\gamma - \mathbf{Q}_{1,1}}, \ldots, \frac{1}{v_\gamma - \mathbf{Q}_{M,M}}\bigg)\mathbf{A}^{\mathrm{T}},$$

multiplying (A5) by $\mathbf{a}_{i_0}$ leads to

$$\hat{\mathbf{\Sigma}}\mathbf{a}_{i_0} = \frac{\mathbf{a}_{i_0}}{v_\gamma - \mathbf{Q}_{M,M}} + \lambda'\mathbf{a}_{i_0} = \bigg(\frac{1}{v_\gamma - \mathbf{Q}_{i_0,i_0}} + \lambda'\bigg)\mathbf{a}_{i_0} \tag{A7}$$

which shows $\mathbf{a}_{i_0}$ is an eigenvector of $\hat{\mathbf{\Sigma}}$. Now suppose $\check{\mathbf{A}}\mathbf{S}\check{\mathbf{A}}^{\mathrm{T}}$ is an eigendecomposition of $\hat{\mathbf{\Sigma}}$ such that $\check{\mathbf{A}} = [\check{\mathbf{a}}_1 \ \dots \ \check{\mathbf{a}}_M]$ is orthonormal and $\check{\mathbf{a}}_i = \mathbf{a}_i$ for all $i \in \mathcal{I}_+$. Let us define $\mathcal{I}' = \{1, 2, \dots, M\} \backslash \mathcal{I}_+ = \{i\colon \mathbf{Q}_{i,i} = 0\}$. Note that

$$\begin{aligned}\mathbf{G}_\gamma &= \sum_{i\in\mathcal{I}_+} \mathbf{Q}_{i,i}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}} + \sum_{i\in\mathcal{I}'} \mathbf{Q}_{i,i}\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}} = \sum_{i\in\mathcal{I}_+} \mathbf{Q}_{i,i}\check{\mathbf{a}}_i\check{\mathbf{a}}_i^{\mathrm{T}} + \sum_{i\in\mathcal{I}'} 0\cdot\mathbf{a}_i\mathbf{a}_i^{\mathrm{T}} \\ &= \sum_{i\in\mathcal{I}_+} \mathbf{Q}_{i,i}\check{\mathbf{a}}_i\check{\mathbf{a}}_i^{\mathrm{T}} + \sum_{i\in\mathcal{I}'} 0\cdot\check{\mathbf{a}}_i\check{\mathbf{a}}_i^{\mathrm{T}},\end{aligned}$$

which implies $\check{\mathbf{A}}$ consists of the eigenvectors of $\mathbf{G}_\gamma$. Therefore, without loss of generality we can assume $\check{\mathbf{A}} = \mathbf{A}$, and write $\hat{\mathbf{\Sigma}} = \mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}$. Comparing this with (A7), we have

$$\mathbf{S}_{i_0,i_0} = \frac{1}{v_\gamma - \mathbf{Q}_{i_0,i_0}} + \lambda', \quad \forall\, i_0 \in \mathcal{I}_+. \tag{A8}$$

Recall that $\mathbf{\Sigma}_\gamma = (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} = (v_\gamma\mathbf{I} - \mathbf{A}\mathbf{Q}\mathbf{A}^{\mathrm{T}})^{-1} = \mathbf{A}\mathbf{H}\mathbf{A}^{\mathrm{T}}$, where $\mathbf{H}_{i,i} = 1/(v_\gamma - \mathbf{Q}_{i,i})$, for $i = 1, \dots, M$. Comparing this expression with (A8) we arrive at

$$\mathbf{H}_{i_0,i_0} = \mathbf{S}_{i_0,i_0} - \lambda', \quad \forall\, i_0 \in \mathcal{I}_+,$$

or more generally

$$\mathbf{H}_{i,i} = \begin{cases} \mathbf{S}_{i,i} - \lambda' & \text{if } \mathbf{Q}_{i,i} > 0, \\ \dfrac{1}{v_\gamma} & \text{otherwise,} \end{cases} \qquad i = 1, \dots, M.$$

Since $\mathbf{H}_{i,i} \geqslant 1/v_\gamma$, to see $\mathbf{H}_{i,i} = \max\{\mathbf{S}_{i,i} - \lambda', 1/v_\gamma\}$, it remains to show $\mathbf{H}_{i,i} \geqslant \mathbf{S}_{i,i} - \lambda'$ for all i. This follows by rearranging (A5)

$$\begin{aligned}&(v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1} - \hat{\mathbf{\Sigma}} + \lambda'\mathbf{I} = \mathbf{\Omega}_\gamma \succeq 0 \\ &\quad\Rightarrow\ \mathbf{A}\mathbf{H}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\lambda'\mathbf{I}\mathbf{A}^{\mathrm{T}} \succeq 0 \ \Rightarrow\ \mathbf{H} - \mathbf{S} + \lambda'\mathbf{I} \succeq 0 \\ &\quad\Rightarrow\ \mathbf{H}_{i,i} \geqslant \mathbf{S}_{i,i} - \lambda', \quad \forall\, i = 1, \dots, M.\end{aligned}$$

Finally, by (A6) we know $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}_\gamma$ share the same trace, and thus $\mathcal{F}_\lambda(\hat{\mathbf{\Sigma}}) = \mathbf{\Sigma}_\gamma$, as desired.

We now prove the reverse direction. Suppose $\tilde{\mathbf{\Sigma}}$ is a matrix in $\mathbb{S}^M_{++}$ that satisfies

$$\mathcal{F}_\lambda(\mathbf{\Sigma}_{\mathrm{SAM}} + \gamma\mathbf{C} \otimes \tilde{\mathbf{D}}) = \tilde{\mathbf{\Sigma}},$$

where $\tilde{\mathbf{D}} = \tilde{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}_{\mathrm{SAM}} - \mathbf{I}$. Let an eigendecomposition of $\mathbf{\Sigma}_{\mathrm{SAM}} + \gamma\mathbf{C} \otimes \tilde{\mathbf{D}}$ be $\tilde{\mathbf{A}}\tilde{\mathbf{S}}\tilde{\mathbf{A}}^{\mathrm{T}}$, where $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1 \ \dots \ \tilde{\mathbf{a}}_M]$ is orthonormal and $\tilde{\mathbf{S}}_{1,1} \geqslant \tilde{\mathbf{S}}_{2,2} \geqslant \cdots \geqslant \tilde{\mathbf{S}}_{M,M}$. By the definition of $\mathcal{F}_\lambda$, we know $\tilde{\mathbf{\Sigma}}$ has an eigendecomposition $\tilde{\mathbf{A}}\tilde{\mathbf{H}}\tilde{\mathbf{A}}$, where the eigenvalues satisfy

$$\begin{aligned}&\sum_{i=1}^M \tilde{\mathbf{H}}_{i,i} = \sum_{i=1}^M \tilde{\mathbf{S}}_{i,i}, \quad \text{and} \\ &\tilde{\mathbf{H}}_{i,i} = \max\left\{\tilde{\mathbf{S}}_{i,i} - \lambda', \frac{1}{\tilde{v}}\right\}, \quad i = 1, \dots, M,\end{aligned}$$

for some scalar $\tilde{v}$. Thus, there exists an integer $\tilde{K} \in [0, M-1]$ such that

$$\begin{aligned}&\tilde{\mathbf{H}}_{i,i} = \tilde{\mathbf{S}}_{i,i} - \lambda' \geqslant \frac{1}{\tilde{v}}, \quad \text{for } i = 1, 2, \dots, \tilde{K}, \quad \text{and} \\ &\tilde{\mathbf{H}}_{i,i} = \frac{1}{\tilde{v}} \geqslant \tilde{\mathbf{S}}_{i,i} - \lambda', \quad \text{for } i = \tilde{K}+1, \dots, M.\end{aligned}$$

Since $\tilde{\mathbf{\Sigma}} \in \mathbb{S}^M_{++}$, we have $\tilde{\mathbf{H}}_{i,i} > 0$ for all i and therefore $\tilde{v} > 0$. Now let $\tilde{\mathbf{G}} = \sum_{i=1}^{\tilde{K}} (\tilde{v} - 1/\tilde{\mathbf{H}}_{i,i})\tilde{\mathbf{a}}_i\tilde{\mathbf{a}}_i^{\mathrm{T}}$ and $\tilde{\mathbf{\Omega}} = \sum_{i=\tilde{K}+1}^M (\tilde{\mathbf{H}}_{i,i} - \tilde{\mathbf{S}}_{i,i} + \lambda')\tilde{\mathbf{a}}_i\tilde{\mathbf{a}}_i^{\mathrm{T}}$. It is easy to see that $\tilde{\mathbf{G}}, \tilde{\mathbf{\Omega}} \in \mathbb{S}^M_+$, and $\mathrm{tr}(\tilde{\mathbf{\Omega}}\tilde{\mathbf{G}}) = 0$. Since $\tilde{\mathbf{\Sigma}} = (\tilde{v}\mathbf{I} - \tilde{\mathbf{G}})^{-1}$, we also have

$$\begin{aligned}&(\tilde{v}\mathbf{I} - \tilde{\mathbf{G}})^{-1} - \gamma\mathbf{C} \otimes \tilde{\mathbf{D}} - \mathbf{\Sigma}_{\mathrm{SAM}} + \lambda'\mathbf{I} - \tilde{\mathbf{\Omega}} \\ &\qquad = \tilde{\mathbf{A}}\tilde{\mathbf{H}}\tilde{\mathbf{A}} - \tilde{\mathbf{A}}\tilde{\mathbf{S}}\tilde{\mathbf{A}} + \lambda'\mathbf{I} - \tilde{\mathbf{\Omega}} = 0, \quad \text{and} \\ &\mathrm{tr}(\gamma\mathbf{C} \otimes \tilde{\mathbf{D}} + \mathbf{\Sigma}_{\mathrm{SAM}} - (\tilde{v}\mathbf{I} - \tilde{\mathbf{G}})^{-1}) = \sum_{i=1}^M \tilde{\mathbf{S}}_{i,i} - \sum_{i=1}^M \tilde{\mathbf{H}}_{i,i} = 0.\end{aligned}$$

In other words, $(\tilde{\mathbf{G}}, \tilde{v}, \tilde{\mathbf{\Omega}})$ satisfy the KKT conditions (A1)–(A4). Since the primal problem is convex, KKT conditions are sufficient for optimality of $(\tilde{\mathbf{G}}, \tilde{v})$, and hence $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_\gamma$. □

To prove Proposition 1, we first prove the following lemma.

LEMMA 1. *If $\mathbf{c}$ is an eigenvector of $\mathbf{\Sigma}_{\mathrm{SAM}}$, then $\mathbf{\Sigma}^{\lambda,\epsilon}_{\mathrm{PEO}}$ and $\mathbf{\Sigma}_{\mathrm{SAM}}$ share the same eigenvectors.*

PROOF. Recall that with appropriate choice of $\gamma \geqslant 0$, we have $\mathbf{\Sigma}^{\lambda,\epsilon}_{\mathrm{PEO}} = (v_\gamma\mathbf{I} - \mathbf{G}_\gamma)^{-1}$, where $(\mathbf{G}_\gamma, v_\gamma)$ is the solution to (13). One standard approach for solving (13) is interior-point method with log barrier. This involves absorbing the $\mathbf{G} \succeq 0$ constraint into a term $(1/t)\log\det(\mathbf{G})$ and iteratively solving

$$\max_{\mathbf{G}, v} \gamma\tilde{g}(\mathbf{\Sigma}) + \log p(\mathscr{X} \mid \mathbf{\Sigma}) - \lambda\,\mathrm{tr}(\mathbf{G}) + \frac{1}{t}\log\det(\mathbf{G}) \tag{A9}$$

for an increasing sequence of $t > 0$. As t goes to infinity, its solution converges to that of (13). Without loss of generality, let us start this iterative procedure with an initial point $(\mathbf{G}_0, v_0) = (\mathbf{I}_M, 2)$. Obviously $\mathbf{G}_0$ has the same eigenvectors as $\mathbf{\Sigma}_{\mathrm{SAM}}$. Let us denote the objective of (A9) by $\mathscr{L}_t(\mathbf{G}, v)$. Then we have

$$\begin{aligned}&\nabla_{\mathbf{G}}\mathscr{L}_t(\mathbf{G}, v) \\ &\quad = -\frac{\gamma N}{2}\left(\mathbf{c}\mathbf{c}^{\mathrm{T}} - \tfrac{1}{2}\mathbf{c}\mathbf{c}^{\mathrm{T}}(v\mathbf{I} - \mathbf{G})\mathbf{\Sigma}_{\mathrm{SAM}} - \tfrac{1}{2}\mathbf{\Sigma}_{\mathrm{SAM}}(v\mathbf{I} - \mathbf{G})\mathbf{c}\mathbf{c}^{\mathrm{T}}\right) \\ &\qquad + \frac{N}{2}(\mathbf{\Sigma}_{\mathrm{SAM}} - (v\mathbf{I} - \mathbf{G})^{-1}) - \lambda\mathbf{I} + \frac{1}{t}\mathbf{G}^{-1}.\end{aligned}$$

Recall that, if two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ share the same eigenvectors, then both $\mathbf{A} + \mathbf{B}$ and $\mathbf{A}\mathbf{B}$ have the same eigenvectors, too. Since $\mathbf{c}\mathbf{c}^{\mathrm{T}}$, $v_0\mathbf{I} - \mathbf{G}_0$, $\mathbf{\Sigma}_{\mathrm{SAM}}$, $(v_0\mathbf{I} - \mathbf{G}_0)^{-1}$, and $\mathbf{G}_0^{-1}$ are all symmetric and have the same eigenvectors, we know $\nabla_{\mathbf{G}}\mathscr{L}_t|_{(\mathbf{G}_0, v_0)}$ also has the same eigenvectors. Therefore, if we update parameter $\mathbf{G}$ by this gradient, the next $\mathbf{G}$ we arrive at will also have the same eigenvectors. By induction, such eigen-structure is invariant over each iteration. In other words, when we solve (A9) by gradient ascent method, the resultant $\mathbf{G}$ will still have the same eigenvectors as $\mathbf{\Sigma}_{\mathrm{SAM}}$.

Since this argument is independent of the value of t, as t goes to infinity, such result still holds. Therefore, $\mathbf{G}_\gamma$ has the same eigenvectors as $\mathbf{\Sigma}_{\mathrm{SAM}}$, and so does $\mathbf{\Sigma}^{\lambda,\epsilon}_{\mathrm{PEO}}$. □

PROPOSITION 1. *If $\mathbf{c} \in \{\mathbf{a}_1, \dots, \mathbf{a}_M\}$, for any $\lambda > 0$ such that $\mathcal{G}(\mathbf{\Sigma}^\lambda_{\mathrm{UTM}}) \neq \sup_{\mathbf{\Sigma}} \mathcal{G}(\mathbf{\Sigma})$, we have*

$$\mathcal{G}(\mathbf{\Sigma}^{\lambda,\epsilon}_{\mathrm{PEO}}) > \mathcal{G}(\mathbf{\Sigma}^\lambda_{\mathrm{UTM}}), \quad \forall\, \epsilon > 0.$$

Furthermore, the gap of this inequality monotonically increases with ϵ.

Proof. By (13), let us rewrite $\mathbf{\Sigma}^{\lambda,\epsilon}_{\text{PEO}}$ as $\mathbf{\Sigma}_\gamma$, where γ is a scalar monotonically increasing with ϵ, and $\gamma = 0$ when $\epsilon = 0$. Since $\mathbf{\Sigma}_{\text{SAM}} = \mathbf{\Sigma}_* = \mathbf{ALA}^{\text{T}}$, by (11) we have $\mathcal{F}_\lambda(\mathbf{\Sigma}_*) = \mathbf{\Sigma}^\lambda_{\text{UTM}}$ and therefore $\mathbf{\Sigma}^\lambda_{\text{UTM}} = \mathbf{AWA}^{\text{T}}$, where $\mathbf{W} \in \mathbb{D}^M_+$ is the soft-thresholded version of $\mathbf{L}$. Furthermore, by Lemma 1 we know $\mathbf{\Sigma}_\gamma$ also has the same eigenvectors $\mathbf{A}$. Let us denote $\mathbf{\Sigma}_\gamma$ by $\mathbf{AH}_\gamma\mathbf{A}^{\text{T}}$, where $\mathbf{H}_\gamma = \text{diag}(h_1(\gamma), \ldots, h_M(\gamma))$, and define $\mathbf{W} = \text{diag}(w_1, \ldots, w_M)$. Obviously $\mathbf{H}_0 = \mathbf{W}$.

Suppose $\mathbf{c} = \mathbf{a}_{i_0}$. By Theorem 1, we know

$$\mathcal{F}_\lambda(\mathbf{\Sigma}_{\text{SAM}} + \gamma\mathbf{C} \otimes \mathbf{D}) = \mathbf{\Sigma}_\gamma,$$

where $\mathbf{C} = \mathbf{cc}^{\text{T}} = \mathbf{a}_{i_0}\mathbf{a}_{i_0}^{\text{T}}$ and $\mathbf{D} = \mathbf{\Sigma}_\gamma^{-1}\mathbf{\Sigma}_{\text{SAM}} - \mathbf{I}$. Since $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{\Sigma}_{\text{SAM}}$ all have the same eigenvectors $\mathbf{A}$, we can remove it and write

$$\mathcal{F}_\lambda\left(\mathbf{L} + \gamma\left(\frac{l_{i_0}}{h_{i_0}(\gamma)} - 1\right)\mathbf{e}_{i_0}\mathbf{e}_{i_0}^{\text{T}}\right) = \mathbf{H}_\gamma.$$

Let $f_\gamma(h)$ be the i_0th diagonal entry of matrix

$$\mathcal{F}_\lambda\left(\mathbf{L} + \gamma\frac{l_{i_0}}{h-1}\mathbf{e}_{i_0}\mathbf{e}_{i_0}^{\text{T}}\right).$$

Then $h_{i_0}(\gamma)$ is a solution to the equation $f_\gamma(h) = h$. Also note that $f_\gamma(h)$ is continuous and monotonically decreases with h.

Now consider three cases:

1. $w_{i_0} < l_{i_0}$: We have

$$f_\gamma(l_{i_0}) = w_{i_0} < l_{i_0}$$

and

$$f_\gamma(w_{i_0}) = \text{the } i_0\text{th diagonal entry of matrix } \mathcal{F}_\lambda\left(\mathbf{L} + \gamma\left(\frac{l_{i_0}}{w_{i_0}} - 1\right)\mathbf{e}_{i_0}\mathbf{e}_{i_0}^{\text{T}}\right) > \text{the } i_0\text{th diagonal entry of matrix } \mathcal{F}_\lambda(\mathbf{L}) = w_{i_0}.$$

This is equivalent to $f_\gamma(w_{i_0}) - w_{i_0} > 0$ and $f_\gamma(l_{i_0}) - l_{i_0} < 0$. Thus, by Intermediate Value Theorem, we know there exists $h' \in (w_{i_0}, l_{i_0})$ that satisfies

$$f_\gamma(h') = h'.$$

Furthermore, by the monotonicity of $f_\gamma(h)$, we know the solution to $f_\gamma(h) = h$ is unique, and therefore $h' = h_{i_0}(\gamma)$.

But for all $h \in (w_{i_0}, l_{i_0})$, we have $(l_{i_0}/h - 1) > 0$, which implies

$$f_{\gamma_1}(h) > f_{\gamma_2}(h), \quad \forall\, \gamma_1 > \gamma_2.$$

Therefore, the root of equation $f_\gamma(h) = h$ monotonically increases with γ. This implies that $h_{i_0}(\gamma)$ monotonically increases from w_{i_0} toward l_{i_0}, as γ increases from 0 toward ∞.

Recall that

$$\frac{1}{2}(\mathbf{\Sigma}^\lambda_{\text{UTM}})^{-1}\mathbf{c} = \frac{\mathbf{a}_{i_0}}{2w_{i_0}}$$

and therefore

$$\mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) = \mathbf{c}^{\text{T}}\left(\frac{\mathbf{a}_{i_0}}{2w_{i_0}}\right) - \left(\frac{\mathbf{a}_{i_0}}{2w_{i_0}}\right)^{\text{T}}\mathbf{\Sigma}_*\left(\frac{\mathbf{a}_{i_0}}{2w_{i_0}}\right) = \frac{1}{2w_{i_0}} - \frac{l_{i_0}}{4w_{i_0}^2} = \frac{1}{4l_{i_0}} - l_{i_0}\left(\frac{1}{2w_{i_0}} - \frac{1}{2l_{i_0}}\right)^2.$$

Similarly,

$$\mathcal{G}(\mathbf{\Sigma}_\gamma) = \frac{1}{4l_{i_0}} - l_{i_0}\left(\frac{1}{2h_{i_0}(\gamma)} - \frac{1}{2l_{i_0}}\right)^2$$

and thus

$$\mathcal{G}(\mathbf{\Sigma}_\gamma) - \mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) = l_{i_0}\left(\left(\frac{1}{2w_{i_0}} - \frac{1}{2l_{i_0}}\right)^2 - \left(\frac{1}{2h_{i_0}(\gamma)} - \frac{1}{2l_{i_0}}\right)^2\right),$$

which is positive and monotonically increases as γ increases from 0 toward ∞ and $h_{i_0}(\gamma)$ increases from w_{i_0} toward l_{i_0}.

2. $w_{i_0} = l_{i_0}$: In this case, $\mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) = \mathcal{G}(\mathbf{\Sigma}_*) = \sup_{\mathbf{\Sigma}} \mathcal{G}(\mathbf{\Sigma})$, which is precluded from the assumption.

3. $w_{i_0} > l_{i_0}$: Similarly to case 1, we can see that $h_{i_0}(\gamma)$ monotonically decreases from w_{i_0} toward l_{i_0} as γ increases from 0 toward ∞, and therefore the desired results follow. □

Proposition 2. *Fixing N, K, and σ_*^2, consider a sequence of covariance matrices $\mathbf{\Sigma}_*^{(M)}$ and objective vectors $\mathbf{c}^{(M)}$, indexed by the dimension M, that satisfy $l_i^{(M)} \in [2\lambda/N - \delta_i, 2\lambda/N + \delta_i]$ for $i = 1, 2, \ldots, K$ and constants $\delta_1, \ldots, \delta_K$. If $\mathbf{c}^{(M)} \in \{\mathbf{a}_1^{(M)}, \mathbf{a}_2^{(M)}, \ldots, \mathbf{a}_K^{(M)}\}$, we have*

$$\lim_{\epsilon\to\infty} \mathcal{G}(\mathbf{\Sigma}^{\lambda,\epsilon}_{\text{PEO}}) - \mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) = \Omega(M).$$

Proof. Suppose $\mathbf{c}^{(M)} = \mathbf{a}^{(M)}_{i_M}$, $i_M \leqslant K$. By the derivation in Proposition 1, we have $\lim_{\epsilon\to\infty} \mathcal{G}(\mathbf{\Sigma}^{\lambda,\epsilon}_{\text{PEO}}) = \lim_{\gamma\to\infty} \mathcal{G}(\mathbf{\Sigma}_\gamma) = 1/(4l^{(M)}_{i_M})$, whereas $\mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}})$ only depends on the i_Mth eigenvalue of $\mathbf{\Sigma}^\lambda_{\text{UTM}}$, for which we now derive an upper bound.

Let $\hat{\sigma}^2_{(M)}$ be the smallest eigenvalue of $\mathbf{\Sigma}^\lambda_{\text{UTM}}$. Since

$$\hat{\sigma}^2_{(M)} \leqslant \frac{1}{M}\,\text{tr}(\mathbf{\Sigma}^{(M)}_*) = \frac{1}{M}\left(\sum_{i=1}^K l_i^{(M)} + (M-K)\sigma_*^2\right) \leqslant \frac{1}{M}\left(\frac{2K\lambda}{N} + \sum_{i=1}^K \delta_i + (M-K)\sigma_*^2\right) = \frac{1}{M}\left(\frac{2KM\sigma_*^2}{N} + M\sigma_*^2 + \text{constant}\right),$$

there exists M_0 such that $\hat{\sigma}^2_{(M)} \leqslant (2K/N+1)\sigma_*^2 + 1$ for all $M > M_0$. Now let the i_Mth eigenvalue of $\mathbf{\Sigma}^\lambda_{\text{UTM}}$ be h_M. By (11), we have

$$h_M = \max\left\{l^{(M)}_{i_M} - \frac{2\lambda}{N}, \hat{\sigma}^2_{(M)}\right\} \leqslant \max\{\delta_{i_M}, \hat{\sigma}^2_{(M)}\}.$$

Letting $h' = \max\{\delta_1, \ldots, \delta_K, (2K/N+1)\sigma_*^2 + 1\}$, we further have $h_M \leqslant h'$, $\forall\, M > M_0$. Since h' is a constant and $l_1^{(M)}, \ldots, l_K^{(M)} = \Omega(M)$, there exists an integer $M_1 > M_0$ such that $l_i^{(M)} > h' > h_M$ for all $i = 1, \ldots, K$ and $M > M_1$. Therefore,

$$\mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) = \frac{1}{4l^{(M)}_{i_M}} - l^{(M)}_{i_M}\left(\frac{1}{2h_M} - \frac{1}{2l^{(M)}_{i_M}}\right)^2 \leqslant \frac{1}{4l^{(M)}_{i_M}} - l^{(M)}_{i_M}\left(\frac{1}{2h'} - \frac{1}{2l^{(M)}_{i_M}}\right)^2, \quad \forall\, M > M_1$$

and

$$\lim_{\epsilon\to\infty} \mathcal{G}(\mathbf{\Sigma}^{\lambda,\epsilon}_{\text{PEO}}) - \mathcal{G}(\mathbf{\Sigma}^\lambda_{\text{UTM}}) \geqslant l^{(M)}_{i_M}\left(\frac{1}{2h'} - \frac{1}{2l^{(M)}_{i_M}}\right)^2, \quad \forall\, M > M_1.$$

The desired result then follows from the fact $l^{(M)}_{i_M} = \Omega(M)$. □

PROPOSITION 3. *Fixing N and σ_*^2, consider a sequence of covariance matrices $\mathbf{\Sigma}_*^{(M)}$ and objective vectors $\mathbf{c}^{(M)}$, indexed by the dimension M, that satisfy $K^{(M)} = \lceil \alpha M \rceil$, where $\alpha \in (0,1)$ is a constant, and $l_i^{(M)} > 2\lambda/N + \sigma_*^2$ for $i = 1, 2, \ldots, K^{(M)}$. If $\mathbf{c}^{(M)} \perp \operatorname{span}\{\mathbf{a}_1^{(M)}, \mathbf{a}_2^{(M)}, \ldots, \mathbf{a}_{K^{(M)}}^{(M)}\}$, we have*

$$\lim_{\epsilon\to\infty} \frac{\mathcal{G}(\mathbf{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon})}{\mathcal{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda})} = \Omega(M).$$

PROOF. Since

$$\mathbf{\Sigma}_*^{(M)} = \mathbf{A}\mathbf{L}\mathbf{A}^{\mathrm{T}} = \sum_{i=1}^{K^{(M)}} (l_i^{(M)} - \sigma_*^2)\mathbf{a}_i^{(M)}\mathbf{a}_i^{(M)\mathrm{T}} + \sigma_*^2\mathbf{I} \quad \text{and}$$

$$\mathbf{c}^{(M)} \perp \operatorname{span}\{\mathbf{a}_1^{(M)}, \mathbf{a}_2^{(M)}, \ldots, \mathbf{a}_K^{(M)}\},$$

we know $\mathbf{c}^{(M)}$ is an eigenvector of $\mathbf{\Sigma}_*^{(M)}$ with eigenvalue σ_*^2. Without loss of generality, let $\mathbf{c}^{(M)} = \mathbf{a}_{i_M}^{(M)}$, $i_M > K$. By the derivation in Proposition 1, we have $\lim_{\epsilon\to\infty} \mathcal{G}(\mathbf{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon}) = \lim_{\gamma\to\infty} \mathcal{G}(\mathbf{\Sigma}_\gamma) = 1/(4\sigma_*^2)$, whereas $\mathcal{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda})$ only depends on the i_Mth eigenvalue of $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}$, for which we now derive a lower bound.

Since $\mathbf{\Sigma}_*^{(M)}$ has $K^{(M)}$ outstanding eigenvalues and the remaining $M - K^{(M)}$ ones equal to σ_*^2, the matrix $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda} = \mathcal{F}_\lambda(\mathbf{\Sigma}_*^{(M)})$ can have at most $K^{(M)}$ outstanding eigenvalues. Let the number of outstanding eigenvalues of $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}$ be $\hat{K}^{(M)}$. Since $i_M \in \{K^{(M)}+1, K^{(M)}+2, \ldots, M\}$, we know the i_Mth eigenvalue of $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}$ will be the smallest eigenvalue of $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}$, and by the trace-preservation property of $\mathcal{F}_\lambda$, the smallest eigenvalue of $\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}$ will be

$$\begin{aligned}
\hat{\sigma}_{(M)}^2 &= \frac{1}{M-\hat{K}^{(M)}}\left(\frac{2\hat{K}^{(M)}\lambda}{N} + \sum_{i=\hat{K}^{(M)}+1}^{K^{(M)}} l_i^{(M)} + (M-K^{(M)})\sigma_*^2\right)\\
&> \frac{1}{M-\hat{K}^{(M)}}\left(\frac{2\hat{K}^{(M)}\lambda}{N} + (K^{(M)}-\hat{K}^{(M)})\left(\frac{2\lambda}{N}+\sigma_*^2\right) + (M-K^{(M)})\sigma_*^2\right)\\
&= \frac{1}{M-\hat{K}^{(M)}}\left(\frac{2\hat{K}^{(M)}\lambda}{N} + (K^{(M)}-\hat{K}^{(M)})\frac{2\lambda}{N}\right) + \sigma_*^2\\
&= \frac{1}{M-\hat{K}^{(M)}}\left(\frac{2K^{(M)}\lambda}{N}\right) + \sigma_*^2\\
&= \frac{1}{M-\lceil\alpha M\rceil}\left(\frac{2\lceil\alpha M\rceil M\sigma_*^2}{N}\right) + \sigma_*^2\\
&\geqslant \frac{1}{M-\alpha M}\left(\frac{2\alpha M^2\sigma_*^2}{N}\right) + \sigma_*^2 = \frac{2\alpha}{N(1-\alpha)}\cdot M\sigma_*^2 + \sigma_*^2.
\end{aligned}$$

Therefore, there exists a constant $\beta > 0$ such that

$$\hat{\sigma}_{(M)}^2 > (\beta M + 1)\sigma_*^2. \tag{A10}$$

Recall that

$$\mathcal{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}) = \frac{1}{2\hat{\sigma}_{(M)}^2} - \frac{\sigma_*^2}{4\hat{\sigma}_{(M)}^4} = \frac{1}{2\hat{\sigma}_{(M)}^2}\left(1 - \frac{\sigma_*^2}{2\hat{\sigma}_{(M)}^2}\right).$$

Plugging (A10) into this we have

$$0 < \mathcal{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda}) < \frac{1}{M}\cdot\frac{1}{2\beta\sigma_*^2},$$

and therefore

$$\lim_{\epsilon\to\infty} \frac{\mathcal{G}(\mathbf{\Sigma}_{\mathrm{PEO}}^{\lambda,\epsilon})}{\mathcal{G}(\mathbf{\Sigma}_{\mathrm{UTM}}^{\lambda})} = \Omega(M). \quad \square$$

Appendix B. Experiment Details

B.1. Implementation of Algorithm 1

Let us denote the objective function of (13) by $f(\mathbf{G}, v)$. Algorithm 3 describes the backtracking line search algorithm we used in our implementation of Algorithm 1. Note that by the construction of $(\Delta\mathbf{G}, \Delta v)$, we know $\mathbf{G} + \Delta\mathbf{G} \in \mathbb{S}_+^M$ and therefore $(\mathbf{G} + \alpha\Delta\mathbf{G}) \in \mathbb{S}_+^M$, $\forall\, \alpha \in [0,1]$. Thus, we do not need to worry about violating the positive-semidefinite constraint in this line search. Algorithm 4 further gives the details of our projected gradient ascent method.

Algorithm 3 (Backtracking Line Search)
Input: Starting point $(\mathbf{G}, v)$ and search direction $(\Delta\mathbf{G}, \Delta v)$
Output: α

```
α ← 1
while f((G, v) + α(ΔG, Δv)) < f(G, v) + 0.1
        · α∇f(G, v)^T(ΔG, Δv) do
  α ← α/2
  if α < 10^−6 then
    return 0   // This update size is too small to be
                  meaningful.
  end if
end while
return α.
```

Algorithm 4 (Projected Gradient Ascent)
Input: Starting point $(\mathbf{G}, v)$
Output: $(\hat{\mathbf{G}}, \hat{v})$

```
α ← 1
repeat
  (G′, v̂) ← (G, v) + α∇f(G, v)
  Let an eigendecomposition of G′ be AQA^T for which
    A = [a_1 ... a_M] is orthonormal
  Ĝ ← Σ_{i=1}^M max{Q_{i,i}, 0} a_m a_m^T
        // Project G′ onto S_+^M by thresholding its eigenvalues
  α ← α/2
until f(Ĝ, v̂) > f(G, v)
return (Ĝ, v̂).
```

For the termination criterion "$(v\mathbf{I} - \mathbf{G})^{-1}$ satisfies (14)", we considered the following heuristics: We say a matrix $\mathbf{\Sigma} \in \mathbb{S}_+^M$ satisfies (14) if

$$\left\|\mathbf{\Sigma} - \mathcal{F}_\lambda(\mathbf{\Sigma}_{\mathrm{SAM}} + \gamma\mathbf{C}\otimes\mathbf{D})\right\|_F < 10^{-3}\|\mathbf{\Sigma}\|_F.$$

B.2. Choosing Regularization Parameters via Cross Validation

For the synthetic data experiment, we select the regularization parameter via the following cross-validation procedure. Let θ be the regularization parameter to be determined and $\mathcal{A}(\mathcal{X}, \theta)$ be the learning algorithm that takes as input $(\mathcal{X}, \theta)$ and returns a covariance matrix estimate. We randomly split $\mathcal{X}$ into a partial training set $\mathcal{X}_{\mathrm{T}}$ and a validation set $\mathcal{X}_{\mathrm{V}}$, whose sizes are roughly 70% and 30% of $\mathcal{X}$, respectively. For each candidate value of θ, $\hat{\mathbf{\Sigma}}_{\mathrm{T}}^{\theta} = \mathcal{A}(\mathcal{X}_{\mathrm{T}}, \theta)$ is computed and the likelihood $p(\mathcal{X}_{\mathrm{V}} \mid \hat{\mathbf{\Sigma}}_{\mathrm{T}}^{\theta})$ of the validation set $\mathcal{X}_{\mathrm{V}}$ conditioned on the solution $\hat{\mathbf{\Sigma}}_{\mathrm{T}}^{\theta}$ is evaluated. The value of θ that maximizes this likelihood is then selected and fed into $\mathcal{A}(\mathcal{X}, \theta)$ along with the full training set $\mathcal{X}$, resulting in our estimate $\hat{\mathbf{\Sigma}}^{\theta}$. In our implementation, the K for URM/EM are selected from $\{0, 1, \ldots, 20\}$, and the λ for UTM/STM are selected from $\{70, 80, \ldots, 300\}$. These ranges are chosen so that the selected values rarely fall on the extremes.

For the real data experiment, we used the sliding-window validation procedure as described in Algorithm 5. For each algorithm $\mathscr{A} \in \{\text{EM}, \text{STM}\}$, its regularization parameter was selected by

$$\hat{\theta} = \arg\max_{\theta} \sum_{j=0}^{4} \mathscr{V}(\mathscr{A}, \theta, N, 1{,}250 + 20j).$$

In our implementation, the K for EM is selected from $\{12, 13, \ldots, 22\}$, and the λ for STM is selected from $\{360, 380, \ldots, 540\}$.

Algorithm 5 (Validation Procedure $\mathscr{V}$)
Input: learning algorithm $\mathscr{A}$, regularization parameter θ, window size N, time point t
Output: log-likelihood of validation set

$$\mathscr{X}_{\text{T}} \leftarrow \left\{ \mathbf{x}_{(\tau)} \,\middle|\, \mathbf{x}_{(\tau)} = \left[\frac{y_{1,\tau}}{\eta_{1,\tau}} \cdots \frac{y_{M,\tau}}{\eta_{M,\tau}} \right]^{\text{T}}, \tau = t - N + 1, \ldots, t \right\}$$
// training set,

$$\mathscr{X}_{\text{V}} \leftarrow \left\{ \mathbf{x}_{(\tau)} \,\middle|\, \mathbf{x}_{(\tau)} = \left[\frac{y_{1,\tau}}{\eta_{1,t}} \cdots \frac{y_{M,\tau}}{\eta_{M,t}} \right]^{\text{T}}, \tau = t + 1, \ldots, t + 20 \right\}$$
// validation set,

$\hat{\boldsymbol{\Sigma}} \leftarrow \mathscr{A}(\mathscr{X}_{\text{T}}, \theta)$
return $\log p(\mathscr{X}_{\text{V}} \mid \hat{\boldsymbol{\Sigma}})$.

B.3. S&P 500 Data Preprocessing

Define November 2, 2001 as trading day 1 and August 9, 2007 as trading day 1,451. After deleting 47 constituent stocks that are not fully defined over this period, we compute for each stock the daily returns as follows:

1. Let $y'_{i,j}$ be the adjusted close price of stock i on day j, $i = 1, \ldots, 453$ and $j = 1, \ldots, 1{,}451$.

2. Compute the raw daily-return of stock i on day j by

$$y''_{i,j} = \frac{y'_{i,j+1}}{y'_{i,j}} - 1, \quad i = 1, \ldots, 453,\ j = 1, \ldots, 1{,}450.$$

3. Let $\bar{y}$ be the smallest number such that at least 99.5% of all $y''_{i,j}$ are less than or equal to $\bar{y}$. Let $\underline{y}$ be the largest number such that at least 99.5% of all $y''_{i,j}$'s are greater than or equal to $\underline{y}$. Clip all $y''_{i,j}$ by the interval $[\underline{y}, \bar{y}]$, and denote the resulting value by $y_{i,j}$.

References

Amini AA, Wainwright MJ (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* 37(5B): 2877–2921.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization*, Princeton Series in Applied Mathematics (Princeton University Press, Princeton, NJ).

Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York).

Besbes O, Philips R, Zeevi A (2010) Testing the validity of a demand model: An operations perspective. *Manufacturing Service Oper. Management* 12(1):162–183.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learn.* 3(1):1–122.

Candès EJ, Li X, Ma Y, Wright J (2009) Robust principal component analysis? *J. ACM* 58(1):1–37.

Chu LY, Shanthikumar JG, Shen Z-JM (2008) Solving operational statistics via a Bayesian analysis. *Oper. Res. Lett.* 36(1):110–116.

D'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GRG (2004) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* 49(3):434–448.

Delage E, Ye Y (2008) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Math. Oper. Res.* 28(1):1–38.

Harman HH (1976) *Modern Factor Analysis*, 3rd ed. (University of Chicago Press, Chicago).

Johnstone IM, Lu AY (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* 104(486):682–693.

Jolliffe IT, Trendafilov NT, Uddin M (2003) A modified principal component technique based on the lasso. *J. Computational and Graphical Statist.* 12(3):531–547.

Kao Y-H, Van Roy B (2013) Estimating a factor model via regularized PCA. *Machine Learn.* 91(3):279–303.

Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. *Math. Programming* 45(3):503–528.

Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Oper. Res. Lett.* 33(4):341–348.

Pison G, Rousseeuw PJ, Filzmoser P, Croux C (2003) Robust factor analysis. *J. Multivariate Anal.* 84(1):145–172.

Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. *Oper. Res.* 55(1):98–112.

Rubin DB, Thayer DT (1982) EM algorithm for ML factor analysis. *Psychometrika* 47(1):69–76.

Steinbach MC (2001) Markowitz revisited: Mean-variance models in financial portfolio analysis. *SIAM Rev.* 43(1):31–85.

Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. *J. Roy. Statist. Soc., Ser. B* 61:611–622.

Xu H, Caramanis C, Mannor S (2010) Principal component analysis with contaminated data: The high dimensional case. *Proc. 23rd Conf. on Learning Theory, COLT 2010* (Omnipress, Madison, WI), 490–502.

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J. Computational and Graphical Statist.* 15(2):265–286.

Yi-Hao Kao received his Ph.D. (2012) in electrical engineering from Stanford University. His research interests include machine learning and optimization, with a particular focus on their applications in operations research. He was the recipient of the honorable mention of INFORMS George Nicholson Student Paper Competition in 2012, and was named a Stanford Graduate Fellow in 2007.

Benjamin Van Roy is a Professor of electrical engineering, management science and engineering, and, by courtesy, computer science, at Stanford University. His current research focusses on the design, analysis, and application of algorithms that learn over time to make effective decisions. He is a member of INFORMS and IEEE.