UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Prediction and Estimation in High Dimensions

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Varlam Kutateladze

June 2021

Dissertation Committee:

Dr. Tae-Hwy Lee, Chairperson
Dr. Jang-Ting Guo
Dr. Aman Ullah

The Dissertation of Varlam Kutateladze is approved:

Committee Chairperson

University of California, Riverside

# Acknowledgments

I would like to express my sincere gratitude to the dissertation committee.

I thank Professor Tae-Hwy Lee for teaching me advanced econometric techniques and forming my foundations in statistics; for inspiring to dive into machine learning; for his consistent guidance on a range of topics, from journal submissions to job market nuances; for always being available to help; for giving me freedom in pursuing fascinating research topics. I also thank Professor Jang-Ting Guo for the witty and memorable classes in macroeconomics, for his advice and support throughout the PhD and of course for admitting me into the program. I thank Professor Ullah for his flawlessly taught classes in nonparametrics and for his continuous support and help; for influencing my taste in econometric models and for his insightful comments on my research.

Finally, my most sincere thanks go to Gary Kuzas for all the big work he is doing for the department.

# ABSTRACT OF THE DISSERTATION

Prediction and Estimation in High Dimensions

by

Varlam Kutateladze

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2021
Dr. Tae-Hwy Lee, Chairperson

This dissertation examines some prediction and estimations problems that arise in "high dimensions", increasingly prevalent settings characterized by the presence of a large number of observations and a large number of variables.

Chapter 1 provides an overview and briefly discusses some challenges in a large-dimensional framework.

Chapter 2 considers factor modeling, an effective tool for extracting information from large panels of data, and extends the classical linear factor analytic approach to accommodate nonlinearities, which is made possible by employing the kernel method. This chapter also establishes the theoretical guarantees, discusses the generality of the proposed approach and considers a forecasting application.

Chapter 3 explores an estimation problem in the context of group testing. It proposes a methodology that is based on $\ell_1$-norm sparse recovery which explicitly leverages the fact that the high-dimensional vector of interest is likely to be sparse in certain applications. The theoretical properties are investigated and extensive numerical simulations are

provided.

Chapter 4 studies estimation of a large-dimensional covariance matrix. This chapter develops an estimator that is suitable for consistent estimation of large matrices with sparse eigenvectors and error components. It also derives the theoretical properties of the proposed method and provides a numerical experiment.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

This work examines prediction and estimation problems arising in “high dimensions”, increasingly prevalent settings characterized by the presence of a large number of observations and a large number of variables. The truth is, large-dimensional datasets are “cursed” in the sense that analyzing them appears to be impossible. In reality, however, most interesting datasets have an inner low-dimensional structure, which is of high value, that often succumbs to modern data scientific treatment. My research exploits this property and studies efficient ways of distilling massive amounts of information into valuable knowledge. Naturally, one must inevitably assume that there is something to extract: Chapter 2 assumes most variation in the data is induced by a few latent variables, while Chapters 3 and 4 assume the high-dimensional space is sparse, i.e. most dimensions are effectively negligible. Luckily, recent developments in machine learning and theoretical discoveries in statistics enable us to tackle such challenges. As large-dimensional datasets are increasingly prevalent in genomics, biomedical imaging, tomography, finance, economics and statistics, the solutions to such challenges would be of high interest to a broad audience.

Chapter 2 "*The Kernel Trick for Nonlinear Factor Modeling*," extends the classical (linear) factor-analytic framework to accommodate nonlinearities. Factor models have become an integral part of multivariate analysis and high-dimensional statistics, and have had a substantial effect on a number of different fields, including psychology ([117]), biology ([62]) and economics ([32]). In econometrics and finance, applications range from portfolio optimization ([50]) and covariance estimation ([51]) to forecasting ([111]). My work demonstrates that one can extract information from a large unstructured panel of data and use it for forecasting a series of interest. The main insight comes from the observation that most data resides in a low-dimensional space and can be described by a handful of variables known as factors. Importantly, my work permits to go beyond a linear factor model by employing the kernel method from machine learning literature. This enables handling the data as big as one could possibly store, let alone superior performance and faster computational speed. The study also establishes the theoretical guarantees and discusses the generality of the proposed approach. Finally, an empirical application to a classical macroeconomic dataset demonstrates that this approach can offer substantial advantages over mainstream methods. This manuscript is accepted for publication in the International Journal of Forecasting.

Chapter 3 "*Sparse Recovery for COVID-19 Group Testing*," examines group testing strategies against the coronavirus. With testing capacity restricted, group testing is an appealing alternative for comprehensive screening and has recently received FDA emergency authorization. We frame this as an estimation problem in high-dimensional context. Simply put, if one represents $N$ subjects' true COVID-19 statuses (positive or negative) in a vector, then it may be possible to infer this vector with $m \ll N$ tests based on pooled samples. We propose a methodology that is based on $\ell_1$-norm sparse recovery which explicitly leverages the fact that the vector of interest is likely to be sparse as most of the subjects will be negative (due to low disease prevalence). This strategy requires fewer testing supplies while providing multiple folds of speedup over individual testing. It is also more efficient than other group testing techniques. COVID-19 testing is not the only possible application, the technique developed in our study can be used in nontesting settings such as coding theory, multiaccess communication, screening for defective items, nonlinear optimization, etc. This paper under the name "*Fast and Efficient Data Science Techniques for COVID-19 Group Testing*," co-authored with Ekaterina Seregina, is published in the Journal of Data Science.

Chapter 4 "*High-Dimensional Covariance Estimation*," focuses on estimation of a central object in statistics – a covariance matrix. Covariance matrix estimates are required in a wide range of applied problems in multivariate data analysis, including principal components analysis ([95]) for dimensionality reduction, linear discriminant analysis ([55]) for classification, spectral clustering ([92]) for community detection. A number of disciplines rely on such estimates to address major challenges, for example portfolio & risk management ([52]) in finance, factor models and testing in economics ([5],[112]), graphical models ([57], [87]) in machine learning and discovering genetic interactions in genomics. In a large dimensional environment, nearly all desirable properties of the classical sample covariance estimator cease to hold. This recently motivated researchers to propose a new family of regularized, or shrinkage ([108]), estimators that essentially average the eigenvalues of the sample covariance matrix with that of a structured matrix. While this approach performs well in certain scenarios, it presumes that the sample eigenvectors remain close to their true counterparts, which in general is implausible. An estimator is proposed that aims to precisely estimate eigenvectors in sparse settings, without requiring strong assumptions on eigenvalues. I derive its rate of convergence in terms of spectral norm and show that it achieves the optimal rate in a sparse setting. I also provide a numerical simulation demonstrating the superior performance of the proposed estimator as compared to other recently proposed high-dimensional estimators.

# Chapter 2

# The Kernel Trick for Nonlinear Factor Modeling

## Abstract

Factor modeling is a powerful statistical technique that permits to capture the common dynamics in a large panel of data with a few latent variables, or factors, thus alleviating the curse of dimensionality. Despite its popularity and widespread use for various applications ranging from genomics to finance, this methodology has predominantly remained linear. This study estimates factors nonlinearly through the kernel method, which allows flexible nonlinearities while still avoiding the curse of dimensionality. We focus on factor-augmented forecasting of a single time series in

a high-dimensional setting, known as diffusion index forecasting in macroeconomics literature. Our main contribution is twofold. First, we show that the proposed estimator is consistent and it nests linear PCA estimator as well as some nonlinear estimators introduced in the literature as specific examples. Second, our empirical application to a classical macroeconomic dataset demonstrates that this approach can offer substantial advantages over mainstream methods.

## 2.1 Introduction

Over the past century, factor models have become an integral part of multivariate analysis and high-dimensional statistics, and have had a substantial effect on a number of different fields, including psychology ([117]), biology ([62]) and economics ([32]). In economics and finance, applications range from portfolio optimization ([50]) and covariance estimation ([51]) to forecasting ([111]). The application that we consider is macroeconomic forecasting, where various forms of factor analysis have become state-of-the-art techniques for prediction.

The general idea behind factor analysis consists of determining a few latent variables, or factors, that drive the dependence of the entire outcomes. Factors are designed to capture the common dynamics in a large panel of data. This feature is crucial in the context of increasing availability of macroeconomic time series coupled with the inability of standard econometric methods to handle many variables. While classical econometric tools break down in such data-rich or “big data” environments, factor models help to compress a large amount of the available information into a few factors, turning the curse of dimensionality into a blessing.

Factor analysis possesses several attractive properties that justify a large amount of literature in its support. First, it effectively handles large dimensions thereby enhancing forecast accuracy in such regimes. This was demonstrated in [111] and [112] who used so-called diffusion indexes, or factors, in forecasting models when dealing with a large number of predictors. More recently, [74] find that factor augmented models nearly always outperform a wide range of big data and machine learning models in terms of predictive power. Second, due to its conceptual simplicity, this methodology found its use beyond academic research. For example, the Federal Reserve Bank of Chicago constructs the Chicago Fed National Activity Index (CFNAI) simply as the first principal component of a large number of time series. Third, factor analysis aligns naturally with the dynamic equilibrium theories as well as the stylized fact of [100] of a small number of variables explaining most of the fluctuations in macroeconomic time series. And finally, factor estimates can also be used to provide efficient instruments for augmenting vector autoregressions (VARs) ([21]) to assist in tracing structural shocks.

Factors, however, are not observable and need to be estimated. Two classical estimation strategies rely on either intertemporal or contemporaneous smoothing. The former casts the model into a state-space representation and estimates it by the maximum likelihood via the Kalman filter. The disadvantages of this approach are that it requires parametric assumptions and that it quickly becomes computationally infeasible as the number of predictor series grows [1]. Contemporaneous smoothing is a more predominant and

[1] Interestingly, there has been some evidence to the contrary, see [46]

computationally simpler way based on principal component analysis (PCA) ([95]), nonparametric least-squares approach for estimating factors.

Forecasts are obtained via a two-step procedure. First, factors estimates are derived from the set of available time series by one of the two methods described above. Once the factors are estimated, run a linear autoregression of the variable of interest onto factor estimates and observed covariates (e.g. lagged values of the dependent variable).

We analyze a factor model that is high-dimensional, static and approximate. High-dimensional framework ([8]), as opposed to classical framework ([5]), allows both time and cross-section dimensions to grow. Static models do not explicitly model time-dependence of factors contrary to more general dynamic counterparts ([56]). Approximate factor structure ([32]) is more flexible compared with a strict version ([97]) as it imposes milder assumptions on the idiosyncratic component.

Factor analysis is closely related to PCA, although the two are not the same ([70]). It is, however, well documented that the two are asymptotically equivalent under suitable conditions (see the pervasiveness assumption in [51] for a recent treatment). There are several results on consistency of PCA estimators of factors ([36], [111], [9] among others) for various forms of factor models. One of the most relevant of the results is established in [8] who derive convergence rates of such estimators for an approximate static factor model of large dimensions.

Despite its widespread use, factor modeling, and diffusion index forecasting methodology in particular, is still fundamentally limited to linear framework. Over the past two decades, leading researchers noted multiple times (e.g. see [112], [10], [113], [35]) that further forecast improvements “will need to come from models with nonlinearities and/or time variation” and that “nonlinear factor-augmented regression should be considered” for forecasting. There have been attempts to incorporate time dependence (see, for example, [91], [89] and [39] among others) as well as some work documenting the superiority of nonlinear models in time series context ([116], [59], [73]). However, the literature on addressing nonlinearity within the factor modeling or diffusion index methodology framework, which arguably remains to be the state-of-the-art technique for macroeconomic prediction ([38]), is scarce.

One of the first such attempts is [126] who assume nonlinear errors-in-variables parametrization of a factor model, which is not designed for forecasting applications. The most prominent work with focus on a prediction exercise is [10]. They either augment the set of predictor time series with their squares and apply standard PC to the augmented set, or use squares of principal components obtained from the original (non-augmented) set. Another closely related work is [49] who substitute the linear second step with kernel ridge regression and discover that this leads to more accurate forecasts of the key economic indicators.

This study adds to the scarce literature on nonlinear factor models. Specifically, the factors are allowed to capture nontrivial functions of predictors. To circumvent the computational difficulties associated with such novelties, we use the kernel trick, or kernel method ([63]), which is discussed in the next section within the diffusion index methodology context.

The rest of the paper is organized as follows. Subsections 2.1 and 2.2 of Section 2 review the diffusion index methodology and the kernel trick; subsections 2.3 and 2.4 derive the estimators and provide the theoretical guarantees. Section 3 outlines the forecasting models, describes the data and provides the empirical results. Section 4 concludes and discusses possible extensions. All proofs are given in the Appendix.

**Notation.** For a vector $\mathrm{v} \in \mathbb{R}^d$, we write its $i$-th element as $v_i$. The corresponding $\ell_p$ norm is $\|\mathrm{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$, which is a norm for $1 \leq p \leq \infty$. An inner product between two vectors of the same dimension is $\langle \mathrm{v_i}, \mathrm{v_j} \rangle = \mathrm{v_i}'\mathrm{v_j}$. For a matrix $A \in \mathbb{R}^{m \times d}$, we write its $(i,j)$-th entry as $\{A\}_{ij} = a_{ij}$ and denote its $i$-th row (transposed) and $j$-th column as column vectors $A_{i\cdot}$ and $A_{\cdot j}$ respectively. Its singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A)$, where $q = \min(m, d)$. The spectral norm is $\|A\|_2 = \max\limits_{\mathrm{v} \neq 0} \frac{\|A\mathrm{v}\|_2}{\|\mathrm{v}\|_2} = \sigma_1(A)$. The $\ell_1$ norm is $\|A\|_1 = \max\limits_{1 \leq j \leq d} \sum_{i=1}^m |u_{ij}|$ and $\ell_\infty$ norm is $\|A\|_\infty = \max\limits_{1 \leq i \leq m} \sum_{j=1}^d |u_{ij}|$. The Frobenius norm is $\|A\|_F = \sqrt{\langle A, A \rangle} = \sqrt{tr(A'A)} = \sqrt{\sum_{i=1}^q \sigma_i^2(A)}$. For a symmetric matrix $W \in \mathbb{R}^{d \times d}$ with eigenvalues $\lambda_1(W) \geq \lambda_2(W) \geq \ldots \geq \lambda_d(W)$, define $eig_r(W) \in \mathbb{R}^{d \times r}$ to be a matrix stacking $r \leq d$ normalized eigenvectors in the order corresponding to $\lambda_1(W), \ldots, \lambda_r(W)$. Finally, for a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of real nonnegative numbers $\{a_n\}_{n=1}^\infty$, denote $X_n = O_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0, \exists M, N > 0$ such that $\forall n > N$, $\mathbb{P}(|X_n/a_n| \geq M) < \epsilon$; and denote $X_n = o_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0$, $\lim\limits_{n \to \infty} \mathbb{P}(|X_n/a_n| \geq \epsilon) = 0$. Finally, let $\mathbf{1}_{1/T}$ be a $T \times T$ matrix of ones divided by $T$.

## 2.2 Methodology

### 2.2.1 Diffusion Index Models

Our goal is to accurately forecast a scalar variable $Y_t$, given a $T \times N$ data matrix $X$ with $t$th row $X'_t$, or $X'_{t\cdot}$. Both the number of observations $T$ and the number of series $N$ are typically large.

Consider the following baseline model, known as a Diffusion Index (DI) model:

$$\underset{1\times 1}{Y_{t+h}} = \underset{1\times r}{\beta'_F}\underset{r\times 1}{F_t} + \underset{1\times p}{\beta'_W}\underset{p\times 1}{W_t} + \underset{1\times 1}{\epsilon_{t+h}}, \tag{2.1}$$

$$\underset{N\times 1}{X_t} = \underset{N\times r}{\Lambda}\underset{r\times 1}{F_t} + \underset{N\times 1}{e_t}. \tag{2.2}$$

Equation 2.1 is a linear forecasting model, where $Y_{t+h}$ is the value of the target variable $h$ periods in the future, $F_t$ is the vector of $r$ factors at time $t$, $W_t$ is a vector of $p$ observed covariates (e.g. an intercept and lags of $Y_{t+h}$), $\epsilon_{t+h}$ is a disturbance term. Equation (2.2) specifies the factor model, where $X_t$ is vector of $N$ candidate predictor series, $\Lambda$ is a loading matrix for $r$ common driving forces in $F_t$, $e_t$ is an idiosyncratic disturbance; and $t = 1, \dots, T$. The latter equation can be rewritten in matrix form

$$\underset{T\times N}{X} = \underset{T\times r}{F}\underset{r\times N}{\Lambda'} + \underset{T\times N}{e}, \tag{2.3}$$

where $X = [X_1, \dots, X_T]'$ and $F = [F_1, \dots, F_T]'$. Throughout the paper it is assumed that all series are weakly stationary, while variables in $X$ have been standardized, meaning that each variable is separately demeaned and set to have unit $\ell_2$-norm.

If the above set of equations is augmented with transition equations for $F_t$, we obtain a dynamic factor model which is estimated by the Kalman filter as discussed above. Let us instead focus on a nonparametric estimation approach as suggested in [111]. The goal at first stage is to solve

$$
\begin{aligned}
&\arg\min_{F,\Lambda} \; \left\| X - F\Lambda' \right\|_F^2 \\
&N^{-1}\Lambda'\Lambda = I_r, \quad F'F \text{ diagonal},
\end{aligned}
\tag{2.4}
$$

where the restrictions are in place for identifying the unique solution (up to a column sign change). It is well known that the estimator of factor loadings $\widehat{\Lambda}$ is given by the $r$ eigenvectors associated with largest eigenvalues of $X'X$, while $\widehat{F} = X\widehat{\Lambda}$. This estimator $\widehat{F}$ is equivalent to principal component (PC) scores derived from the matrix $X$. Once we have an estimate of $F$, the second stage involves least squares estimation of equation 2.1 with $F$ substituted with its estimate.

It is clear that the standard PC estimator reduces the dimensionality of $X$ linearly: $F_t$ represents the projection of $X_t$ onto $r$ eigenvector directions exhibiting the most variation. However, if there is a nonlinearity in $X$, that is if the true lower dimensional representation is a nonlinear submanifold in the original space, such linear projections will be inaccurate. There are several ways to take into account a possible nonlinearity. For example, [10] propose a squared principal components (SPC) procedure, which applies the standard PCA algorithm to the matrix $X$ augmented by its square, that is $[X, X^2]$. Although this procedure supposedly leads to additional forecasting gains, it is limited by the second-order features of the data.

Other nonlinear dimension reduction techniques include Laplacian eigenmaps ([19]), Locally-Linear Embeddings ([99]), Isomaps ([115]) and a number of others ([61], [77], [120]). In this paper we use the approach that applies the kernel trick to the standard PCA, so-called kernel PCA (kPCA) ([101]). This algorithm can be shown to contain a number of widely used dimensionality reduction methods, including the ones listed above ([63]). While it permits modeling a set of nonlinearities rich enough for successful applications in nontrivial pattern recognition tasks such as face recognition ([75]), the algorithm does not involve any iterative optimization.

### 2.2.2 Kernel Method

In this subsection we review the kernel trick methodology and illustrate its usefulness. The kernel method implicitly maps the original data nonlinearly into a high-dimensional space, known as a feature space, $\varphi(\cdot) : \mathcal{X} \to \mathcal{F}$. This space can in fact be infinite-dimensional which would seemingly prohibit any calculations. However, the trick is precisely in avoiding such calculations. The focus is instead on similarities between any two transformed data points $\varphi(\mathrm{x_i})$ and $\varphi(\mathrm{x_j})$ in the feature space[2] as measured by $\varphi(\mathrm{x_i})'\varphi(\mathrm{x_j})$, calculating which would at first sight require the knowledge of the functional form of $\varphi(\cdot)$. The solution is to use a kernel function $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which would output the inner product in the feature space without ever requiring the explicit functional form of $\varphi(\cdot)$. Moreover, a valid kernel function guarantees the existence of a feature mapping $\varphi(\cdot)$ although its analytic form may be unknown. The only requirement for this

[2] Formally, the feature space is thus a Hilbert space, that is a vector space with a dot product defined on it.

is positive-definiteness of the kernel function (Mercer's condition, see 2.H), specifically, $\int\int f(\mathrm{x_i})k(\mathrm{x_i},\mathrm{x_j})f(\mathrm{x_j})d\mathrm{x_i}d\mathrm{x_j} \geq 0$, for any square-integrable function $f(\cdot)$.

This kernel function forms a Gram matrix $K$, known as a Kernel matrix, elements of which are inner products between transformed training examples, that is $\{K\}_{ij} = k(\mathrm{x_i},\mathrm{x_j}) = \varphi(\mathrm{x_i})'\varphi(\mathrm{x_j})$. What makes the kernel trick useful is the fact that many models can be written exclusively in terms of dot products between data points. For example, the ridge regression coefficient estimator can be formulated as $X'(XX' + \lambda I_T)^{-1}\mathrm{y}$[3], and hence the prediction for a test example $\mathrm{x}_*$ is $\hat{y} = \mathrm{x}_*{}'X'(XX' + \lambda I_T)^{-1}\mathrm{y}$, where the dependence is exclusively on inner products between the covariates. This property allows us to apply the kernel method by substituting dot products between original variables with their nonlinear kernel evaluations, that is dot products between transformed variables. Hence, an alternative form is $k_*'(K + \lambda I_T)^{-1}\mathrm{y}$, where $k_*$ with $\{k_*\}_i = \{\varphi(\mathrm{x}_*)'\varphi(X)'\}_i$ is a vector of similarities between the test example and training examples in the feature space. In terms of the time complexity, the algorithm needs to invert a $T \times T$ matrix instead of inverting an $N \times N$ matrix.

The key advantage of the kernel method is that it effectively permits using a linear model in a high-dimensional nonlinear space, which amounts to applying a nonlinear technique in the original space. As a toy example, consider a classification problem shown in Figure 2.1, where the true function separating the two classes is a circle of radius .5 around the origin. The left panel depicts observations from two classes which are not linearly separable in the original two-dimensional space. Applying a simple polynomial kernel of

[3] This can be derived by solving the dual of the ridge least squares optimization problem, however a simpler approach would be to apply the matrix identity $(B'C^{-1}B + A^{-1})^{-1}B'C^{-1} = AB'(BAB' + C)^{-1}$ to the usual ridge estimator $(X'X + \lambda I_N)^{-1}X'\mathrm{y}$.

degree 2, $k(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j})^2$, implicitly corresponds to working in the feature space depicted on the right panel, since for $\varphi(\mathrm{x_i}) = (x_{i1}^2,\ \sqrt{2}x_{i1}x_{i2},\ x_{i2}^2)'$ we have $\varphi(\mathrm{x_i})'\varphi(\mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j})^2$. In this toy example, a linear classifier could perfectly separate the observations in the right panel of Figure 2.1.

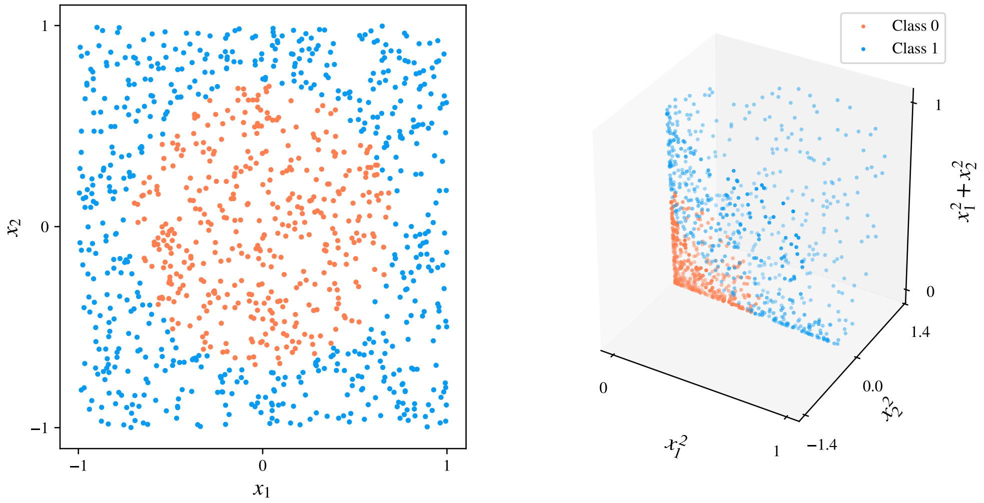


Figure 2.1: Kernel trick illustration for a toy classification example.

*Left*: observations from two classes in the original space, not linearly separable. *Right*: observations in the feature space, linearly separable. See the details in the text.

While all valid kernel functions are guaranteed to have the corresponding feature space, in many cases it is implicit and infinite-dimensional, as, for instance, for the radial basis function (RBF) kernel $k(\mathrm{x_i}, \mathrm{x_j}) = e^{-\gamma\|\mathrm{x_i}-\mathrm{x_j}\|_2^2}$ (see 2.A). Again, luckily, the knowledge of the feature mapping is not required.

### 2.2.3 Nonlinear Modeling and kPCA

Suppose there is a nonlinear function $\varphi(\cdot) : \mathbb{R}^N \rightarrow \mathbb{R}^M$, where $M \gg N$ is very large (often infinitely large), mapping each observation to a high-dimensional feature space,

$X_t \to \varphi(X_t)$. For now we consider $M$ to be finite for simplicity of exposition, so the original $T \times N$ data matrix $X$ can be represented as a $T \times M$ matrix $\Phi = [\varphi(X_1), \ldots, \varphi(X_T)]'$ in the transformed space, which may not be observable. Infinite-dimensional case induces several complications and is considered later.

Both the original $X$ and its transformation $\Phi$ are assumed to be demeaned. The latter requirement is simple to incorporate in the kernel matrix despite the mapping being unobserved. Specifically, supposing the original (non-demeaned) transformation is $\widetilde{\Phi}$, the kernel associated with demeaned features is

$$K = (I_T - \mathbf{1}_{1/T})\widetilde{\Phi}\widetilde{\Phi}'(I_T - \mathbf{1}_{1/T})' = \widetilde{K} - \mathbf{1}_{1/T}\widetilde{K} - \widetilde{K}\mathbf{1}_{1/T} + \mathbf{1}_{1/T}\widetilde{K}\mathbf{1}_{1/T}, \tag{2.5}$$

where $\widetilde{K} = \widetilde{\Phi}\widetilde{\Phi}'$ is based on the original $\widetilde{\Phi}$.

Our modeling of nonlinearity is through the feature mapping $\varphi(\cdot)$. This function replaces the original variables of interest in equation (2.2) with their transformations,

$$\underset{M\times 1}{\varphi(X_t)} = \underset{M\times r}{\Lambda_\varphi}\, \underset{r\times 1}{F_{\varphi,t}} + \underset{M\times 1}{e_{\varphi,t}}\,, \tag{2.6}$$

where the subscript $\varphi$ indicates the association with the transformation. By stacking these into a $T \times M$ matrix $\Phi$ we can rewrite the minimization problem (2.4) as

$$\begin{gathered} \underset{F_\varphi, \Lambda_\varphi}{\arg\min} \; \left\| \Phi - F_\varphi \Lambda_\varphi' \right\|_F^2 \\ N^{-1}\Lambda_\varphi'\Lambda_\varphi = I_r, \quad F_\varphi' F_\varphi \text{ diagonal.} \end{gathered} \tag{2.7}$$

Note that solving this directly through the eigendecomposition of $\Phi'\Phi$ is generally infeasible, since $\Phi'\Phi$ is $M \times M$ dimensional. Even if the dimension $M$ was not prohibitive, the map $\varphi(\cdot)$ is unknown for interesting problems rendering any computation dependent on $\Phi$ or $\Phi'\Phi$ alone impossible. Fortunately, it is possible to reformulate this problem in terms of the $T \times T$ Gram matrix $K = \Phi\Phi'$.

While we are assuming $M$ to be prohibitively large but finite, the following decomposition generalizes to infinite dimensions. Starting from the "infeasible" eigendecomposition of the unknown covariance matrix of $\Phi$

$$\frac{\Phi'\Phi}{T} V_\varphi^{[i]} = \lambda_i^c V_\varphi^{[i]}, \qquad i = 1, \ldots, M, \tag{2.8}$$

where the eigenvalues $\lambda_i^c = \lambda_i(\frac{\Phi'\Phi}{T})$ satisfy $\lambda_1^c \geq \lambda_2^c \geq \ldots \geq \lambda_T^c$ and $\lambda_j^c = 0$ for $j > T$ (assuming $M \geq T$) and $V^{[i]}$ is an $M$-dimensional eigenvector associated with the $i$th eigenvalue $\lambda_i^c$.

The key is to observe that each $V_\varphi^{[i]}$ can be expressed as a linear combination of features

$$V_\varphi^{[i]} = \frac{\Phi'\Phi}{\lambda_i^c T} V_\varphi^{[i]} \equiv \Phi' A^{[i]}, \qquad i = 1, \ldots, M, \tag{2.9}$$

where $A^{[i]} = \frac{\Phi V_\varphi^{[i]}}{\lambda_i^c T} = \left[\alpha_1^{[i]}, \ \ldots \ , \alpha_T^{[i]}\right]'$ is a vector of weights which is determined next. Plugging this back into (2.8) yields

$$\lambda_i^c \Phi' A^{[i]} = \frac{\Phi'\Phi}{T} \Phi' A^{[i]}, \qquad i = 1, \ldots, M. \tag{2.10}$$

Finally, premultiplying equation (2.10) to the left by $\Phi$ and removing $K = \Phi\Phi'$ from both sides we obtain

$$\frac{K}{T}A^{[i]} = \lambda_i^c A^{[i]}, \qquad i = 1, \ldots, M, \tag{2.11}$$

hence the $i$th vector of weights $A^{[i]}$ corresponds to an eigenvector of a finite-dimensional Gram matrix $K$ associated with the $i$th largest eigenvalue $\lambda_i(\frac{K}{T}) = \lambda_i^c$ with $\lambda_j(\frac{K}{T}) = 0$ for $j > T$.

Notice that while solving the eigenvalue problem of $\frac{K}{T}$ allows to compute $A^{[i]}$, we are still unable to obtain the vector $V^{[i]} = \Phi' A^{[i]}$ since $\Phi$ may not be known. However, the main object of interest is recoverable: to calculate principal component projections, we project the data onto (unknown) eigenspace,

$$\underset{T\times 1}{\widehat{F}_{\varphi}^{[i]}} = \Phi V^{[i]} = \Phi\Phi' A^{[i]} = K A^{[i]}, \qquad i = 1, \ldots, M. \tag{2.12}$$

Stacking estimated factors corresponding to the first $r$ eigenvalues, define a $T \times r$ matrix $\widehat{F}_{\varphi} = \left[\widehat{F}_{\varphi}^{[1]}, \ \ldots \ , \widehat{F}_{\varphi}^{[r]}\right]$, where the subindex $r$ is dropped to simplify the notation. We refer to the factors constructed this way as *kernel factors.*

A similar alternative solution that only involves the Gram matrix has been known in econometrics since at least [37]. In particular, for a given $r$ the optimization problem in (2.7) with the identification constraints $T^{-1}F_{\varphi}'F_{\varphi} = I_r$ and diagonal $\Lambda_{\varphi}'\Lambda_{\varphi}$, has the solution $\widetilde{F}_{\varphi} = \sqrt{T} eig_r(\Phi\Phi') = \sqrt{T}A_r$, where $A_r = \left[\widehat{A}^{[1]}, \ \ldots \ , \widehat{A}^{[r]}\right]$. Hence, the kPCA estimator is equivalent to the latter premultiplied by $\frac{\Phi\Phi'}{\sqrt{T}} = \frac{K}{\sqrt{T}}$. Note, however, that both estimators yield the same predictions when passed to the main forecasting equation (2.1) as they have identical column spaces. This idea is summarized in the following proposition.

**Proposition 1** *Estimators $\widehat{F}_{\varphi} = \Phi\Phi' eig_r(\Phi\Phi')$ and $\widetilde{F}_{\varphi} = \sqrt{T} eig_r(\Phi\Phi')$ produce the same projection matrix.*

Importantly, we also establish that certain commonly used kernels allow the kernel factor estimator to incorporate the usual PC estimator. The following proposition demonstrates that RBF and sigmoid kernels allow to nest (a constant multiple of) the PC estimator for limiting values of the hyperparameter.

**Proposition 2** *For a column-centered matrix $X \in \mathbb{R}^{T\times N}$, let $\widehat{F}_{\varphi} = K eig_r(K)$ be the kernel factor estimator and $\widehat{F} = X eig_r(X'X)$ be the usual linear PCA factor estimator. Then*

$$\exists s = \pm 1, \text{ such that } \lim_{\gamma\to 0} c\gamma^{-1}\widehat{F}_{\varphi}L^{-1/2} = s\widehat{F}, \quad \forall r = 1,\ldots,\min\{T,N\},$$

*where $K = \widetilde{K} - \mathbf{1}_{1/T}\widetilde{K} - \widetilde{K}\mathbf{1}_{1/T} + \mathbf{1}_{1/T}\widetilde{K}\mathbf{1}_{1/T}$, $L$ is a diagonal matrix of $r$ largest eigenvalues of $XX'$ sorted in nonincreasing order. Furthermore,*

***(a)*** *for RBF kernel $\widetilde{k}_{ij} = e^{-\gamma\|X_{i\cdot}-X_{j\cdot}\|_2^2}$ we have $c = 2^{-1}$,*

***(b)*** *for sigmoid kernel $\widetilde{k}_{ij} = \tanh(c_0 + \gamma X_{i\cdot}'X_{j\cdot})$ we have $c = (1-\tanh^2(c_0))^{-1}$, where $c_0$ is an arbitrary (hyperparameter) constant.*

**Proof.** See appendix 2.B. ■

Adjustment by $\gamma^{-1}$ is necessary as the entries of $\widehat{F}_{\varphi} = K eig_r(K)$ approach 0 as $\gamma \to 0$. This is due to the nature of $K = \widetilde{K} - \mathbf{1}_{1/T}\widetilde{K} - \widetilde{K}\mathbf{1}_{1/T} + \mathbf{1}_{1/T}\widetilde{K}\mathbf{1}_{1/T}$, where $k_{ij}$

converge to 0 for both kernels for all $i, j = 1, \ldots, T$. Luckily $\gamma^{-1}K$ has a nice non-degenerate limiting value (shown in the proof) that ensures the limit in proposition 2.B holds.

Proposition 2 states that the (properly scaled) kernel factor matrix converges pointwise to its PCA analog (up to a sign flip) as the value of the hyperparameter $\gamma$ nears zero. That is, in the limit the two factor estimators are constant multiples of one another and hence produce the same forecasts. This ability to mimic the linear estimator is important for certain applications. For example, in macroeconomic forecasting it is notoriously difficult to beat linear models in a short horizon prediction exercise.

Figure 2.2 illustrates the implicit procedure for obtaining $r$ kernel factors. The selected kernel function induces nonlinearity $\varphi(\cdot)$ on each element of the input layer, $N$-dimensional observations $X_{1\cdot}, \ldots, X_{T\cdot}$. Next, pairwise similarities between high-dimensional vectors $\varphi(X_{1\cdot}), \ldots, \varphi(X_{T\cdot})$ are computed, with a kernel function given as

$$k(X_{i\cdot}, \cdot) = \Big[\varphi(X_{i\cdot})'\varphi(X_{1\cdot}),\ \ldots\ ,\varphi(X_{i\cdot})'\varphi(X_{T\cdot})\Big]'.$$

Finally, each factor is obtained as a linear combination of these inner products with the weights given as a solution to the eigenvalue problem discussed above. Of course, the kernel PCA algorithm does not explicitly nonlinearize the inputs as the kernel trick allows us to immediately calculate similarities and avoid the expensive high-dimensional computation. However, as mentioned earlier, the existence of such implicit nonlinearities is guaranteed by Mercer's theorem (2.H). As opposed to standard feedforward neural networks, there is no iterative training involved and no peril of being trapped in local optima. On the other hand, it has been noted that certain kernels, including RBF and sigmoid, allow extracting features of the same type as the ones extracted by neural networks ([101]). The two necessary steps involve evaluation of similarities in the kernel matrix and solving its eigenvalue problem; the complexity is thus dependent only on the sample size.

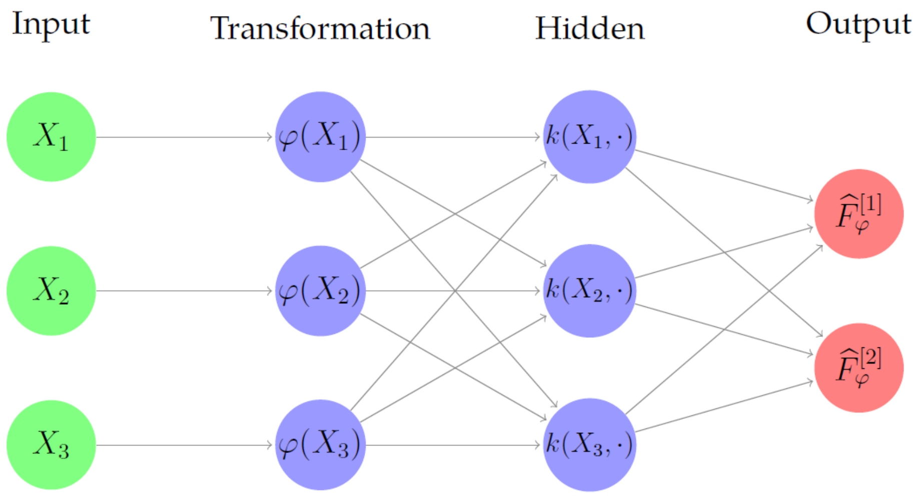


Figure 2.2: Neural network interpretation of kernel PCA, $T = 3$, $r = 2$.

Each observation is nonlinearly transformed and the inner products are computed. The output units are kernel factors with $\widehat{F}_{\varphi}^{[j]} = \sum_{i=1}^{T} \alpha_i^{[j]} k(X_i, \cdot)$ which linearly combine these dot products, with the weight estimate calculated as the eigenvector of the kernel matrix $K$.

### 2.2.4 Theory

Depending on the choice of the kernel, the induced feature space could be either finite or infinite. A polynomial kernel considered earlier generates a finite-dimensional feature space, consisting of a set of polynomial functions over the inputs. In this simple case, the eigenspace associated with the kernel factor estimator in equation (2.12) can generally be consistently estimated within the framework of [7]. Particularly, proposition 1 and the following theorem in [7] immediately imply $\sqrt{M}$-consistency of $\widetilde{F}_{\varphi}$.

For the model associated with equation (2.6):

**Assumption A**: There exists a constant $c_1 < \infty$ independent of $M$ and $T$, such that

(a) $\mathbb{E}\left\|F_{\varphi,t}\right\|_F^4 \leq c_1$ and $T^{-1}F'_\varphi F_\varphi \xrightarrow{p} \Sigma_F > 0$, where $\Sigma_F$ is a non-random positive definite matrix;

(b) $\mathbb{E}\left\|\Lambda_{\varphi,i\cdot}\right\|_F \leq c_1$ and $N^{-1}\Lambda'_\varphi \Lambda_\varphi \xrightarrow{p} \Sigma_\Lambda > 0$, where $\Sigma_\Lambda$ is a non-random positive definite matrix.

(c1) $\mathbb{E}(e_{\varphi,it}) = 0$, $\mathbb{E}\left|e_{\varphi,it}\right|^8 \leq c_1$;

(c2) $\mathbb{E}(e'_{\varphi,s}e_{\varphi,t}/M) = \gamma_M(s,t)$, $|\gamma_M(s,s)| \leq c_1$ $\forall s$, $T^{-1}\sum_{s=1}^T\sum_{t=1}^T|\gamma_M(s,t)| \leq c_1$, $\sum_{s=1}^T\gamma_M(s,t)^2 \leq M$ $\forall t, T$;

(c3) $\mathbb{E}(e_{\varphi,it}e_{\varphi,jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some $\tau_{ij}$ and $\forall t$; and $M^{-1}\sum_{i=1}^M\sum_{j=1}^M |\tau_{ij}| \leq c_1$;

(c4) $\mathbb{E}(e_{\varphi,it}e_{\varphi,js}) = \tau_{ij,ts}$, $(MT)^{-1}\sum_{i=1}^M\sum_{j=1}^M\sum_{t=1}^T\sum_{s=1}^T|\tau_{ij,ts}| \leq c_1$;

(c5) $\mathbb{E}\left|M^{-1/2}\sum_{i=1}^M\left(e_{\varphi,is}e_{\varphi,it} - \mathbb{E}(e_{\varphi,is}e_{\varphi,it})\right)\right|^4 \leq c_1$ $\forall t, s$;

(d) $\mathbb{E}\left(\frac{1}{M}\sum_{i=1}^M\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T F_{\varphi,t}e_{\varphi,it}\right\|_F^2\right) \leq c_1$;

(e) $T^{-1}F'_\varphi F_\varphi = I_r$ and $\Lambda'_\varphi\Lambda_\varphi$ is diagonal with distinct entries.

Part (a) is standard in factor model literature. Part (b) ensures pervasiveness of factors in the sense that each factor has non-negligible effect on the variability of covariates. This is crucial for asymptotic identification of the common and idiosyncratic components. Parts (c$\cdot$) partially permit time-series and cross-section dependence in the idiosyncratic component, as well as heteroskedasticity in both dimensions. Possible correlation of $e_{\varphi,it}$ across $i$ sets up the model to have an approximate factor structure. Part (d) allows weak dependence between factors and idiosyncratic errors and (e) is for identification.

The following theorem establishes consistency of kernel factors, for kernels inducing finite nonlinearities, in a large-dimensional framework.

**Theorem 3 (Theorem 1 [8], adapted)** *Suppose the kernel function induces finite dimensional nonlinearity, i.e. for* $N < \infty$*,* $\varphi(\cdot) : \mathbb{R}^N \to \mathbb{R}^M$*, where* $M := M(N)$ *is such that* $N \leq M(N) < \infty$*, and Assumption A holds. Then for any fixed* $r \geq 1$*, as* $M, T \to \infty$

$$\delta_{NT}^2 \left\| \widetilde{F}_{\varphi,t} - H' F_{\varphi,t} \right\|_F^2 = O_p(1), \quad \forall t = 1, \ldots, T,$$

*where* $\delta_{NT} = \min\{\sqrt{M}, \sqrt{T}\}$*,* $\underset{r\times r}{H} = \frac{\Lambda_\varphi^{0\prime}\Lambda_\varphi^0}{M}\frac{F_\varphi^{0\prime}\widetilde{F}_\varphi}{T}V_{MT}^{-1}$*,* $\underset{r\times r}{V_{MT}}$ *is a diagonal matrix of* $r$ *largest eigenvalues of* $\frac{\Phi\Phi'}{MT}$*,* $\widetilde{F}_{\varphi,t}$ *and* $F_{\varphi,t}$ *are* $t$*-th rows in* $\widetilde{F}_\varphi = \sqrt{T} eig_r(\Phi\Phi')$ *and* $F_\varphi$*, respectively.*

**Proof.** See appendix 3.B. ■

Some comments are in order. First, the theorem states that the squared differences between the proposed factor estimator and (a rotation of) the true factor vanish as $M, T \to \infty$. While true factors themselves are not identifiable unless additional assumptions are imposed (see [11]), identification of the latent space spanned by factors is just as good as exact identification for forecasting purposes. Second, since for any given number of original variables $N$ the dimension of the transformed space $M$ is fixed and finite, the growth in $M$ is only possible through $N$ and thus the limit on $M$ implies one on $N$. Third, this result does not imply uniform convergence in $t$. Lastly, the results suggest the possibility of $\sqrt{T}$-consistent estimation of the forecasting equation with respect to its conditional mean.

Unfortunately, this result does not generalize to the most interesting kernels (e.g. RBF) inducing infinite-dimensional Hilbert spaces, rendering the traditional approach unsuitable for establishing theoretical properties. Hence we turn to a functional analytic framework which allows rigorous treatment of infinite-dimensional spaces and which has been the classical framework for analyzing statistical properties of functional PCA, and kernel PCA in particular, in machine learning literature.

As will be shown later, it turns out that we can still show that the estimator concentrates around its population counterpart. We briefly present the necessary terms for understanding this result, without aiming to be exhaustive. A sufficiently detailed introduction to the analysis in Hilbert spaces can be found in [24], while a classical reference for operator perturbation theory is [72].

One of the first major investigations of the statistical properties of kernel PCA can be found in [103]. The study provides concentration bounds on the sum of eigenvalues of the kernel matrix towards that of corresponding (infinite-dimensional) kernel operators. This permits to characterize the accuracy of kPCA in terms of the reconstruction error, that is the ability to preserve the information about a high-dimensional input in low dimensions. This is of significant interest for certain types of applications, such as pattern recognition. [24] further extend the results of the aforementioned study and improve the bounds on eigenvalues using tools of perturbation theory.

Although the theoretical discussion and the mathematical approaches developed in this literature are extremely valuable, our interest is not in kPCA's ability to reconstruct a given observation. The kernel factor estimator only reduces the dimensionality and passes it to the next stage, without ever going through the reconstruction phase. As the dimensionality is reduced by projecting observations onto the eigenspace – the space spanned by eigenfunctions of the true covariance operator with largest eigenvalues – the interest is in convergence of empirical eigenfunctions towards the true counterparts. Importantly, the proximity of eigenvalues does not guarantee that underlying eigenspaces will also be close.

We now briefly introduce the technical background for understanding our result. Let $\mathcal{H}$ be an inner product space, that is a linear vector space endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, commonly denoted together as $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$. An inner product space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a Hilbert space if and only if the norm induced by the inner product $\|\cdot\|_{\mathcal{H}} = \langle\cdot,\cdot\rangle_{\mathcal{H}}^{1/2}$ is complete[4]. The results below rely on this norm, so we drop the subscript $\mathcal{H}$ to simplify the notation.

Let $X$ be a random variable taking values on a general probability space $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a real-valued Hilbert space and a measurable feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X},\ k(x, x') = \langle \phi(x), \phi(x') \rangle$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), if $k$ satisfies (i) $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{X}$, and the reproducing property (ii) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle = f(x)$. An important result from [6] guarantees the existence of a unique RKHS for every positive definite $k$.

Assume that $\mathbb{E}\,\phi(X) = 0$ and $\mathbb{E}\,\|\phi(X)\|^2 < \infty$. A unique covariance operator on $\phi(X)$, $\Sigma = \mathbb{E}\,\phi(X) \otimes \phi(X)$, satisfying $\langle g, \Sigma h\rangle_{\mathcal{H}} = \mathbb{E}\,\langle h, \phi(X)\rangle_{\mathcal{H}} \langle g, \phi(X)\rangle_{\mathcal{H}}$, $\forall g, h \in \mathcal{H}$, always exists (Theorem 2.1 in [24]) and is a positive, self-adjoint trace-class operator. Denote

[4] Specifically, the limits of all Cauchy sequences of functions must be in the Hilbert space.

$\widehat{\Sigma} = \frac{1}{T}\sum_{i=1}^{T}\phi(X_i)\otimes\phi(X_i)$ to be its empirical counterpart. Finally, an orthogonal projector in $\mathcal{H}$ onto a closed subspace $V$ is an operator $\Pi_V$ such that $\Pi_V^2 = \Pi_V$ and $\Pi_V = \Pi_V^*$.

We now lay out two lemmas which lead to the result. Lemma (10) bounds the difference between the true and empirical covariance operators.

**Lemma 4 (Difference between sample and true covariance operators)** *Assume random variables $X_1, \ldots, X_T \in \mathcal{X}$ are independent and $\sup_{x\in\mathcal{X}} k(x,x) \leq \bar{k}$, then*

$$\mathbb{P}\left(\left\|\widehat{\Sigma} - \Sigma\right\| \geq \left(1 + \sqrt{\frac{\epsilon}{2}}\right)\frac{2\bar{k}}{\sqrt{T}}\right) \leq e^{-\epsilon}.$$

**Proof.** See appendix 2.E. ■

Note that both sigmoid and RBF kernels are bounded and hence satisfy the requirement of the above lemma. Lemma (5) is an operator perturbation theory result and is adapted from [76] and [131]:

**Lemma 5** *Let $A, B$ be two symmetric linear operators. Denote the distinct eigenvalues of $A$ as $\mu_1 > \ldots > \mu_k > 0$ and let $\Pi_i$ be the orthogonal projector onto the i-th eigenspace. For a positive integer $p \leq k$ define $\delta_p(A) := \min\{|\mu_i - \mu_j| : 1 \leq i < j \leq p+1\}$. Assuming $\|B\| < \delta_p(A)/4$, then*

$$\|\Pi_i(A) - \Pi_i(A+B)\| \leq \frac{4\|B\|}{\delta_i(A)}.$$

**Proof.** See appendix 2.F. ■

Finally, the following theorem bounds the difference between empirical and true eigenvectors.

**Theorem 6** *Denote the i-th eigenvectors of $\widehat{\Sigma}$ and $\Sigma$ as $\widehat{\psi}_i$ and $\psi_i$ respectively. Then, under the assumptions of Lemma 10 and 5, as $T \to \infty$ we have*

$$\left\| \widehat{\psi}_i - \psi_i \right\| = o_p(1)$$

**Proof.** See appendix 2.G. ■

Some comments are in order. First, note that we require the eigenvalues to be distinct, a well-known restriction (similar to $sin(\theta)$ theorem of [41]), since it is impossible to identify eigenspaces with the same eigenvalues. Second, Theorem 6 suggests that the eigenspace estimated by kernel PCA will concentrate close to the true eigenspace. Our kernel factor estimator simply projects onto that eigenspace and hence the precision is expected to increase as $T \to \infty$. Third, this rate does not address the case when variables exhibit dependence, although the exercise in the next section is indicative of some form of concentration, which suggests that the theoretical assumptions might be too conservative. Lastly, it may be possible to obtain a sharper bound since the proof relies on crude inequalities (e.g. triangle inequality).

## 2.3 Empirical Evaluation

### 2.3.1 Forecasting Models

This subsection discusses specific forms of equations (2.1) and (2.2) that are used for forecasting. Autoregressive Diffusion Index (ARDI) model is specified as

$$Y_{t+h} = \beta_0^h + \sum_{p=1}^{P_t^h} \beta_{Y,p}^h Y_{t-p+1} + \sum_{m=1}^{M_t^h} \underset{1\times K_t^h}{\beta_{F,m}^h}{}' \underset{K_t^h\times 1}{F_{t-m+1}} + \epsilon_{t+h}, \tag{2.13}$$

where superscript $h$ indicates dependence on the time horizon. Note, $P_t^h, M_t^h, K_t^h$ are the number of lags of the target variable, the number of lags of factors, the number of factors respectively. These three parameters are estimated simultaneously for each time period and time horizon using BIC. Since the true factors are unknown, we instead plug in the estimates from the factor equation, which is discussed next.

Three factor equation specifications are considered,

$$\underset{N\times 1}{X_t} = \Lambda F_t + e_t, \tag{2.14}$$

$$\underset{2N\times 1}{X_{*,t}} = \Lambda_* F_{*,t} + e_{*,t}, \tag{2.15}$$

$$\underset{M\times 1}{\varphi(X_t)} = \Lambda_\varphi F_{\varphi,t} + e_{\varphi,t}. \tag{2.16}$$

Factors in equation (2.14) are estimated by PCA. Equation (2.15) is similar, replacing the left-hand side with an augmented vector $X_{*,t} = [X_t, X_t^2]$. This procedure was dubbed as squared principal components (SPC) in [10]. Finally, the last equation applies nonlinearity induced by the selected kernel and is estimated by kPCA.

Forecasts using PCA and SPC are produced in three steps. First, we extract three factors from the set transformed and standardized predictors using one of the two methods. Second, three parameters are determined according to BIC for each out-of-sample forecasting period and each prediction horizon: the number of lags of the target variable $P_t^h$, the number of lags of factors $M_t^h$, the number of factors $K_t^h$. Third, the forecasting equation is estimated by least squares and forecasts are produced. The procedure for predicting with kPCA is similar, except there is an additional step where the value of the hyperparameter is specified, and the estimation is instead made in accordance with Algorithm 1.

**Algorithm 1:** kPCA Algorithm

```
Input: Observations $X_1, \dots, X_T \in \mathbb{R}^N$, kernel function $k(\cdot,\cdot)$, dimension $r$.
for $i = 1, \dots, T$ do
    for $j = 1, \dots, T$ do
        $K_{ij} = k(X_i, X_j)$ ;                                  /* Compute similarities */
    end
end
$K = K - \mathbf{1}_{1/T}K - K\mathbf{1}_{1/T} + \mathbf{1}_{1/T}K\mathbf{1}_{1/T}$ ;          /* Standardization */
$[A, \Lambda] = K/T$ ;                                          /* Eigendecomposition */
$\widehat{F}_r = KA_r$ ;                                           /* Compute factors */
Output: $T \times r$ matrix $\widehat{F}_r$.
```

The algorithm starts by computing similarities via a given kernel. This induces transformed observation that need to be demeaned according to equation (2.5). As in ordinary PCA, one then needs to compute the eigenvectors. The difference here is that our object is a Gram matrix, not a covariance matrix. The kernel factors are then simply derived as projections of the kernel matrix onto $r$ eigenvectors associated with the largest eigenvalues.

### 2.3.2 Data and Forecast Construction

As an empirical investigation, we examine whether using kernel factors leads to improved performance in forecasting several key macroeconomic indicators. We use a large dataset from FRED-MD ([85]), which has become one of the classical datasets for empirical analysis of big data. Its latest release consists of 128 monthly US variables running from $1959 : 01$ through $2020 : 04$, 736 observations in total. Following previous studies, we set $1960 : 01$ as the first sample, leaving 724 observations. Since the models presented in this study require stationary series, each of the variables undergo a transformation to achieve stationarity. The decision on a particular form of transformation is generally dependent on the outcome of a unit root test, which is known to lack power in finite samples. So instead, following [85], all interest and unemployment are assumed to be $I(1)$, while price indexes are assumed to be $I(2)$. The transformations applied to each series are described in supplemental materials.

We aim to predict a single time series from this dataset by utilizing the remaining variables. The series to be predicted include 8 variables characterizing different aspects of the economy. Specifically, we take one series from each of the eight variable "groups" in the dataset. The summary is provided in Table A1 in Appendix.

Forecasts are constructed for $h = 1, 3, 6, 9, 12, 18, 24$ months ahead with a rolling time window, the size of which is taken to be $120 - h$. Thus, the pseudo-out-of-sample forecast evaluation period is $1970 : 01$ to $2020 : 04$, which is 604 months. We estimate 6 variants of autoregressive diffusion index models. The first model, taken as a benchmark, is a classical ARDI with PC estimates. Several studies have documented a strong performance of this model (see for example, [38]). The second and third take SPC and so-called PC-squared ($PC^2$) estimates ([10]) respectively, where the latter is identical to the first model with squares of factor estimates added in the forecasting equation. The remaining models are based on kPCA estimates with three different kernels: a sigmoid $k(\mathrm{x_i}, \mathrm{x_j}) = \tanh(\gamma(\mathrm{x_i}'\mathrm{x_j}) + 1)$, a radial basis function (RBF) $k(\mathrm{x_i}, \mathrm{x_j}) = e^{-\gamma\|\mathrm{x_i}-\mathrm{x_j}\|^2}$ and a quadratic[5] polynomial (poly(2)) kernel $k(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j} + 1)^d,\ d = 2$.

Optimal parameters for each model at each step, $P_t^h, M_t^h, K_t^h$ and the kernel hyperparameter, are determined within the rolling window period, that is our setup only permits the information set that would be available at the moment of making a prediction. Specifically, $P_t^h, M_t^h, K_t^h$ are selected by BIC (maximum value allowed for each is set equal to 6) for each out-of-sample period, while $\gamma$ is determined over a grid of values by so-called time series cross-validation. Specifically, we consecutively predict the latest 5 available observations and select the hyperparameter that minimizes the average error. The standard cross-validation may not theoretically be fully adequate due to the presence of serial correlation in the data and several approaches were suggested to correct it ([96]).

### 2.3.3 Results

The main empirical findings are presented in Table 2.1. The subtitles of series indicate the names of dependent variables, while each value in the table represents the ratio of out-of-sample MSPE of a given estimation method to out-of-sample MSPE of the autoregression augmented by diffusion indexes estimated by PCA. An asterisk indicates the best performer for each horizon (no asterisk indicates the superiority of the baseline

[5] Polynomial kernels of lower and higher order demonstrated poor forecasting ability and are not included.

method). Values printed in boldface suggest statistical significance of the Diebold-Mariano test of equal predictive ability at 90%, when the corresponding method is compared against the autoregression with PCA estimates. The results range for 8 variables across 7 different forecast horizons. Our results are reproducible: in supplemental materials we provide the script written in Python 3.6 that generates all results within a few hours.

Table 2.1: Relative MSPEs for 8 variables across 7 different prediction horizons.

| | $h = 1$ | $h = 3$ | $h = 6$ | $h = 9$ | $h = 12$ | $h = 18$ | $h = 24$ |
|---|---|---|---|---|---|---|---|
| Real Personal Income | | | | | | | |
| SPC | **1.0327** | **1.3195** | 1.5310 | 1.8579 | 1.3750 | 1.1647 | 1.1979 |
| PC$^2$ | 1.0130 | 1.1736 | 1.2709 | 1.0859 | 1.2503 | **38.1598** | **19.4460** |
| kPCA poly(2) | **1.0769** | **2.2967** | **3.1602** | **2.5349** | 1.9702 | 2.2061 | 1.3281 |
| kPCA sigmoid | 1.0013 | 0.9886$^*$ | **0.9650**$^*$ | 0.8835$^*$ | 0.8891$^*$ | 0.8377$^*$ | **0.8688**$^*$ |
| kPCA RBF | 0.9995$^*$ | 0.9972 | 0.9977 | 0.9318 | **0.9367** | 0.9269 | **0.9890** |
| Civilian Employment | | | | | | | |
| SPC | 1.0257 | **1.4758** | 1.6383 | 2.2547 | 1.9158 | 1.3148 | 1.3624 |
| PC$^2$ | 0.9871$^*$ | 1.1357 | 1.2180 | 1.2806 | **1.4784** | **20.8259** | **386.2726** |
| kPCA poly(2) | **1.4307** | **1.6269** | 2.7371 | **2.0439** | 1.7979 | 1.6266 | **1.6697** |
| kPCA sigmoid | 1.0017 | 0.9830 | **0.9245**$^*$ | **0.9213**$^*$ | **0.9084**$^*$ | **0.9171**$^*$ | 0.9138$^*$ |

*Continued on next page*

Table 2.1 – *continued from previous page*

| | $h = 1$ | $h = 3$ | $h = 6$ | $h = 9$ | $h = 12$ | $h = 18$ | $h = 24$ |
|---|---|---|---|---|---|---|---|
| kPCA RBF | 1.0004 | **0.9758**$^*$ | **0.9448** | **0.9317** | 0.9381 | 1.0028 | 0.9696 |
| | | | Housing Starts: Privately Owned | | | | |
| SPC | 1.0102 | 1.5529 | 2.3532 | 1.9179 | 1.4604 | 1.2191 | **1.7948** |
| PC$^2$ | **1.0978** | 1.1887 | 1.7617 | 1.8405 | **1.3927** | **67.1838** | **23.4097** |
| kPCA poly(2) | 1.0684 | 1.3471 | 2.8672 | **1.6391** | 2.2092 | 2.2018 | 2.8618 |
| kPCA sigmoid | 0.9953 | 0.9742 | 0.9786$^*$ | **0.9583** | 0.9576 | **0.9453** | 0.9771 |
| kPCA RBF | 0.9950$^*$ | 0.9588$^*$ | 0.9990 | 0.9225$^*$ | **0.9412**$^*$ | **0.8944**$^*$ | **0.9409**$^*$ |
| | | | Real personal consumption | | | | |
| SPC | 1.0790 | 1.3103 | 1.6410 | 2.0345 | 1.5595 | 1.1362 | 1.3404 |
| PC$^2$ | **1.0456** | 1.1577 | 1.2311 | 1.3181 | **1.1786** | **11.5388** | **19.6892** |
| kPCA poly(2) | **1.1722** | 2.0990 | **1.8980** | 1.8964 | 2.5677 | 2.8196 | 1.7777 |
| kPCA sigmoid | 1.0037 | 0.9901 | **0.9549**$^*$ | **0.9310**$^*$ | **0.9072**$^*$ | **0.9131**$^*$ | **0.9477**$^*$ |
| kPCA RBF | 0.9919$^*$ | 0.9894$^*$ | **0.9609** | **0.9363** | **0.9596** | **0.9622** | 0.9799 |
| | | | M1 Money Stock | | | | |
| SPC | 1.0660 | 1.5959 | 1.3715 | **1.5438** | 1.7104 | 1.4845 | 1.3175 |
| PC$^2$ | **1.3453** | 1.4641 | 1.1331 | 1.1987 | 2.5199 | **27.0868** | **157.3719** |

*Continued on next page*

Table 2.1 – *continued from previous page*

| | $h = 1$ | $h = 3$ | $h = 6$ | $h = 9$ | $h = 12$ | $h = 18$ | $h = 24$ |
|---|---|---|---|---|---|---|---|
| kPCA poly(2) | 1.9656 | 3.9986 | 2.6646 | 2.6261 | 2.5258 | 2.1058 | 1.6008 |
| kPCA sigmoid | 1.0074 | **0.9631** | **0.9275*** | **0.9183*** | **0.9311*** | 0.8977* | **0.8295*** |
| kPCA RBF | 0.9892* | 0.9994* | **0.9663** | **0.9569** | **0.9756** | **0.9521** | 0.9827 |
| Effective Federal Funds Rate | | | | | | | |
| SPC | 0.9261* | 1.4816 | 3.1258 | 2.6737 | 2.4173 | 1.7712 | 2.2228 |
| PC$^2$ | 1.2278 | 1.3466 | 1.3976 | 2.3356 | 1.2200 | **9.0571** | **23.7870** |
| kPCA poly(2) | 1.3172 | 2.2581 | 3.8848 | 2.8236 | 1.3301 | 2.8415 | 1.6740 |
| kPCA sigmoid | 0.9921 | 0.9250* | 0.8363* | 0.8503* | 0.8654* | **0.8094*** | 0.7950* |
| kPCA RBF | 1.0388 | 0.9782 | 0.9544 | **0.9119** | **0.9619** | 0.9335 | 0.8907 |
| CPI: All Items | | | | | | | |
| SPC | 1.0445 | 2.0341 | 2.2723 | 1.3606 | 1.3226 | **1.3522** | **1.3168** |
| PC$^2$ | 1.1932 | 1.6602 | 1.2048 | 1.9320 | 1.5478 | **12.5806** | **34.1029** |
| kPCA poly(2) | 1.6292 | 3.0142 | 3.9154 | 1.9549 | 1.4750 | 1.3192 | 1.2334 |
| kPCA sigmoid | 0.9901* | 1.0189 | **0.9418*** | 0.9659* | **0.9540*** | **0.9518** | **0.9538*** |
| kPCA RBF | 0.9967 | 1.0596 | 1.0101 | 1.0138 | 0.9803 | **0.9504*** | **0.9564** |
| S&P 500 Index | | | | | | | |

*Continued on next page*

Table 2.1 – *continued from previous page*

| | $h = 1$ | $h = 3$ | $h = 6$ | $h = 9$ | $h = 12$ | $h = 18$ | $h = 24$ |
|---|---|---|---|---|---|---|---|
| SPC | **1.1235** | 1.3328 | 2.0309 | 2.0885 | **2.0879** | 1.5848 | 2.0441 |
| PC$^2$ | **1.1131** | 1.2793 | 1.4291 | 1.6809 | 1.9813 | 41.7948 | **43.1356** |
| kPCA poly(2) | **1.6190** | **1.7477** | **2.5330** | 2.3499 | 2.0381 | 1.4199 | **1.3398** |
| kPCA sigmoid | 0.9890* | **0.9675*** | 0.9336* | 0.8708* | 0.8634* | 0.8125* | **0.8749*** |
| kPCA RBF | 1.0001 | 0.9877 | **0.9701** | 0.9683 | 0.9450 | 0.8319 | 0.9441 |

Some comments are in order. First, sigmoid and RBF kernel approaches do lead to improved forecasting accuracy, especially at medium- and long-term time horizons. One of these two approaches dominates others in nearly 95% of the cases considered. Additionally, Diebold-Mariano test ([42]) of equal predictive ability suggests that only these two methods are capable of often significantly outperforming the baseline PCA autoregression across a range of variables and horizons, while never being dominated. The kernel method is least advantageous for one-step-ahead forecasting. The phenomenon that linearity is hard to beat in a very short horizon is rather well known in the literature. Luckily, as was shown in Proposition 2, kPCA is capable of mimicking a linear PCA by adjusting the kernel hyperparameter closer to 0, which often leads to the parity of the two methods in near-term forecasting. While the gains are not pronounced at $h = 1$, they become apparent at longer horizons. Results vary across variables, but the improvement is prevailing at medium-term horizons and is uniform in one-year and longer predictions. The superiority exhibited by kPCA in many cases is remarkable for macroeconomic forecasting literature.

Second, both SPC and PC$^2$ perform substantially worse than a simple PCA. This result contradicts to [10], but is consistent with a recent empirical comparison of [49]. Besides, PC$^2$ is extremely unreliable for long-term predictions. Similar to SPC, poly(2) kernel seeks to model the second-order features of the data and, as a result, often performs on par with SPC. As pointed out by the referee, the additional squared terms might often be leading to inefficiency.

Ultimately, note that kPCA's computational complexity is dependent on the number of time periods for estimation $120 - h$, making kPCA slightly faster in this particular exercise. Most importantly, kPCA's advantage would grow in a macroeconomic setting, where "bigger" data (i.e. larger $N$) is becoming the norm.

## 2.4 Concluding Remarks

In this study we have introduced a nonlinear extension of factor modeling based on the kernel method. Although our exposition mainly focused on a feature mapping $\varphi(\cdot)$ enforcing nonlinearity, it is also convenient to think of this approach as kernel smoothing in an inner product space. That is, the kernel factor estimator implicitly relies on the weighted distances between original observations. This alternative viewpoint presumes that analyzing the variation in the inner product space, rather than the original space, may be more beneficial. This idea had a profound impact on machine learning and pattern recognition fields, especially as regards to support vector machines (SVMs). By using a positive definite kernel, one can be very flexible in the original space while effectively retaining the simplicity of the linear case in the high-dimensional feature space.

We have demonstrated that constructing factor estimates nonlinearly can be beneficial for macroeconomic forecasting. Specifically, the nonlinearity induced by the sigmoid and RBF kernels leads to considerable gains at medium- and long-term time horizons. This gain in performance comes at no substantial sacrifice, the algorithm remains scalable and computationally fast.

There are several possible extensions. First, it is interesting to see how the performance would change have we pre-selected the variables (targeting) before reducing the dimensionality. As shown in [10] and [26] this generally leads to better precision. Second, the forecasting accuracy can be compared with other nonlinear dimension reduction techniques mentioned earlier, such as autoencoders. For the latter, however, one must be aware of the possibility of implicit overfitting by tuning the network architecture. This is not an issue in the current framework as there are a lot fewer parameters to specify. Third, the static factors considered here could possibly be extended to dynamic factors ([56]), by explicitly incorporating the time domain, or “efficient” factors, by weighing observations by the inverse of the estimated variance.

# Appendices

## 2.A Decomposition of RBF kernel

Let $\mathrm{x}, \mathrm{z} \in \mathbb{R}^k$ and $k(\mathrm{x}, \mathrm{z}) = e^{-\gamma\|\mathrm{x}-\mathrm{z}\|^2}$. Then through the Tailor expansion we can write

$$\begin{aligned}
k(\mathrm{x}, \mathrm{z}) &= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} e^{2\gamma \mathrm{x}'\mathrm{z}} \\
&= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} \sum_{j=0}^{\infty} \frac{(2\gamma)^j}{j!} (\mathrm{x}'\mathrm{z})^j \\
&= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} \sum_{j=0}^{\infty} \frac{(2\gamma)^j}{j!} \sum_{\sum_{i=1}^{k} n_i = j} j! \prod_{i=1}^{k} \frac{(x_i y_i)^{n_i}}{n_i!} \\
&= \sum_{j=0}^{\infty} \sum_{\sum_{i=1}^{k} n_i = j} \left( (2\gamma)^{j/2} e^{-\gamma\|\mathrm{x}\|^2} \prod_{i=1}^{k} \frac{x_i^{n_i}}{\sqrt{n_i!}} \right) \left( (2\gamma)^{j/2} e^{-\gamma\|\mathrm{z}\|^2} \prod_{i=1}^{k} \frac{y_i^{n_i}}{n_i!} \right) \\
&= \varphi(\mathrm{x})' \varphi(\mathrm{z})
\end{aligned}$$

That is, $\varphi_j(\mathrm{x}) = \sum_{\sum_{i=1}^{k} n_i = j} (2\gamma)^{j/2} e^{-\gamma\|\mathrm{x}\|^2} \prod_{i=1}^{k} \frac{x_i^{n_i}}{\sqrt{n_i!}}, \quad j = 0, \ldots, \infty.$

## 2.B Proof of Proposition 2

**Proof.** (a) First show that $\lim_{\gamma\to 0}(2\gamma)^{-1}K = XX'$, or $\lim_{\gamma\to 0}(2\gamma)^{-1}k_{ij} = X_i'X_j$, $\forall i,j = 1\ldots T$, where $k_{ij} = \widetilde{k}_{ij} - \frac{1}{T}\sum_{l=1}^{T}\widetilde{k}_{il} - \frac{1}{T}\sum_{s=1}^{T}\widetilde{k}_{sj} - \frac{1}{T^2}\sum_{m,p=1}^{T}\widetilde{k}_{mp}$ and $\widetilde{k}_{ij} = e^{-\gamma\|X_i - X_j\|_2^2}$. By L'Hopital's rule we can write $\lim_{\gamma\to 0}(2\gamma)^{-1}k_{ij}$ as

$$-\frac{1}{2}\left\|X_i - X_j\right\|_2^2 + \frac{1}{2T}\sum_{l=1}^{T}\left\|X_i - X_l\right\|_2^2 + \frac{1}{2T}\sum_{l=1}^{T}\left\|X_l - X_j\right\|_2^2 - \frac{1}{2T^2}\sum_{l,m=1}^{T}\left\|X_l - X_m\right\|_2^2.$$

Next, use the fact that $X$ is centered, that is $X'\mathbf{1} = \mathbf{0}$ (zero column means) and hence $\sum_{l=1}^{T} X_i'X_l = \sum_{l=1}^{T} X_l'X_i = 0$, $\forall i = 1\ldots T$. This allows to simplify the above as
$$-\frac{1}{2}X_i'X_i + X_i'X_j - \frac{1}{2}X_j'X_j + \frac{1}{2}X_i'X_i + \frac{1}{T}\sum_{l=1}^{T}X_l'X_l + \frac{1}{2}X_j'X_j - \frac{1}{T}\sum_{l=1}^{T}X_l'X_l = X_i'X_j,$$
$\forall i,j = 1\ldots T$, completing the first step. Hence, since the eigenvectors are normalized, we have $\lim_{\gamma\to 0}(2\gamma)^{-1}Keig_r(K) = sXX'eig_r(XX')$ for $s$ equal either to $+1$ or $-1$. Second, given the SVD decomposition of $X = UDV'$, we have $XX' = VD^2V'$ and $X'X = UD^2U'$, with $D^2 = L$. Thus, $XX'eig_r(XX')L^{-1/2} = UD = Xeig_r(X'X)$.

(b) First show that $\lim_{\gamma\to 0}(\gamma(1 - \tanh^2(c_0))^{-1}k_{ij} = X_i'X_j$, $\forall i,j = 1\ldots T$, where $k_{ij} = \widetilde{k}_{ij} - \frac{1}{T}\sum_{l=1}^{T}\widetilde{k}_{il} - \frac{1}{T}\sum_{s=1}^{T}\widetilde{k}_{sj} - \frac{1}{T^2}\sum_{m,p=1}^{T}\widetilde{k}_{mp}$ and $\widetilde{k}_{ij} = \tanh(c_0 + \gamma X_i'X_j)$. By L'Hopital's rule

$$\lim_{\gamma\to 0}(\gamma(1 - \tanh^2(c_0))^{-1}k_{ij} = X_i'X_j - T^{-1}\sum_{l=1}^{T}X_i'X_l - T^{-1}\sum_{l=1}^{T}X_l'X_j + T^{-2}\sum_{l,m=1}^{T}X_l'X_m,$$

which immediately leads to the result once mean-zero property is taken into account. The second step is analogous to that in (a). ■

## 2.C Proof of Theorem 3

**Proof.** The proof of Theorem 3 for a linear case, $\varphi(X_t) = X_t$, is available in [8]. For a general finite-dimensional $\varphi(\cdot)$ the result follows by applying the original theorem to a vector $\varphi(X_t)$ instead. ■

## 2.D Bounded differences inequality

**Theorem 7 ([86])** *Given independent random variables $X_1, \ldots, X_n \in \mathcal{X}$ and a mapping $f : \mathcal{X}^n \to \mathbb{R}$ satisfying*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} \left| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n) \right| \leq c_i,$$

*then for all $\epsilon > 0$,*

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}(f(X_1, \ldots, X_n)) \geq \epsilon) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}.$$

## 2.E Proof of Lemma 10

**Proof.** Let $\Sigma_x := \varphi(x) \otimes \varphi(x)$. Note that $\|\Sigma_x\| = k(x,x) \leq \bar{k}$, and hence

$$\sup_{x_1,\ldots,x_T,x_i' \in \mathcal{X}} \left| \left\| \frac{1}{T} \sum_{x_1 \ldots x_i \ldots x_T} \Sigma_{x_i} - \mathbb{E}\,\Sigma_X \right\| - \left\| \frac{1}{T} \sum_{x_1 \ldots x_i' \ldots x_T} \Sigma_{x_i} - \mathbb{E}\,\Sigma_X \right\| \right| \leq$$

$$\sup_{x_i \in \mathcal{X}} \frac{1}{T} \left| \|\Sigma_{x_i} - \mathbb{E}\,\Sigma_X\| \right| \leq \frac{2\bar{k}}{T}.$$

Thus, by bounded difference inequality ([86]) we have

$$\mathbb{P}\left(\left\|\widehat{\Sigma}-\Sigma\right\|-\mathbb{E}\left(\left\|\widehat{\Sigma}-\Sigma\right\|\right)\geq 2\bar{k}\sqrt{\frac{\epsilon}{2T}}\right)\leq e^{-\epsilon}.$$

Finally,

$$\mathbb{E}\left(\left\|\widehat{\Sigma}-\Sigma\right\|\right)\leq \mathbb{E}\left(\left\|\widehat{\Sigma}-\Sigma\right\|^2\right)^{1/2}=T^{-1/2}\,\mathbb{E}\left(\|\Sigma_X-\mathbb{E}(\Sigma_X)\|^2\right)^{1/2}\leq\frac{2\bar{k}}{\sqrt{T}},$$

since $\mathbb{E}\left(\|\Sigma_X-\mathbb{E}(\Sigma_X)\|^2\right)=\langle\Sigma_X-\mathbb{E}\left(\Sigma_X\right),\Sigma_X-\mathbb{E}\left(\Sigma_X\right)\rangle\leq 4\bar{k}^2$. ■

## 2.F Proof of Lemma 5

**Proof.** See the proof of Lemma 5.2 in [76]. ■

## 2.G Proof of Theorem 6

**Proof.** Since $\psi_i, \widehat{\psi}_i$ are standardized to be unit length, we have $\left\langle\psi_i,\widehat{\psi}_i\right\rangle^2\leq 1$ by Cauchy-Schwarz inequality. Choosing eigenvector signs so that $\left\langle\psi_i,\widehat{\psi}_i\right\rangle>0$, we have

$$\left\|\psi_i-\widehat{\psi}_i\right\|^2=2-2\left\langle\psi_i,\hat{\psi}_i\right\rangle\leq 2-2\left\langle\psi_i,\hat{\psi}_i\right\rangle^2=\left\|\Pi_i(\Sigma)-\Pi_i(\widehat{\Sigma})\right\|^2.$$

Using Lemma 5,

$$\left\|\Pi_i(\Sigma)-\Pi_i(\widehat{\Sigma})\right\|\leq 4\delta_i^{-1}(\Sigma)\left\|\widehat{\Sigma}-\Sigma\right\|,$$

and hence through Lemma 10 we have

$$\mathbb{P}\left(\left\|\widehat{\psi}_i - \psi_i\right\| \geq \left(1 + \sqrt{\frac{\epsilon}{2}}\right) \frac{8\bar{k}}{\sqrt{T}\delta_i(\Sigma)}\right) \leq e^{-\epsilon},$$

which implies the result. ■

## 2.H Mercer's Theorem

**Theorem 8 ([88])** *Given compact $\mathcal{X} \subseteq \mathbb{R}^d$ and continuous $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, satisfying*

$$\int_{\mathrm{y}} \int_{\mathrm{x}} K^2(\mathrm{x}, \mathrm{y}) d\mathrm{x} d\mathrm{y} < \infty \quad \textit{and} \quad \int_{\mathrm{y}} \int_{\mathrm{x}} f(\mathrm{x}) K(\mathrm{x}, \mathrm{y}) f(\mathrm{y}) d\mathrm{x} d\mathrm{y} \geq 0, \quad \forall f \in L^2(\mathcal{X}),$$

*where $L^2(\mathcal{X}) = \{f : \int f^2(\mathrm{x}) d\mathrm{x} < \infty\}$, then there exist $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and functions $\{\psi_i(\cdot) \in L^2(\mathcal{X}), i = 1, 2, \ldots\}$ forming an orthonormal system in $L^2(\mathcal{X})$, i.e. $\langle \psi_i, \psi_j \rangle_{L^2(\mathcal{X})} = \int \psi_i(\mathrm{x}) \psi_j(\mathrm{x}) d\mathrm{x} = \mathbb{1}_{\{i=j\}}$, such that*

$$K(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathrm{x}) \psi_i(\mathrm{y}), \quad \forall \mathrm{x}, \mathrm{y} \in \mathcal{X}.$$

## 2.I Time Series

Table 2.I.1: Target variables from FRED-MD dataset.

| Group | Fred-code | Description |
|---|---|---|
| Output & income | `RPI` | Real Personal Income |
| Labor market | `CE16OV` | Civilian Employment |
| Housing | `HOUST` | Housing Starts: Privately Owned |
| Consumption & inventories | `DPCERA3M086SBEA` | Real personal consumption |
| Money & credit | `M1SL` | M1 Money Stock |
| Interest & exchange rates | `FEDFUNDS` | Effective Federal Funds Rate |
| Prices | `CPIAUCSL` | CPI: All Items |
| Stock Market | `S&P 500` | S&P's Common Stock Price Index |

# Chapter 3

# Sparse Recovery for COVID-19 Group Testing

## Abstract

[1] Researchers and public officials tend to agree that until a vaccine is readily available, stopping SARS-CoV-2 transmission is the name of the game. Testing is the key to preventing the spread, especially by asymptomatic individuals. With testing capacity restricted, group testing is an appealing alternative for comprehensive screening and has recently received FDA emergency authorization. This technique tests pools of individual samples, thereby often requiring fewer testing resources

[1] This paper is published in the Journal of Data Science under the name "*Fast and Efficient Data Science Techniques for COVID-19 Group Testing*," co-authored with Ekaterina Seregina

while potentially providing multiple folds of speedup. We approach group testing from a data science perspective and offer two contributions. First, we provide an extensive empirical comparison of modern group testing techniques based on simulated data. Second, we propose a simple one-round method based on $\ell_1$-norm sparse recovery, which outperforms current state-of-the-art approaches at certain disease prevalence rates.

## 3.1 Introduction

There is broad consensus among epidemiologists that massive testing a key to preventing the spread of COVID-19. However, large-scale testing is not realistic due to substantial restrictions in testing kits, chemical reagents, skilled personnel and time. Group testing is an appealing alternative to individual testing that suggests to combine a set of individual specimens into a common pool, and test the pool rather than each individual sample. As long as the disease prevalence is not too large, testing pooled samples permits to considerably reduce the total number of tests required for diagnosing the population.

First experiments with pooling samples trace back to dilution studies in 1915 ([65]), which attempted to determine the presence or absence of organisms in a fluid based on pooled information. Researchers cultured samples of the fluid to let the bacteria, if they were present, grow, which served as a test. The results were then gathered across the samples to infer the bacterial density in the original fluid.

Many academics, however, attribute the invention of group testing to a Harvard economist Robert Dorfman, whose influential work ([45]) proposed a simple pooling method for weeding out syphilitic men called up for induction. Instead of analyzing individual blood samples for the presence or absence of a ”syphilitic antigen”, it is suggested to examine pooled samples combining the individual blood sera into groups of five. If the corresponding men are healthy, the pooled test should be negative. On the other hand, if at least one of the patients is syphilitic, the pool will contain antigen, which the test is supposed to reveal. In that case, all associated patients need to be retested individually. Putting aside possible dilution concerns, it is clear that such strategy leads to savings of chemical reagents and higher overall testing capacity in a population with low disease prevalence. The idea is illustrated in Figure 3.1.1.

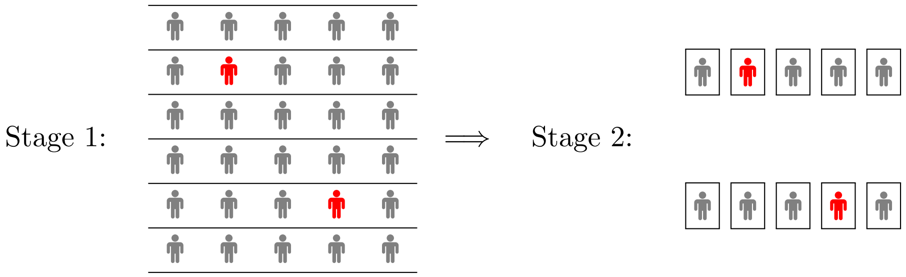


Figure 3.1.1: Dorfman pooling illustration.

“User” icons represent individuals, red are infected and grey are healthy. In the first stage, all $N = 30$ specimen are pooled into $N/n = 6$ groups (rows) of $n = 5$, which are then tested. In the second stage, everyone in infected groups (rows two and four) is tested individually. As a result, it is possible to detect $k = 2$ positives with $6 + 10 < N = 30$ tests.

Due do its simplicity, Dorfman’s two-stage approach has found widespread use in medicine. Many of its properties are also readily available. Suppose we collect $N$ individual samples and pool them into $N/n$ groups of size $n$. Given the disease prevalence rate per hundred, $p$ (which is also the probability of a randomly selected individual being positive), the expected number of tests for diagnosing the population is

$$\mathbb{E}(T) = N/n + \underbrace{(1-(1-p)^n)}_{\mathbb{P}(\text{at least one positive})} n\frac{N}{n}. \tag{3.1}$$

The first term on the right-hand side corresponds to the number of tests in the first stage, the second term is $n$ times the expectation of a random variable distributed as $\textsc{Bi}(N/n, (1-(1-p)^n))$, characterizing the number of positive groups in the second stage. Clearly, this method is more beneficial at a lower prevalence rate $p$. For fixed $\mathbb{E}(T)$ and $p$, one could also optimize over the pool size $n$ to get the largest possible coverage $N$,

$$n^* = \frac{2W(-\frac{1}{2}\sqrt{-\ln(1-p)})}{\ln(1-p)}, \tag{3.2}$$

where $W(\cdot)$ is the Lambert $W$ function. Notice that this also is decreasing in $p$ and, interestingly, is independent of $\mathbb{E}(T)$. The expected number of tests per person is approximately minimized when the group size is $n = 1/\sqrt{p}$ and hence the expected number of tests per person is $2\sqrt{p}$. Graphs illustrating the above relationships are provided in Appendix 3.A.

[109] proposed a modification to Dorfman's second stage: instead of testing every individual, one would only do so until a positive sample is found, after which continue with group testing. In that case, if the prevalence is low, it is likely that the new sub-pool will be negative. This leads to an increase in savings of tests from Dorfman's 80% (compared to individual testing) to 86% at 1% rate. There have been other alternatives as well, e.g. [107],

halving techniques in [83] and others. These methods generally trade off higher efficiency with more complexity and longer wait times.

In group testing literature, Dorfman's approach and its modifications are classified as adaptive (or hierarchical), in a sense that they "adapt" to the results of preceding stages. An alternative approach, known as non-adaptive (or non-hierarchical), designs a single-stage experiment, results of which should allow to infer (often in a probabilistic manner) the original assignment of positives and negatives. This should generally, although not necessarily, come at the cost of having to run more tests overall since the sequential approach takes in more information. A distinctive feature of the non-adaptive approach is in assigning a single individual to multiple groups, i.e. groups are overlapping. How to design such assignment is a crucial question that is considered later. In a scenario such as the current SARS-CoV-2 pandemic, with the disease spreading fast and standard testing kits not showing results immediately, such non-adaptive approaches would have a clear advantage.

Furthermore, with multiple stages of testing, adaptive techniques also bear the risk of running out of tests before learning the outcomes. Given a limited number of tests, Dorfman's approach may require more test than are available. In contrast, single-round designs do not suffer from such indeterminacy.

Hence, we focus on such "fast" single-stage techniques. We consider several recent combinatorial and probabilistic algorithms. Importantly, we propose a simple method based on $\ell_1$-norm sparse recovery, which outperforms the above algorithms.

Pooling strategies help resolve two kinds of problems, namely estimation and classification. The first seeks to estimate the prevalence of positive individuals in a population. The second, which may or may not rely on the information on estimated prevalence, aims to identify the infected individuals. The performance is typically gauged by the expected number of tests required for a given specificity or sensitivity, or conversely, based on predictive accuracy for a given number of tests. We focus on classification.

### 3.1.1 RT-qPCR test

The two main ways of determining whether an individual has a SARS-CoV-2 virus are (1) to check for the presence of antibodies to the virus, (2) to check for the presence of the virus RNA itself. The former, although capable of uncovering whether a recovered individual had the virus in the past, is less widespread; the latter includes so-called reverse transcription quantitative polymerase chain reaction (RT-qPCR), the gold standard for COVID-19 testing recommended by the Centers for Disease Control and Prevention. Though popular, massive individual PCR testing is not possible due to serious constraints in equipment, chemical reagents and skilled personnel. The resulting readouts of RT-qPCR are of key interest to this study so we briefly describe the testing process.

To perform this test, nasopharyngeal swabs from subjects are collected and diluted in a fluid medium. The first stage, reverse transcription, then transforms the virus RNA to complementary DNA (cDNA). This eventually allows to start the next stage, polymerase chain reaction, which aims to exponentially amplify the viral cDNA molecules through the process that involves up to nearly 40 cycles of heating and cooling. To trace this increase, the virus-specific sequences are marked by fluorescent. The testing machine then measures the amount of fluorescent signal in real time and displays it as a function of cycles. This information we are interested in is when (if at all) the fluorescence exceeds the critical level associated with a positive subject. This is given by a cycle threshold ($C_t$), the number of cycles completed before crossing the threshold. The subject is then declared positive if the threshold is exceeded before about 40 cycles. The $C_t$ indicator is (negatively) correlated with the original viral load, with larger initial viral loads leading to sooner crossing of the threshold and thus shorter cycle thresholds. The entire process takes up to approximately 4 hours.

The information on cycle thresholds of pooled samples is the key input to group testing algorithms. Surprisingly, many known group testing algorithms do not take this quantitative information into account and instead work with degenerate binary transformations. The algorithm proposed in this study is capable of not only incorporating the quantitative information, but also producing corresponding quantitative predictions measuring the original individual viral loads.

### 3.1.2 Biomedical Considerations

One of the major concerns with pooling approaches is dilution. Fortunately, there is growing evidence that pooling of SARS-CoV-2 with negative samples does not lead to substantial dilution of the virus DNA. In a recent study ([127]), Israeli researchers discovered that it is possible to detect a single positive SARS-CoV-2 sample in pools of up to 64 samples with reasonably high accuracy. That is, the fluorescent signal of a pooled positive sample, diluted with up to 63 negative samples, amplifies sufficiently to cross the required threshold. Other investigations ([1], [64], [90]) tend to agree with such claims.

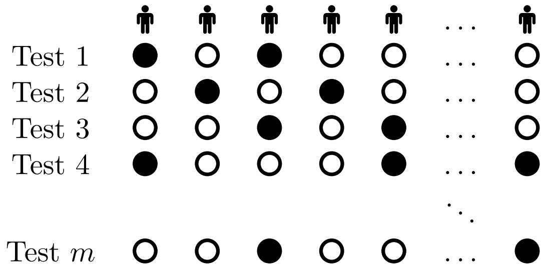
Pooled testing for SARS-CoV-2 has been conducted in a number of countries, including the United States (Stanford Health Care Clinical Virology Laboratory and Nebraska's Public Health Laboratory), Germany (University Hospital Frankfurt at Goethe University), China and Israel (Rambam Health Care Campus).

Group testing has been used for detecting the HIV ([48]); in fact, it is now a routine option in blood screening. Pooling not only decreases the cost but also the probability of making an error in low disease prevalence populations. Pooling has also been deployed against malaria ([114]), influenza ([119]) and a few other diseases. It has also found its use in non-medical settings, for example detecting defective units in manufacturing, computer fault diagnosis or testing collections of documents in data forensics.

## 3.2 Non-adaptive Group Testing

We first briefly review some of the recent non-adaptive algorithms introduced in the literature ([34]). As all are one-round approaches, it is required to construct an $m \times N$ pooling matrix A [2], which would assign each of the $N$ individuals to one or more of the $m$ groups. As opposed to adaptive testing, we may have the same subject sample split among several groups. Figure 3.2.1 illustrates the idea. The corresponding pooling matrix has its $(i, j)$th entry equal to one if an $i$th individual is assigned to group $j$, $i \leq N, j \leq m$, and zero otherwise. One would typically also normalize the matrix, but this does not raise any substantial challenges so we omit this issue.

[2] This matrix is not to be confused with the $(N/n) \times n$ Dorfman "matrix" in Figure 3.1.1. Dorfman approach does not involve allocating individuals to groups.



|  | 🧍 | 🧍 | 🧍 | 🧍 | 🧍 | … | 🧍 |
|---|---|---|---|---|---|---|---|
| Test 1 | ● | ○ | ● | ○ | ○ | … | ○ |
| Test 2 | ○ | ● | ○ | ● | ○ | … | ○ |
| Test 3 | ○ | ○ | ● | ○ | ● | … | ○ |
| Test 4 | ● | ○ | ○ | ○ | ● | … | ● |
|  |  |  |  |  |  | ⋱ |  |
| Test $m$ | ○ | ○ | ● | ○ | ○ | … | ● |

Figure 3.2.1: Illustration of a pooling matrix assigning $N$ individuals to $m$ pooled tests. A black circle indicates that the corresponding individual (column) has been assigned to the given test (row). If an individual is assigned to several tests, his sample is split accordingly.

Once the pooling matrix is specified, one then observes the results of $m$ pooled tests via an $m \times 1$ vector y and the goal is to identify which of the $N \gg m$ individuals are truly positive. We briefly discuss novel algorithms and provide an intuitive explanations behind their principles. Exact algorithmic formulations can be found in our code.

Combinatorial Basis Pursuit (CBP) is a simple algorithm that is based on the following idea: declare all individuals that were included in negative pools as negative (since if at least one sample was positive, the entire group would have been positive), and declare the remaining individuals as positives. Since this strategy would identify healthy individuals for certain, it will not produce any false negatives.

Instead of looking at each test (row), one may instead try to decode the matrix "column"-wise, i.e. by going over the individuals. This is what the combinatorial orthogonal matching pursuit (COMP) algorithm does: if all tests an individual participated in turn out to be positive, then the individual is considered to be infected, and negative otherwise. This deciphering method never produces false negatives, only false positives. A false positive would only occur if a healthy individual happened to always participate in tests that contained at least one infected sample; the probability of this happening decreases with $m$.

Definite defectives (DD) algorithm starts with COMP to leverage its ability of identifying true negatives. Once COMP is over, DD switches to “row”-wise search by looking at positive tests and seeks to determine individuals that are “definite defectives” (positives). All remaining subjects are declared negative. This reversal in the algorithm leads to DD producing only false negatives, and no false positives. This leads to greater accuracy in sparse settings: because there are a lot more healthy subjects, one should by default assume that an individual is not infected, all else equal.

Finally, sequential COMP (SCOMP) further attempts to improve DD by modifying its last step of labeling remaining subjects as negatives. The key is to observe that if the current set of individuals that are declared positive cannot explain all of the positive pooled tests, one can do better by sequentially declaring potential positives as positives until the set of positives accounts for all positive tests. From a list of potential candidates, the algorithm picks the one that would account for the largest number of unexplained tests. SCOMP has been shown to perform close to the information-theoretic bound.

## 3.3 Problem Formulation

We now turn to our algorithm that leverages recent advancements in the field of compressed sensing in engineering and statistics literature ([30, 43]). Our main goal is to efficiently infer x, the $N$-dimensional sparse vector of individual viral loads, from $m \ll N$ available group test results stacked in a vector y. In general, we have $\mathrm{y} = g(\mathrm{Ax}) + \epsilon$, however we will restrict our attention to the simplest case when $\mathrm{y} = \mathrm{Ax} + \epsilon$. Hence, $\mathrm{A} \in \mathbb{R}^{m \times N}$ represents a set of linear measurements on the variable of interest x. This formulation has a crucial difference with regression type of problems: in our setting one gets to choose how to design the pooling matrix A, while in regression problems A is pre-determined by the data. Thus, there are two steps to solving such problems.

The first step is to encode the sparse signal, by designing a proper pooling matrix A. This matrix provides the assignments for each individual specimen to the corresponding groups and must satisfy certain desirable conditions pertaining to the group testing problem.

The second step attempts to decipher the first step with fewest test measurements. For a large-dimensional vector x finding the corresponding sparsest vector that would be consistent with $m$ pooled observations is an NP-hard problem. However, recent advancements in engineering allow to transition this problem to a convex domain where exact decoding is feasible with high probability.

While x is generally a vector of quantitative measurements of all individuals, one can equivalently think of it as a sparse vector, i.e. with most entries equal to 0 (associated with healthy individuals) and very few 1's, without loss of generality. What matters is that the vector needs to be sparse in some transformed coordinate system.

### 3.3.1 Pooling Matrix Design

We first focus on the design of a pooling (also known as sensing or measurement) matrix A. Due to the nature of our primary application, we only consider sparse pooling matrices with few nonzero elements. When properly formed, this should ensure there is not too much dilution: a single sample is not split into too many subsamples and any one group sample does not contain too many specimens. The simplest approach would be to generate a random Bernoulli matrix with entries, which together with a normally distributed random matrix, has been shown to satisfy desirable properties, mainly the null space condition (NSC) and the restricted isometry property (RIP) discussed later, that guarantee a precise recovery of the original vector of interest with high probability.

We instead use a pooling matrix that was proposed in a different branch of group testing literature. Known as a constant column weight design ([2]), it was shown to outperform simple Bernoulli matrices in terms of its encoding capabilities. The initial approach outlined in [2] constructs A by inserting up to an $L$ of ones into each column. Concretely, $L$ indices of each column are sampled with replacement and ones are inserted in the unique positions. This complication seems to be necessary for their proofs, however the real performance does not depend on whether one bootstraps or simply permutes a fixed number of ones. Hence, we focus on a simpler, permutation version. One example of such matrix with $N = 6$, $L = 2$ and $m = 4$ is

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

As discussed earlier, the $i$th individual is assigned to group $j$ only if the $(i, j)$th entry is 1. This design avoids too much dilution as long as $0 < L < m \ll N$, and outperforms Bernoulli design. For simplicity, in our experiments we set $L = \lceil m/2 \rceil$ (rounded up), but this value could be theoretically and practically optimized as is done in [2] and [66]. Importantly, we prove that this design is RIP which has immediate theoretical implications, which are discussed in the next section.

**Theorem 9** *A random matrix* $\mathrm{A} \in \mathbb{R}^{m\times N}$ *of constant column weight design with* $L$ *ones in each column satisfies the restricted isometry property with high probability, specifically there exists* $\delta \in (0,1)$ *such that*

$$(1-\delta)\left\|\mathrm{x}\right\|_2^2 \leq \left\|\mathrm{Ax}\right\|_2^2 \leq (1+\delta)\left\|\mathrm{x}\right\|_2^2$$

*holds with high probability for any* $\mathrm{x} \in \mathbb{R}^N$ *and* $0 < L < m$.

**Proof.** See Appendix **??**. ■

The $\ell_p$ norm of a vector $\mathrm{v} \in \mathbb{R}^d$ is $\left\|\mathrm{v}\right\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$, which is a norm for $1 \leq p \leq \infty$.

### 3.3.2 $\ell_1$-norm Sparse Recovery

Once $m$ measurements in y are formed, one can employ several strategies for decoding the original signal. A direct, brute force approach to tackle the problem would be to find the sparsest vector of viral loads x that is consistent with the linear measurements, that is

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \left\|\mathrm{x}\right\|_0 \quad \text{s.t.} \quad \left\|\mathrm{Ax}-\mathrm{y}\right\|_2 \leq \epsilon.$$

Unfortunately, this problem is NP-hard as its solution requires an exhaustive search over all possible combinations in x, although this may still be feasible for low-dimensional problems. Luckily, a convenient convex relaxation is available, which has been proven to yield accurate solutions as long as the sensing matrix A satisfies RIP ([30],[43]). This is a sufficient condition, which Theorem 9 shows to hold with high probability. In practice one could generate a random matrix and attempt to verify whether RIP holds for a given matrix, although doing so is by itself NP-hard ([17]). The corresponding convex alternative is

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \|\mathrm{x}\|_1 \quad \text{s.t.} \quad \|\mathrm{Ax}-\mathrm{y}\|_2 \leq \epsilon. \tag{3.3}$$

This is known as Basis Pursuit Denoising ([102]), although many statisticians are more familiar with its equivalent formulation, Lasso,

$$\min_{\mathrm{x}\in\mathbb{R}^N} \quad \|\mathrm{Ax}-\mathrm{y}\|_2^2 + \lambda \|\mathrm{x}\|_1$$

The two problems are identical for certain choices of $\epsilon$ and $\lambda$. We simply add a nonnegativity constraint $\mathrm{x} \geq 0$ which reflects the inherent characteristic of the problem and also improves the empirical performance.

This type of $\ell_1$-norm recovery uses $m = O(k\log(N))$ tests while standard group testing algorithms require $m = O(k^2\log(N))$ tests. Another advantage of this approach is its ability to handle real-valued quantitative readouts; many group testing algorithms are only capable of dealing with binary measurements. Furthermore, the output is also a real-valued number estimating individual's viral load.

## 3.4 Application

This section compares the performance of the above algorithms in simple numerical experiments with no noise. For clarity of exposition, the vector of interest x is generated to be a binary 0-1 vector instead of a real-numbered qPCR-like measurements. More general treatments can be found in our code.

We consider a case with $N = 100$ specimens where there are $k = 2$ true positive cases. This is a conservative estimate in a sense that this share of positives is larger than the share of active cases in the United States as of December 1, 2020 ([124]). In Appendix 3.C we additionally report cases with $k = 1, 3, 4, 5$.

To illustrate, we generate a 100-dimensional binary vector with 2 ones and the sensing matrices as described above to obtain $m = 20$ linear measurements. We then apply the decoding algorithms to try to infer the original binary vector (both the number of positive $k$ and their positions) with only 20 measurements. Figure 3.4.1 demonstrates a particular realization where only the proposed algorithm, denoted as SR (for sparse recovery), is capable of correctly identifying the positions. Other algorithms produce either false positives, false negatives or both.

Specifically, to obtain SR estimates we first solve

$$\tilde{\mathrm{x}} = \underset{\mathrm{x} \in \mathbb{R}^N}{\arg\min} \quad \|\mathrm{Ax} - \mathrm{y}\|_2^2 + \lambda \|\mathrm{x}\|_1 , \quad \mathrm{x} \geq 0, \tag{3.4}$$

which generally would not produce 0-1 estimates. Hence we simply round the estimates at the threshold value of $\tau = .5$, that is $\widehat{x}_i = \begin{cases} 1 & \text{if } \tilde{x}_i \geq \tau \\ 0 & \text{if } \tilde{x}_i < \tau. \end{cases}$

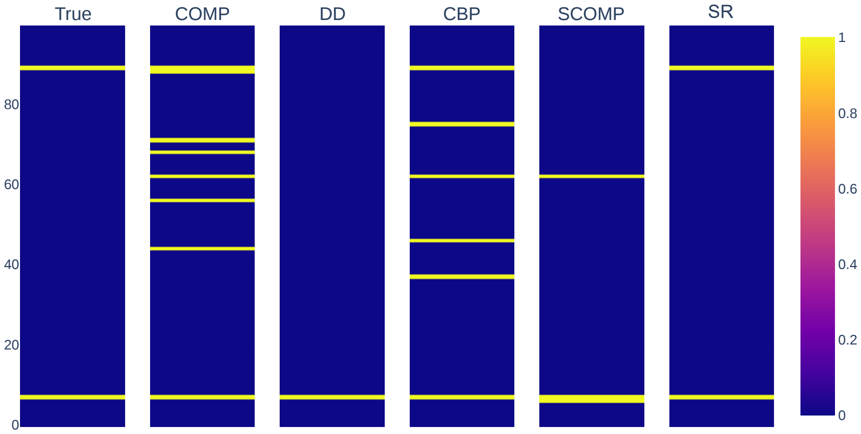


Figure 3.4.1: Identification of negative and positive positions.

Only the proposed (SR) algorithm is able to successfully determine which samples correspond to positive/negative specimens.

Next, we repeat this process for 1000 iterations across different group sizes $m$ and report the average root mean square error (RMSE) plotted against the number of test measurements $m$ in Panel A of Figure 3.4.2. RMSE is defined as $\frac{\|\mathrm{x}-\widehat{\mathrm{x}}\|_2}{\|\widehat{\mathrm{x}}\|_2}$, where x is a true binary vector and $\widehat{\mathrm{x}}$ is one of the estimates. We still keep $N = 100$ and $k = 2$, but Appendix 3.C reports cases for $k = 1, 3, 4, 5$.

As can be seen, the proposed method makes approximately the same error with $m = 15$ tests as the best alternative (SCOMP) with $m = 30$ tests. For comparison, Dorfman approach would require approximately $m = 30$ tests and two testing stages. As the more detailed comparison in Appendix 3.C shows, SR is still superior for $k \geq 2$, but loses dominance to SCOMP for $k = 1$.

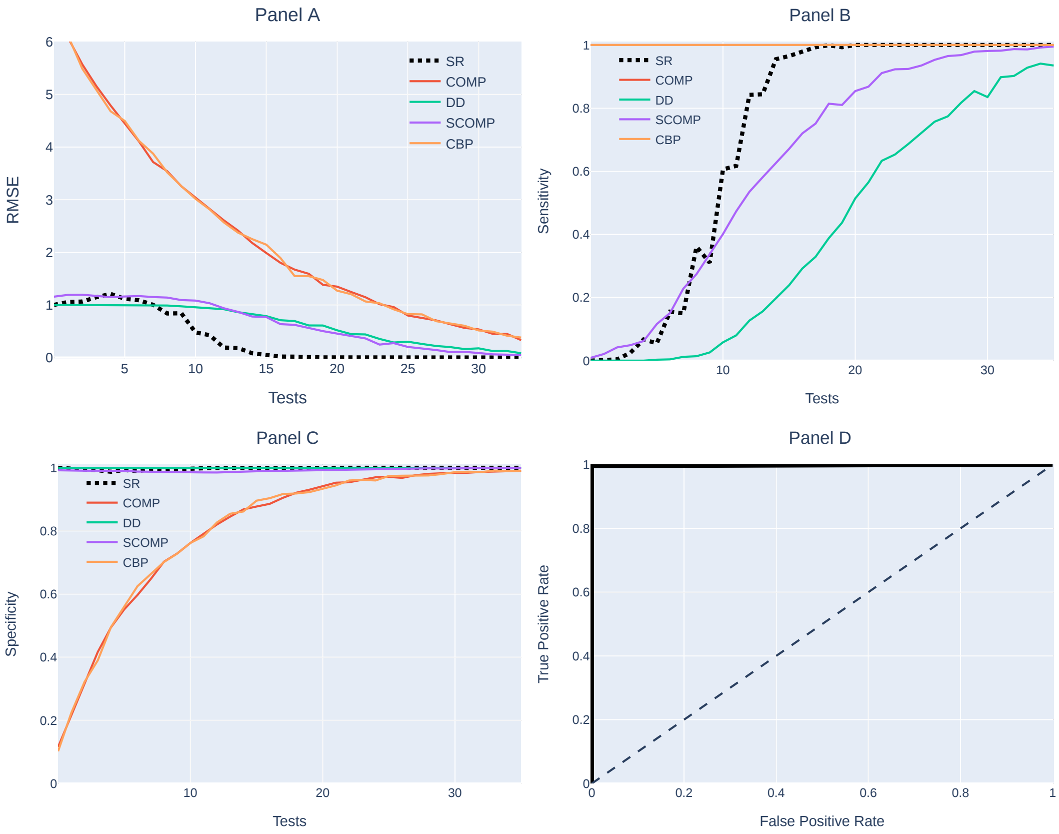


Figure 3.4.2: RMSE, Sensitivity, Specificity and ROC for $N = 100, k = 2$.

*Panel A*: RMSE of each approach as a function of test measurements $m$. SR outperforms standard non-adaptive group testing algorithms. *Panel B*: Sensitivity of each approach as a function of test measurements $m$. *Panel C*: Specificity of each approach as a function of test measurements $m$. *Panel D*: ROC, thresholding SR estimates. AOC = .9995.

However, RMSE does not tell the whole story. One is also interested in sensitivity (or true positive rate) and specificity (or true negative rate). These are defined as the ratio of identified positives to all true positives and the ratio of identified negatives to all true negatives respectively, and reported in Panel B and C of Figure 3.4.2.

Notice that CBP and COMP report perfect sensitivity. This is a sanity check since these algorithms should not produce any false negatives. Among the other algorithm SR is again a clear winner. Naturally, the relationship between the two groups reverses for specificity: we have SR, COMP and DD achieving ideal (or almost ideal) specificity with a minimum number of tests, while COMP and CBP slowly catch up.

Additionally, we plot the receiver operating characteristic curve (ROC) for SR in Panel D of Figure 3.4.2, where we keep the same parameter values $N = 100$, $m = 20$, $k = 2$. The corresponding area under the curve (AUC) is .9995. This curve traces the true and false positive rates for different values of the threshold $\tau$. The figure is indicative of a strong classification ability of the proposed method.

Finally, we report so-called improvement factors in Table 3.4.1, given as the ratio of the number of specimens $N$ to the expected number of tests required for achieving at least 95% in specificity & sensitivity. When needed, the expected number is computed through Monte Carlo averaging. An improvement factor measures the effectiveness of a given method by computing how many more tests standard individual testing would need compared to a group testing algorithm. It essentially provides an estimate of how many individuals one group test effectively "covers". For the five prevalence ratios, SR dominates both the non-adaptive algorithms and Dorfman approach for $k/N \geq 2\%$, while both CBP and SCOMP seem to be more efficient for $k/N = 1\%$.

| | $\frac{k}{N} = 1\%$ | $\frac{k}{N} = 2\%$ | $\frac{k}{N} = 3\%$ | $\frac{k}{N} = 4\%$ | $\frac{k}{N} = 5\%$ |
|---|---|---|---|---|---|
| Dorfman | 4.02 | 3.37 | 2.93 | 2.60 | 2.34 |
| COMP | 8.10 | 4.69 | 3.63 | 2.80 | 2.21 |
| DD | 6.25 | 2.86 | 2.39 | 1.99 | 1.99 |
| CBP | 9.42 | 4.49 | 3.60 | 2.81 | 2.19 |
| SCOMP | 9.83 | 3.95 | 3.06 | 2.48 | 2.10 |
| SR | 8.69 | 7.14 | 5.54 | 4.42 | 3.07 |

Table 3.4.1: Improvement factors, $N/\,\mathbb{E}(\#\text{ of tests})$, for different prevalence rates.

## 3.5 Related Work

A closely related work by [128] also considers the techniques based on compressed sensing. However, there are differences in both encoding and decoding steps. While their pooling matrix is also sparse, its 0-1 entries are generated from Bernoulli distribution with probability .5. [2] and [66] compared the encoding capabilities of Bernoulli and constant column weight design matrices and documented substantial theoretical and empirical superiority of the latter in nonadaptive testing settings. The superiority is even more pronounced for sparse settings; in fact, such design "*in sparse cases is the best proven performance of any practical algorithm*" ([2]). Furthermore, Bernoulli design is not suitable for biomedical considerations as it is prone to undesirable extremes: one may have columns with all zeroes (or all ones) which corresponds to the case when the sample is not used at all (or is being split into too many subsamples); this is excluded by constant column weight design with $0 < L < m$. Another design [128] consider are expander matrices, which perform on par with Bernoulli in their experiments. However, the precise structure (for example, the number of ones per column) and theoretical fitness of such matrices are not discussed. Their decoding algorithm is non-negative BPDN (Eq. (3.3) coupled with a nonnegativity constraint) which has a less readily available software implementation than non-negative Lasso. We were not able to obtain the code of [128] to make a direct comparison.

Another extensive study by [58] also focuses on sparse nonadaptive methods. Researchers used so-called Kirkman triple matrices, which as in our work assign a finite number of $1'$s per column. However, this design suffers from a major restriction on the dimensions of the sensing matrix, $m$ and $N$, the number of groups and individuals respectively. Specifically, $m$ is only allowed to be an integer multiple of 3, while $N$ needs to be an integer multiple of $m/3$. This makes direct Monte Carlo comparison with classical approaches, as is done e.g. in Figure 3.4.2, difficult, since $m$ cannot vary freely. To resolve this by sequentially "truncating" their pooling matrix, i.e. only taking the first $m$ rows from a fixed size $24 \times 60$ Tapestry matrix. The results are plotted in Appendix 3.D. Our approach is more precise for lower disease prevalence rates, approximately for $k/N \leq 8\%$. Finally, although their matrix is claimed to satisfy RIP, we have not seen the proof.

## 3.6 Concluding Remarks

Pooled testing has been around for more than 70 years and has been successfully employed against a number of diseases. There are reasons to believe that pooling can also be effective against SARS-CoV-2. First, low prevalence of the virus is crucial to making group testing effective. Second, the recent evidence with dilution experiments suggests that pooling can be a viable method. Third, pooling is also compatible with the widely used testing kits such as RT-qPCR. Finally, group testing has been authorized by the FDA ([54]), which claimed it to be "especially important as infection rates decline and we begin testing larger portions of the population."

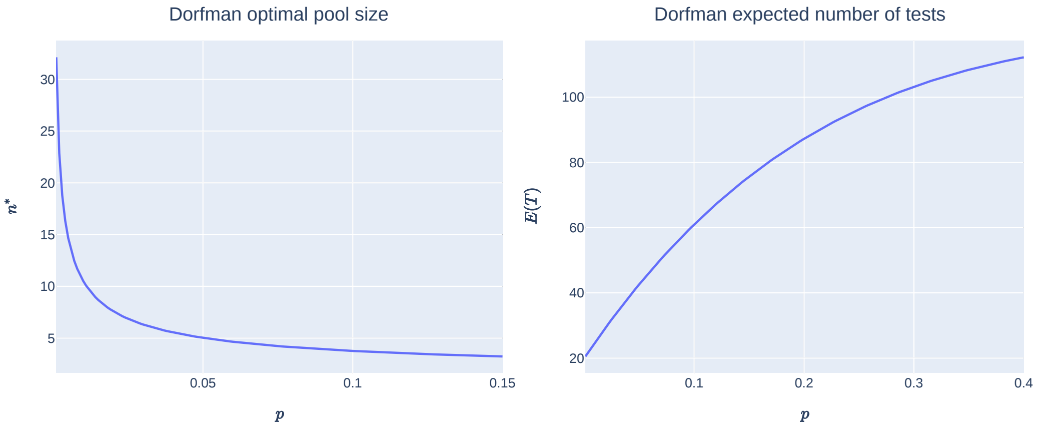
To this end, we considered a simple one-stage group testing method that is able to diagnose a large number of specimens with the fewest number of tests and thus substantially increase the throughput of testing. Our approach does not require to know the number of positive samples in population to run and compares favorably based on the experiments on synthetic data. It produces very few false positives and false negatives, and is also capable of predicting viral loads. Compared to widely used adaptive strategies it minimizes latency in delivering test results, while compared with non-adaptive strategies it only requires $m \sim O(k \log n)$ tests. The numerical results suggest this approach is appealing for a wide range of prevalence rates; particularly, it outperforms standard non-adaptive methods at prevalence rates greater than 1%, and performs better than Kirkman matrices ([58]) for prevalence rates lower than 8%.

## Supplemental Material

The code supplement [78] is available in Google Colab environment. It is written in Python and readily allows to replicate all the graphs provided, as well as produce additional exercises.

# Appendices

## 3.A Dorfman group testing figures


Dorfman optimal pool size

Dorfman expected number of tests

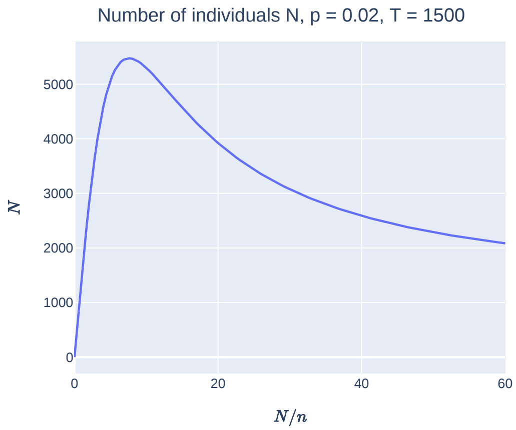


Figure A.1: Dorfman group testing

*Upper left*: Dorfman theoretical optimal pool size, $n^*$, plotted over the prevalence rate, $p$. Larger pools sizes are preferred at a lower disease prevalence. *Upper right*: Dorfman expected number of tests, $\mathbb{E}(T)$, plotted over the prevalence rate, $p$. The benefits of classical approach diminish at higher prevalence rates. *Lower*: Number of individuals, $N$, Dorfman strategy can cover at 2% prevalence with $T = 1500$ tests, plotted over the number of groups, $N/n$. The maximum is achieved at $n = n^*$.

## 3.B Proof of Theorem (9)

$\mathrm{A} = \{a\}_{ij}$ is an $m \times N$ matrix where each column randomly permutes $0 < L < m$ ones among zeros. Without loss of generality, let us assume that each column has been demeaned and normalized to be of unit length, i.e. divided by $\sqrt{L\left(1 - \frac{L}{m}\right)^2 + (m-L)\left(\frac{L}{m}\right)^2} = \sqrt{L\left(1 - \frac{L}{m}\right)}$. It is then evident that $\mathbb{E}(a_{ij}) = 0$ and $\mathbb{E}(a_{ij}^2) = \frac{1}{m}$.

First, we want to show that for any fixed $\mathrm{x} \in \mathbb{R}^N$, the random variable $\|\mathrm{Ax}\|_2^2$ concentrates around its mean, i.e.

$$\Pr\left(\left|\|\mathrm{Ax}\|_2^2 - \|\mathrm{x}\|_2^2\right| \geq \epsilon \|\mathrm{x}\|_2^2\right) \leq 2e^{-m(\epsilon^2/4 - \epsilon^3/6)}. \tag{3.5}$$

For $i = 1, \ldots, m$, denote $c_i$ the $i$th entry of Ax, i.e. $c_i = \sum_{j=1}^{N} a_{ij} x_j$, then

$$\begin{aligned}
\mathbb{E}\, c_i = \mathbb{E} \left( \sum_{j=1}^{N} a_{ij} x_j \right) &= \sum_{j=1}^{N} \mathbb{E}(a_{ij}) x_j = 0, \\
\mathbb{E}(c_i^2) = \mathbb{E} \left( \left( \sum_{j=1}^{N} a_{ij} x_j \right)^2 \right) &= \mathbb{E} \left( \sum_{j=1}^{N} (a_{ij} x_j)^2 + 2 \sum_{l=1}^{N} \sum_{m=1}^{N} a_{lj} a_{mj} x_l x_m \right) \\
= \sum_{j=1}^{N} \mathbb{E}(a_{ij}^2) x_j^2 + 2 \sum_{l=1}^{N} \sum_{m=1}^{N} \mathbb{E}(a_{lj})\, \mathbb{E}(a_{mj}) x_l x_m &= \frac{1}{m} \|\mathrm{x}\|_2^2,
\end{aligned}$$

and hence $\mathbb{E}(\|\mathrm{Ax}\|_2^2) = \mathbb{E}\left(\sum_{i=1}^{m} c_i^2\right) = \sum_{i=1}^{m} \mathbb{E}(c_i^2) = \|\mathrm{x}\|_2^2$.

Since $\|\mathrm{Ax}\|_2^2$ is proportional to $\|\mathrm{x}\|_2^2$, it is sufficient to demonstrate the concentration for arbitrary unit vectors. For all fixed unit vectors $\mathrm{x} \in \mathbb{R}^N$,

$$\mathbb{P}\left( \|\mathrm{Ax}\|_2^2 > 1 + \epsilon \right) = \mathbb{P}\left( e^{t\|\mathrm{Ax}\|_2^2} > e^{t(1+\epsilon)} \right) \tag{3.6}$$

$$< \mathbb{E}\left( e^{t\|\mathrm{Ax}\|_2^2} \right) e^{-t(1+\epsilon)}, \tag{3.7}$$

where (3.6) and (3.7) simply apply the Chernoff technique for $t > 0$. Now, because the columns $c_i$ are i.i.d. we have $\mathbb{E}\left(e^{t\|\mathrm{Ax}\|_2^2}\right) = \mathbb{E}\left(e^{t\sum_{i=1}^{m} c_i^2}\right) = \left(\mathbb{E}\left(e^{tc_1^2}\right)\right)^m$, leading to

$$\mathbb{P}\left( \|\mathrm{Ax}\|_2^2 > 1 + \epsilon \right) < \left( \mathbb{E}\left( e^{tc_1^2} \right) \right)^m e^{-t(1+\epsilon)} \tag{3.8}$$

$$\leq (1 - 2t/m)^{-m/2} e^{-t(1+\epsilon)}, \tag{3.9}$$

where (3.9) follows from Lemma 10. Optimizing this bound with respect to $t$, $t^* = \frac{m\epsilon}{2(1+\epsilon)}$, we can write

$$\begin{aligned}
\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 > 1+\epsilon\right) &< ((1+\epsilon)e^{-\epsilon})^{m/2} && (3.10)\\
&< e^{-m(\epsilon^2/4-\epsilon^2/6)}, && (3.11)
\end{aligned}$$

where the last inequality comes from truncating the Taylor approximation of (3.10). Similarly, for the other bound,

$$\begin{aligned}
\mathbb{P}\left(\|\mathrm{Ax}\|_2^2 < 1-\epsilon\right) &< \left(\mathbb{E}\left(e^{-tc_1^2}\right)\right)^m e^{t(1-\epsilon)} && (3.12)\\
&< \left(\mathbb{E}\left(1 - tc_1^2 + t^2c_1^4/2\right)\right)^m e^{t(1-\epsilon)} && (3.13)\\
&\leq \left(1 - \frac{t}{m} + \frac{3t^2}{2m^2}\right)^m e^{t(1-\epsilon)} && (3.14)\\
&= \left(1 - \frac{\epsilon}{2(1+\epsilon)} + \frac{3\epsilon^2}{8(1+\epsilon)^2}\right)^m e^{\frac{m\epsilon(1-\epsilon)}{2(1+\epsilon)}} && (3.15)\\
&< e^{-m(\epsilon^2/4-\epsilon^3/6)}, && (3.16)
\end{aligned}$$

where (3.13) and (3.16) is a Taylor approximation, (3.14) uses the fact $\mathbb{E}(c_1^4) = \frac{1}{m^2} \leq \frac{3}{m^2}$ and (3.15) plugs in the earlier value of $t^*$.

**Lemma 10** *For $m \geq 1$ and all $\mathrm{x} \in \mathbb{R}^N$ s.t. $\|\mathrm{x}\|_2^2 = 1$, $\mathbb{E}\left(e^{tc_1^2}\right) \leq (1 - 2t/m)$, $\forall t \in [0, m/2]$.*

**Proof.** Let $W \sim \mathcal{N}(0, \frac{1}{m})$, then

$$\begin{aligned}
\mathbb{E}\left(e^{tc_1^2}\right) =& \sum_{i=1}^{\infty} \frac{t^i}{i!} \, \mathbb{E}\left(c_1^{2i}\right) && (3.17) \\
\leq& \sum_{i=1}^{\infty} \frac{t^i}{i!} \, \mathbb{E}\left(W^{2i}\right) && (3.18) \\
=& \, \mathbb{E}\left(e^{tW^2}\right) && (3.19) \\
=& (1 - 2t/m)^{-1/2}. && (3.20)
\end{aligned}$$

Observe that for $t \in [0, m/2]$ the expectations in (3.17) and (3.19) are bounded, allowing to push the expectation inside the limiting sums in (3.17) and (3.18). Inequality in (3.18) holds since $\mathbb{E}(c_1^{2i}) = m^{-i} \leq \mathbb{E}\left(W^{2i}\right) = m^{-i} \frac{(2i)!}{i! 2^i}$ holds for each $i = 0, 1, 2, \ldots$. ■

Given the concentration of $\|\mathrm{Ax}\|_2^2$ around its mean, RIP follows from Lemma 5.1 in [18], which is adapted and reiterated below for completeness.

**Lemma 11** *Let a random matrix $A \in \mathbb{R}^{m \times N}$ satisfy the concentration inequality in* (3.5)*. Then, for any set $T$ with $q = \#(T) < m$ and any $0 < \delta < 1$, we have*

$$\Pr\left((1-\delta)\|\mathrm{x}\|_2^2 \leq \|\mathrm{Ax}\|_2^2 \leq (1+\delta)\|\mathrm{x}\|_2^2\right) \geq 1 - 2(12/\delta)^q e^{-(\delta/2)m(\epsilon^2/4 - \epsilon^3/6)}$$

**Proof.** See Lemma 5.1 in [18]. ■

## 3.C Additional experiments for $k = 1, 3, 4, 5$

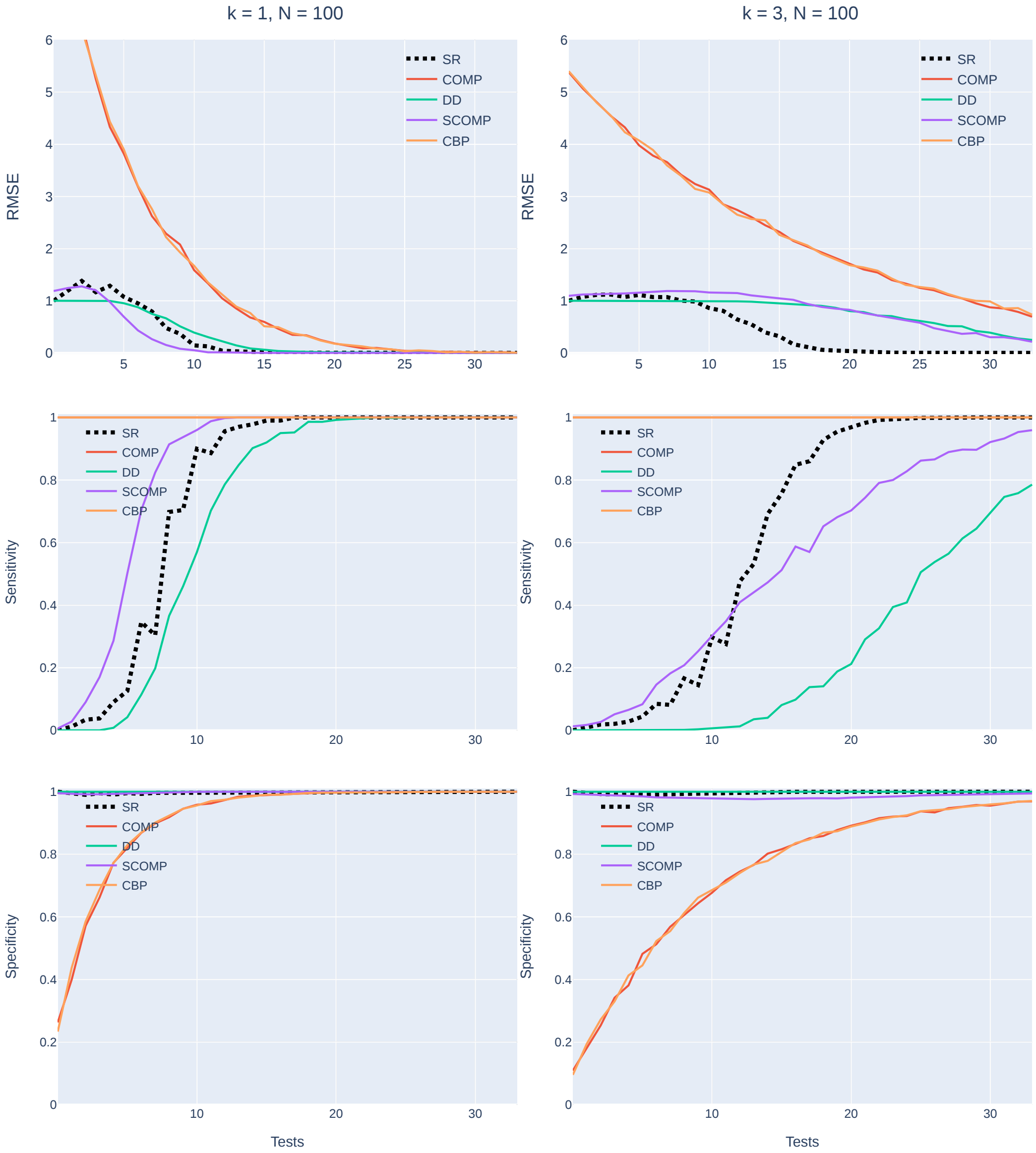


Figure A.2: Additional experiments for $N = 100$ and $k = 1, 3$.

Rows represent RMSE, sensitivity and specificity (top to bottom), columns correspond to $k = 1$ and $k = 3$ (left to right) infected individuals in population of $N = 100$. The proposed approach is second-best for $k = 1$ and the best for $k = 3$.

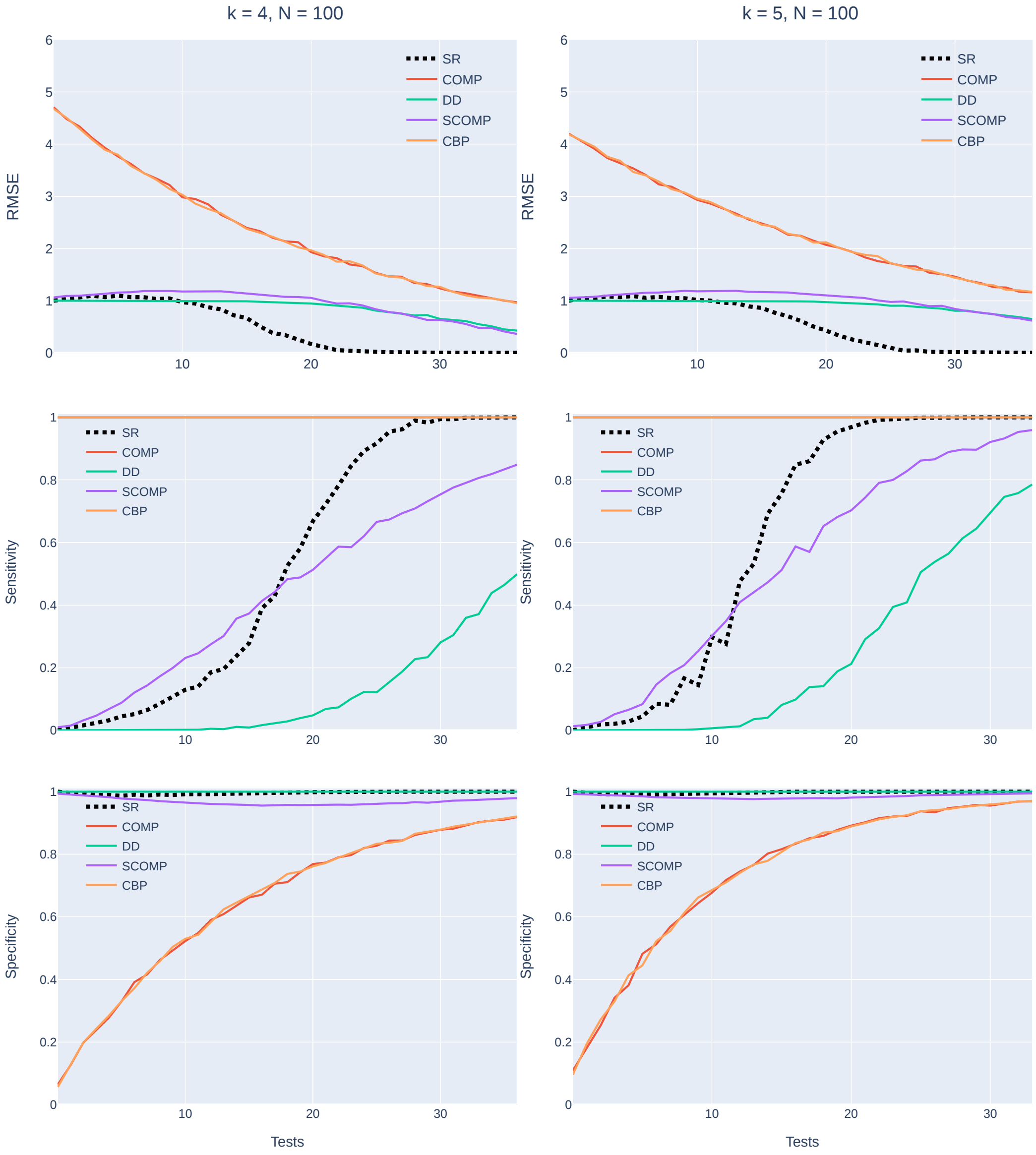


Figure A.3: Additional experiments for $N = 100$ and $k = 4, 5$.

Rows represent RMSE, sensitivity and specificity (top to bottom), columns correspond to $k = 4$ and $k = 5$ (left to right) infected individuals in population of $N = 100$. The proposed approach retains its desirable properties.

## 3.D Comparison with Tapestry ([58])

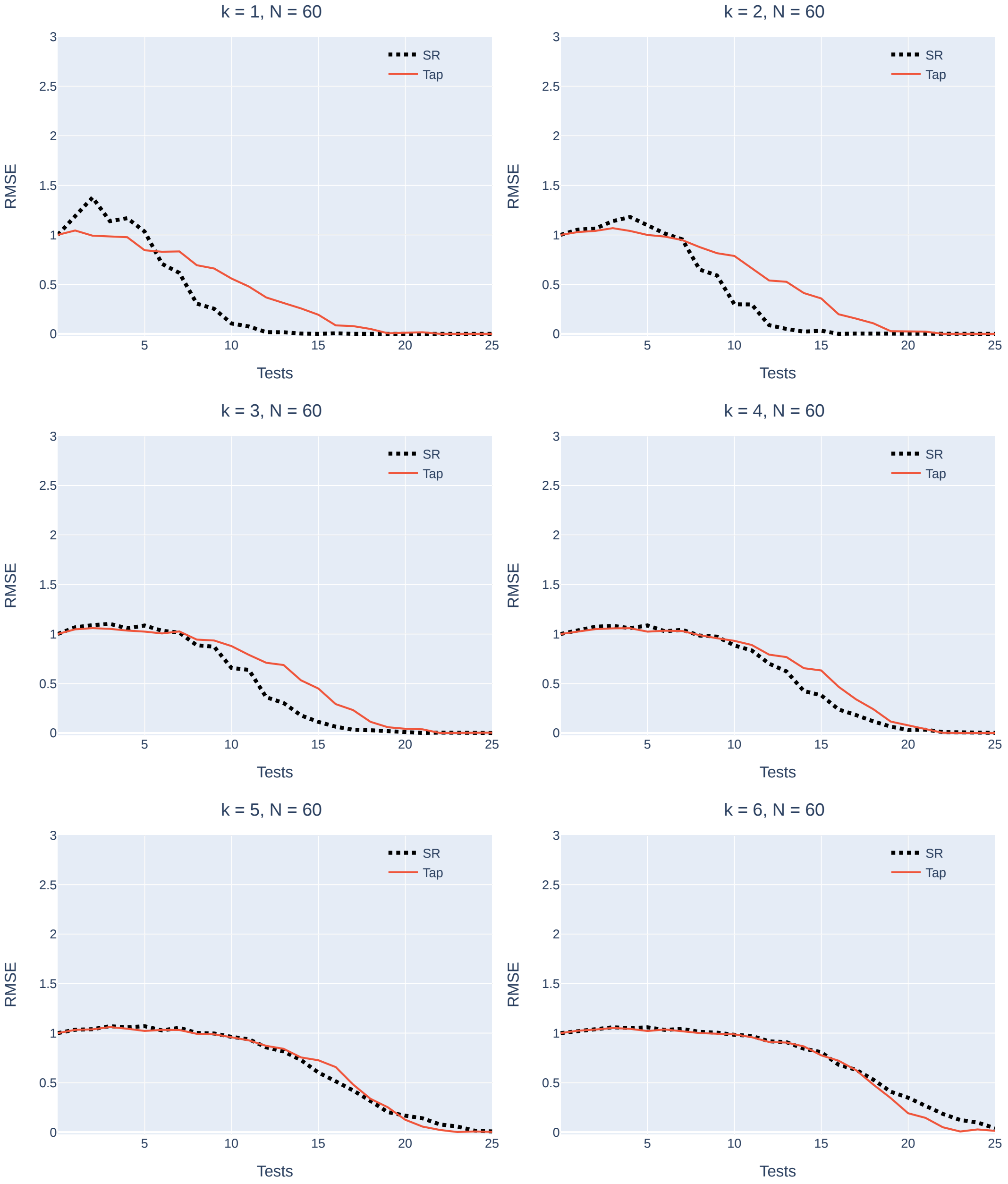


Figure A.4: RMSE comparison for $N = 60, k = 1, 2, 3, 4, 5, 6$.

The proposed SR method dominates Tapestry ([58]) for $k = 1, 2, 3, 4$, but seems to be less precise for $k = 5, 6$.

# Chapter 4

# High-Dimensional Covariance Estimation

## Abstract

[1]Covariance matrix estimates are required in a wide range of applied problems in multivariate data analysis, including portfolio and risk management in finance, factor models and testing in economics, and graphical models and classification in machine learning. In modern applications, where often the model dimensionality is comparable or even larger than the sample size, the classical sample covariance estimator lacks desirable properties, such as consistency, and suffers from eigenvalue

[1] This paper is co-authored with Ekaterina Seregina.

spreading. In recent years, improved estimators have been proposed based on the idea of regularization. Specifically, such estimators, known as rotation-equivariant estimators, shrink the sample eigenvalues, while keeping the eigenvectors of the sample covariance estimator. In high dimensions, however, the sample eigenvectors will generally be strongly inconsistent, rendering eigenvalue shrinkage estimators suboptimal. We consider an estimator that goes beyond mere eigenvalue shrinkage and aims at precise estimation of eigenvectors in sparse settings, without requiring eigenvalues to diverge. The rate of convergence is provided in terms of spectral norm and it achieves the optimal rate under reasonable assumptions. We also provide a numerical simulation demonstrating the superior performance of the proposed estimator as compared to the competition.

## 4.1 Introduction

Covariance matrix estimation is fundamental to multivariate statistical analysis. In statistics and machine learning, some of its applications include graphical modeling, clustering, classification by linear or quadratic discriminant analysis and dimensionality reduction by PCA. In finance, covariance estimates play a central role in portfolio optimization and risk management. In economics, its uses include Kalman filtering, factor analysis, hypothesis testing, GLS and GMM. Covariance estimation is also key in many application in signal processing, bioinformatics and several other fields.

With rapidly increasing availability of data, the analysis of covariance matrices in the low-dimensional (or classical) regime quickly becomes obsolete. This paper considers a large-dimensional framework, where the number of variables $p$ is comparable or even larger than the sample size $n$. In such settings, the sample covariance estimator $S$ loses desirable properties and its classical theoretical foundations break down. For instance, if $p > n$, $S$ is not full rank, so the inverse does not exist. Even when $S$ is invertible, its inverse is highly biased for the theoretical inverse when $p$ and $n$ are comparable. In portfolio optimization this may lead to imprecise and highly volatile weights. Several other major issues such as eigenvalue spreading and eigenvector inconsistency are considered in this manuscript.

On the other hand, while a generic high-dimensional analysis is complicated, sparsity may be a reasonable simplifying assumption to resort to. Furthermore, in applications ranging from genomics to finance, sparsity-inducing approach may be preferred to unrestricted estimation. For example, assets weights in eigenportfolio methodology are proportional to the corresponding eigenvector entries; hence, sparse estimation of an eigenvector leads to more parsimonious allocations with less associated transaction costs. In addition, sparse estimates are easier to interpret.

The literature on covariance estimation proposes some remedies to tackle these challenges. One approach is to shrink a high-variance sample estimator to some structured matrix which may be highly biased and thus produce a better estimator which would achieve optimal bias-variance trade-off. Another approach is to assume a low-dimensional structure in the data as in factor models and consider the implied covariance. Statisticians and mathematicians have also looked into estimation based on the behavior of random matrices, where the analysis is primarily driven by theoretical advancements in random matrix theory.

This paper proposes an estimator that is suitable for the high-dimensional regime, analyzes its theoretical properties and provides a numerical experiment comparing it with the alternative methods. The manuscript is structured as follows. Section 2 describes some of the phenomena and challenges that arise in settings where the number of dimensions is large, reviews some existing approaches and, in this context, motivates the proposed estimator. Section 3 described the model setup, introduces the estimator and point out the similarities with the factor model framework. Section 4 considers a numerical simulation and Section 5 concludes and mentions possible extensions.

**Notation.** For a vector $\mathrm{v} \in \mathbb{R}^d$, we write its $i$-th element as $v_i$. The corresponding $\ell_p$ norm is $\|\mathrm{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$. For a matrix $A \in \mathbb{R}^{m\times d}$, we write its $(i,j)$-th entry as $\{A\}_{ij}$ and denote its $i$-th row (transposed) and $j$-th column as column vectors $A_{i\cdot}$ and $A_{\cdot j}$ respectively. Its singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A)$, where $q = \min(m,d)$. The spectral norm is a matrix operator norm induced by the Euclidean norm, $\|A\|_2 = \max\limits_{\mathrm{v}\neq 0} \frac{\|A\mathrm{v}\|_2}{\|\mathrm{v}\|_2} = \sigma_1(A)$. The max and Frobenius norms are given as $\|A\|_{\max} = \max\limits_{i,j}|a_{ij}|$ and $\|A\|_F = \sqrt{tr(A'A)} = \sqrt{\sum_{i=1}^q \sigma_i^2(A)}$ respectively. Finally, for a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of real nonnegative numbers $\{a_n\}_{n=1}^\infty$, denote $X_n = O_{\mathbb{P}}(a_n)$ if $\forall\epsilon > 0, \exists M, N > 0$ such that $\forall n > N$, $\mathbb{P}(|X_n/a_n| \geq M) < \epsilon$; and denote $X_n = o_{\mathbb{P}}(a_n)$ if $\forall\epsilon > 0$, $\lim\limits_{n\to\infty} \mathbb{P}(|X_n/a_n| \geq \epsilon) = 0$. Finally, let $\mathbb{1}(\cdot)$ be an indicator function and $I_d$ is a $d \times d$ identity matrix.

## 4.2 Background & Related Literature

We observe a data matrix $X \in \mathbb{R}^{n \times p}$, where $n$ and $p$ are the number of observations and the number of variables respectively. Denote the population covariance matrix by $\Sigma \in \mathbb{R}^{p \times p}$ and write its eigendecomposition as

$$\Sigma = ULU' = \sum_{j=1}^{p} \ell_j \mathrm{u}_j \mathrm{u}_j',$$

where $U = [\mathrm{u}_1 \ \cdots \ \mathrm{u}_p]$ is an orthonormal matrix of eigenvectors, and $L = diag(\ell_1, \ldots, \ell_p)$ is a diagonal matrix of eigenvalues with $\ell_1 \geq \ldots \geq \ell_p$. The sample covariance estimator is given as $S := \frac{1}{n} X'X$ (demeaned data) and we write its eigendecomposition as

$$S = V\Lambda V' = \sum_{j=1}^{p} \lambda_j \mathrm{v}_j \mathrm{v}_j',$$

with $V = [\mathrm{v}_1 \ \cdots \ \mathrm{v}_p]$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$.

### 4.2.1 Sample Eigenvalues

It is well-known that $S$ is unbiased and consistent in a classical regime, i.e. when $p$ is fixed and $n$ diverges. Furthermore, it is generally invertible and has an asymptotically normal spectral distribution centered around the true value ([4]), $\sqrt{n}(\lambda_i - \ell_i) \xrightarrow{d} \mathcal{N}(0, 2\ell_i^2), \ j \leq p$.

However, it was observed that many of the desirable properties cease to hold once $p$ also grows, specifically when $\gamma := \lim \frac{p}{n} \in (0, \infty)$. In fact, consistent estimation of the entire spectrum becomes a lot more problematic as both sample eigenvalues and eigenvectors tend to concentrate beyond their true destination.

Specifically, [84] derived the empirical distribution of sample eigenvalues, which became known as Marčenko-Pastur distribution. In its simple formulation when $\Sigma = I_p$, the empirical distribution of sample eigenvalues of a random matrix $F_p(x) := \frac{1}{p}\#\{\lambda_j \leq x\}$ approaches a limiting distribution for which the density is given as

$$f^{MP}(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x\gamma}, \quad \lambda_- = (1-\sqrt{\gamma})^2, \quad \lambda_+ = (1+\sqrt{\gamma})^2,$$

where $\lambda_-, \lambda_+$ are the lower and upper bounds of the support. This result illustrates how sample eigenvalues spread out away from their true values, in this case $\ell_i = 1,\ \forall i$. Moreover, there is a positive bias in the largest sample eigenvalues and a negative bias in the smallest eigenvalues. Furthermore, the magnitude of the bias increases with $p$.

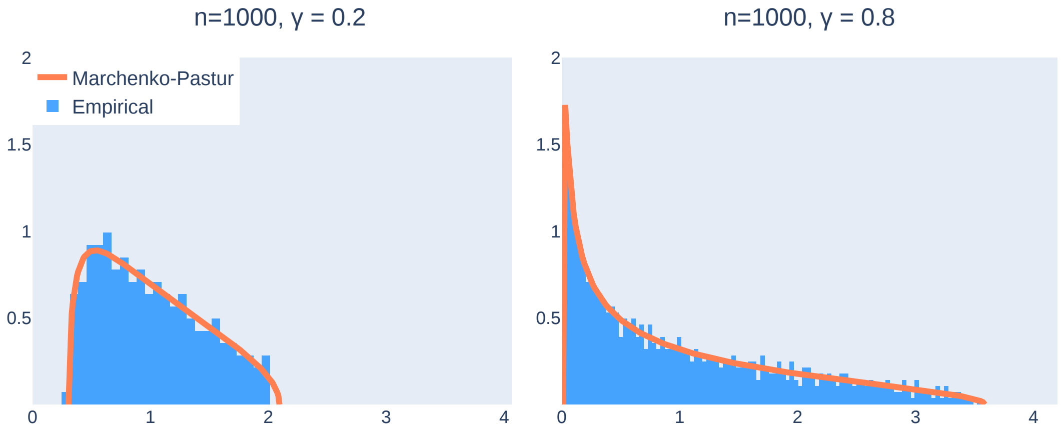


Figure 4.2.1: Marčenko-Pastur and empirical sample eigenvalue distributions

For empirical distribution both panels have $n = 1000$, while $\gamma := p/n$ is different. *Left*: $\gamma = .2$; *Right*: $\gamma = .8$.

Figure 4.2.1 plots the theoretical density on top of the empirical sample eigendistribution for $n$ datapoints sampled from $\mathcal{N}_p(0, I_p)$ for different values of $\gamma := p/n$. It demonstrates two phenomena. First, the sample eigenvalues are spread out asymmetrically around the true value of 1. Second, the smallest and largest eigenvalues concentrate around $\lambda_-$ and $\lambda_+$ respectively. In fact, Bai-Yin's law ([14]) states that for matrices with bounded fourth moments the extreme sample eigenvalues land almost surely on these edges, i.e. $\lambda_p \overset{a.s.}{\to} \lambda_-$ and $\lambda_1 \overset{a.s.}{\to} \lambda_+$. Excellent treatment is given in [12].

This paper focuses on the case when a few population eigenvalues are larger than the bulk, in other words top eigenvalues are spiked ([67]),

$$\Sigma = diag(\ell_1, \ldots, \ell_r, 1, \ldots, 1), \; \ell_r > 1.$$

The convergence of the sample eigenvalues in this case depends on the magnitude of the true spikes in comparison to the so-called BBP transition point $\lambda_+^{1/2}$, named after its discoverers [15]. Specifically, we have for $j \leq r$,

$$\lambda_j \overset{a.s.}{\to} \begin{cases} \lambda_+, & \ell_j < \lambda_+^{1/2}, \\ \ell_j + \gamma \frac{\ell_j}{\ell_j - 1}, & \ell_j > \lambda_+^{1/2}, \end{cases}$$

as $n, p \to \infty$. That is, there is an upward bias in leading sample eigenvalues and the amount of bias is asymptotically known. [16] establish the almost sure limits of the eigenvalues of large sample covariance matrices in a spiked population model framework. Moreover, the exact asymptotic distribution of the largest and smallest eigenvalues is also known ([118]).

Hence, if the true spikes are not large enough, the sample eigendistribution will follow the Marčenko-Pastur distribution. In the opposite case, the spiked sample eigenvalues will overshoot the true counterparts and lie above the Marčenko-Pastur sea. In general this knowledge can be used as a heuristic for inferring the number of principal components or factors.

In view of the above phenomena, statisticians have proposed to construct rotationally invariant estimators (RIE), which would correct the sample eigenvalues while assuming the sample and true eigenvectors coincide. This includes many popular estimators, e.g. linear shrinkage of the sample covariance with a structured matrix (typically, an identity) proposed in [81] or nonlinear extensions as in [82], which in essence seek to pull upward and downward biased sample estimates towards the center. Another approach is to set all eigenvalues inside the Marčenko-Pastur sea to some constant, as these are deemed as noise, while keeping the spikes unaltered; this strategy is known as eigenvalue clipping ([25]). [44] do not address eigenvector inconsistency but partially address fix this issue by proposing an RIE that accounts for the non-vanishing angle between population and sample eigenvectors. They find a univariate function $\nu$ that when applied to eigenvalues would optimally (for a given loss) shrink it such that eigenvector estimation inaccuracy is taken into consideration.

### 4.2.2 Sample Eigenvectors

However, the assumption that sample and population eigenvectors coincide in high dimensions is unrealistic as sample eigenvectors are generally also inconsistent in high dimensions. This motivates us to develop an estimator that primarily aims at accurate eigenvector estimation. We seek to carefully characterize the assumptions this will require and the trade-offs between different sets of assumptions. Before that, we briefly review some of the related work without aiming to be exhaustive.

As described in the previous section, there has been a lot of work done examining the features of sample eigenvalues. [13] points out that this may partly be explained by the quantum mechanics origins of the random matrix theory, where the sample eigenvalues are associated with energy levels of particles. Many applications, however, require precise estimates of eigenvectors and the research in this direction is gradually being recognized.

[68] propose an (adaptive) sparse PCA for settings when $p, n \to \infty$ in a single factor model framework. They show that plain PCA leads to consistent estimates if and only if $p/n \to 0$, however it is possible to recover consistency even when $p \gg n$ if some preselection of variables, possibly in an alternative sparse basis, is made in advance. Their Theorems 1,2 and 3 characterize the inconsistency of PCA when $\lim p/n = \gamma > 0$ in terms of an angle between the true leading eigenvector and its estimate. Theorem 5 suggest a solution for cases when the true principal eigenvector satisfies $\ell_q$-ball sparsity assumption.

In a closely related work, [69] provide a similar characterization of inconsistency in Theorem 1 but in terms of the normalized inner product between the true and estimated principal eigenvectors, while Theorem 2 proves that consistency is recovered as long as the PCA is performed after the proposed variable selection algorithm.

[94] characterize the asymptotic behavior of sample eigenvectors and its dependence on the eigenvalue phase transition. Specifically, under mild conditions on a spiked covariance model, as $p/n \to \gamma \in (0,1)$ we have

$$\langle \mathrm{v}_j, \mathrm{u}_j \rangle^2 \overset{a.s.}{\to} \begin{cases} 0, & \ell_j < \lambda_+^{1/2}, \\ \frac{1-\gamma/(\ell_j-1)^2}{1+\gamma/(\ell_j-1)}, & \ell_j > \lambda_+^{1/2}. \end{cases}$$

That is, the sample eigenvector $\mathrm{v}_j$ is asymptotically orthogonal to the true vector $\mathrm{u}_j$ when the corresponding eigenvalue is small. On the other hand, consistent estimation of eigenvectors with sufficiently strong signals requires $\gamma \to 0$, , i.e. $n$ grows faster than $p$. The conventional PCA becomes confused in the presence of large number of variables. Sparsity, either in the original or some transformed domain, becomes crucial for consistent estimation of principal component directions.

[104] further investigate consistency and asymptotic behavior of sample eigenvalues and eigenvectors in a more general multiple-component spike covariance framework with $r$ spikes, $\ell_1 > \ldots > \ell_r \gg \ell_{r+1} \to \ldots \to \ell_p \to 1$. They also consider a more general asymptotic framework with $\frac{p}{n\ell_j} \to c_j > 0,\ j \leq r$, where $0 < c_1 < \ldots < c_r < \infty$; allowing $\ell_j$ to potentially diverge turns out to be crucial. Their Theorem 3 states that

$$\begin{cases} \frac{\lambda_j}{\ell_j} \overset{a.s.}{\to} 1 + c_j, \quad 1 \leq j \leq r, \\ \\ \frac{n\lambda_j}{p\ell_j} \overset{a.s.}{\to} 1, \quad r+1 \leq j \leq n \wedge p, \end{cases} \tag{4.1}$$

and

$$\begin{cases} |\langle v_j, u_j \rangle| \overset{a.s.}{\to} (1 + c_j)^{-1/2}, \quad 1 \leq j \leq r, \\ \\ |\langle v_j, u_j \rangle| \overset{a.s.}{\to} O_{a.s.}\big((n/p)^{1/2}\big), \quad r+1 \leq j \leq n \wedge p, \\ \\ \angle\left(v_j, \text{span}(u_k : k = r+1, \ldots, p)\right) \overset{a.s.}{\to} 0, \quad r+1 \leq j \leq n \wedge p. \end{cases} \tag{4.2}$$

Equation (4.1) formalizes the idea that sample eigenvalues with stronger signals will be less biased, but still almost surely biased when $c_j \neq 0$. Notice that there is no bias if the eigenvalues grow linearly with the dimension. Equation (4.2) reveals that leading sample eigenvectors lie in a cone along the true eigendirections as long as the ratio of the dimension to the product of the sample size and the spike size, $\frac{p}{n\lambda_j} \to c_j > 0,\ j \leq r$. This shows that sample eigenvectors might still be consistent in the high-dimensional regime when $\gamma > 0$ as long as their corresponding eigenvalues are large. Intuitively, a strong signal helps identify the direction of most variation.

Observe that $\frac{p}{n\ell_j} \to c_j > 0$ can contain three cases: (i) $p, n, \ell_j \to \infty$, (ii) $p, \ell_j \to \infty$, while $n < \infty$, (iii) $p, n \to \infty$ and $\ell_j < \infty$. The result stated in equation (4.2) refers to the first case, [104] also covers the second case. We focus on the third case, where leading eigenvalues are bounded.

A natural extension to the work of [94] and [104] is the study by [53] who analyze the asymptotic distributions of the sample eigenstructure and derive the precise rates of convergence in a similar setup with a high-dimensional spiked covariance and $p, n, \ell_j \to \infty$

with $\frac{p}{n\ell_j} < \infty, \ j \leq r$. Although this comes at the cost of a sub-Gaussianity assumption on the data. In particular, they establish that the normalized spiked part of the sample eigenvector converges to a vector of ones. [51] also consider a diverging eigenvalue setup.

Instead of assuming increasing eigenvalues, a simple alternative approach which permits efficient estimation in high dimensions is sparsity. [23] propose regularizing a large covariance matrix by hard thresholding, which leads to a consistent (in the operator norm) estimator as long as the true covariance matrix is sparse. However, imposing sparsity on a high-dimensional covariance directly may be unjustified in certain applications, e.g. in portfolio theory covariance of asset returns is not sparse. [51] consider conditional sparsity, i.e. sparsity after estimating and accounting for the factor structure. To distinguish the signal and noise components, they assume diverging (with $p$) signal eigenvalues. This assumption may be "misleading in many economic and financial applications", as pointed out some researchers ([51], discussion on the paper).

On the other hand, the eigenvectors themselves are often sparse in many high-dimensional applications or may be required to be sparse in certain scenarios, e.g. in capital allocation problems. This also leads to better interpretability since eigenvectors are only linear combinations of a subset of variables. Most importantly, sparsity can help even in cases when $\gamma > 0$ and the signal is bounded, $\ell_j = O(1)$. A similar idea is considered in [3], however they focus on sparse eigenvector support recovery.

Besides, precise estimation of eigenvectors in high dimensions has its own benefit. For example, the top eigenvectors of a covariance identify the directions of the most variation, while the bottom eigenvectors of a graph's Laplacian provide insights into its cluster structure.

## 4.3 Model

Given a $n \times p$ data matrix $X$ of $p$ i.i.d. mean-zero variables with the population covariance

$$\Sigma = \mathbb{E}(X'X) = \sum_{i=1}^{r} \ell_i \mathrm{u}_i \mathrm{u}_i' + \sum_{i=r+1}^{p} \ell_i \mathrm{u}_i \mathrm{u}_i', \tag{4.3}$$

and the sample covariance estimator

$$S = \frac{1}{n} X'X = \sum_{i=1}^{r} \lambda_i \mathrm{v}_i \mathrm{v}_i' + \sum_{i=r+1}^{p} \lambda_i \mathrm{v}_i \mathrm{v}_i',$$

where $r$ is the number of signal eigenvalues (assumed to be known and fixed), our primary goal is accurate estimation of $\Sigma$ in a high-dimensional setting under a spiked covariance framework ([67]). Throughout this paper we measure the estimation error in terms of spectral (operator) norm $\|\cdot\|$. By Weyl's and Davis-Kahan Theorems (see Appendix 4.D) the $\|\widehat{\Sigma} - \Sigma\| \to 0$ implies the convergence of the corresponding eigenvalues and eigenvectors as well as the convergence of PCA loadings.

**Assumption 1 (Spiked covariance)** *There are $r \ll p \wedge n$ spikes in eigenvalues $\ell_1 > \ldots > \ell_r > 1$, independent of $p$ and $n$, with $\Delta := \ell_r - \ell_{r+1} \gg 0$. All spiked eigenvalues are distinct.*

**Remark.** In particular, while we need not have diverging signals, the eigengap $\ell_r - \ell_{r+1}$ should be large enough for identification purposes. This also inherently relates to the eigenvector instability demonstrated in the following example.

**Example.** (**[121]**) Consider a perturbation of a diagonal $A$ by another diagonal matrix $\epsilon P$,

$$A_\epsilon = A + \epsilon P = \begin{pmatrix} 1 & 0 \\ 0 & 1.01 \end{pmatrix} + \epsilon \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Clearly, the eigenvalues of unperturbed $A$ are $\{1,\ 1.01\}$ and the eigenvalues of the perturbed $A_n$ are

$$\left\{ \frac{1}{2}(2.01 + \sqrt{.0001 + 4\epsilon^2}),\ \frac{1}{2}(2.01 - \sqrt{.0001 + 4\epsilon^2}) \right\},$$

satisfying Weyl's theorem

$$\max_{i=1,2} |\ell_i(A) - \ell_i(A_\epsilon)| = \frac{1}{2}|.01 - \sqrt{.0001 + 4\epsilon^2}| \leq \|\epsilon P\|_2 = \epsilon,$$

and thus displaying resilience to small perturbations. On the other hand, the maximal eigenvector of $A$ changes its direction substantially from $\mathrm{u}_1(A) = (0\ ,1)'$ to $\mathrm{u}_1(A_\epsilon) \approx (.53\ ,.85)'$, so that $\|\mathrm{u}_1(A) - \mathrm{u}_1(A_\epsilon)\|_2 \gg \epsilon$. The problem arises due to small eigengap, and hence a large enough eigengap is needed to ensure the stability.

A simple example of a spiked model that was widely considered in the literature is of the form,

$$\Sigma = \ell_1 \mathrm{u}_1 \mathrm{u}_1' + I_p,$$

with $\ell_1 > 0, \|\mathrm{u}_1\|_2 = 1$, where $\mathrm{u}_1$ is a unique maximal eigenvector with eigenvalue $1+\ell_1$ and all other eigenvalues are 1. That is, we have a low-rank perturbation of a sparse matrix. [22] examine the possibility of detection of the low-rank component and propose a minimax optimal test based on an eigenvalue statistic.

Spiked models are also inherently related to factor models considered in financial econometrics and we examine the similarity in Section 4.3.2. The differences mainly arise due to different assumptions placed on the behavior of the eigenstructure.

**Assumption 2 (High-dimensional asymptotics)** *$n, p \to \infty$ and $\ell_j = O(1),\ j = 1, \ldots, p$.*

**Remark.** In particular, we need not have strong (pervasive) signals, so it is not necessary that $\ell_i = O(p),\ i \leq r$. In fact, the pervasiveness assumption ([51]) can make consistent estimation impossible in terms of spectral norm, as discussed in the following example.

**Example.** Suppose we know the entire spectrum of $\Sigma$ except for the first eigenvector, for which suppose we have a good estimator with $\|\mathrm{v}_1 - \mathrm{u}_1\| = O_p(n^{-1/2})$. Then we can construct a sample covariance $S^*$ with this population information in its spectrum, then

$$\|S^* - \Sigma\| = \|\ell_1(\mathrm{v}_1\mathrm{v}_1' - \mathrm{u}_1\mathrm{u}_1')\| = \ell_1 O_p(\|\mathrm{v}_1 - \mathrm{u}_1\|) = O_p(\ell_1 n^{-1/2}),$$

which does not converge if $\ell_1$ grows linearly in $p$ and $n = O(p^2)$, so under certain conditions $\Sigma$ may not be estimated consistently in terms of spectral norm in the presence of diverging spiked eigenvalues.

In accordance with Equation (4.3) we can decompose the true covariance into two parts,

$$\Sigma = \Sigma_s + \Sigma_e, \tag{4.4}$$

where the two matrices on the right-hand side represent signal and noise (error) components. In particular, one can view the above equation as low-rank plus sparse matrix structure. This idea is formalized in the following assumption.

**Assumption 3 (Low-rank plus sparse)** *In equation* (4.4)*, rank*$(\Sigma_s) = r$ *and* $\Sigma_e$ *is (approximately) sparse with bounded eigenvalues. Moreover,* $\Sigma_s$ *has a fixed number* $r$ *sparse unit norm eigenvectors,* $\|\mathrm{u}_j\|_0 = s,\ j \leq r, \|\mathrm{u}_j\|_2 = 1,\ \forall j$.

**Remark.** The low-rank plus sparse structure has been thoroughly studied, see e.g. [125] or [31] on the possibility of identification of the two matrices, low-rank and sparse, only from the sum alone. This structure is also implied by approximate factor models ([33]) where $\Sigma = \Lambda\Lambda' + \Omega$.

**Remark.** In general, the results throughout this paper can be adapted to cases with approximate sparsity. For example, if $\widetilde{\mathrm{u}}_1$ is an approximation of exactly $s$-sparse vector $\mathrm{u}_1$ of $\Sigma_s$, we can instead analyze a slightly different perturbation, $S = \Sigma + E = \widetilde{\Sigma} + (E - \widetilde{\Sigma} + \Sigma) = \widetilde{\Sigma} + \widetilde{E}$, where $\widetilde{E} := E - \widetilde{\Sigma} + \Sigma$.

**Remark.** Notice that this formulation with sparse eigenvectors does not necessarily imply that $\Sigma$ is sparse. A similar framework with sparse eigenstructure was analyzed by [3] with the focus on support recovery.

Notice that $\Sigma_e$ is approximately sparse; we characterize sparsity as in [23],

$$m := \max_{i \leq p} \sum_{j \leq p} |\{\Sigma_e\}_{i,j}|^q,$$

so that $m$ stays bounded for some $0 \leq q < 1$, although for simplicity we will focus on the case with $q = 0$ corresponding to exact sparsity, $m = \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\{\Sigma_e\}_{i,j} \neq 0)$. The fact that eigenvalues are thus bounded can be seen from

$$\|\Sigma_e\| \leq \|\Sigma_e\|_1 \leq \max_{i \leq p} \sum_{j \leq p} |\{\Sigma_e\}_{i,j}|^q (\{\Sigma_e\}_{i,i}\{\Sigma_e\}_{j,j})^{(1-q)/2} = O(m),$$

when $\{\Sigma_e\}_{i,i}$ are bounded. This assumption is not very restrictive and corresponds to weak correlation between idiosyncratic components in factor models.

### 4.3.1 Estimator

In what follows we consider a simpler version of Equation (4.4) with $r = 1$, namely

$$\Sigma = \ell_1 \mathrm{u}_1 \mathrm{u}_1' + \Sigma_e, \tag{4.5}$$

where the single signal eigenvector is $s$-sparse, i.e. $\|\mathrm{u}_1\|_0 = s$, and $\Sigma_e$ is approximately sparse in a sense that $\|\Sigma_e\|$ has bounded eigenvalues as $p \to \infty$. In other words, the covariance is a rank-1 perturbation of a sparse matrix.

As a first step, we seek to estimate the sparse eigenvectors of $\Sigma_s$. To recover the first eigenvector, one could proceed by explicitly imposing the sparsity restriction in a rank-one approximation problem by solving

$$\begin{gathered} \min_{\nu,\xi} \|X'X - \nu\xi\xi'\|_F^2 \\ \nu \geq 0,\ \|\xi\|_0 \leq s,\ \|\xi\|_2 = 1, \end{gathered} \tag{4.6}$$

which is equivalent to solving

$$\begin{gathered} \min_{\xi} \xi' X' X \xi \\ \|\xi\|_0 \leq s,\ \|\xi\|_2 = 1. \end{gathered} \tag{4.7}$$

Clearly, both are NP-hard problems due to the presence of $\|\cdot\|_0$-norm. Efficient convex relaxations with $\|\cdot\|_1$-norm substitution have been studied ([40]) as well as alternative formulations imposing sparse structure (e.g. [71],[130], [123]). The theoretical underpinnings for the above algorithms and formulations are less developed.

Luckily, it is possible to directly approximate the solution via the principle of power iteration. [129] propose a truncated power method which tackles the optimization problem in Equation (4.7) and analyze its theoretical properties. This approach achieves an optimal bound ([28]) and is guaranteed to converge under mild technical conditions. The goal is to recover a sparse eigenvector given a perturbation of an original matrix.

The truncated power iteration procedure is described in Algorithm 2. The only difference with a conventional power method is in the truncation step, which forces the $p - r$ smallest entries of a vector to zero in each iteration thus naturally inducing sparse estimates.

The following perturbation formulation is useful,

$$S = \Sigma + E, \tag{4.8}$$

where $S$ is a sample covariance matrix, and $E := S - \Sigma$ is an error. The theorem of [129] is adapted in 12 and assures the recovery of $s$-sparse eigenvectors so long as the spectral norm of $\hat{s} \times \hat{s}$ principal submatrix of $E$, denoted as $\underline{\mathrm{E}}_{\hat{s}}$, is sufficiently small for some initial estimate of sparsity $\hat{s}$. Notice that this norm can be a substantially smaller than the norm of the full matrix $E$.

**Theorem 12 (Sparse recovery)** *Given Assumptions 1, 3 and the initial vector* $\widehat{\mathrm{v}}_1^{(0)}$ *with* $\|\widehat{\mathrm{v}}_1^{(0)}\|_0 \leq \widehat{s}$, $\|\widehat{\mathrm{v}}_1^{(0)}\|_2 = 1$, $\widehat{s} \geq s$, $|\widehat{\mathrm{v}}_1^{(0)\prime}\mathrm{u}_1| - \delta \geq \theta$, *where* $0 < \theta < 1$ *and*

$$\delta := \frac{\sqrt{2}\|\underline{E}_{\hat{s}}\|}{\sqrt{\|\underline{E}_{\hat{s}}\|^2 + (\Delta - 2\|\underline{E}_{\hat{s}}\|)^2}},$$

*we have*

$$\sqrt{1 - |\widehat{\mathrm{v}}_1^{(t)\prime}\mathrm{u}_1|} \leq c_1^t\sqrt{1 - |\widehat{\mathrm{v}}_1^{(0)\prime}\mathrm{u}_1|} + \sqrt{10}\delta(1 - c_1)^{-1}, \quad \forall t \geq 0, \tag{4.9}$$

*where* $c_1$ *can be chosen to be less than* 1.

**Proof.** See [129] Theorem 4. ■

Since the convergence of the algorithm is guaranteed, let us denote $\lim_{t \to \infty} \widehat{\mathrm{v}}_1^{(t)} = \widehat{\mathrm{v}}_1$.

**Corollary 13** *Given the assumptions of Theorem 12, we have*

$$\|\widehat{\mathrm{v}}_1 - \mathrm{u}_1\| = O_{\mathbb{P}}(\|\underline{E}_{\hat{s}}\|).$$

**Corollary 14** *Given the assumptions of Theorem 12 and assuming entries in $E$ are Gaussian iid and $\hat{s} = O(s)$, we have*

$$\|\widehat{\mathrm{v}}_1 - \mathrm{u}_1\| = O_{\mathbb{P}}\left(\sqrt{\frac{s \log p}{n}}\right).$$

Theorem 12 bounds the angle between the $t$-th iteration of sparse eigenvector estimate $\widehat{\mathrm{v}}_1^{(t)}$ and its population counterpart $\mathrm{u}_1$. Corollary 14 is an immediate consequence and states that the error depends on the norm of $s$-dimensional principal submatrix which could be much smaller than the norm of the entire perturbation implied by standard perturbation inequalities. Corollary 14 follows when entries of the perturbation are normally distributed; this is a standard result in random matrix theory. The arguments of the proof are based on eigenvector perturbation inequalities provided in the appendix. Initialization is also discussed.

Hence, this iterative approach can recover the leading sparse eigenvector even from noisy observations. The remaining sparse eigenvector estimates could be obtained greedily, i.e. for the second iteration one would optimize over an unexplained component

$X'X - (\hat{\xi}'X'X\hat{\xi})\hat{\xi}\hat{\xi}'$, where $\hat{\xi}$ solves equation (4.7). This procedure is known as iterative deflation.

Suppose that we have obtained eigenvector estimates $\{\widehat{\mathrm{v}}\}_{i=1}^{r}$. The corresponding weight estimates for $r$ top matrices can be estimated consistently by least squares under the standard assumptions with usual parametric rates, i.e. $|\widehat{\lambda}_j - \ell_j| = O(n^{-1/2})$. This completes the estimation of the signal part of the covariance matrix.

Once the signal component is estimated we turn to the error part $\Sigma_e$. Consistent with the assumption of conditionally sparse covariance, after removing the estimated low-rank part we threshold the remainder. In a general case one can obtain an estimate of the remainder as

$$\widehat{S}_e = S - \sum_{j=1}^{r} \widehat{\lambda}_j \widehat{\mathrm{v}}_j \widehat{\mathrm{v}}_j{}', \tag{4.10}$$

where we subtract the estimated low-rank component from the sample covariance. Next we apply entry-wise adaptive hard thresholding similar to [29] to obtain $\widehat{S}_e^{\tau}$, where each entry is set as

$$\{\widehat{S}_e^{\tau}\}_{i,j} = \begin{cases} \{\widehat{S}_e\}_{i,i}, \quad i = j, \\ \{\widehat{S}_e\}_{i,j}\mathbb{1}(|\{\widehat{S}_e\}_{i,j}| \geq \tau\sqrt{\{\widehat{S}_e\}_{i,i}\{\widehat{S}_e\}_{i,j}}), \quad i \neq j, \end{cases} \tag{4.11}$$

for a given $\tau > 0$, which amounts to thresholding the corresponding correlation matrix. This approach yields an optimal rate of convergence ([28]) for $\Sigma_e$.

**Theorem 15 (Error component)** *Under assumptions of Theorem 12, Assumption 2 and given* $\|S - \Sigma\|_{max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right)$, *for a large enough* $\tau > 0$ *we have*

$$\|\widehat{S}_e^{\tau} - \Sigma_e\| = O_p\left(m\sqrt{\frac{s \log p}{n}}\right).$$

**Proof.** See Appendix 4.B. ■

Optimal estimation of a sparse component is discussed in detail in [28], [51]. The assumptions are not unusual and are easy to verify.

Thus the final estimator is given as

$$\widehat{S} = \sum_{j=1}^{r} \widehat{\ell}_j \widehat{\mathrm{v}}_j \widehat{\mathrm{v}}_j{}' + \widehat{S}_e^{\tau}, \tag{4.12}$$

The estimation involves two stages, one for estimating each of the components. The full algorithm is provided in Algorithm 3.

**Theorem 16** *Suppose the assumptions of Theorem 15 hold. Then*

$$\|\widehat{S} - \Sigma\| = O_p\left(m\sqrt{\frac{s \log p}{n}}\right).$$

**Proof.** See Appendix 4.C. ■

Notice that this estimator achieves the optimal bound [28] for high-dimensional covariance estimators in sparse settings. The proof is provided for rank-1 perturbations as in Equation (4.5), however it can be generalized to multi-spike covariances.

### 4.3.2 Factor Model Framework

This subsection demonstrates that the covariance structure considered in the previous sections is implied by weak (non-pervasive) factors with approximately sparse loading matrices.

A latent factor model is given as

$$\underset{p\times 1}{X_i} = \Lambda \underset{r\times 1}{F_i} + e_i, \tag{4.13}$$

$\Lambda$ is an approximately sparse loading matrix for the $r$ factors in $F_i$, $e_i$ is an idiosyncratic disturbance; and $i = 1, \ldots, n$. Only $X_i$ are observable. The latter equation can be rewritten in matrix form

$$X = F\Lambda' + e, \tag{4.14}$$

where $X = [X_1 \ \cdots \ X_n]'$ and $F = [F_1 \ \cdots \ F_n]'$.

For the above factor model specification, we have the corresponding population covariance matrix of $X_i$,

$$\Sigma := \mathbb{E}(X'X) = \Lambda\Lambda' + \Omega, \tag{4.15}$$

where $\Omega = \mathbb{E}(e'e)$ is assumed to be approximately sparse. Hence Equation (4.15) admits low-rank plus sparse representation. The two components can be asymptotically identified only when the eigengap $\Delta$ is sufficiently large while $\Omega$ has bounded eigenvalues as a consequence of sparsity. This is crucial for consistent estimation since if nonzero eigenvalues of $\Lambda\Lambda'$ are smaller than $\|\Omega\|$, then it is impossible to distinguish signal from noise. In practice, factors are expected to exhibit sufficiently strong signal while the remaining part normally has weak correlation. The leading $r$ eigenvectors of $\Sigma$ should be nearly aligned with the corresponding columns of $\Lambda$. One can also view $\Sigma$ in Equation (4.15) as a perturbation of $\Lambda\Lambda'$ by $\Omega$. Further, denote the eigendecomposition of $\Sigma$ as in Equation (4.3).

In vanilla factor modeling, one can obtain factors and loading estimates via PCA by solving

$$\begin{gathered}\arg\min_{F,\Lambda} \left\|X - F\Lambda'\right\|_F^2 \\ p^{-1}\Lambda'\Lambda = I_r, \quad F'F \text{ diagonal.}\end{gathered}$$

We can formulate a similar optimization problem in a penalized PCA fashion, similar to Equation (4.7), that would correspond to a sparse setting considered in this paper. Specifically, to obtain an approximate solution for the first loading column $\Lambda_1$ one can solve

$$\begin{gathered}\arg\min_{\Lambda_1} \Lambda_1' X' X \Lambda_1 \\ \|\Lambda_1\|_0 \leq s, \ \|\Lambda_1\|_2 = 1.\end{gathered} \tag{4.16}$$

Clearly, the truncated power iteration method described earlier can be used. Its solution $\widehat{\Lambda}$ will be the first $r$ sparse eigenvectors of $X'X$. The corresponding factor estimate can simply be calculated as $\widehat{F}_i = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'X_i = \widehat{\Lambda}'X_i$. Hence, the technique and the theory considered above are applicable when a sparsity assumption on the loadings is justifiable. This may also be beneficial for constructing interpretable factor models; a similar setup from the Bayesian viewpoint is considered in [93].

## 4.4 Numerical Experiment

Generate the covariance as follows

$$\Sigma = ULU' = U_r L_r U_r' + \Sigma_e,$$

where $U_r$ is a $p \times r$ matrix of eigenvectors corresponding to top eigenvalues and $L_r$ is an $r \times r$ diagonal matrix of eigenvalues in descending order. Specifically, we set $r = 2$ across all simulations are generate the two columns in $U_r = (\mathrm{u}_1 \quad \mathrm{u}_2)$ as follows,

$$\{\mathrm{u}_1\}_i = \begin{cases} \frac{1}{\sqrt{s}}, & i \in [1,\ s] \\ 0, & \text{otherwise} \end{cases}, \quad \{\mathrm{u}_2\}_i = \begin{cases} \frac{1}{\sqrt{s}}, & i \in [s+1,\ 2s] \\ 0, & \text{otherwise} \end{cases},$$

We set $\ell_1, \ell_2$ to either $(200, 100)$ or $(500, 300)$. The entries of the error component $\Sigma_e$ are generated in a block-diagonal fashion with

$$\{\Sigma_e\}_{i,j} = \rho^{|i-j|} \mathbb{1}(|i-j| \leq 1),$$

and set $\rho = .5$. This design ensures sparsity and is sometimes referred to as MA(1).

The data $X \in \mathbb{R}^{n \times p}$ is generated by drawing $n$ samples from $X \sim \mathcal{N}_p(0, \Sigma)$. We run 100 Monte Carlo simulations. The proposed method is compared against POET ([51]), a hard-thresholding estimator of [23] and a linear shrinkage estimator of [81].

We vary $p \in \{100, 300, 500\}$ and the sparsity $s \in \{5, 15, 25\}$. The results report the ratios of a spectral (or Frobenius) norm of the covariance estimation error of a given method with respect to POET's corresponding norm. Tables 4.4.1 and 4.4.2 consider cases with $\ell_1 = 200, \ell_2 = 100$ and $\ell_1 = 500, \ell_2 = 300$ respectively. The proposed method is dubbed as "DSCE" for doubly sparse covariance estimator.

Table 4.4.1: Ratios to POET error $n = 300, \ell_1 = 200, \ell_2 = 100$.

| $p$ | Method | Spectral | | | Frobenius | | |
|---|---|---|---|---|---|---|---|
| | | $s = 5$ | $s = 15$ | $s = 25$ | $s = 5$ | $s = 15$ | $s = 25$ |
| 100 | DSCE | 0.7462 | 0.7308 | 0.7800 | 0.7587 | 0.7776 | 0.7717 |
| 100 | B&L | 0.8701 | 0.9139 | 0.9287 | 0.9072 | 0.9246 | 0.9501 |
| 100 | L&W | 1.0297 | 1.0122 | 0.9598 | 1.0317 | 1.0066 | 0.9804 |
| 300 | DSCE | 0.7076 | 0.7382 | 0.7551 | 0.7380 | 0.7733 | 0.7706 |
| 300 | B&L | 0.8844 | 0.9253 | 0.9331 | 0.9073 | 0.9336 | 0.9481 |
| 300 | L&W | 1.0318 | 0.9989 | 0.9887 | 1.0518 | 1.0377 | 0.9894 |
| 500 | DSCE | 0.6943 | 0.7100 | 0.7362 | 0.6964 | 0.7390 | 0.7477 |
| 500 | B&L | 0.9127 | 0.9286 | 0.9463 | 0.9286 | 0.9483 | 0.9623 |
| 500 | L&W | 1.0937 | 1.0626 | 1.0215 | 1.1254 | 1.0750 | 1.0532 |

Table 4.4.2: Ratios to POET error $n = 300, \ell_1 = 500, \ell_2 = 300$.

| $p$ | Method | Spectral | | | Frobenius | | |
|---|---|---|---|---|---|---|---|
| | | $s = 5$ | $s = 15$ | $s = 25$ | $s = 5$ | $s = 15$ | $s = 25$ |
| 100 | DSCE | 0.7108 | 0.7135 | 0.7776 | 0.7292 | 0.7415 | 0.7677 |
| 100 | B&L | 0.8999 | 0.9257 | 0.9280 | 0.9183 | 0.9569 | 0.9454 |
| 100 | L&W | 1.0446 | 0.9944 | 0.9720 | 1.0603 | 1.0266 | 0.9937 |
| 300 | DSCE | 0.6919 | 0.7038 | 0.7395 | 0.7008 | 0.7299 | 0.7474 |
| 300 | B&L | 0.9144 | 0.9257 | 0.9239 | 0.9409 | 0.9542 | 0.9544 |
| 300 | L&W | 1.0358 | 1.0209 | 0.9789 | 1.0847 | 1.0125 | 0.9800 |
| 500 | DSCE | 0.6560 | 0.6886 | 0.6955 | 0.6877 | 0.7167 | 0.7115 |
| 500 | B&L | 0.9504 | 0.9685 | 0.9605 | 0.9593 | 0.9733 | 0.9714 |
| 500 | L&W | 1.0873 | 1.0594 | 1.0239 | 1.1203 | 1.0807 | 1.0673 |

The tables are indicative of high estimation accuracy of the proposed method in sparse settings as compared to the competition. The precision of DSCE seems to generally increase with dimension $p$ and decreased in sparsity $s$. As evident from Table 4.4.2, stronger signal makes both DSCE and POET perform better compared to the other methods.

## 4.5 Concluding Remarks

We consider estimation high-dimensional covariance estimation in sparse settings. Our approach goes beyond mere eigenvalue shrinkage by taking into consideration the behavior of eigenvectors in a large dimensional framework. Furthermore, we do not require diverging (pervasive) signals, but instead assume sparsity, which is a feature of many large datasets and a desirable characteristic to require in a number of applications. Our model consists of low-rank and sparse components, where the former also has sparse eigenvectors.

Our analysis shows that accurate estimation is possible, with the rate of convergence proportional to the optimal rate for sparse estimation as in [28]. The numerical experiment also reveals that the proposed algorithm can accurately estimate the induced covariance. It would also be worth considering an empirical application where nonzero coefficients are associated with loss, e.g. transaction costs in finance, so that it is desirable to have sparse estimates.

Our empirical application also demonstrates that the proposed method may offer substantial advantages over other high-dimensional estimation techniques. One of the possible extensions is a closer examination of a multi-spike covariance model and the issues arising therein, e.g. whether sparsity should vary across eigenvectors and how. Another extension would be the consideration of linear shrinkage (of eigenvalues) of a sparse estimator discussed here with a structured estimator. Finally, it is valuable to explore the properties of the inverse (precision) estimator induced by the proposed method.

# Appendices

## 4.A Auxiliary lemmas

**Lemma 17** *For unit vectors* u *and* v*,*

$$\|\mathrm{uu}' - \mathrm{vv}'\|_F^2 = 2 - 2(\mathrm{u}'\mathrm{v})^2$$

**Proof.**

$$\begin{aligned}\|\mathrm{uu}' - \mathrm{vv}'\|_F^2 &= \mathrm{tr}((\mathrm{uu}' - \mathrm{vv}')(\mathrm{uu}' - \mathrm{vv}')) \\ &= 1 - \mathrm{tr}(\mathrm{uu}'\mathrm{vv}) - \mathrm{tr}(\mathrm{vv}'\mathrm{uu}') + 1 \\ &= 2 - 2(\mathrm{u}'\mathrm{v})^2\end{aligned}$$

■

**Lemma 18** *For a rank-1 matrix* $A = \mathrm{x}_1\mathrm{x}_2'$ *we have*

$$\|A\| = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

**Proof.** The equality is trivial if $\mathrm{x}_2 = 0$, so consider $\mathrm{x}_2 \neq 0$. For a vector $\mathrm{u} = \frac{\mathrm{x}_2}{\|\mathrm{x}_2\|_2}$ we have

$$\|A\| \geq \|A\mathrm{u}\|_2 = \left\|\mathrm{x}_1\mathrm{x}_2' \frac{\mathrm{x}_2}{\|\mathrm{x}_2\|_2}\right\|_2 = \frac{1}{\|\mathrm{x}_2\|_2}\|\mathrm{x}_1\mathrm{x}_2'\mathrm{x}_2\|_2 = \frac{\|\mathrm{x}_2\|_2^2}{\|\mathrm{x}_2\|_2}\|\mathrm{x}_1\|_2 = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

On the other hand,

$$\|A\| = \|\mathrm{x}_1\mathrm{x}_2'\|_2 \leq \|\mathrm{x}_1\|_2\|\mathrm{x}_2'\|_2 = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

■

## 4.B Proof of Theorem 15

**Proof.** It suffices to prove that the max error is bounded

$$\|\widehat{S}_e - \Sigma_e\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right).$$

The desired rate on adaptive threshold estimator would follow immediately as discussed in [29] and [98].

Recall from Equation 4.10 that

$$\widehat{S}_e = S - \widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r',$$

where $\widehat{V}_r$ is a $p \times r$ matrix of sparse eigenvector estimates and $\widehat{\Lambda}_r$ is an $r \times r$ diagonal matrix of the corresponding eigenvalue estimates in descending order.

$$\begin{aligned}\|\widehat{S}_e - \Sigma_e\|_{\max} &= \|S - \widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - (\Sigma - U_rL_rU_r')\|_{\max} \\ &\leq \|S - \Sigma\|_{\max} + \|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\max}\end{aligned}$$

By assumption, $\|S - \Sigma\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right)$, so we want to show $\|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right)$. Then since the eigenvalues are bounded by assumption, we have

$$\begin{aligned}&\|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\max} \\ &\leq \|\widehat{V}_r(\widehat{\Lambda}_r - L_r)\widehat{V}_r'\|_{\max} + \|(\widehat{V}_r - U_r)L_r(\widehat{V}_r - U_r)'\|_{\max} + 2\|U_rL_r(\widehat{V}_r - U_r)'\|_{\max} \\ &= O_{\mathbb{P}}\left(\|\widehat{\Lambda}_r - L_r\|_{\max} + \|\widehat{V}_r - U_r\|_{\max}\right) \\ &= O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right).\end{aligned}$$

■

## 4.C Proof of Theorem 16

**Proof.**

$$\|\widehat{S} - \Sigma\| = \|\widehat{S}_s + \widehat{S}_e^\tau - \Sigma_s - \Sigma_e\|$$

$$\le \|\ell_1 \mathrm{u}_1\mathrm{u}_1' - \widehat{\lambda}_1 \widehat{\mathrm{v}}_1\widehat{\mathrm{v}}_1'\| + \|\widehat{S}_e^\tau - \Sigma_e\|$$

$$= \|\ell_1(\mathrm{u}_1\mathrm{u}_1' - \widehat{\mathrm{v}}_1\widehat{\mathrm{v}}_1') + (\ell_1 - \widehat{\lambda}_1)\widehat{\mathrm{v}}_1\widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1}s\log p}\right)$$

$$= \ell_1 O_{\mathbb{P}}(\|\mathrm{u}_1 - \widehat{\mathrm{v}}_1\|) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1\widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1}s\log p}\right) \tag{4.17}$$

$$= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}\|) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1\widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1}s\log p}\right) \tag{4.18}$$

$$= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1\|_2^2 + O_p\left(m\sqrt{n^{-1}s\log p}\right) \tag{4.19}$$

$$= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}) + O_{\mathbb{P}}(n^{-1/2}) + O_p\left(m\sqrt{n^{-1}s\log p}\right), \tag{4.20}$$

where (4.17) follows from Lemma 17 and the fact that $\|\cdot\| \le \|\cdot\|_F$ for a square matrices; (4.19) follows from Corrolary 13; (4.19) holds by Lemma 18.

To complete the proof observe that in Equation 4.20 the first term becomes $O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right)$ by Corrolary 14, while the second term is asymptotically negligible under the Assumption 2. ■

## 4.D Weyl and Davis-Kahan Theorems

Denote eigenvectors and eigenvalues as $\xi_i(\cdot)$ and $\lambda_i(\cdot)$ respectively. For Hermitian $p \times p$ matrices $A, \widehat{A}$,

**Proposition 19**

$$\max_{i=1,\ldots,p} |\lambda_i(\widehat{A}) - \lambda_i(A)| \leq \|\widehat{A} - A\|.$$

**Proposition 20**

$$\left\|\xi_i(\widehat{A}) - \xi_i(A)\right\| \leq \frac{\sqrt{2}\|\widehat{A} - A\|}{\min(|\lambda_{i-1}(\widehat{A}) - \lambda_i(A)|, |\lambda_{i+1}(\widehat{A}) - \lambda_i(A)|)}.$$

## 4.E T-Power Algorithm

```
Algorithm 2: Truncated Power Method [129]
    Input: $p \times p$ PSD matrix $A$, initial estimate $x_0 \in \mathbb{R}^p$, dimension $r$.
    t = 1
    while not converged do
        $x'_t = Ax_{t-1}/\|Ax_{t-1}\|$;                    /* Power iteration */
        $F_t = \mathrm{supp}(x'_t, r)$ ;                   /* Support of indices */
        $\hat{x}_t = \mathrm{Truncate}(x'_t, F_t)$;        /* Truncate */
        $x_t = \hat{x}_t/\|\hat{x}_t\|$ ;                  /* Standardize */
        $t = t + 1$
    end
```

## 4.F DSCE Algorithm

```
Algorithm 3: Proposed DSCE Algorithm
    Input: standardized $p \times n$ data matrix $X$, dimension $r$.
    $S = \frac{1}{n}X'X$ ;                                  /* Sample covariance */
    $\widehat{S}_s = \text{T-Power}(S)$ ;                   /* Rank-$r$ Truncuted power method */
    $\widehat{S}_e^\tau = \tau(S - \widehat{S}_s)$ ;        /* Adaptive Thresholding,(4.10) */
    Output: $p \times p$ matrix $\widehat{S} = \widehat{S}_s + \widehat{S}_e^\tau$.
```

# Bibliography

[1] Baha Abdalhamid, Christopher R Bilder, Emily L McCutchen, Steven H Hinrichs, Scott A Koepsell, and Peter C Iwen. Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources. *American Journal of Clinical Pathology*, 153(6):715–718, 04 2020.

[2] M. Aldridge, O. Johnson, and J. Scarlett. Improved group testing rates with constant column weight designs. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1381–1385, 2016.

[3] Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877 – 2921, 2009.

[4] T. W. Anderson. Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, 34(1):122 – 148, 1963.

[5] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* A Wiley publication in mathematical statistics. Wiley, 1984.

[6] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[7] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[8] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[9] Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[10] Jushan Bai and Serena Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317, 2008. Honoring the research contributions of Charles R. Nelson.

[11] Jushan Bai and Serena Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

[12] Z. Bai and Jack Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. 01 2010.

[13] Z. D. Bai, B. Q. Miao, and G. M. Pan. On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532 – 1572, 2007.

[14] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

[15] Jinho Baik, Gerard Ben Arous, and Sandrine Peche. Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Annals of Probability*, 33, 09 2005.

[16] Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

[17] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450, 2013.

[18] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, December 2008.

[19] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 585–591, Cambridge, MA, USA, 2001. MIT Press.

[20] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[21] Ben S. Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*. *The Quarterly Journal of Economics*, 120(1):387–422, 02 2005.

[22] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780 – 1815, 2013.

[23] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008.

[24] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294, 2006.

[25] Jean-Philippe Bouchaud and Marc Potters. Financial applications of random matrix theory: a short review. *arXiv.org, Quantitative Finance Papers*, 10 2009.

[26] Guido Bulligan, Massimiliano Marcellino, and Fabrizio Venditti. Forecasting economic activity with targeted predictors. *International Journal of Forecasting*, 31(1):188 – 206, 2015.

[27] T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074 – 3110, 2013.

[28] T. Tony Cai and Harrison H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389 – 2420, 2012.

[29] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[30] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[31] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), June 2011.

[32] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

[33] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

[34] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama. Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1832–1839, 2011.

[35] Xu Cheng and Bruce E. Hansen. Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280 – 293, 2015. High Dimensional Problems in Econometrics.

[36] Gregory Connor and Robert Korajczyk. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3):373–394, 1986.

[37] Gregory Connor and Robert Korajczyk. A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4):1263–91, 1993.

[38] Philippe Goulet Coulombe, Dalibor Stevanovic, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting? Cirano working papers, CIRANO, 2019.

[39] Stevanovic Dalibor. Common time variation of parameters in reduced-form macroeconomic models. *Studies in Nonlinear Dynamics & Econometrics*, 20(2):159–183, April 2016.

[40] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[41] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[42] Francis Diebold and Roberto Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20:134–44, 02 2002.

[43] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[44] David Donoho, Matan Gavish, and Iain Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742 – 1778, 2018.

[45] Robert Dorfman. The detection of defective members of large populations. *Ann. Math. Statist.*, 14(4):436–440, 12 1943.

[46] Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, 94(4):1014–1024, November 2012.

[47] Ding-Zhu Du and Frank K Hwang. *Combinatorial Group Testing and Its Applications.* WORLD SCIENTIFIC, 2nd edition, 1999.

[48] J C Emmanuel, M T Bassett, H J Smith, and J A Jacobs. Pooling of sera for human immunodeficiency virus (hiv) testing: an economical method for use in developing countries. *Journal of Clinical Pathology*, 41(5):582–585, 1988.

[49] Peter Exterkate, Patrick J.F. Groenen, Christiaan Heij, and Dick van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, 32(3):736 – 753, 2016.

[50] Eugene F. Fama and Kenneth R. French. The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465, 1992.

[51] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(4):603–680, September 2013.

[52] Jianqing Fan and Jinchi Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B*, 70:849–911, 11 2008.

[53] Jianqing Fan, Lingzhou Xue, and Jiawei Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306, 2017.

[54] FDA. Emergency Authorization for Sample Pooling, 2020. `https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-\issues-first-emergency-authorization-sample-pooling-diagnostic`.

[55] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[56] Mario Forni, Lucrezia Reichlin, Marc Hallin, and Marco Lippi. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82:540–554, 02 2000.

[57] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9:432–41, 08 2008.

[58] Sabyasachi Ghosh, Rishi Agarwal, Mohammad Rehan, Shreya Pathak, Pratyush Agarwal, Yash Gupta, Sarthak Consul, Nimay Gupta, Ritika Goyal, Ajit Rajwade, and Manoj Gopalkrishnan. A compressed sensing approach to group-testing for covid-19 detection. 05 2020.

[59] Bruno Giovannetti. Nonlinear forecasting using factor-augmented models. *Journal of Forecasting*, 32(1):32–40, 2013.

[60] Lisa Goldberg, Alex Papanicolaou, and Alex Shkolnik. The dispersion bias, 2020.

[61] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[62] A. H. Hirzel, J. Hausser, D. Chessel, and N. Perrin. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7):2027–2036, 2002.

[63] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 06 2008.

[64] Catherine A. Hogan, Malaya K. Sahoo, and Benjamin A. Pinsky. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. *JAMA*, 323(19):1967–1969, 05 2020.

[65] Jacqueline M. Hughes-Oliver. *Pooling Experiments for Blood Screening and Drug Discovery*, pages 48–68. Springer New York, New York, NY, 2006.

[66] O. Johnson, M. Aldridge, and J. Scarlett. Performance of group testing algorithms with near-constant tests per item. *IEEE Transactions on Information Theory*, 65(2):707–723, 2019.

[67] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327, 2001.

[68] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis, 2004.

[69] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. PMID: 20617121.

[70] I. T. Jolliffe. *Principal Component Analysis and Factor Analysis*, pages 115–128. Springer New York, New York, NY, 1986.

[71] Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[72] Tosio Kato. On the perturbation theory of closed linear operators. *J. Math. Soc. Japan*, 4(3-4):323–337, 12 1952.

[73] Hyun Hak Kim and Norman R. Swanson. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178:352 – 367, 2014. Recent Advances in Time Series Econometrics.

[74] Hyun Hak Kim and Norman R. Swanson. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2):339–354, 2018.

[75] Kwang In Kim, Keechul Jung, and Hang Joon Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2):40–42, 2002.

[76] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 02 2000.

[77] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[78] Varlam Kutateladze and Ekaterina Seregina. Code supplement to "Fast and Efficient Data Science Techniques for Covid-19 Group Testing, 2020. `https://tinyurl.com/y4vo86sb`.

[79] D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

[80] Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151, 11 2009.

[81] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[82] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012.

[83] Eugene Litvak, X. M. Tu, and Marcello Pagano. Screening for the presence of a disease by pooling sera samples. 1994.

[84] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967.

[85] Michael W. McCracken and Serena Ng. FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

[86] Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, page 148–188, 1989.

[87] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006.

[88] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

[89] Jakob Guldbæk Mikkelsen, Eric Hillebrand, and Giovanni Urga. Maximum Likelihood Estimation of Time-Varying Loadings in High-Dimensional Factor Models. CREATES Research Papers 2015-61, Department of Economics and Business Economics, Aarhus University, 12 2015.

[90] Leon Mutesa, Pacifique Ndishimye, Yvan Butera, Jacob Souopgui, Annette Uwineza, Robert Rutayisire, Emile Musoni, Nadine Rujeni, Thierry Nyatanyi, Edouard Ntagwabira, Muhammed Semakula, Clarisse Musanabaganwa, Daniel Nyamwasa, Maurice Ndashimye, Eva Ujeneza, Ivan Emile Mwikarago, Claude Mambo Muvunyi, Jean Baptiste Mazarati, Sabin Nsanzimana, Neil Turok, and Wilfred Ndifon. A strategy for finding people infected with sars-cov-2: optimizing pooled testing at low prevalence. *medRxiv*, 2020.

[91] Marco Del Negro and Christopher Otrok. Dynamic factor models with time-varying parameters: measuring changes in international business cycles. Staff Reports 326, Federal Reserve Bank of New York, 2008.

[92] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pages 849–856. MIT Press, 2002.

[93] Debdeep Pati, Anirban Bhattacharya, Natesh S. Pillai, and David Dunson. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102 – 1130, 2014.

[94] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[95] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[96] Jeff Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39 – 61, 2000.

[97] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341 – 360, 1976.

[98] Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[99] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[100] Thomas Sargent and Christopher Sims. Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis, 1977.

[101] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Advances in kernel methods - Support vector learning*, pages 327–352. MIT Press, 1999.

[102] Shaobing Chen and D. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1, 1994.

[103] John Shawe-Taylor, Chris Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In Nicolò Cesa-Bianchi, Masayuki Numao, and Rüdiger Reischuk, editors, *Algorithmic Learning Theory*, pages 23–40, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[104] Dan Shen, Haipeng Shen, Hongtu Zhu, and J. S. Marron. The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747–1770, 2016.

[105] Jack W Silverstein. Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. *Journal of Multivariate Analysis*, 15(3):295–324, 1984.

[106] Nasa Sinnott-Armstrong, Daniel Klein, and Brendan Hickey. Evaluation of group testing for sars-cov-2 rna. *medRxiv*, 2020.

[107] Milton Sobel and Phyllis A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal*, 38(5):1179–1252, 1959.

[108] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif., 1956. University of California Press.

[109] Andrew Sterrett. On the detection of defective members of large populations. *The Annals of Mathematical Statistics*, 28(4):1033–1036, 1957.

[110] James H Stock and Mark W Watson. Business cycle fluctuations in u.s. macroeconomic time series. Working Paper 6528, National Bureau of Economic Research, April 1998.

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

[112] James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

[113] James H. Stock and Mark W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, 2012.

[114] Steve M. Taylor, Jonathan J. Juliano, Paul A. Trottman, Jennifer B. Griffin, Sarah H. Landis, Paluku Kitsa, Antoinette K. Tshefu, and Steven R. Meshnick. High-throughput pooling and real-time pcr-based strategy for malaria detection. *Journal of Clinical Microbiology*, 48(2):512–519, 2010.

[115] Joshua Tenenbaum, Vin Silva, and John Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 01 2000.

[116] Timo Teräsvirta, Dag Tjøstheim, and Clive W.J. Granger. Chapter 48 aspects of modelling nonlinear time series. In *Handbook of Econometrics*, volume 4, pages 2917 – 2957. Elsevier, 1994.

[117] Godfrey H. Thompson. Methods of estimating mental factors. *Nature*, 141(3562):246–246, Feb 1938.

[118] Craig A. Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727 – 754, 1996.

[119] Tam T. Van, Joseph Miller, David M. Warshauer, Erik Reisdorf, Daniel Jernigan, Rosemary Humes, and Peter A. Shult. Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by pcr. *Journal of Clinical Microbiology*, 50(3):891–896, 2012.

[120] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[121] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[122] Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342 – 1374, 2017.

[123] Daniela Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10:515–34, 05 2009.

[124] Worldometer. US SARS-CoV-2 cases, 2020. `https://www.worldometers.info/coronavirus/country/us/`.

[125] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[126] Ilker Yalcin and Yasuo Amemiya. Nonlinear factor analysis as a statistical method. *Statist. Sci.*, 16(3):275–294, 08 2001.

[127] Idan Yelin, Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagam Gandali, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, and Roy Kishony. Evaluation of covid-19 rt-qpcr test in multi-sample pools. *medRxiv*, 2020.

[128] Jirong Yi, Raghu Mudumbai, and Weiyu Xu. Low-cost and high-throughput testing of covid-19 viruses and antibodies via compressed sensing: System concepts and computational experiments. 04 2020.

[129] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14, 12 2011.

[130] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

[131] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press, 2006.