2017

# Topics In Multivariate Statistics

Xin Lu Tan
*University of Pennsylvania*, xinlutan92@gmail.com

# Topics In Multivariate Statistics

**Abstract**
Multivariate statistics concerns the study of dependence relations among multiple variables of interest. Distinct from widely studied regression problems where one of the variables is singled out as a response, in multivariate analysis all variables are treated symmetrically and the dependency structures are examined, either for interest in its own right or for further analyses such as regressions. This thesis includes the study of three independent research problems in multivariate statistics.

The first part of the thesis studies additive principal components (APCs for short), a nonlinear method useful for exploring additive relationships among a set of variables. We propose a shrinkage regularization approach for estimating APC transformations by casting the problem in the framework of reproducing kernel Hilbert spaces. To formulate the kernel APC problem, we introduce the Null Comparison Principle, a principle that ties the constraint in a multivariate problem to its criterion in a way that makes the goal of the multivariate method under study transparent. In addition to providing a detailed formulation and exposition of the kernel APC problem, we study asymptotic theory of kernel APCs. Our theory also motivates an iterative algorithm for computing kernel APCs.

The second part of the thesis investigates the estimation of precision matrices in high dimensions when the data is corrupted in a cellwise manner and the uncontaminated data follows a multivariate normal distribution. It is known that in the setting of Gaussian graphical models, the conditional independence relations among variables is captured by the precision matrix of a multivariate normal distribution, and estimating the support of the precision matrix is equivalent to graphical model selection. In this work, we analyze the theoretical properties of robust estimators for precision matrices in high dimensions. The estimators we analyze are formed by plugging appropriately chosen robust covariance matrix estimators into the graphical Lasso and CLIME, two existing methods for high-dimensional precision matrix estimation. We establish error bounds for the precision matrix estimators that reveal the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution, and also analyze the breakdown point of both estimators. We also discuss implications of our work for Gaussian graphical model estimation in the presence of cellwise contamination.

The third part of the thesis studies the problem of optimal estimation of a quadratic functional under the Gaussian two-sequence model. Quadratic functional estimation has been well studied under the Gaussian sequence model, and close connections between the problem of quadratic functional estimation and that of signal detection have been noted. Focusing on the estimation problem in the Gaussian two-sequence model, in this work we propose optimal estimators of the quadratic functional for different regimes and establish the minimax rates of convergence over a family of parameter spaces. The optimal rates exhibit interesting phase transition in this family. We also discuss the implications of our estimation results on the associated simultaneous signal detection problem.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

TOPICS IN MULTIVARIATE STATISTICS

Xin Lu Tan

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

**Supervisor of Dissertation**

Andreas Buja
Liem Sioe Liong/First Pacific Company
Professor of Statistics

**Co-Supervisor of Dissertation**

Zongming Ma
Associate Professor of Statistics

**Graduate Group Chairperson**

Catherine Schrand
Celia Z. Moh Professor of Accounting

**Dissertation Committee**

Lawrence D. Brown
Miers Busch Professor of Statistics

Abba M. Krieger
Robert Steinberg Professor of Statistics

Po-Ling Loh
Assistant Professor of Electrical and Computer Engineering
University of Wisconsin-Madison

TOPICS IN MULTIVARIATE STATISTICS



Xin Lu Tan

# Acknowledgments

I would like to start by thanking my advisors, Professor Andreas Buja and Professor Zongming Ma. Andreas instilled within me the virtues of deep and intuitive understanding of problems as well as the importance of healthy skepticism in research. With his patience, guidance, and generous sharing of life wisdoms, Andreas guided me through a lot of hurdles in my Ph.D. study and helped me embrace my innate strengths as a researcher. In addition, Andreas inspired me with his immense curiosity and openness in learning new subjects, and his tenacity in pursuing formulation and solution of problems that are both elegant and fundamental. His high standards in writing and publishing had also time and again pushed me to go above and beyond my self-presumed limits. Andreas also has an exuberant and playful persona that I personally enjoy. On the other hand, Zongming piqued my interest in functional analysis and the theory of multivariate statistics. He encouraged me to make several moves that had since became keys in shaping my current Ph.D. portfolio. He is also generous in sharing perspectives and in offering advice when I am in doubt. Andreas and Zongming, thank you so much for willing to come together and work with me on topics that may be new to you as well. Thank you so much for encouraging me to explore broadly rather than to specialize prematurely in research. I am grateful

for the time we spent together studying functional principal component analysis and reproducing kernel Hilbert spaces, and for the many fun-filled conversations we had on a diverse range of topics. The time and memories we shared will always remain in my mind.

I would next like to thank Professor Tony Cai for introducing me to the minimax framework of statistical thinking, and for guiding me throughout our study of quadratic functional estimation. His keen acumen, extensive knowledge and dedication to research have made working with him an enjoyable learning experience.

My special thanks go to Professor Po-Ling Loh, who has been a friend, a mentor, and a role model to me. She gave me invaluable support, encouragement and companionship during our overlapping years. Po-Ling, the opportunity to attend your wedding has been an extremely illuminating experience as it prompted me to start contemplating who I want to be as I transition into adulthood. You are a truly inspiring person and I am very grateful for our friendship and for our research collaboration in robust statistics.

I would also like to thank the rest of my committee members, Professor Abba Krieger and Professor Larry Brown. Abba, thank you so much for working with me on the model selection probability project over one of my summers at Penn. Having TA'ed for you twice, I am also inspired by the deep care you showed on students learning. Larry, thank you so much for tossing out the problem of model selection probability estimation. Your relentless enthusiasm for research and devotion to the development of the statistical community are great exemplifications of scholarly spirit and are nothing short of inspirational.

Additionally, I would like to thank Professor Mark Low for admitting me to the program, without which this journey would not have been possible.

My thanks also go to the entirety of the Wharton Statistics Department. To the

many professors with whom I have interacted, thank you so much for your friendliness, time, and advice. To our wonderful staff, thank you so much for the boundless support you provided, including assistance in room scheduling, package tracking, computing, conferences reimbursement, doctoral program logistics, and so much more. To the friends whom I had the pleasure of meeting during my Ph.D. study, including my cohort-mates (Zijian, Sam, Peichao, Dan, Tengyuan), and other students with whom I have overlapped, thank you for your camaraderie and friendship. Your companionship has made my life during Ph.D. study so much more enjoyable.

Last, but not least, I would like to thank my parents, Chiam Hua Tan and Ai Sian Teo, and my aunt, Sarah Teo, for believing in me and for unconditionally loving me. Thank you for being supportive of my education decision, and for always being there for me through times of joy and times of hardship. No matter what life may throw at me, I know I will always have your unwavering love and support.

# ABSTRACT

# TOPICS IN MULTIVARIATE STATISTICS

Xin Lu Tan

Andreas Buja

Zongming Ma

Multivariate statistics concerns the study of dependence relations among multiple variables of interest. Distinct from widely studied regression problems where one of the variables is singled out as a response, in multivariate analysis all variables are treated symmetrically and the dependency structures are examined, either for interest in its own right or for further analyses such as regressions. This thesis includes the study of three independent research problems in multivariate statistics.

The first part of the thesis studies additive principal components (APCs for short), a nonlinear method useful for exploring additive relationships among a set of variables. We propose a shrinkage regularization approach for estimating APC transformations by casting the problem in the framework of reproducing kernel Hilbert spaces. To formulate the kernel APC problem, we introduce the Null Comparison Principle, a principle that ties the constraint in a multivariate problem to its criterion in a way that makes the goal of the multivariate method under study transparent. In addition to providing a detailed formulation and exposition of the kernel APC problem, we study asymptotic theory of kernel APCs. Our theory also motivates an iterative algorithm for computing kernel APCs.

The second part of the thesis investigates the estimation of precision matrices in high dimensions when the data is corrupted in a cellwise manner and the uncontaminated data follows a multivariate normal distribution. It is known that in the setting of Gaussian graphical models, the conditional independence relations among variables is captured by the precision matrix of a multivariate normal distribution, and estimating the support of the precision matrix is equivalent to graphical model selection. In this work, we analyze the theoretical properties of robust estimators for precision matrices in high dimensions. The estimators we analyze are formed by plugging appropriately chosen robust covariance matrix estimators into the graphical Lasso and CLIME, two existing methods for high-dimensional precision matrix estimation. We establish error bounds for the precision matrix estimators that reveal the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution, and also analyze the breakdown point of both estimators. We also discuss implications of our work for Gaussian graphical model estimation in the presence of cellwise contamination.

The third part of the thesis studies the problem of optimal estimation of a quadratic functional under the Gaussian two-sequence model. Quadratic functional estimation has been well studied under the Gaussian sequence model, and close connections between the problem of quadratic functional estimation and that of signal detection have been noted. Focusing on the estimation problem in the Gaussian two-sequence model, in this work we propose optimal estimators of the quadratic functional for different regimes and establish the minimax rates of convergence over a family of parameter spaces. The optimal rates exhibit interesting phase transition in this family. We also discuss the implications of our estimation results on the associated simultaneous signal detection problem.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

In the past decades, advances in technology have enabled collection of massive amounts of data, opening the door to a new approach to understanding the world and making decisions. Despite the wealth of data available, the ability to unlock the value in data rests on our ability to summarize the data and provide interpretation of the summary quantities computed. Such summaries and corresponding interpretations can rarely be produced by just looking at the raw data, and a careful scientific scrutiny and statistical analysis are crucial for the generation of valuable insights from data.

Often times, the data collected involves measurements of multiple variables on the same unit, rendering the variables correlated and univariate analyses insufficient for deriving conclusion and guiding next steps. In these cases, a statistical analysis of the dependencies structure of the variables is essential. The study of dependence relations among multiple variables of interest is at the heart of multivariate statistics, and is the focus of this thesis. There are three main chapters within the body of this thesis, each of which is a single, self-contained paper. While the topics studied in these chapters fall under the general realm of multivariate statistics, they also come with interesting twists by having connections to nonlinear statistics, robust statistics, as well as high-dimensional statistics.

A brief summary of the contents in subsequent chapters is provided below.

# Kernel Additive Principal Components

In Chapter 2, we study additive principal components (APCs for short), a nonlinear generalization of linear principal components. We focus on smallest APCs to describe additive nonlinear constraints that are approximately satisfied by the data. Thus, an APC analysis fits data with implicit equations that treat the variables symmetrically, as opposed to regression analyses which fit data with explicit equations that treat the variables asymmetrically by singling out a response.

APCs were initially proposed by Donnell et al. (1994), where a subspace restriction regularization approach was introduced for estimating APC transformations. In this chapter, we cast APCs in the context of penalized least squares and reproducing kernel Hilbert spaces (RKHSs), and take advantage of the extensions offered by kernelizing. In contrast to the existing subspace restriction approach, kernelizing approaches achieve regularization through shrinkage and therefore grant distinctive flexibility in APCs estimation by allowing the use of infinite-dimensional function spaces while retaining computational feasibility. Furthermore, the interpretation of regularization kernels as similarity measures makes possible the exploration of implicit additive redundancies in *non-Euclidean* data, a flexibility not available in the original APC proposal.

Introducing kernelizing into a multivariate method is not a mechanical exercise. We motivate our formulation of kernel APCs by the Null Comparison Principle, a principle that ties the constraint in a multivariate problem to its criterion in a way that makes the goal of the multivariate method under study transparent. This simple yet powerful principle is potentially useful for devising generalizations of other multivariate methods and thus can be of independent interest.

On the other hand, kernel canonical correlation analysis (CCA) is a special case of kernel APCs with two variables, and the statistical convergence of kernel CCA was

first established in Fukumizu et al. (2007). In this chapter, we establish the statistical convergence of kernel APCs under a decay rate for regularization parameters involved that is less stringent than that in Fukumizu et al. (2007). Our proof of convergence is built on an elegant RKHS-based theory we develop for APCs, which covers general RKHSs not studied in Fukumizu et al. (2007) and do not require the population targets to lie in RKHSs a priori. Our theory also motivates an iterative algorithm for computing finite-sample kernel APCs. Lastly, we provide data examples, simulated and real, to illustrate the kernel APC methodology. Supplementary materials for this chapter can be found in Appendix A.

This chapter is joint work with Andreas Buja and Zongming Ma.

# High-dimensional Robust Precision Matrix Estimation: Cellwise Corruption under $\epsilon$-Contamination

In Chapter 3, we analyze theoretical properties of robust estimators for precision matrices, when data are contaminated in a cellwise manner: each element of the data matrix is independently corrupted according to a certain proportion. Such contamination mechanisms may be used to model various phenomena in real-world scientific data, including measurement error in DNA microarray analysis and dropouts in sensor arrays.

When data follows an uncontaminated multivariate normal distribution, the graphical Lasso (GLasso) (Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) and the constrained $\ell_1$-minimization for inverse matrix estimation (CLIME) (Cai et al., 2011) estimators are known to possess rigorous theoretical guarantees for the estimation of precision matrices in high dimensions; however, their performance may be compromised severely when data are contaminated by even a single outlier.

The estimators we study are inspired by techniques in robust statistics and are constructed by plugging appropriately chosen robust covariance matrix estimators into the GLasso and CLIME. We derive high-dimensional error bounds that reveal the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution, and also analyze the breakdown point of both estimators. Our results show that although the graphical Lasso and CLIME estimators perform equally well from the point of view of statistical consistency, the breakdown property of the graphical Lasso is superior to that of CLIME. We also discuss implications of our work for gaussian graphical model estimation in the presence of contamination, where the goal is to estimate the support of the graph associated with the clean distribution. Our results apply to arbitrary contaminating distributions and allow for a nonvanishing fraction of cellwise contamination. Finally, we examine the performance of our estimators in comparison to that of other (possibly non-robust) estimators through simulation studies. Supplementary materials for this chapter can be found in Appendix B.

This chapter is joint work with Po-Ling Loh.

# Optimal Estimation of A Quadratic Functional under the Gaussian Two-Sequence Model

While Chapters 2 and 3 focus on the analysis of *covariance structure* of multiple variables, Chapter 4 involves an analysis of *mean structure* of the variables. Specifically, we study in Chapter 4 the problem of optimal estimation of the quadratic functional $Q(\mu, \theta) = \frac{1}{n}\sum_{i=1}^{n} \mu_i^2\theta_i^2$ under the gaussian two-sequence model. The mean vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\theta = (\theta_1, \ldots, \theta_n)$ are assumed to be sparse.

In addition to being of significant theoretical interest in its own right, this es-

timation problem is motivated by the problem of simultaneous signal detection in integrative genomics, which, under our simplified framework, is equivalent to the detection of locations $i$ where $\mu_i$ and $\theta_i$ are simultaneously non-zero. We propose optimal estimators of $Q(\mu, \theta)$ and establish the minimax rates of convergence over a family of parameter spaces. Interestingly, the optimal rates exhibit different phase transitions in three regimes, each characterized by the sparsity of simultaneous non-zero means relative to that of non-zero entries in individual mean vectors. Along with the establishment of the minimax rates of convergence, we explain the intuition behind the construction of the optimal estimators in each regime. A simulation study is included to complement the theoretical results in the chapter. We also give a brief discussion on the application of quadratic functional estimators to the problem of simultaneous signal detection. Supplementary materials for this chapter can be found in Appendix C.

This chapter is joint work with T. Tony Cai, and will appear in *Statistica Sinica*.

# 2 Kernel Additive Principal Components*

## 2.1 Introduction

Linear principal component analysis (PCA) is a tool commonly used to reduce the dimensionality of data sets consisting of several interrelated variables $X_1, X_2, \ldots, X_p$. PCA amounts to finding linear functions of the variables, $\sum a_j X_j$, whose variances are maximal or, more generally, large and stationary under a unit norm constraint, $\sum a_j^2 = 1$. These linear combinations, called *largest linear principal components* (largest LPCs for short), are thought to represent low-dimensional linear structure of the data. The reader is referred to Jolliffe (2002) for a comprehensive review of PCA.

One can similarly define the *smallest linear principal component* (smallest LPCs) as linear functions of the variables whose variances are minimal or small and stationary subject to a unit norm constraint on the coefficients. If these variances are near zero, $\mathrm{Var}\left(\sum a_j X_j\right) \approx 0$, the interpretation is that the data lie near the hyperplane defined by the linear constraint $\sum a_j X_j = 0$ (assuming that the variables $X_j$ are centered). Thus the purpose of performing PCA on the lower end of the princi-

*Joint work with Andreas Buja and Zongming Ma

pal components spectrum is quite different from that of performing it on the upper end: largest principal components are concerned with structure of *low dimension*, whereas smallest principal components are concerned with structure of *low codimension*. Largest LPCs provide *projections* to low dimensions, whereas smallest LPCs provide *implicit equations* to approximate linear dependencies among variables.

The topic of this chapter is a generalization of smallest linear principal components in function spaces, called "smallest additive principal components" ("APCs" for short). APCs were initially proposed by Donnell et al. (1994) before kernelizing became a well-understood methodology. The goal of this chapter is to cast APCs in the context of penalized least squares and reproducing kernel Hilbert spaces (RKHSs), and take advantage of the extensions offered by kernelizing.

Before proceeding, here is a brief summary of additive approaches to multivariate function fitting: The step from a linear method to an additive method consists of replacing linear terms $a_j X_j$ with nonlinear terms $\phi_j(X_j)$, thereby allowing nonlinear marginal transformations of the coordinate variables $X_j$, each to be estimated by some nonlinear fitting method. It is known that additive approaches avoid the curse of dimensionality that fully nonlinear function fitting $\phi(X_1, X_2, ..., X_p)$ would entail. Historically the generalization from linear to additive approaches first appeared in the context of regression, where fitting linear equations $Y \sim \sum_j a_j X_j$ was extended to fitting additive equations $Y \sim \sum_j \phi_j(X_j)$ to a response $Y$, as documented by Breiman & Friedman (1985), Buja et al. (1989), culminating in the classical book by Hastie & Tibshirani (1990). Additive extensions were enabled at the time by the emergence of fast smoothing technology that allows estimation and computation of suitably regularized transformations $\phi_j(X_j)$ with an iterative algorithm called "backfitting", whereby each $\phi_j(X_j)$ is updated in turn by a smoothing step of partial residuals on $X_j$: $Y - \sum_{k \neq j} \phi_k(X_k) \sim \phi_j(X_j)$. The main output is a series of plots, $\phi_j(X_j)$ against

$X_j$, that reveal the nonlinearities graphically, while relative variable importances are measured by the standard deviations of the transforms $\phi_j(X_j)$.

Similar to the additive extension of linear regression, the additive extension of LPCs implies the replacement of the linear terms $a_j X_j$ with nonlinear terms $\phi_j(X_j)$, hence an additive principal component is of the form $\sum \phi_j(X_j)$. In additive regression it is approximation of the response variable that produces non-trivial transformations; in additive principal components it is a normalizing constraint resulting in an eigenvalue problem that achieves the same. In generalizing LPCs to APCs, one therefore needs to find a suitable way to generalize the LPC constraint $\sum a_j^2 = 1$. Donnell et al. (1994) proposed to use the constraint $\sum \mathrm{Var}\, \phi_j(X_j) = 1$, their justification being that for $\phi_j(X_j) = a_j X_j$ we have $\mathrm{Var}\,(\phi_j(X_j)) = a_j^2$ for real-valued $X_j$ with $\mathrm{Var}\,(X_j) = 1$, resulting in the conventional constraint $\sum a_j^2 = 1$. A *smallest APC* can then be defined as a $p$-tuple of marginal transformations $\phi_1, \phi_2, \ldots, \phi_p$ that minimizes $\mathrm{Var}\,(\sum \phi_j(X_j))$ subject to $\sum \mathrm{Var}\,(\phi_j(X_j)) = 1$.

The interpretation of a smallest APC is that the additive constraint represented by the implicit additive equation $\sum \phi_j(X_j) = 0$ defines a nonlinear or, more precisely, an additive manifold that approximates the data. Smallest APCs can have multiple methodological uses:

- APCs can be used as a generalized collinearity diagnostic for additive regression models. Just as approximate collinearities $\sum \alpha_j X_j \approx 0$ destabilize inference in linear regression $Y \sim \sum \beta_j X_j$, additive approximate "concurvities" (Donnell et al., 1994) of the form $\sum \phi_j(X_j) \approx 0$ destabilize inference in additive regression $Y \sim \sum \psi_j(X_j)$. Such concurvities can be found by applying APC analysis to the predictors of an additive regression.

- APCs can also be used as a symmetric alternative to additive regression as well as to ACE regression (Breiman & Friedman, 1985) when it is not possible or not

desirable to single out any one of the variables as a response. Additive implicit equations estimated with APCs will then freely identify the variables that have strong additive associations with each other.

- Even when there is a specific response variable of interest in the context of an additive regression, an APC analysis of all variables, including predictors as well as response, can serve as an indicator of the strength of the regression, depending on whether the response variable has a strong presence in the smallest APC. If the response shows up only weakly, it follows that the predictors have stronger additive associations among each other than with the response.

Examples of applications of smallest APCs will be given in Section 2.9, and simulation examples in Section 2.10.

Estimation of APCs and their transforms $\phi_j(X_j)$ from finite data requires some form of regularization. There exist two broad classes of regularization in nonparametric function estimation, namely, subspace regularization and shrinkage regularization. Subspace regularization restricts the function estimates $\hat{\phi}_j$ to finite-dimensional function spaces on $X_j$. Shrinkage regularization produces function estimates by adding a penalty to the goodness-of-fit measure in order to impose the spatial structure of $X_j$ on $\hat{\phi}_j$. Commonly used are generalized ridge-type quadratic penalties (also called the "kernelizing approach") and lasso-type $\ell_1$-penalties. The original APC proposal in Donnell et al. (1994) only uses subspace regularization for estimation, and it does not provide asymptotic theory for it. In the present chapter we investigate APCs based on shrinkage/kernelizing regularization and provide some asymptotic consistency theory.

It should be pointed out that introducing a shrinkage/kernelizing approach into a multivariate method is not a mechanical exercise. It is not a priori clear where and how the penalties should be inserted into a criterion of multivariate analysis, which in the case of PCA is variance subject to a constraint. The situation differs from

regression where there is no conceptual difficulty in adding a regularization penalty to a goodness-of-fit measure. In a PCA-like method such as APC analysis, however, it is not clear whether penalties should be added to, or subtracted from, the variance, or somehow added to the constraint, or both. An interesting and related situation occurred in functional multivariate analysis where the same author (B. Silverman) co-authored two different approaches to the same PCA regularization problem (Rice & Silverman, 1991; Silverman, 1996), differing in where and how the penalty is inserted. Our approach, if transposed to functional multivariate analysis, agrees with neither of them. One reason for our third way is that neither of the approaches in Rice & Silverman (1991) or Silverman (1996) generalize to the low end of the PCA spectrum. In contrast, the regularized criterion proposed in this chapter can be applied to the high and the low end of the spectrum, and hence to the discovery of low dimension as well as low co-dimension. Our more specific interest is in the latter.

An immediate benefit of injecting penalty regularization into multivariate analysis stems from recent methodological innovations in kernelizing. These include the possibility of using infinite-dimensional function spaces, the interpretation of regularization kernels as positive definite similarity measures, and the kernel algebra with the freedom of modeling it engenders. Two decades ago, when Donnell et al. (1994) was written, it would have been harder to make the case for penalty regularization.

In what follows we first describe the mathematical structure of APCs and give a review on population APCs that constitute our targets of estimation (Section 2.2). Section 2.3 introduces the Null Comparison Principle that guides the derivation of our kernel APC problem in Section 2.4. Section 2.5 poses the kernel APC problem in the framework of reproducing kernel Hilbert spaces. Although our focus on the lower end of the spectrum seems to have found little precedence in the literature, the criterion we use for kernel APC turns out to be equivalent to that of kernel canonical

correlation analysis (kernel CCA) (Bach & Jordan, 2003), a nonlinear extension of canonical correlation analysis, when there are only two variables of interest. The statistical convergence of kernel CCA was first established in Fukumizu et al. (2007). In Section 2.6, we establish the statistical convergence of kernel APCs under a decay rate for regularization parameters involved that is less stringent than that in Fukumizu et al. (2007). Our proof of convergence is built on an elegant RKHS-based theory we develop for APCs in Section 2.5, which covers general RKHSs not studied in Fukumizu et al. (2007) and do not require the population targets to lie in RKHSs a priori. Section 2.7 presents the power algorithm for computing kernel APCs, whereas Section 2.8 contains a brief discussion on the selection of penalty parameters. In Section 2.9 we present the kernel APC methodology in terms of two data examples. Section 2.10 contains simulation studies to complement our theoretical results. A discussion on the relation of kernel APC with kernel PCA (Schölkopf et al., 1998; Schölkopf & Smola, 2002) and kernel CCA is given in Section 2.11. Section 2.12 concludes. To deal with the generality of RKHSs considered in Section 2.5, we need some technical results whose proofs are collected in Appendix A.1. Proofs of the consistency results stated in Section 2.6 are given in Appendix A.2, whereas proofs related to the power algorithm of Section 2.7 are given in Appendix A.3. Appendix A.4 contains implementation details for the power algorithm, while Appendix A.5 contains an alternative linear algebra method for computing sample kernel APCs. Details on the comparison of kernel APC with kernel PCA is given in Appendix A.6.

The following notations and concepts in functional analysis are useful for the discussion that follows.

**Notation:** Let $\mathcal{H}$, $\mathcal{H}_1$, $\mathcal{H}_2$ be Hilbert spaces. In this chapter, a Hilbert space always means a separable Hilbert space. We denote the norm of a bounded linear operator $\mathbf{T} : \mathcal{H}_1 \to \mathcal{H}_2$ by $\|\mathbf{T}\| := \sup_{\|\phi\|_{\mathcal{H}_1} \leq 1} \|\mathbf{T}\phi\|_{\mathcal{H}_2}$. The null space and the range of $\mathbf{T}$

are denoted by $\mathcal{N}(\mathbf{T})$ and $\mathcal{R}(\mathbf{T})$, respectively, where $\mathcal{N}(\mathbf{T}) = \{\phi \in \mathcal{H}_1 : \mathbf{T}\phi = 0\}$ and $\mathcal{R}(\mathbf{T}) = \{\mathbf{T}\phi \in \mathcal{H}_2 : \phi \in \mathcal{H}_1\}$. We denote by $\mathbf{T}^*$ the Hilbert space adjoint of $\mathbf{T}$. We say that $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ is self-adjoint if $\mathbf{T}^* = \mathbf{T}$, and that a bounded linear self-adjoint operator $\mathbf{T}$ is positive if $\langle \phi, \mathbf{T}\phi \rangle \geq 0$ for all $\phi \in \mathcal{H}$. We write $\mathbf{T} \succeq 0$ if $\mathbf{T}$ is positive, and $\mathbf{T}_1 \succeq \mathbf{T}_2$ if $\mathbf{T}_1 - \mathbf{T}_2$ is positive. If $\mathbf{T}$ is positive, we denote by $\mathbf{T}^{1/2}$ the unique positive operator $\mathbf{B}$ satisfying $\mathbf{B}^2 = \mathbf{T}$. On the other hand, a bounded linear operator $\mathbf{T} : \mathcal{H}_1 \to \mathcal{H}_2$ is compact if $\mathbf{T}$ takes bounded sets in $\mathcal{H}_1$ into precompact sets in $\mathcal{H}_2$. One nice property of a compact operator is the availability of singular value decomposition: for some $N \in \mathbb{N} \cup \{\infty\}$, there exist (not necessarily complete) orthonormal sets $\{\phi_\nu\}_{\nu=1}^N \subset \mathcal{H}_1$ and $\{\psi_\nu\}_{\nu=1}^N \subset \mathcal{H}_2$ and positive real numbers $\{\lambda_\nu\}_{\nu=1}^N$ called singular values, such that

$$\mathbf{T} = \sum_{\nu=1}^{N} \lambda_\nu \langle \phi_\nu, \cdot \rangle_{\mathcal{H}_1} \psi_\nu.$$

If $N = \infty$, then $\lambda_\nu \to 0$ and the infinite series in the equation above converges in norm. We say that a bounded linear operator $\mathbf{T} : \mathcal{H}_1 \to \mathcal{H}_2$ is Hilbert-Schmidt if $\sum_{k=1}^\infty \sum_{l=1}^\infty \langle \psi_l, \mathbf{T}\phi_k \rangle_{\mathcal{H}_2}^2 = \sum_{k=1}^\infty \|\mathbf{T}\phi_k\|_{\mathcal{H}_2}^2 < \infty$ for a complete orthonormal basis system (CONS) $\{\phi_k\}_{k=1}^\infty$ of $\mathcal{H}_1$ and $\{\psi_l\}_{l=1}^\infty$ of $\mathcal{H}_2$. It is known that this sum is independent of the choices of CONS. For two Hilbert-Schmidt operators $\mathbf{T}_1$ and $\mathbf{T}_2$, the Hilbert-Schmidt inner product is defined by

$$\langle \mathbf{T}_1, \mathbf{T}_2 \rangle_{\mathrm{HS}} = \sum_{k=1}^\infty \sum_{l=1}^\infty \langle \psi_l, \mathbf{T}_1\phi_k \rangle_{\mathcal{H}_2} \langle \psi_l, \mathbf{T}_2\phi_k \rangle_{\mathcal{H}_2} = \sum_{k=1}^\infty \langle \mathbf{T}_1\phi_k, \mathbf{T}_2\phi_k \rangle_{\mathcal{H}_2},$$

with which the set of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ form a Hilbert space. The Hilbert-Schmidt norm $\|\mathbf{T}\|_{\mathrm{HS}}$ is again given by $\|\mathbf{T}\|_{\mathrm{HS}}^2 = \langle \mathbf{T}, \mathbf{T} \rangle_{\mathrm{HS}} = \sum_{k=1}^\infty \|\mathbf{T}\phi_k\|_{\mathcal{H}_2}^2$. Obviously, if $\mathbf{T}$ is Hilbert-Schmidt, then $\|\mathbf{T}\| \leq \|\mathbf{T}\|_{\mathrm{HS}}$. Moreover, a Hilbert-Schmidt operator is compact, whereas a compact operator is Hilbert-Schmidt

iff the singular values satisfy $\sum \lambda_\nu^2 < \infty$. For other standard functional analysis concepts, see Reed & Simon (1980).

## 2.2 Population APCs

In this section, we give a review on population APCs (Donnell et al., 1994) which forms the foundation for RKHS-based theory of APCs in later sections.

### 2.2.1 Transformations and Their Interpretations

Let $X_1, \ldots, X_p$ be random variables taking on values in arbitrary measurable spaces $(\mathcal{X}_1, \mathcal{B}_{\mathcal{X}_1})$, ..., $(\mathcal{X}_p, \mathcal{B}_{\mathcal{X}_p})$, each of which can be continuous or discrete, temporal or spatial, high- or low-dimensional. The only assumption at this point is that they have a joint distribution $P_{1:p}(dx_1, \ldots, dx_p)$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_p$. Quantitative random variables $\phi_j(X_j)$ can be obtained by applying real-valued functions $\phi_j : \mathcal{X}_j \to \mathbb{R}$ to the arbitrarily-valued $X_j$. The functions $\phi_j$ are often interpreted as "scorings" or "scalings" or "quantifications" of the underlying spaces $\mathcal{X}_j$. If $X_j$ is already real-valued, then $\phi_j$ is interpreted as a variable transformation.

Donnell et al. (1994) considers functions $\phi_j$ that belong to some closed subspace $H_j$ of square-integrable functions with regard to their marginal distributions $P_j(dx_j)$:

$$\phi_j \in H_j \subset L^2(\mathcal{X}_j, P_j) := \{\phi_j : E(\phi_j^2(X_j)) < \infty\}.$$

The role of the coefficient vector $\mathbf{a} = (a_1, \ldots, a_p)^T$ in LPCs is taken on by a vector of transformations:

$$\mathbf{\Phi} := (\phi_1, \ldots, \phi_p) \in \mathbf{H} := H_1 \times \cdots \times H_p. \tag{2.1}$$

Similarly, the role of the linear combination $\sum a_j X_j$ in LPCs is taken on by an additive function $\sum \phi_j(X_j)$. APCs contain LPCs as a special case when all $X_j$ are real-valued with unit variances and $H_j = \{\phi_j : \phi_j(x_j) = a_j x_j,\ a_j \in \mathbb{R}\}$. A smallest APC (associated with $\mathbf{H}$) is now defined as a solution to

$$\min_{\mathbf{\Phi} \in \mathbf{H}} \mathrm{Var}\,(\sum_{j=1}^{p} \phi_j(X_j)) \qquad \text{subject to} \qquad \sum_{j=1}^{p} \mathrm{Var}\,(\phi_j(X_j)) = 1. \tag{2.2}$$

When $\mathbf{H} = L^2(\mathcal{X}_1, P_1) \times \cdots \times L^2(\mathcal{X}_p, P_p)$, a solution to (2.2), if it exists, is said to be a ***population APC***. We will use population APCs as targets of estimation, and in this we differ, for example, from Fukumizu et al. (2007) who assume their targets of estimation to be in RKHSs. In the present work, the role of RKHS theory is to provide regularization devices for estimation, but the targets of estimation may fall outside and will be reached in the limit in the $L^2$ sense. RKHS theory appropriate for APCs is the subject of Sections 2.4−2.6.

### 2.2.2 A Note on the Role of Constants

A particular nuisance in the context of APCs is the non-identifiability of constants in additive functions $\sum \phi_j$. For example, $\tilde{\phi}_k = \phi_k + c$, $\tilde{\phi}_l = \phi_l - c$ for some $k \neq l$ (and $\tilde{\phi}_j = \phi_j$ else) result in the same additive function, $\sum \tilde{\phi}_j = \sum \phi_j$. Donnell et al. (1994) deal with this issue by taking $H_j$ to be closed subspaces of centered transformations, $H_j = L^2_c(\mathcal{X}_j, P_j) := \{\phi_j : E(\phi_j(X_j)) = 0,\ E(\phi_j^2(X_j)) < \infty\}$. This approach raises unnecessary questions because strictly speaking estimates $\hat{\phi}_j$ of the transformations $\phi_j$ cannot be centered at the population mean (which is not known) and hence cannot be in $H_j$. Yet it is obvious that this should a non-issue if viewed appropriately.

Our preferred solution is to consider $L^2(\mathcal{X}_j, P_j)$ as consisting of equivalence classes of functions where two elements are equivalent if they differ almost surely by a con-

stant. This may be expressed as $L^2(\mathcal{X}_j, P_j)/\mathbb{R}$, but for notational simplicity we continue writing $L^2(\mathcal{X}_j, P_j)$ with the understanding that its elements are intended modulo constants. It is then straightforward to check that $L^2(\mathcal{X}_j, P_j)$ is a Hilbert space wrt covariance as the inner product:

$$\langle \phi_j, \psi_j \rangle_{P_j} \;:=\; \mathrm{Cov}(\phi_j(X_j), \psi_j(X_j)), \tag{2.3}$$

where $\phi_j$ and $\psi_j$ are any functions in their respective equivalence classes modulo constants. Our framework therefore says that differences by constants are irrelevant and should be ignored. We will have to make sure that quantities of interest defined on $L^2(\mathcal{X}_j, P_j)$ are invariant under $\phi_j \mapsto \phi_j + c_j$.

### 2.2.3 Population APCs — Review

We adapt a few facts about population APCs from Donnell et al. (1994) which prefigure some of the steps that will be required for RKHS-based theory of APCs. The first fact is the reformulation of APCs in terms of function spaces and operators between them. The second fact is the existence of APC solutions under suitable assumptions, here chosen a little stronger than in Donnell et al. (1994), namely, the Hilbert-Schmidt property rather than compactness of operators. The operator representation was inspired by a natural power algorithm (Section 2.7) which in turn was inspired by the ACE algorithm of Breiman & Friedman (1985).

We first introduce the natural inner product and associated norm for $p$-tuples of functions, $\mathbf{\Phi}, \mathbf{\Psi} \in \mathbf{H}^* := L^2(\mathcal{X}_1, P_1) \times \cdots \times L^2(\mathcal{X}_p, P_p)$, turning $\mathbf{H}^*$ into a Hilbert space:

$$\langle \mathbf{\Phi}, \mathbf{\Psi} \rangle_P \;:=\; \sum_{j=1}^{p} \langle \phi_j, \psi_j \rangle_{P_j} \;=\; \sum_{j=1}^{p} \mathrm{Cov}(\phi_j(X_j), \psi_j(X_j)),$$

$$\|\mathbf{\Phi}\|_P^2 \;:=\; \sum_{j=1}^{p} \|\phi_j\|_{P_j}^2 \;=\; \sum_{j=1}^{p} \mathrm{Var}\,(\phi_j(X_j)).$$

The APC constraint can now be expressed by $\|\mathbf{\Phi}\|_P^2 = 1$. To do likewise for the APC criterion, we introduce operators to express

$$\mathrm{Var}\,(\textstyle\sum_j \phi_j(X_j)) \;=\; \sum_j \mathrm{Var}\,(\phi_j(X_j)) + \sum_{i,j} \mathrm{Cov}(\phi_i(X_i), \phi_j(X_j))$$

in terms of inner products $\langle \cdot, \cdot \rangle_{P_i}$. Let $\psi_i(X_i) = E(\phi_j(X_j)|X_i)$. We note that

$$\mathrm{Cov}(\phi_i(X_i), \phi_j(X_j)) \;=\; \mathrm{Cov}(\phi_i(X_i),\, E(\phi_j(X_j)|X_i)) \;=\; \langle \phi_i, \psi_i \rangle_{P_i}.$$

Thus the required operators are the conditional expectations between the $L_2$ spaces:

$$\mathbf{P}_{ij} : L^2(\mathcal{X}_j, P_j) \to L^2(\mathcal{X}_i, P_i), \quad \phi_j \mapsto \mathbf{P}_{ij}\phi_j \;=\; \psi_i.$$

These are also the orthogonal projections between the respective subspaces: $\mathbf{P}_{ij}\phi_j = \mathrm{argmin}_{f \in L^2(\mathcal{X}_i, P_i)} \mathrm{Var}\,(\phi_j(X_j) - f(X_i))$ (leaving constants undetermined; see Section 2.2.2.) Finally, we collect the operators $\mathbf{P}_{ij}$ in a matrix to act as an operator on $\mathbf{H}^*$: $\mathbf{P} = (\mathbf{P}_{ij})_{i,j}$, where the $i^{th}$ component mapping is given by

$$(\mathbf{P}\mathbf{\Phi})_i \;:=\; \textstyle\sum_j \mathbf{P}_{ij}\phi_j \in L^2(\mathcal{X}_i, P_i). \tag{2.4}$$

Thus the population APC problem can be stated as

$$\boxed{\min_{\mathbf{\Phi} \in \mathbf{H}^*} \langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Phi} \rangle_P \qquad \text{subject to} \qquad \|\mathbf{\Phi}\|_P^2 = 1.} \tag{2.5}$$

This statement is suggestive of power algorithms based on the operator matrix $\mathbf{P}$.

The existence of solutions to (2.5) can be granted under certain conditions. We are not striving for generality but for simplicity, hence we adopt the technically convenient condition that the conditional expectation operators $\mathbf{P}_{ij}$ $(i \neq j)$ have the Hilbert-Schmidt property. Assuming that the spaces $L^2(\mathcal{X}_i, P_i)$ and $L^2(\mathcal{X}_j, P_j)$ are separable and hence have countable orthonormal bases $(\phi_{ik})_k$ and $(\phi_{jl})_l$, the Hilbert-Schmidt property can be stated as the following requirement, which can be shown to be independent of the particular bases:

$$\|\mathbf{P}_{ij}\|^2_{\mathrm{HS}} \;:=\; \textstyle\sum_{k,l} \langle\, \phi_{ik},\, \mathbf{P}_{ij}\phi_{jl} \,\rangle^2_{P_i} \;<\; \infty.$$

Such *Hilbert-Schmidt operators* form a Hilbert space with $\|\cdot\|_{\mathrm{HS}}$ as the norm. For $\mathbf{P}_{ij}$ the property amounts to a condition on the covariance functional on $L^2(\mathcal{X}_i, P_i) \times L^2(\mathcal{X}_j, P_j)$:

$$\|\mathbf{P}_{ij}\|^2_{\mathrm{HS}} \;=\; \textstyle\sum_{k,l} \mathrm{Cov}(\, \phi_{ik}(X_i),\, \phi_{jl}(X_j)\,)^2 \;<\; \infty,$$

which is equivalent to the following condition on the joint distribution:

$$\iint \frac{p^2_{X_i,X_j}(x_i,x_j)}{p_{X_i}(x_i)\, p_{X_j}(x_j)}\, dx_i dx_j \;=\; E_{P_i \otimes P_j}\left( \frac{p^2_{X_i,X_j}(x_i,x_j)}{p^2_{X_i}(x_i)\, p^2_{X_j}(x_j)} \right) \;<\; \infty.$$

The Hilbert-Schmidt property limits the strength of the association between $X_i$ and $X_j$ by limiting how far the actual joint distribution $p_{X_i,X_j}(x_i,x_j)$ can be from independence, $p_{X_i}(x_i)\, p_{X_j}(x_j)$. It precludes, for example, $X_1 = \cdots = X_p$. See Buja (1990) for context.

To calculate the Hilbert-Schmidt norm for operator matrices such as $\mathbf{P}$, we embed the bases $(\phi_{jl})_l$ of $L^2(\mathcal{X}_j, P_j)$ in $\mathbf{H}^*$ through $\phi_{jl} \mapsto \mathbf{\Phi}_{j,l} = (0, \ldots, 0, \phi_{jl}, 0, \ldots, 0)'$, so $(\mathbf{\Phi}_{j,l})_{j,l}$ forms an orthonormal basis of $\mathbf{H}^*$. Now, the Hilbert-Schmidt norm of $\mathbf{P}$ is infinite because $\mathbf{P}_{jj} = \mathbf{Id}_{L^2(\mathcal{X}_j,P_j)}$, but $\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*}$ is Hilbert-Schmidt if all $\mathbf{P}_{ij}$ for $i \neq j$

are Hilbert-Schmidt:

$$\|\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*}\|_{\mathrm{HS}}^2 = \sum_{i\neq j;k,l} \langle \mathbf{\Phi}_{i,k}, \mathbf{P}\mathbf{\Phi}_{j,l}\rangle_P^2 \;=\; \sum_{i\neq j}\sum_{k,l} \langle \phi_{ik}, \mathbf{P}_{ij}\phi_{jl}\rangle_{P_i}^2$$
$$= \sum_{i\neq j} \|\mathbf{P}_{ij}\|_{\mathrm{HS}}^2 \;<\; \infty.$$

Because $\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*}$ is Hilbert-Schmidt and self-adjoint wrt $\langle\cdot,\cdot\rangle_P$ (Donnell et al. (1994), Lemma 4.1), it has an eigen expansion:

$$(\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*})\mathbf{\Phi} \;=\; \textstyle\sum_\nu \lambda'_\nu \langle \mathbf{\Phi}, \mathbf{\Phi}_\nu\rangle_P \, \mathbf{\Phi}_\nu, \qquad \sum_\nu \lambda'^2_\nu \;=\; \|\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*}\|_{\mathrm{HS}}^2 \;<\; \infty,$$

where $(\mathbf{\Phi}_\nu)_\nu$ form a complete orthonormal system of eigenvectors for $\mathbf{P} - \mathbf{Id}_{\mathbf{H}^*}$, and $(\lambda'_\nu)_\nu$ is the set of corresponding eigenvalues with 0 as the only possible accumulation point. This translates to an eigen expansion of $\mathbf{P}$:

$$\boxed{\lambda_\nu := \lambda'_\nu + 1 \quad \Rightarrow \quad \mathbf{P}\mathbf{\Phi} \;=\; \textstyle\sum_\nu \lambda_\nu \langle \mathbf{\Phi}, \mathbf{\Phi}_\nu\rangle_P \, \mathbf{\Phi}_\nu, \quad \sum_\nu (\lambda_\nu - 1)^2 < \infty.} \tag{2.6}$$

It can be shown that $0 \leq \lambda_\nu \leq p$ (Donnell et al., 1994). Since the only possible accumulation points of $\lambda_\nu$ is $+1$, we will use $+1$ as a natural dividing lines between small and large APCs. To relate the expansion (2.6) back to the population APC problem (2.5), form the inner product with $\mathbf{\Phi}$ assuming unit norm:

$$\|\mathbf{\Phi}\|_P^2 = 1 \quad \Rightarrow \quad \langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Phi}\rangle_P \;=\; \textstyle\sum_\nu \lambda_\nu \langle \mathbf{\Phi}, \mathbf{\Phi}_\nu\rangle_P^2 \quad \text{and} \quad \sum_\nu \langle \mathbf{\Phi}, \mathbf{\Phi}_\nu\rangle_P^2 \;=\; 1. \tag{2.7}$$

From (2.7) follows that the APC minimization problem (2.5) has the following solution:

$$\boxed{\min_{\|\mathbf{\Phi}\|_P^2 = 1} \langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Phi}\rangle_P \;=\; \min_\nu \lambda_\nu.} \tag{2.8}$$

Any eigenvector $\mathbf{\Phi}_\nu$ with minimizing eigenalue $\lambda_\nu$ is therefore a smallest population APC.

For an understanding of APCs, it is important to know the situation in which APCs are unable to discover association among variables. The following equivalent statements from Donnell et al. (1994), Proposition 4.8, characterize the "null situation" for APCs:

$$\min_\nu \lambda_\nu = 1 \quad \Leftrightarrow \quad \max_\nu \lambda_\nu = 1 \quad \Leftrightarrow \quad \lambda_\nu = 1 \ \forall \nu \quad \Leftrightarrow \quad \mathbf{P}_{ij} = \mathbf{0} \ \forall i \neq j$$

$$\Leftrightarrow \quad L^2(\mathcal{X}_i, P_i) \perp L^2(\mathcal{X}_j, P_j) \ \forall i \neq j \quad \Leftrightarrow \quad X_i, X_j \text{ independent } \forall i \neq j$$

Pairwise independence is not the same as full independence. Thus APCs can only find association that is detectable through pairwise association, which is natural because APCs rely on covariances $\mathrm{Cov}(\phi_i(X_i), \phi_j(X_j))$. This, however, should be a "limited limitation" as in practice multivariate associations are unlikely to hide behind pairwise independence.

## 2.3 Criterion and Constraint — A Null Comparison Principle

Donnell et al. (1994) chose the constraint $\sum \mathrm{Var}\,(\phi_j) = 1$ for APCs because it generalizes the contraint of LPCs. Generalization is a convenient justification but, as will be seen, it is insufficient to guide us in kernelizing APCs. Without a guiding principle, attempts at kernelizing multivariate methods end up relying on ad hoc proposals, some of which we discuss in Section 2.4.2. Even for LPCs we may ask: what is it that makes $\sum a_j^2 = 1$ "natural" as a constraint? When variables are heterogeneous with incompatible units, one tends to standardize the variables before using the constraint

$\sum a_j^2 = 1$. This, on the other hand, is equivalent to using $\sum a_j^2 \operatorname{Var}(X_j) = 1$ as the constraint on the unstandardized variables. Thus practioners have been aware of issues surrounding the constraint since the inception of LPCs. Constraints seem like separate choices, detached from the criteria. To show that this is not so and that there exists a tight coupling between criteria and constraints, we introduce the following:

> ***Null Comparison Principle for multivariate analysis:*** *The quadratic form to be used for the constraint is the optimization criterion evaluated under the null assumption of vanishing correlations of interest.*

Here are a number of illustrations of the principle, three for extant linear multivariate methods, and three for their additive analogs.

- For LPCs the criterion is $\operatorname{Var}(\sum a_j X_j)$, and the null assumption of interest is

$$\operatorname{Cov}(X_j, X_k) = 0 \quad \forall j \neq k.$$

  The evaluation of the criterion under the null assumption results in

$$\operatorname{Var}(\sum a_j X_j) = \sum \operatorname{Var}(a_j X_j) = \sum a_j^2 \operatorname{Var}(X_j),$$

  which evaluates to the familiar $\sum a_j^2$ if the variables are standardized.

- For Canonical Correlation Analysis (CCA), one divides the variables into two blocks, $X_1, \ldots, X_p$ and $Y_1, \ldots, Y_q$. The criterion is still the variance of a linear combination of all variables: $\operatorname{Var}(\sum a_i X_i + \sum b_j Y_j)$. The correlations of interest are only those between $X_i$ and $Y_j$ variables:

$$\operatorname{Cov}(X_i, Y_j) = 0 \quad \forall i, j.$$

Under this "null assumption" the criterion evaluates to $\text{Var}\,(\sum a_i X_i)+\text{Var}\,(\sum b_j Y_j)$. Thus the CCA problem is seen to be

$$\max_{a_i,b_j} \text{Var}\,(\sum a_i X_i + \sum b_j Y_j) \quad \text{subject to} \quad \text{Var}\,(\sum a_i X_i) + \text{Var}\,(\sum b_j Y_j) = 1,$$

which is algebraically equivalent to the more familiar form

$$\max_{a_i,b_j} \text{Cov}(\sum a_i X_i, \sum b_j Y_j) \quad \text{subject to} \quad \text{Var}\,(\sum a_i X_i) = \text{Var}\,(\sum b_j Y_j) = 1.$$

- Multi-block versions called "Generalized Canonical Analysis" (GCA) can be obtained by expanding from two to three or more blocks. Here is for three blocks of variables, $X_1, \ldots, X_p$, $Y_1, \ldots, Y_q$ and $Z_1, \ldots, Z_r$: The criterion is $\text{Var}\,(\sum a_i X_i + \sum b_j Y_j + \sum c_k Z_k)$, and the null assumtion is vanishing correlations between the blocks, that is,

$$\text{Cov}(X_i, Y_j) = \text{Cov}(X_i, Z_k) = \text{Cov}(Y_j, Z_k) = 0 \quad \forall i, j, k.$$

Under this null assumption the criterion evaluates in the familiar way, and the three-block GCA problem can be stated as

$$\max_{a_i,b_j,c_k} \text{Var}\,(\sum a_i X_i + \sum b_j Y_j + \sum c_k Z_k) \quad \text{subject to}$$
$$\text{Var}\,(\sum a_i X_i) + \text{Var}\,(\sum b_j Y_j) + \text{Var}\,(\sum c_k Z_k) = 1.$$

LPC is then GCA with $p$ blocks and every block containing only one variable.

- Turning from linear to additive methods, for APCs the criterion is $\text{Var}\,(\sum \phi_i(X_i))$,

and the null assumption of interest is

$$\mathrm{Cov}(\phi_i(X_i), \phi_j(X_j)) = 0 \quad \forall \phi_i, \phi_j,\ i \neq j. \tag{2.9}$$

The evaluation of the criterion results in $\mathrm{Var}\left(\sum \phi_j(X_j)\right) = \sum \mathrm{Var}\left(\phi_j(X_j)\right)$, hence the null comparison principle leads to the familiar form of the APC problem:

$$\min_{\phi_j} \mathrm{Var}\left(\sum \phi_j(X_j)\right) \quad \text{subject to} \quad \sum \mathrm{Var}\left(\phi_j(X_j)\right) = 1.$$

- To further illustrate the null comparison principle we show how additive CCA can be devised, without further pursuing it later on: Again, the variables are divided into two blocks as in linear CCA, but the criterion is $\mathrm{Var}\left(\sum \phi_i(X_i) + \sum \psi_j(Y_j)\right)$. The null assumption is

$$\mathrm{Cov}(\phi_i(X_i), \psi_j(Y_j)) = 0 \quad \forall \phi_i, \psi_j$$

The evaluation of the criterion under the null assumption leads to the following:

$$\max_{\phi_i, \psi_j} \mathrm{Var}\left(\sum \phi_i(X_i) + \sum \psi_j(Y_j)\right) \quad \text{subject to}$$
$$\mathrm{Var}\left(\sum \phi_i(X_i)\right) + \mathrm{Var}\left(\sum \psi_j(Y_j)\right) = 1.$$

When the $Y$-block contains just one variable, additive CCA amounts to the ACE method of Breiman & Friedman (1985).

- It is now obvious how a multi-block version of additive GCA can be devised, and we may simply skip to its final form for three blocks:

$$\max_{\phi_i, \psi_j, \xi_k} \mathrm{Var}\left(\sum \phi_i(X_i) + \sum \psi_j(Y_j) + \sum \xi_k(Z_k)\right) \quad \text{subject to}$$

$$\mathrm{Var}\left(\sum \phi_i(X_i)\right) + \mathrm{Var}\left(\sum \psi_j(Y_j)\right) + \mathrm{Var}\left(\sum \xi_k(Z_k)\right) = 1.$$

Again, APC amounts to additive GCA with $p$ blocks, each block with just one variable.

These examples illustrate how the null comparison principle ties the constraint to the criterion, thereby making it less an arbitrary choice. The choice is no longer that of a constraint but of a null assumption that identifies the correlations of interest and assumes them to vanish. The constraint is then derived by evaluating the criterion under the null assumption. We thus arrive at a powerful and principled way of devising generalizations of multivariate methods, a way whose real power will be revealed when we introduce penalized APCs.

## 2.4 Penalized APCs

In this section, we derive the penalized APC problem using the null comparison principle introduced previously. We also give a brief discussion on alternative approaches to penalizing APCs.

### 2.4.1 Introducing Penalties in APCs Using the Null Comparison Principle

Estimation of APCs from finite data requires some form of regularization. The estimation procedure of Donnell et al. (1994) can be characterized as using finite-dimensional subspaces $H_j$ (possibly adapted to the data, as for regression splines with knots placed at empirical quantiles) and replacing the population distribution with the empirical distribution of the data. Regularization necessary for estimation is achieved by choosing a suitably low dimensionality of the spaces $H_j$.

In this chapter we will consider APC estimation based on kernelizing whereby regularization is achieved through additive quadratic penalties $J_j(\phi_j)$ that are scaled versions of (squared) semi-norms derived from reproducing kernels. While estimation is again based on the empirical distribution, a regularized population version based on the actual distribution $P_{1:p}$ exists also and is useful for bias-variance calculations. For simplicity of notation we continue the discussion using the population case. The natural optimization criterion for kernel APCs is penalized variance:

$$\mathrm{Var}\,(\sum_{j=1}^{p}\phi_j) + \sum_{j=1}^{p} J_j(\phi_j). \tag{2.10}$$

This choice forces the transformations $\phi_j$ not only to generate small variance but also regularity in the sense of the penalties. A concrete example is the cubic spline penalty $J_j(\phi_j) = \alpha_j \int (\phi_j''(x_j))^2 dx_j$ for a quantitative variable $X_j$ (where we absorbed the tuning constant $\alpha_j$ in $J_j$), but the reader versed in kernelizing will recognize the generality of modeling offered by penalties derived from general reproducing kernels.

The question is next what the natural constraint should be. Informed by the null comparison principle of Section 2.3, we will not naively carry $\sum \mathrm{Var}\,(\phi_j) = 1$ over to the kernelized problem. Instead we evaluate the criterion (2.10) under the assumption of absent correlations between the transformations $\phi_j$, resulting in

$$\sum_{j=1}^{p} \mathrm{Var}\,(\phi_j) + \sum_{j=1}^{p} J_j(\phi_j) \;=\; 1. \tag{2.11}$$

As it turns out, this formulation produces meaningful results both for minimization *and* maximization. It therefore serves both for estimating smallest APCs, hence implicit additive equations (structure of *low co-dimension*), and for estimating largest APCs, hence additive dimension reduction (structure of *low dimension*). In the present chapter we pursue the former goal, but we take the well-posedness of both

the minimization and maximization problems as evidence that the approach based on the null comparison principle is sound. As is shown in Sections 2.4.2 and 2.11.1, some generalizations of PCA are not sound in this regard, one of them being kernel PCA.

On data we will replace the population quantities in equations (2.10) and (2.11) with their sample counterparts. As is usual, the penalties will be expressed in terms of quadratic forms of certain kernel matrices.

### 2.4.2 Alternative Approaches to Penalized APCs

A brief historic digression is useful to indicate the conceptual problem solved by the null comparison principle: As mentioned in the introduction, in the related but different field of functional multivariate analysis, Silverman co-authored two different approaches to the same PCA regularization problem where largest principal components are sought for dimension reduction. These can be transposed to the APC problems as follows:

$$\max_{\phi_j} \operatorname{Var}\left(\sum \phi_j\right) - \sum J_j(\phi_j) \qquad \text{subject to} \qquad \sum \operatorname{Var}\left(\phi_j\right) = 1, \tag{2.12}$$

$$\max_{\phi_j} \operatorname{Var}\left(\sum \phi_j\right) \qquad \text{subject to} \qquad \sum \operatorname{Var}\left(\phi_j\right) + \sum J_j(\phi_j) = 1, \tag{2.13}$$

where (2.12) is due to Rice & Silverman (1991) and (2.13) is due to Silverman (1996). The first approach (2.12) subtracts the penalty from the criterion, which does what it should do for regularized variance *maximization*. It is unsatisfactory for reasons of mathematical aesthetics: a difference of two quadratic forms can result in negative values, which may not be a practical problem but "does not seem right". The second approach (2.13) solves this issue by adding a penalty to the constraint rather than subtracting it from the criterion, which again does what it should do for variance

maximization. Both approaches can be criticized for resulting in non-sense when the goal is regularized variance *minimization*. Here the first approach (2.12) is more satisfying because it is immediately clear how to modify it to work for regularized variance minimization:

$$\min_{\phi_1,\ldots,\phi_p} \operatorname{Var}\left(\sum \phi_j\right) + \sum J_j(\phi_j) \qquad \text{subject to} \qquad \sum \operatorname{Var}\left(\phi_j\right) = 1,$$

whereas for the approach (2.13) it is not clear how it could be modified to work in this case. Subtracting the penalty from the constraint variance, $\sum \operatorname{Var}\left(\phi_j\right) - \sum J_j(\phi_j) = 1$, is clearly not going to work.

Eschewing these problems, we propose the following kernel APC problem:

$$\min_{\phi_1,\ldots,\phi_p} \operatorname{Var}\left(\sum \phi_j\right) + \sum J_j(\phi_j) \qquad \text{subject to} \qquad \sum \operatorname{Var}\left(\phi_j\right) + \sum J_j(\phi_j) = 1. \tag{2.14}$$

The merits of this proposal are that (1) it has no aesthetic issues, (2) it works for both ends of the variance spectrum, and (3) it derives from a more fundamental principle rather than a mathematical ad hoc choice.

## 2.5 Penalized APCs in Reproducing Kernel Hilbert Spaces

A preliminary note on vocabulary: Because there will be many occasions to use the clumsy term "*squared norm*", we will simplify by using sloppy language whereby the term "*norm*" stands for both "norm" and "squared norm" according to the context.

In this section, we introduce suitable RKHSs for APCs, one per variable. We then formalize the statement of the kernel APC problem (2.14) and establish the existence of solutions. The complications addressed in this section have to do with

the frequent occurrence of penalties that are not norms but semi-norms, such as the cubic spline penalty $J(\phi) = \alpha \int \phi''(x)^2 dx$. The relevant quadratic form needed for APC constraints, however, is $\mathrm{Var}\,(\phi) + J(\phi)$. In Section 2.5.1, we establish conditions under which $\mathrm{Var}\,(\phi) + J(\phi)$ is an RKHS norm. These conditions would be unnecessary if interest were limited to penalties that are actual RKHS norms, but the practical importance of penalties that are semi-norms mandates the mundane elaborations of the present section.

### 2.5.1 RKHS for APC Variables

Let $\mathcal{X}$ be a non-empty set, and let $\mathcal{H}$ be a Hilbert space of functions $\phi : \mathcal{X} \to \mathbb{R}$, endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The space $\mathcal{H}$ is a (real-valued) *reproducing kernel Hilbert space* if all evaluation functionals (the maps $\delta_x : f \mapsto f(x)$, where $x \in \mathcal{X}$) are bounded. Equivalently, $\mathcal{H}$ is an RKHS if there exists a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies (a) $\forall x \in \mathcal{X}$, $k_x = k(x, \cdot) \in \mathcal{H}$, (b) the reproducing property: $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, $\langle f, k_x \rangle_{\mathcal{H}} = f(x)$. Such a $k$ is called the *reproducing kernel* of $\mathcal{H}$. There is a one-to-one correspondence between an RKHS $\mathcal{H}$ and its reproducing kernel $k$. Thus, specifying $k$ is equivalent to specifying $\mathcal{H}$, and we may write $\langle \cdot, \cdot \rangle_k$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_k$ for $\| \cdot \|_{\mathcal{H}}$. Also, $\|k_x\|_k^2 = k(x, x)$.

In principle, regularization through kernelizing can be achieved by taking $J(\phi) = \alpha \|\phi\|_k^2$ after having specified a kernel (and hence, the corresponding RKHS $\mathcal{H}$). On the other hand, textbook examples of RKHS include those based on Sobolev type norms such as $\|\phi\|_k^2 = \phi(a)^2 + \phi'(a)^2 + \int \phi''(x)^2 dx$ ($a \in \mathbb{R}$ fixed). A pecularity here is that the finite-rank part of the norm, $\phi(a)^2 + \phi'(a)^2$, is arbitrary and not used for penalization; only the infinite-rank part is: $J(\phi) = \alpha \int \phi''(x)^2 dx$, which is the cubic spline penalty. Characteristically, this penalty alone is not an RKHS norm, only a semi-norm. To accommodate this situation we introduce the following definitions:

**Definitions:** *Let* $\|\phi\|_1^2 = J(\phi)$ *on* $\mathcal{H}$ *be a non-negative semi-definite quadratic form derived from a bilinear form* $\langle \cdot, \cdot \rangle_1$ *defined on a function space* $\mathcal{H}$*, and let*

$$\mathcal{H}^0 \ := \ \{\phi \in \mathcal{H} \,|\, \|\phi\|_1 = 0\}$$

*be its null space. We say* $\|\cdot\|_1^2$ *is a* ***kernel semi-norm*** *if there exists a complement* $\mathcal{H}^1 \subset \mathcal{H}$ *of* $\mathcal{H}^0$ *that is an RKHS with regard to the restriction of* $\|\cdot\|_1^2$ *to* $\mathcal{H}^1$*. We call* $\mathcal{H}^1$ ***an RKHS complement*** *for* $\|\cdot\|_1^2$*.*

We now combine the RKHS structure for regularization with the distributional structure of the data, focusing still on one variable. Consider a measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and a random variable $X$ with values in $\mathcal{X}$ and distribution $P(dx)$ on $\mathcal{X}$. We assume a space $\mathcal{H}$ of functions that are $\mathcal{B}_\mathcal{X}$-measurable, with a kernel semi-norm $\|\cdot\|_1$ with RKHS complement $\mathcal{H}^1 \subset \mathcal{H}$. The structure that expresses kernelized APCs is given by a combination of the $L^2(\mathcal{X}, P)$ inner product and the RKHS inner product based on the penalty kernel $k^1$:

$$\langle \phi, \psi \rangle_\alpha \ = \ \mathrm{Cov}(\phi, \psi) + \alpha \langle \phi, \psi \rangle_1, \qquad \|\phi\|_\alpha^2 \ = \ \mathrm{Var}\,(\phi) + \alpha \|\phi\|_1^2 \qquad (\alpha > 0), \quad (2.15)$$

where of course $\mathrm{Cov}(\phi, \psi)$ is understood to mean $\mathrm{Cov}(\phi(X), \psi(X))$. To avoid confusion in notation, we denote the alpha inner product and norm for the special case $\alpha = 1$ by $\langle \phi, \psi \rangle_\star$ and $\|\phi\|_\star^2$, respectively. For (2.15) to represent an RKHS we will make the following more restrictive assumptions, which, however, suffice to cover the case of Sobolev semi-norms:

**Lemma 1.** *On the linear space* $\mathcal{H} \subset L^2(\mathcal{X}, P_X)$*, let* $\|\phi\|_1^2$ *be a kernel semi-norm with null space* $\mathcal{H}^0$ *and RKHS complement* $\mathcal{H}^1$*. Suppose that* $\mathcal{H}^0$ *is finite-dimensional and the covariance matrix of a basis of* $\mathcal{H}^0$ *is of full-rank, so that* $\mathrm{Var}\,(\cdot)$ *turns* $\mathcal{H}^0$ *into an RKHS. Assume further that the reproducing kernel* $k^1$ *of* $\mathcal{H}^1$ *satisfies* $E(k^1(X, X)) <$

$\infty$. *Then the alpha inner products and norms of* (2.15) *turn* $\mathcal{H}$ *into an RKHS.*

A concern left over by the lemma is that the construction of the RKHS structure on $\mathcal{H}$ depends on the specific choice of the space $\mathcal{H}^1$. While the null space $\mathcal{H}^0$ is unique, the complement $\mathcal{H}^1$ is not, as evidenced again by the example of cubic splines: $\mathcal{H}^1$ can be defined by a host of different conditions, such as $\phi(a) = \phi'(a) = 0$ for an arbitrary location $a$, or $\phi(a) = \phi(b) = 0$ for two arbitrary locations $a < b$, or $\int_a^b \phi(x)dx = 0$ and $\phi(a) = \phi(b)$ again for two arbitrary locations $a < b$. Different choices of $a$ and $a < b$ result in different spaces $\mathcal{H}^1$. Now the question is how two different RKHS complements $\mathcal{H}^1$ and $\tilde{\mathcal{H}}^1$ for the same kernel semi-norm somehow affect the construction of $\|\phi\|_\alpha^2$. It is evident that the choice of $\mathcal{H}^1$ does not affect the construction as such:

**Lemma 2.** *Let* $\|\cdot\|_1$ *be a semi-norm, and let* $\mathcal{H}^1$ *and* $\tilde{\mathcal{H}}^1$ *be two complements of its null space* $\mathcal{H}^0$. *Then there exists an isometry between the two complements wrt* $\|\cdot\|_1$.

**Proof**: Because $\tilde{\mathcal{H}}^1$ is a complement of $\mathcal{H}^0$, there exists for any $\phi^1 \in \mathcal{H}^1$ unique $\phi^0 \in \mathcal{H}^0$ and $\tilde{\phi}^1 \in \tilde{\mathcal{H}}^1$ such that $\phi^1 = \tilde{\phi}^1 + \phi^0$. Then $\phi^1 \mapsto \tilde{\phi}^1$ defines a linear bijection $\mathcal{H}^1 \to \tilde{\mathcal{H}}^1$. It is an isometry, $\|\phi^1\|_1 = \|\tilde{\phi}^1\|_1$ because $\|\phi^1 - \tilde{\phi}^1\|_1 = \|\phi^0\|_1 = 0$. $\square$

The lemma is about two arbitrary algebraic complements without requiring them to be RKHS. This, however, is of little help for the issues on hand:

- The RKHS property of bounded evaluation functionals does not transfer from $\mathcal{H}^1$ to arbitrary algebraic complements $\tilde{\mathcal{H}}^1$.
- The property $E(k^1(X, X)) < \infty$ does not transfer to arbitrary choices of algebraic complements $\tilde{\mathcal{H}}^1$ either.

Both points can be understood by analyzing the proof of Lemma 2: In $\tilde{\phi}^1 = \phi^1 - \phi^0$ the term $\phi^0$ prohibits us from controlling evaluations $\tilde{\phi}^1(x)$ as well as $\mathrm{Var}\,(\tilde{\phi}^1)$ without further assumptions.

Some intuitions can be gained by working through the example of cubic splines: If the original constraint to form $\mathcal{H}^1$ was $\phi(a)=\phi'(a)=0$ and the new constraint to form $\tilde{\mathcal{H}}^1$ is $\phi(a)=\phi(b)=0$ $(a\neq b)$, then there is a simple mapping $\phi^1 \mapsto \tilde{\phi}^1 = \phi^1 - \phi^0$, where in this instance $\phi^0(x) = a\phi^1(b)/(a-b) - (\phi^1(b)/(a-b))x$, which is a linear function and hence an element of $\mathcal{H}^0$. Thus the change of space from $\mathcal{H}^1$ to $\tilde{\mathcal{H}}^1$ is obtained through a mapping $T_0 : \mathcal{H}^1 \to \mathcal{H}^0$, $\phi^1 \mapsto \phi^0$ that produces the new subspace $\tilde{\mathcal{H}}^1 = \{\phi^1 - T_0(\phi^1) : \phi^1 \in \mathcal{H}^1\}$. Important in the cubic spline example is that the linear forms $\phi^1 \mapsto \phi^1(a)$ and $\phi^1 \mapsto \phi^1(b)$ are both continuous. This observation provides the critical condition for forming alternative RKHS complements:

**Lemma 3.** *Under the same assumptions as in Lemma 1, let $T_0 : \mathcal{H}^1 \to \mathcal{H}^0$ be a linear map that is bounded with regard to the norms $\|\cdot\|_1^2$ on $\mathcal{H}^1$ and $\mathrm{Var}\,(\cdot)$ on $\mathcal{H}^0$. Then the space $\tilde{\mathcal{H}}^1 = \{\phi^1 - T_0(\phi^1) : \phi^1 \in \mathcal{H}^1\}$ is an RKHS under $\|\cdot\|_1^2$ that is isometric to $\mathcal{H}^1$ and its reproducing kernel $\tilde{k}^1$ satisfies $E(\tilde{k}^1(X,X)) < \infty$.*

The proof is in Appendix A.1. The next lemma shows that there always exists a canonical orthogonal RKHS complement of the null space $\mathcal{H}^0$:

**Lemma 4.** *Under the same assumptions as in Lemma 1, the orthogonal complement of $\mathcal{H}^0$ wrt $\langle\cdot,\cdot\rangle_\alpha$ is an RKHS complement. This complement is also the orthogonal complement of $\mathcal{H}^0$ wrt $\mathrm{Cov}(\cdot,\cdot)$ and hence independent of any $\alpha > 0$.*

The usefulness of Lemma 4 is that orthogonal decompositions $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$ of an RKHS allow additive decompositions of kernels: $k = k^0 + k^1$. This is worth a definition:

**Definition:** *Under the assumptions of Lemma 1 we call the orthogonal RKHS complement $\tilde{\mathcal{H}}^1$ of Lemma 4 the* ***canonical complement*** *for $\|\cdot\|_1^2$ wrt $\mathrm{Cov}(\cdot,\cdot)$.*

Finally, we have the following fact which is useful for establishing the main results in Sections 2.6 and 2.7.

**Lemma 5.** *Suppose that the conditions in Lemma 1 hold. Then the reproducing kernel $k$ of $(\mathcal{H}, \langle\cdot,\cdot\rangle_\alpha)$ satisfies $E(k(X,X)) < \infty$.*

**Remark 1.** Using covariances we made implicit use of the convention that all functions are really equivalence classes of functions modulo constants. This applies to Sobolev-type RKHS for which constants are in the null space $\mathcal{H}^0$. RKHS based on Gaussian kernels do not contain non-zero constants in the first place (Steinwart & Christmann, 2008). In order to make $\mathcal{H}^0$ an RKHS with variance as a kernel norm, one has to select a subspace of co-dimension 1 under a restriction such as $\phi(a) = 0$ or $E(\phi(X)) = 0$ in order to remove dependence on irrelevant constants. Lemma 3 can be leveraged to imply that if a change of restriction stems from a continuous mapping $T_0 : \mathcal{H}^1 \mapsto \mathcal{H}^0$ with regard to the alpha norm (2.15), then the RKHS structure is not affected.

### 2.5.2 Penalized APCs based on RKHS

The definition of penalized/kernelized APCs requires a product structure for tuples of functions $\phi_j$ in spaces that follow the framework of the preceding subsection.

Let $(\mathcal{X}_1, \mathcal{B}_{\mathcal{X}_1})$, ..., $(\mathcal{X}_p, \mathcal{B}_{\mathcal{X}_p})$ be measurable spaces, and consider the random vector $(X_1, \ldots, X_p) : \Omega \to \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ with joint distribution $P = P_{1:p}(dx_1, \ldots, dx_p)$. For $1 \leq j \leq p$, the marginal distribution of $X_j$ is denoted by $P_j(dx_j)$. Associated are the space $L^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_p, P)$ of functions $\phi(x_1, \ldots, x_p)$ and the spaces $L^2(\mathcal{X}_j, P_j)$ of functions $\phi_j(x_j)$. The former contains, but is not limited to, additive functions:

$$\sum \phi_j \;\in\; L^2(\mathcal{X}_1, P_1) + \cdots + L^2(\mathcal{X}_p, P_p) \;\subset\; L^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_p, P).$$

Assume spaces $\mathcal{H}_1, \ldots, \mathcal{H}_p$ which are RKHSs under respective alpha norms $\|\phi_j\|^2_{\alpha_j,j} = \mathrm{Var}\,(\phi_j) + \alpha_j \|\phi_j\|^2_{1,j}$, where $\alpha_j > 0$ and $\|\phi_j\|^2_{1,j}$ is a kernel semi-norm on $\mathcal{H}_j = \mathcal{H}^0_j + \mathcal{H}^1_j$

with finite-dimensional null space $\mathcal{H}_j^0$ and RKHS complement $\mathcal{H}_j^1$. Further, bases of the null spaces $\mathcal{H}_j^0$ have full-rank covariance matrices and the reproducing kernel $k_j^1$ of $\mathcal{H}_j^1$ satisfies $E(k_j^1(X_j, X_j)) < \infty$.

For any $\alpha > 0$, we have $\alpha_j \|\phi_j\|_{1,j}^2 = \alpha \|\phi_j\|_{1',j}^2$, where $\|\phi_j\|_{1',j}^2 = \frac{\alpha_j}{\alpha} \|\phi_j\|_{1,j}^2$ induces on $\mathcal{H}_j$ a topology that is equivalent to that induced by $\|\phi_j\|_{1,j}^2$. Without loss of generality, we will set $\alpha_j$ at a common level $\alpha$ in the remainder of this section and in Sections 2.6 and 2.7. The search space for kernel APCs is now the product of the spaces $\mathcal{H}_j$:

$$\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_p) \in \boldsymbol{\mathcal{H}} \;:=\; \mathcal{H}_1 \times \cdots \times \mathcal{H}_p. \tag{2.16}$$

Following Section 2.4, the population kernel APC problem of (2.10) and (2.11) can be stated in the RKHS framework as follows:

$$\boxed{\begin{array}{ll} \min_{\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}} & \operatorname{Var}\left(\sum_{j=1}^p \phi_j\right) \;+\; \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2 \\ \text{subject to} & \sum_{j=1}^p \operatorname{Var}(\phi_j) \;+\; \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2 \;=\; 1. \end{array}} \tag{2.17}$$

A solution to (2.17), if it exists, is said to be a ***population kernel APC***. For a discussion on the existence of population kernel APCs, see Section 2.5.5.

To obtain the second-smallest as well as higher-order smallest kernel APCs, we require an orthogonality constraint and hence an inner product on the space $\boldsymbol{\mathcal{H}}$. A natural inner product and squared norm derives from the product structure of $\boldsymbol{\mathcal{H}}$:

$$\langle \boldsymbol{\Phi}, \boldsymbol{\Psi} \rangle_\alpha \;:=\; \sum_j \langle \phi_j, \psi_j \rangle_{\alpha,j}\,, \qquad \|\boldsymbol{\Phi}\|_\alpha^2 \;:=\; \sum_j \|\phi_j\|_{\alpha,j}^2. \tag{2.18}$$

Observe that the constraint in (2.17) can be expressed as $\|\boldsymbol{\Phi}\|_\alpha^2 = 1$, hence the natural inner product is given by (2.18). Therefore, in order to recursively define the $l$'th smallest penalized APC, assume that $\boldsymbol{\Phi}_\ell = (\phi_{\ell,1}, \ldots, \phi_{\ell,p})$ encompass all the

previous kernel APCs obtained so far ($\ell = 1, \ldots, l-1$); then solve (2.17) subject to the additional orthogonality constraint $\langle \mathbf{\Phi}, \mathbf{\Phi}_\ell \rangle_\alpha = 0$:

$$\boxed{\sum_{j=1}^{p} \mathrm{Cov}(\phi_{\ell,j}, \phi_j) + \alpha \sum_{j=1}^{p} \langle \phi_{\ell,j}, \phi_j \rangle_{1,j} = 0.} \tag{2.19}$$

for all $\ell = 1, \ldots, l-1$.

Similar as before, we denote the inner products and norms in (2.18) for the special case $\alpha = 1$ by $\langle \phi_j, \psi_j \rangle_{\star,j}$, $\langle \mathbf{\Phi}, \mathbf{\Psi} \rangle_\star$ and $\|\phi_j\|^2_{\star,j}$, $\|\mathbf{\Phi}\|^2_\star$, respectively. We will use $\langle \mathbf{\Phi}, \mathbf{\Psi} \rangle_\star$ as the reference inner product when restating the kernel APC problem (2.17) in quadratic forms, as will be detailed out in Section 2.5.5.

### 2.5.3 A Subspace Interpretation of Penalized APCs

The spaces $\mathcal{H}_j$ can be canonically embedded in $\boldsymbol{\mathcal{H}}$ by

$$\begin{aligned} &\phi_j \mapsto \mathbf{\Phi}_j = (0, \ldots, 0, \phi_j, 0, \ldots, 0) \qquad (\phi_j \in \mathcal{H}_j \text{ in the } j'\text{th position}), \\ &\boldsymbol{\mathcal{H}}_j \;=\; \{\mathbf{\Phi}_j \,|\, \phi_j \in \mathcal{H}_j\}. \end{aligned}$$

The spaces $\boldsymbol{\mathcal{H}}_j$ are mutually orthogonal with regard to the inner product (2.18). If we abbreviate the penalized APC criterion as

$$\mathbf{Q}(\mathbf{\Phi}) \;:=\; \mathrm{Var}\,(\sum_{j=1}^{p} \phi_j) + \alpha \sum_{j=1}^{p} \|\phi_j\|^2_{1,j},$$

then the squared norm $\|\cdot\|^2_\alpha$ can be written as

$$\|\mathbf{\Phi}\|^2_\alpha \;=\; \sum_j \mathbf{Q}(\mathbf{\Phi}_j).$$

The penalized APC problem becomes a subspace APC problem as follows:

$$\min_{\mathbf{\Phi}\in\boldsymbol{\mathcal{H}}} \mathbf{Q}(\mathbf{\Phi}) \qquad \text{subject to} \qquad \sum_j \mathbf{Q}(\mathbf{\Phi}_j) = 1,\ \sum_j \mathbf{\Phi}_j = \mathbf{\Phi},\ \mathbf{\Phi}_j \in \boldsymbol{\mathcal{H}}_j.$$

This is a generalizable geometric version of the null comparison principle: Given a non-negative definite quadratic form of interest, $\mathbf{Q}(\mathbf{\Phi})$, defined on a space decomposable into subspaces of interest, $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}_1 + \cdots + \boldsymbol{\mathcal{H}}_p$ ($\boldsymbol{\mathcal{H}}_j \cap \boldsymbol{\mathcal{H}}_k = \mathbf{0},\ \forall j \neq k$), we ask how orthogonal the subspaces $\boldsymbol{\mathcal{H}}_j$ are as measured by the quadratic form $\mathbf{Q}(\mathbf{\Psi})$. If they were mutually orthogonal, there would hold the Pythagorean identity, $\mathbf{Q}(\mathbf{\Psi}) \equiv \sum_j \mathbf{Q}(\mathbf{\Psi}_j)$. The subspace APC problem finds the directions $\mathbf{\Phi}$ of strongest deviation from hypothetical orthogonality.

If $\mathbf{B}(\mathbf{\Phi}, \mathbf{\Psi}) := \frac{1}{2}[\mathbf{Q}(\mathbf{\Phi}+\mathbf{\Psi}) - \mathbf{Q}(\mathbf{\Phi}) - \mathbf{Q}(\mathbf{\Psi})]$ is the bilinear form induced by $\mathbf{Q}(\cdot)$, we have the decomposition $\mathbf{Q}(\mathbf{\Phi}) = \sum_j \mathbf{Q}(\mathbf{\Phi}_j) + 2\sum_{i<j} \mathbf{B}(\mathbf{\Phi}_i, \mathbf{\Phi}_j)$, and orthogonality as implied by the null assumption is equivalent to $\mathbf{B}(\mathbf{\Phi}_i, \mathbf{\Phi}_j) = 0$ for all $\mathbf{\Phi}_i \in \boldsymbol{\mathcal{H}}_i$, $\mathbf{\Phi}_j \in \boldsymbol{\mathcal{H}}_j$ and $i \neq j$.

For the penalized APC criterion the Pythagorean identity $\mathbf{Q}(\mathbf{\Phi}) \equiv \sum_j \mathbf{Q}(\mathbf{\Phi}_j)$ holds iff all $\phi_j(X_j)$ and $\phi_k(X_k)$ are uncorrelated for $j \neq k$, which is the null assumption of the null comparison principle of Section 2.3.

### 2.5.4 Estimation of Kernel APCs

For estimation we assume that data are given as i.i.d. random vectors $\{(X_{\ell 1}, \ldots, X_{\ell p}) : 1 \leq \ell \leq n\}$ drawn from $P = P_{1:p}$. The role of the data is to allow empirical estimation of the variance of additive functions, $\widehat{\mathrm{Var}}\,(\sum_{j=1}^p \phi_j)$ and transformations $\widehat{\mathrm{Var}}\,(\phi_j)$. Estimation of kernel APCs is therefore by plug-in in the population kernel

APC problem (2.17):

$$\boxed{\begin{array}{ll} \min_{\boldsymbol{\Phi}\in\boldsymbol{\mathcal{H}}} & \widehat{\mathrm{Var}}\left(\sum_{j=1}^{p}\phi_j\right) \;+\; \alpha\sum_{j=1}^{p}\|\phi_j\|_{1,j}^2 \\ \text{subject to} & \sum_{j=1}^{p}\widehat{\mathrm{Var}}\left(\phi_j\right) \;+\; \alpha\sum_{j=1}^{p}\|\phi_j\|_{1,j}^2 \;=\; 1. \end{array}} \tag{2.20}$$

Similarly, higher-order smallest kernel APCs are obtained by solving (2.20) subject to a plug-in orthogonality constraint:

$$\boxed{\sum_{j=1}^{p}\widehat{\mathrm{Cov}}\left(\hat{\phi}_{\ell,j},\phi_j\right)+\alpha\sum_{j=1}^{p}\langle\hat{\phi}_{\ell,j},\phi_j\rangle_{1,j}=0.} \tag{2.21}$$

A solution to (2.20), if it exists, is said to be a ***sample kernel APC***. The existence and consistency of sample kernel APCs will be discussed in Sections 2.5.5 and 2.6, respectively. Details on computing these estimators will be given in Section 2.7.

### 2.5.5 Existence of Kernelized APCs

In this section, we establish the existence of solutions to the population kernel APC problem (2.17) and the sample kernel APC problem (2.20). For this, we need a reference RKHS inner product for our kernel APC search space $\boldsymbol{\mathcal{H}}$. Section 2.5.2 introduces a family of RKHS inner products $\{\langle\cdot,\cdot\rangle_\alpha : \alpha>0\}$ for $\boldsymbol{\mathcal{H}}$. In the following, we take $\langle\cdot,\cdot\rangle_\star$, the inner product corresponding to $\alpha=1$, as our reference inner product.

By Lemma 5, the reproducing kernel $k_j$ of $(\mathcal{H}_j,\langle\cdot,\cdot\rangle_{\star,j})$ satisfies $E(k_j(X_j,X_j))<\infty$. Under such a condition, the RKHS $\mathcal{H}_j$ is continuously embedded in $L^2(\mathcal{X}_j,P_j)$. Note, however, that $\mathcal{H}_j$ is generally not a closed subspace of $L^2(\mathcal{X}_j,P_j)$ and hence not a Hilbert space with regard to the inner product $\langle\cdot,\cdot\rangle_{P_j}$. Following Fukumizu et al.

(2007), under the condition $E(k_j(X_j, X_j)) < \infty$, we can define the mean element $\mu_j \in \mathcal{H}_j$ with respect to a random variable $X_j$ as

$$\langle \phi_j, \mu_j \rangle_{\star,j} = E(\langle \phi_j, k_{X_j} \rangle_{\star,j}) = E(\phi_j(X_j)) \qquad \forall \phi_j \in \mathcal{H}_j. \tag{2.22}$$

On the other hand, we define the cross-covariance operator of $(X_i, X_j)$ as a bounded linear operator from $\mathcal{H}_j$ to $\mathcal{H}_i$ given by

$$\begin{aligned} \langle \phi_i, \mathbf{C}_{ij} \phi_j \rangle_{\star,i} &= E(\langle \phi_i, k_{X_i} - \mu_i \rangle_{\star,i} \langle \phi_j, k_{X_j} - \mu_j \rangle_{\star,j}) \\ &= \mathrm{Cov}\left(\phi_i(X_i), \phi_j(X_j)\right) \qquad \forall \phi_i \in \mathcal{H}_i, \phi_j \in \mathcal{H}_j. \end{aligned}$$

The existence and uniqueness of both $\mu_j$ and $\mathbf{C}_{ij}$ are proved by the Riesz Representation Theorem. It is immediate that $\mathbf{C}_{ij} = \mathbf{C}_{ji}^*$, and it can be verified that $\mathbf{C}_{ij}$ is Hilbert-Schmidt (Fukumizu et al., 2007). When $i = j$, the positive, self-adjoint operator $\mathbf{C}_{jj}$ is called the covariance operator.

Let $\{(X_{\ell 1}, \ldots, X_{\ell p}) : 1 \leq \ell \leq n\}$ be i.i.d. random vectors on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ with joint distribution $P_{1:p}(dx_1, \ldots, dx_p)$. The empirical cross-covariance operator $\hat{\mathbf{C}}_{ij}^{(n)}$ is defined as the cross-covariance operator wrt the empirical distribution $\frac{1}{n}\sum_{\ell=1}^n \delta_{X_{\ell i}} \delta_{X_{\ell j}}$, in which case

$$\begin{aligned} \langle \phi_i, \hat{\mathbf{C}}_{ij}^{(n)} \phi_j \rangle_{\star,i} &= \frac{1}{n}\sum_{\ell=1}^n \left\langle \phi_i, k_{X_{\ell i}} - \frac{1}{n}\sum_{a=1}^n k_{X_{ai}} \right\rangle_{\star,i} \left\langle \phi_j, k_{X_{\ell j}} - \frac{1}{n}\sum_{b=1}^n k_{X_{bj}} \right\rangle_{\star,j} \\ &= \widehat{\mathrm{Cov}}(\phi_i, \phi_j), \qquad \forall \phi_i \in \mathcal{H}_i, \phi_j \in \mathcal{H}_j. \end{aligned}$$

Since $\mathcal{R}(\hat{\mathbf{C}}_{ij}^{(n)})$ and $\mathcal{N}(\hat{\mathbf{C}}_{ij}^{(n)})^\perp$ are included in $\mathrm{span}\{k_{X_{\ell i}} - \frac{1}{n}\sum_{a=1}^n k_{X_{ai}} : 1 \leq \ell \leq n\}$ and $\mathrm{span}\{k_{X_{\ell j}} - \frac{1}{n}\sum_{b=1}^n k_{X_{bj}} : 1 \leq \ell \leq n\}$, respectively, $\hat{\mathbf{C}}_{ij}^{(n)}$ is of finite rank.

It is known (Baker, 1973, Theorem 1) that $\mathbf{C}_{ij}$ has a representation

$$\mathbf{C}_{ij} = \mathbf{C}_{ii}^{1/2} \mathbf{V}_{ij} \mathbf{C}_{jj}^{1/2}, \tag{2.23}$$

where $\mathbf{V}_{ij} : \mathcal{H}_j \to \mathcal{H}_i$ is a unique bounded linear operator with $\|\mathbf{V}_{ij}\| \leq 1$.

In what follows, we establish the existence of solutions to the population kernel APC problem (2.17) and the sample kernel APC problem (2.20). As a first step, we rewrite (2.17) and (2.20) in terms of quadratic forms with respect to the RKHS inner product $\langle \cdot, \cdot \rangle_\star$. To this end, using the cross-covariance operators introduced previously and setting $\alpha = \alpha_n$ that depends on the sample size $n$, we can rewrite (2.17) and (2.20) as follows:

$$\min_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} \langle \mathbf{\Phi}, (\mathbf{C} + \mathbf{J}^{(n)}) \mathbf{\Phi} \rangle_\star \quad \text{subject to} \quad \langle \mathbf{\Phi}, (\operatorname{diag}(\mathbf{C}) + \mathbf{J}^{(n)}) \mathbf{\Phi} \rangle_\star = 1, \tag{2.24a}$$

$$\min_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} \langle \mathbf{\Phi}, (\hat{\mathbf{C}}^{(n)} + \mathbf{J}^{(n)}) \mathbf{\Phi} \rangle_\star \quad \text{subject to} \quad \langle \mathbf{\Phi}, (\operatorname{diag}(\hat{\mathbf{C}}^{(n)}) + \mathbf{J}^{(n)}) \mathbf{\Phi} \rangle_\star = 1, \tag{2.24b}$$

where

$$\mathbf{C} = (\mathbf{C}_{ij})_{i,j}, \quad \hat{\mathbf{C}}^{(n)} = (\hat{\mathbf{C}}_{ij}^{(n)})_{i,j} \quad \text{and} \quad \mathbf{J}^{(n)} = \operatorname{diag}(\alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))_j.$$

We denote the solutions to (2.24a) and (2.24b), when they exist, as $\tilde{\mathbf{\Phi}}^{(n)}$ and $\hat{\mathbf{\Phi}}^{(n)}$, respectively.

Consider the following changes of variables

$$f_j = (\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{1/2} \phi_j, \tag{2.25a}$$

$$f_j = (\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{1/2} \phi_j, \tag{2.25b}$$

for $1 \leq j \leq p$ in (2.24a) – (2.24b), respectively. Then (2.24a) – (2.24b) can be further

rewritten as

$$\min_{\mathbf{f}\in\boldsymbol{\mathcal{H}}} \langle \mathbf{f}, \tilde{\mathbf{V}}^{(n)}\mathbf{f}\rangle_\star \qquad \text{subject to} \qquad \langle \mathbf{f}, \mathbf{f}\rangle_\star = 1, \tag{2.26a}$$

$$\min_{\mathbf{f}\in\boldsymbol{\mathcal{H}}} \langle \mathbf{f}, \hat{\mathbf{V}}^{(n)}\mathbf{f}\rangle_\star \qquad \text{subject to} \qquad \langle \mathbf{f}, \mathbf{f}\rangle_\star = 1, \tag{2.26b}$$

respectively, with

$$\tilde{\mathbf{V}}^{(n)} = (\tilde{\mathbf{V}}_{ij}^{(n)})_{i,j}, \quad \hat{\mathbf{V}}^{(n)} = (\hat{\mathbf{V}}_{ij}^{(n)})_{i,j}, \tag{2.27}$$

where $\tilde{\mathbf{V}}_{jj}^{(n)} = \hat{\mathbf{V}}_{jj}^{(n)} = \mathbf{Id}_{\mathcal{H}_j}$ for any $j = 1, \ldots, p$ and

$$\tilde{\mathbf{V}}_{ij}^{(n)} = (\mathbf{C}_{ii} + \alpha_n(\mathbf{Id}_{\mathcal{H}_i} - \mathbf{C}_{ii}))^{-1/2}\mathbf{C}_{ij}(\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{-1/2}, \tag{2.28a}$$

$$\hat{\mathbf{V}}_{ij}^{(n)} = (\hat{\mathbf{C}}_{ii}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_i} - \mathbf{C}_{ii}))^{-1/2}\hat{\mathbf{C}}_{ij}^{(n)}(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{-1/2}, \tag{2.28b}$$

for $1 \leq i, j \leq p$, $i \neq j$. We need to ensure the operators $(\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{-1/2}$ and $(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{-1/2}$ in (2.28a) and (2.28b) are well-defined with high probability. This is guaranteed by the following lemma, the proof of which is given in Appendix A.2.1.

**Lemma 6.** *Suppose that $\alpha_n \to 0$. Then*

$$\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \succeq \alpha_n \mathbf{Id}_{\mathcal{H}_j}, \qquad \textit{for } 1 \leq j \leq p, \tag{2.29}$$

*for sufficiently large values of $n$. Moreover, with probability at least $1 - d\alpha_n^{-1}n^{-1/2}$,*

$$\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \succeq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j}, \qquad \textit{for } 1 \leq j \leq p,$$

*where $d$ is a constant not depending on $n$.*

Next, we show that the solutions to (2.26a)−(2.26b) exist. To this end, we are to show that the operators $\tilde{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ and $\hat{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ are both compact with high probability. Using the fact that the product of a bounded linear operator and a compact operator is compact, and that both $(\mathbf{C}_{ii} + \alpha_n(\mathbf{Id}_{\mathcal{H}_i} - \mathbf{C}_{ii}))^{-1/2}$ and $(\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))^{-1/2}$ are bounded and $\mathbf{C}_{ij}$ is compact, we see that $\tilde{\mathbf{V}}_{ij}^{(n)}$ is also compact. Moreover, on the event that it is well-defined, $\hat{\mathbf{V}}_{ij}^{(n)}$ is compact since it is of finite-rank. In summary, we have the following result.

**Corollary 1.** *On the event that the conclusions of Lemma 6 hold, the operators* $\tilde{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ *and* $\hat{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ *are well-defined and compact.*

Note that compactness implies the spectra of $\tilde{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ and $\hat{\mathbf{V}}^{(n)} - \mathbf{Id}_{\boldsymbol{\mathcal{H}}}$ are countable with 0 as the only possible accumulation point. Consequently, the spectra of $\tilde{\mathbf{V}}^{(n)}$ and $\hat{\mathbf{V}}^{(n)}$ are countable with $+1$ as the only possible accumulation point. It follows that the solutions to (2.26a)−(2.26b) can be obtained as the eigenvectors $\tilde{\mathbf{f}}^{(n)}$ and $\hat{\mathbf{f}}^{(n)}$ corresponding to the smallest eigenvalues of $\tilde{\mathbf{V}}^{(n)}$ and $\hat{\mathbf{V}}^{(n)}$, respectively. We can then obtain the population kernel APC $\tilde{\boldsymbol{\Phi}}^{(n)}$ and the sample kernel APC $\hat{\boldsymbol{\Phi}}^{(n)}$ by inverse transforming $\tilde{\mathbf{f}}^{(n)}$ and $\hat{\mathbf{f}}^{(n)}$ following (2.25a)−(2.25b).

In summary, we rewrite the kernel APC problems as in (2.24a)−(2.24b) and (2.26a)−(2.26b), and we know that under the assumptions introduced in Section 2.5.2 on $\boldsymbol{\mathcal{H}}$, the solutions to (2.24a)−(2.24b) and (2.26a)−(2.26b) exist (with high probability for $n$ sufficiently large).

**Remark 2.** It may seem natural to rewrite the population kernel APC problem (2.17) in terms of quadratic forms wrt $\langle \cdot, \cdot \rangle_\alpha$, so that we obtain an eigenproblem (rather than a generalized eigenproblem as in (2.24a)). Indeed, the resulting expression is what motivates the power algorithm presented in Section 2.7 for computation of kernel APCs. Unfortunately, this approach does not extend nicely to the sample kernel APC problem (2.20), as the definition of $\langle \cdot, \cdot \rangle_\alpha$ involves probability measures. Rewriting the

kernel APC problems in quadratic forms wrt $\langle \cdot, \cdot \rangle_\star$ is not only useful for establishing existence of kernel APC solutions, but also essential for establishing consistency of sample kernel APCs in Section 2.6.

## 2.6 Consistency

In this section, we establish the existence and uniqueness of the population APC (denoted by $\mathbf{\Phi}^*$ hereinafter) and also the consistency of a sample kernel APC $\hat{\mathbf{\Phi}}^{(n)}$ as an estimator of $\mathbf{\Phi}^*$ under mild conditions.

### 2.6.1 Main Assumptions

The following conditions guarantee the existence and uniqueness [2] of the population APC.

**Assumption 1.** *Let* $(\mathcal{X}_1, \mathcal{B}_{\mathcal{X}_1})$, ..., $(\mathcal{X}_p, \mathcal{B}_{\mathcal{X}_p})$ *be measurable spaces, and consider the random vector* $(X_1, \ldots, X_p) : \Omega \to \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ *with joint distribution* $P = P_{1:p}(dx_1, \ldots, dx_p)$. *Assume that*

*(a) the conditional expectation operators* $\mathbf{P}_{ij}$ *are Hilbert-Schmidt for all* $i \neq j$;

*(b)* $\mathbf{P}_{ij} \neq \mathbf{0}$ *for some* $i \neq j$;

*(c) the smallest eigenvalue of the operator* $\mathbf{P}$, $\lambda_1$, *is simple.*

Based on the discussion in Section 2.2.3, Assumption 1(a) ensures the existence of population APCs, whereas Assumption 1(b) rules out the uninteresting case where $X_1, \ldots, X_p$ are pairwise independent and there exists no non-trivial additive relationship among them. Moreover, under Assumption 1(b), $\lambda_1 < 1$, so $\lambda_1$ is an isolated

[2] Throughout this chapter, uniqueness of any eigenvector means uniqueness up to a sign change and the equivalence relation in the norm $\|\cdot\|_P$.

eigenvalue. Assumption 1(c) ensures that $\mathbf{\Phi}^*$ is uniquely defined as the eigenvector of $\mathbf{P}$ corresponding to $\lambda_1$.

We impose the following assumptions on the kernel APC search space $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$:

**Assumption 2.** *For $j = 1, \ldots, p$, let $\mathcal{H}_j \subset L^2(\mathcal{X}_j, P_j)$ be a linear space consisting of real-valued functions with domain $\mathcal{X}_j$, and let $\|\phi_j\|_{1,j}^2$ be a kernel semi-norm on $\mathcal{H}_j = \mathcal{H}_j^0 + \mathcal{H}_j^1$ with null space $\mathcal{H}_j^0$ and RKHS complement $\mathcal{H}_j^1$. Assume that*

*(a)* $\mathcal{H}_j^0 = \text{span}\{q_{j,1}, \ldots, q_{j,m_j}\}$ *with* $\dim(\mathcal{H}_j^0) = m_j < \infty$;

*(b)* $\text{rank}(\text{Var}\,(q_{1,j}(X_j), \ldots, q_{j,m_j}(X_j))) = m_j$;

*(c) the reproducing kernel $k_j^1$ of $\mathcal{H}_j^1$ satisfies $E(k_j^1(X_j, X_j)) < \infty$;*

*(d) $\mathcal{H}_j$ is dense in $L^2(\mathcal{X}_j, dP_j)$.*

As discussed in Section 2.5 (see, in particular, Lemma 1), Assumptions 2(a)$-$(c) guarantee that $\mathcal{H}_j$ is an RKHS wrt $\langle \cdot, \cdot \rangle_{\star,j}$. This then allows us to establish the existence of kernel APC solutions in Section 2.5.5. On the other hand, Assumption 2(d) is needed for consistent estimation of arbitrary functions in $L^2(\mathcal{X}_j, P_j)$. However, only denseness wrt $\|\cdot\|_{P_j}$ (as opposed to the usual $L^2$-norm) is required/of interest. When $\mathcal{X}_j$ is a compact subset of $\mathbb{R}^{\text{d}}$, Assumption 2(d) is satisfied if $\mathcal{H}_j$ is the RKHS associated with the Gaussian kernels (which do not contain non-zero constants in the first place) or the Sobolev-type kernels (after removing irrelevant constants from the null space, so that it does not contradict Assumption 2(c). See Remark 1 for more details.).

### 2.6.2 Statement of Main Theorem

Our main results shows the convergence of individual sample kernel APC transformation to the corresponding population APC transformation in the $\|\cdot\|_{P_j}$ norm of

$L^2(\mathcal{X}_j, P_j)$, for $j = 1, \ldots, p$.

**Theorem 1.** *Suppose that Assumptions 1 and 2 hold. Consider estimation of APC transformations according to* (2.20)*, where the penalty parameter $\alpha = \alpha_n$ depends on the training sample size $n$. Let $(\alpha_n)_{n=1}^{\infty}$ be a sequence of positive numbers such that*

$$\lim_{n\to\infty} \alpha_n = 0, \qquad \lim_{n\to\infty} \frac{n^{-1/2}}{\alpha_n} = 0. \tag{2.30}$$

*Then, with probability tending to one, there exists solution $\hat{\boldsymbol{\Phi}}^{(n)} = (\hat{\phi}_1^{(n)}, \ldots, \hat{\phi}_p^{(n)})$ to* (2.20)*. Moreover, the sequence $(\hat{\boldsymbol{\Phi}}^{(n)})_{n=1}^{\infty}$ satisfies*

$$\sum_{j=1}^{p} \mathrm{Var}\left(\hat{\phi}_j^{(n)}(X_j) - \phi_j^*(X_j)\right) \xrightarrow{P} 0 \quad \textit{and} \tag{2.31}$$

$$|\hat{\lambda}_1^{(n)} - \lambda_1| \xrightarrow{P} 0 \tag{2.32}$$

*where $\lambda_1 = \mathrm{Var}\left(\sum \phi_j^*\right)$ and $\hat{\lambda}_1^{(n)} = \widehat{\mathrm{Var}}\left(\sum \hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{1,j}^2$.*

Note that in (2.31), for each $1 \le j \le p$, the variance $\mathrm{Var}\left(\hat{\phi}_j^{(n)}(X_j) - \phi_j^*(X_j)\right)$ integrates only over future observations $X_j$ but not over the past training data from which the estimates $\hat{\phi}_j^{(n)}(\cdot)$ are obtained. As a function of the training data, the variance terms are random variables. Essentially, the convergence in (2.31) says that the $j^{th}$ component of a sample kernel APC converges to the $j^{th}$ component of the population APC in the norm of $L^2(\mathcal{X}_j, P_j)$ in probability, for $j = 1, \ldots, p$, while the convergence in (2.32) says that the optimal value of the sample kernel APC criterion converges in probability to the optimal value of the population APC criterion.

Theorem 1 parallels the consistency results for kernel CCA in Fukumizu et al. (2007), but generalizes to $p \ge 2$ and concerns the lower end of the eigenspectrum. More importantly, our results hold for more general RKHSs with finite-dimensional null spaces under the more relaxed condition $\alpha_n^{-1} = o(n^{1/2})$, and do not require the

target of estimation to lie in the RKHS a priori. Our proof techniques are inspired by ideas in Leurgans et al. (1993) and Silverman (1996), and are much simpler and clearer than those used in Fukumizu et al. (2007), enabling us to improve upon the $\alpha_n^{-1} = o(n^{1/3})$ rate in Fukumizu et al. (2007).

### 2.6.3 Proof of Main Theorem

We now turn to establishing the consistency of sample kernel APCs. We first define the Rayleigh quotients

$$R_\alpha(\mathbf{\Phi}) := \frac{\mathrm{Var}\left(\sum_{j=1}^p \phi_j(X_j)\right) + \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2}{\sum_{j=1}^p \mathrm{Var}\left(\phi_j(X_j)\right) + \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2}, \tag{2.33a}$$

$$\hat{R}_\alpha(\mathbf{\Phi}) := \frac{\widehat{\mathrm{Var}}\left(\sum_{j=1}^p \phi_j(X_j)\right) + \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2}{\sum_{j=1}^p \widehat{\mathrm{Var}}\left(\phi_j(X_j)\right) + \alpha \sum_{j=1}^p \|\phi_j\|_{1,j}^2}. \tag{2.33b}$$

Note that if $\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}$ satisfies $R_0(\mathbf{\Phi}) \leq 1$, then for $\alpha \leq \alpha'$, we have $R_\alpha(\mathbf{\Phi}) \leq R_{\alpha'}(\mathbf{\Phi})$. In other words, $R_\alpha(\mathbf{\Phi})$ is monotonically increasing wrt $\alpha$ when $R_0(\mathbf{\Phi}) \leq 1$. This trivial observation turns out to be very useful in the establishment of the consistency proof below.

Under conditions in Theorem 1, we know that the population APC $\mathbf{\Phi}^*$ exists and is unique, while the population kernel APCs $\tilde{\mathbf{\Phi}}^{(n)}$ and the sample kernel APCs $\hat{\mathbf{\Phi}}^{(n)}$ exist with high probability for sufficiently large values of $n$. It follows that the infimum of the Rayleigh quotients are attained at the corresponding APC solutions (with high probability for $n$ sufficiently large):

$$\lambda_1 = \inf_{\mathbf{\Phi} \in \mathbf{H}^*} R_0(\mathbf{\Phi}) = R_0(\mathbf{\Phi}^*),$$
$$\tilde{\lambda}_1^{(n)} = \inf_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} R_{\alpha_n}(\mathbf{\Phi}) = R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}),$$

$$\hat{\lambda}_1^{(n)} = \inf_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} \hat{R}_{\alpha_n}(\mathbf{\Phi}) = \hat{R}_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}).$$

The following three key lemmas are key steps in the proof of our main theorems. The first lemma deals with the difference between $\hat{\lambda}_1^{(n)}$ and $\tilde{\lambda}_1^{(n)}$ which constitutes the stochastic error.

**Lemma 7.** *Suppose that Asusmptions 1 and 2 hold, and $(\alpha_n)_{n=1}^{\infty}$ is a sequence of positive numbers satisfying* (2.30)*. Then, for any $\epsilon > 0$,*

$$\lim_{n \to \infty} P\left( \sup_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} |\hat{R}_{\alpha_n}(\mathbf{\Phi}) - R_{\alpha_n}(\mathbf{\Phi})| > \epsilon \right) = 0.$$

The second lemma deals with the *deterministic* difference between $\tilde{\lambda}_1^{(n)}$ and $\lambda_1$ which can be viewed as approximation error.

**Lemma 8.** *Suppose that Assumptions 1 and 2(d) hold. Then for any $\epsilon \in (0, 1)$, there exists $\alpha(\epsilon) > 0$ and $\mathbf{\Psi} \in \boldsymbol{\mathcal{H}}$ such that*

$$R_{\alpha(\epsilon)}(\mathbf{\Psi}) < \lambda_1 + \epsilon. \tag{2.34}$$

The third lemma asserts the convergence of any sequence $(\mathbf{\Phi}^{(n)})_{n=1}^{\infty}$ to $\mathbf{\Phi}^*$ in the form of (2.35) provided that $R_0(\mathbf{\Phi}^{(n)}) \to \lambda_1$.

**Lemma 9.** *Suppose that Assumption 1 holds, and that $\mathbf{\Phi}^{(n)} = (\phi_1^{(n)}, \ldots, \phi_p^{(n)})$ satisfies $\lim_{n \to \infty} R_0(\mathbf{\Phi}^{(n)}) = \lambda_1$. Then*

$$\frac{\left( \sum \mathrm{Cov}(\phi_j^{(n)}, \phi_j^*) \right)^2}{\left( \sum \mathrm{Var}\,(\phi_j^{(n)}) \right) \left( \sum \mathrm{Var}\,(\phi_j^*) \right)} \to 1, \tag{2.35}$$

*as $n \to \infty$.*

We are now ready to present the proof of Theorem 1. For the reason of space, we

defer the details of the proof for the three key lemmas to Appendix A.2.2.

*Proof.* The proof of Theorem 1 can be divided into the following four parts, which we prove successively:

$$\text{(i) } \hat{\lambda}_1^{(n)} \xrightarrow{P} \lambda_1; \qquad \text{(ii) } \frac{\alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{1,j}^2}{\sum \mathrm{Var}\,(\hat{\phi}_j^{(n)}) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{1,j}^2} \xrightarrow{P} 0;$$

$$\text{(iii) } \frac{\left(\sum \mathrm{Cov}(\hat{\phi}_j^{(n)}, \phi_j^*)\right)^2}{\left(\sum \mathrm{Var}\,(\hat{\phi}_j^{(n)})\right)\left(\sum \mathrm{Var}\,(\phi_j^*)\right)} \xrightarrow{P} 1; \quad \text{(iv) } \sum_{j=1}^{p} \mathrm{Var}\,(\hat{\phi}_j^{(n)} - \phi_j^*) \xrightarrow{P} 0.$$

(i) We first show that

$$\tilde{\lambda}_1^{(n)} \to \lambda_1. \tag{2.36}$$

Following the remark after Assumption 1, we know that $0 \leq \lambda_1 < 1$ under Assumption 1. Consider $\epsilon > 0$ with $\lambda_1 + \epsilon < 1$. Since $\mathcal{H}_j$ is dense in $L^2(\mathcal{X}_j, P_j)$, by Lemma 8, there exist $\alpha(\epsilon) > 0$ and $\boldsymbol{\Psi} \in \boldsymbol{\mathcal{H}}$ sufficiently close to $\boldsymbol{\Phi}^*$ such that

$$R_{\alpha(\epsilon)}(\boldsymbol{\Psi}) < \lambda_1 + \epsilon < 1. \tag{2.37}$$

On the other hand, $\alpha_n \to 0$ implies that there exists $n(\epsilon)$ such that for all $n \geq n(\epsilon)$, $\alpha_n \leq \alpha(\epsilon)$, in which case

$$\begin{aligned} \tilde{\lambda}_1^{(n)} = \inf_{\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}} R_{\alpha_n}(\boldsymbol{\Phi}) \quad &\overset{(*)}{\leq} R_{\alpha_n}(\boldsymbol{\Psi}) \quad \overset{(**)}{\leq} R_{\alpha(\epsilon)}(\boldsymbol{\Psi}) \quad < \lambda_1 + \epsilon \quad < 1, \\ \tilde{\lambda}_1^{(n)} = R_{\alpha_n}(\tilde{\boldsymbol{\Phi}}^{(n)}) \quad &\overset{(**)}{\geq} R_0(\tilde{\boldsymbol{\Phi}}^{(n)}) \quad \overset{(*)}{\geq} \inf_{\boldsymbol{\Phi} \in \mathbf{H}^*} R_0(\boldsymbol{\Phi}) \quad = \lambda_1. \end{aligned} \tag{2.38}$$

In (2.38), the inequalities (*) hold trivially, while the inequalities (**) hold due to monotonicity of $R_\alpha(\boldsymbol{\Psi})$ and $R_\alpha(\tilde{\boldsymbol{\Phi}}^{(n)})$ wrt $\alpha$. From (2.38), we conclude that (2.36) holds.

To this end, it suffices to show that

$$|\tilde{\lambda}_1^{(n)} - \hat{\lambda}_1^{(n)}| \xrightarrow{P} 0 \tag{2.39}$$

to complete the proof. By Lemma 7, under condition (2.30), for any $\epsilon > 0$ and $\delta > 0$, there exists $n(\epsilon, \delta)$ such that for all $n \geq n(\epsilon, \delta)$, with probability at least $1 - \delta$,

$$\sup_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} |\hat{R}_{\alpha_n}(\mathbf{\Phi}) - R_{\alpha_n}(\mathbf{\Phi})| < \epsilon. \tag{2.40}$$

It follows that

$$\begin{aligned} \hat{\lambda}_1^{(n)} &= \hat{R}_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) \leq \hat{R}_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) < R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) + \epsilon = \tilde{\lambda}_1^{(n)} + \epsilon, \\ \tilde{\lambda}_1^{(n)} &= R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) \leq R_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) < \hat{R}_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) + \epsilon = \hat{\lambda}_1^{(n)} + \epsilon. \end{aligned} \tag{2.41}$$

Equivalently, $|\hat{\lambda}_1^{(n)} - \tilde{\lambda}_1^{(n)}| < \epsilon$. On both lines in (2.41), the first inequality holds trivially, whereas the second inequality holds according to (2.40). This completes the proof.

(ii) Consider again $\epsilon > 0$ with $\lambda_1 + \epsilon < 1$. By Lemma 7, with probability at least $1 - \delta$,

$$\sup_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} |\hat{R}_{\alpha_n}(\mathbf{\Phi}) - R_{\alpha_n}(\mathbf{\Phi})| < \frac{\epsilon}{3}.$$

for sufficiently large values of $n$. It follows that with probability at least $1 - \delta$,

$$R_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) \leq \hat{R}_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) + \frac{\epsilon}{3} \leq \hat{R}_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) + \frac{\epsilon}{3} \leq R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) + \frac{2\epsilon}{3} \leq \lambda_1 + \epsilon \tag{2.42}$$

holds for $n$ sufficiently large. The last inequality in (2.42) comes from $R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) = \tilde{\lambda}_1^{(n)} < \lambda_1 + \epsilon/3$ , which holds as a consequence of (2.36) for $n$ sufficiently large.

Thus, with probability at least $1-\delta$,

$$\lambda_1 \leq R_0(\hat{\mathbf{\Phi}}^{(n)}) \leq R_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) \leq \lambda_1 + \epsilon < 1 \tag{2.43}$$

for sufficiently large values of $n$. In (2.43), the first inequality is trivial while the second inequality follows from monotonicity of $R_\alpha(\hat{\mathbf{\Phi}}^{(n)})$ wrt $\alpha$. From (2.43), we conclude that

$$\begin{aligned} R_0(\hat{\mathbf{\Phi}}^{(n)}) &= \frac{\operatorname{Var}\left(\sum \hat{\phi}_j^{(n)}\right)}{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right)} \xrightarrow{P} \lambda_1, \\ R_{\alpha_n}(\hat{\mathbf{\Phi}}^{(n)}) &= \frac{\operatorname{Var}\left(\sum \hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2}{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2} \xrightarrow{P} \lambda_1. \end{aligned} \tag{2.44}$$

Since $\lambda_1 < 1$, it follows that from (2.44) that (ii) holds.

(iii) From (2.44), $R_0(\hat{\mathbf{\Phi}}^{(n)}) \xrightarrow{P} \lambda_1$, so (iii) follows directly from Lemma 9.

(iv) One can show (i.e., by applying Lemma 12 in Appendix A.2) that

$$\frac{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2}{\sum \widehat{\operatorname{Var}}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2} \xrightarrow{P} 1.$$

It then follows from (ii) that

$$\begin{aligned} &\frac{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right)}{\sum \widehat{\operatorname{Var}}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2} \\ &\quad = \left(1 - \frac{\alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2}{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2}\right) \cdot \frac{\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2}{\sum \widehat{\operatorname{Var}}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2} \\ &\quad \xrightarrow{P} 1. \end{aligned} \tag{2.45}$$

By definition, the sample kernel APC $\hat{\mathbf{\Phi}}^{(n)}$ satisfies $\sum \widehat{\operatorname{Var}}\left(\hat{\phi}_j^{(n)}\right) + \alpha_n \sum \|\hat{\phi}_j^{(n)}\|_{j,1}^2 = 1$, so (2.45) implies that $\sum \operatorname{Var}\left(\hat{\phi}_j^{(n)}\right) \xrightarrow{P} 1$. Combining this and $\sum \operatorname{Var}\left(\phi_j^*\right) = 1$

(again, by definition) with (iii), we obtain $(\sum \mathrm{Cov}(\hat{\phi}_j^{(n)}, \phi_j^*))^2 \xrightarrow{P} 1$. It follows that with an appropriate choice of sign for $\hat{\mathbf{\Phi}}^{(n)}$, we have

$$\sum_{j=1}^{p} \mathrm{Var}\,(\hat{\phi}_j^{(n)} - \phi_j^*) = \sum_{j=1}^{p} \mathrm{Var}\,(\hat{\phi}_j^{(n)}) - 2\sum_{j=1}^{p} \mathrm{Cov}(\hat{\phi}_j^{(n)}, \phi_j^*) + \sum_{j=1}^{p} \mathrm{Var}\,(\phi_j^*)$$
$$\xrightarrow{P} 1 - 2 + 1 = 0.$$

The proof is complete.

□

**Remark 3.** Equation (2.36) established that $R_{\alpha_n}(\tilde{\mathbf{\Phi}}^{(n)}) = \tilde{\lambda}_1^{(n)} \to \lambda_1$. On the other hand, equation (2.38) reveals that $R_0(\tilde{\mathbf{\Phi}}^{(n)}) \to \lambda_1$. Thus, by applying arguments similar to that in the proof of part (ii), (iii) and (iv), we can also conclude that the population kernel APC $\tilde{\mathbf{\Phi}}^{(n)}$ satisfies $\sum \mathrm{Var}\,(\tilde{\phi}_j^{(n)} - \phi_j^*) \to 0$.

## 2.7 Estimation and Computation

In this section, we motivate an iterative method for computing kernel APCs. This involves the use of power algorithm, an iterative algorithm for extracting the first few largest (or smallest) eigenvectors of a bounded linear operator. In addition to detailing out the algorithm, we provide theoretical justification of the use of power algorithm in the RKHS framework.

Consider a matrix $\mathbf{M}$ with the eigen-decomposition $\mathbf{M} = \sum_{i=1}^{m} \lambda_i \mathbf{M}_i$, where $\mathbf{M}_i = \mathbf{u}_i \mathbf{u}_i^T$ and the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ are distinct. The power algorithm allows us to compute the eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue $\lambda_1$ (see, e.g., Golub & Van Loan (2013)) by forming normalized powers $\mathbf{M}^t\mathbf{u}_0/\|\mathbf{M}^t\mathbf{u}_0\|$ which can be shown to converge to $\mathbf{u}_1$ as long as $\mathbf{u}_0$ is not orthogonal to $\mathbf{u}_1$.

To compute the eigenvector $\mathbf{u}_m$ corresponding to the smallest eigenvalue $\lambda_m$, the spectrum needs to be flipped and shifted by replacing $\mathbf{M}$ with $\gamma\mathbf{I} - \mathbf{M}$ in the power algorithm. If $0 \leq \lambda_1 < \lambda_m \leq B$ for some $B > 0$, then using $\gamma = (B+1)/2$, we have

$$\begin{aligned} -\frac{B-1}{2} \leq & \gamma - \lambda_i \leq \frac{B-1}{2} && \text{if } 1 \leq \lambda_i \leq B, \\ \frac{B-1}{2} \leq & \gamma - \lambda_i \leq \frac{B-1}{2} + 1 && \text{if } 0 \leq \lambda_i \leq 1. \end{aligned}$$

In this case, the large eigenvalues of $\mathbf{M}$, $\{\lambda : \lambda > 1\}$, are mapped to an interval centered at 0, while the small eigenvalues $\{\lambda : \lambda < 1\}$ are affixed to the right end of this interval.

### 2.7.1 Eigen-characterization of Kernel APCs

To relate power algorithm to kernel APCs, we first show that the kernel APC problem (2.17) can be reformulated as an eigenvalue problem wrt the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\alpha$ defined on $\boldsymbol{\mathcal{H}}$. As a consequence, the smallest kernel APC can be obtained as the eigenvector corresponding to the smallest eigenvalue of an operator $\tilde{\mathbf{S}}^{(\alpha)}$ defined on $\boldsymbol{\mathcal{H}}$. Then, computation of sample kernel APC reduces to an application of power algorithm on an empirical version of $\tilde{\mathbf{S}}^{(\alpha)}$.

Consider the following smoothing operator $\mathbf{S}_{ij}^{(\alpha)}$, defined through a "generalized" regularized population regression problem:

$$\begin{aligned} \mathbf{S}_{ij}^{(\alpha)} : (\mathcal{H}_j, \langle \cdot, \cdot \rangle_{\alpha,j}) \to (\mathcal{H}_i, \langle \cdot, \cdot \rangle_{\alpha,i}), & \qquad (2.46) \\ \phi_j \mapsto \underset{f \in \mathcal{H}_i}{\operatorname{argmin}} \left\{ \operatorname{Var} \left( \phi_j(X_j) - f(X_i) \right) + \alpha \|f\|_{i,1}^2 \right\}. & \end{aligned}$$

Note that (2.46) reduces to the population version of the usual regularized regression problem, when $\phi_j$ and $f$ are both required to have mean zero. With the establishment

of existence and uniqueness of solution to the problem, the smoothing operator $\mathbf{S}_{ij}^{(\alpha)}$ in (2.46) mapping $\phi_j$ to its "smoothed" version in $\mathcal{H}_i$ is well-defined. In addition, it enjoys some nice properties:

**Theorem 2.** *For $j = 1, \ldots, p$, let $(\mathcal{H}_j, \langle \cdot, \cdot \rangle_{\alpha,j})$ be RKHS as defined in Assumptions 2(a)−(c). Then $\mathbf{S}_{ij}^{(\alpha)}$ is well-defined. In fact, $\mathbf{S}_{ij}^{(\alpha)}$ is the cross-covariance operator from $(\mathcal{H}_j, \langle \cdot, \cdot \rangle_{\alpha,j})$ to $(\mathcal{H}_i, \langle \cdot, \cdot \rangle_{\alpha,i})$:*

$$\langle \phi_i, \mathbf{S}_{ij}^{(\alpha)} \phi_j \rangle_{\alpha,i} = \mathrm{Cov}\left(\phi_i(X_i), \phi_j(X_j)\right), \qquad \forall \phi_i \in \mathcal{H}_i, \phi_j \in \mathcal{H}_j, \tag{2.47}$$

*and it follows that $\mathbf{S}_{ij}^{(\alpha)}$ is compact. Moreover,*

$$\|\mathbf{S}_{ij}^{(\alpha)} \phi_j\|_{\alpha,i} \leq \left(\mathrm{Var}\left(\phi_j(X_j)\right)\right)^{1/2} \leq \|\phi_j\|_{\alpha,j}, \qquad \forall \phi_j \in \mathcal{H}_j. \tag{2.48}$$

Theorem 2 says that the operator $\mathbf{S}_{ij}^{(\alpha)}$ is not only well-defined, but also is the cross-covariance operator from $(\mathcal{H}_j, \langle \cdot, \cdot \rangle_{\alpha,j})$ to $(\mathcal{H}_i, \langle \cdot, \cdot \rangle_{\alpha,i})$. In particular, we have $\mathbf{S}_{ij}^{(1)} = \mathbf{C}_{ij}$, where $\mathbf{C}_{ij}$ is the cross-covariance operator from $(\mathcal{H}_j, \langle \cdot, \cdot \rangle_{\star,j})$ to $(\mathcal{H}_i, \langle \cdot, \cdot \rangle_{\star,i})$ given in Section 2.5.5. Equation (2.48) states that $\mathbf{S}_{ij}^{(\alpha)}$ is a contraction operation. We are now ready to restate the kernel APC problem as an eigenvalue problem wrt the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\alpha$.

**Theorem 3.** *Let $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$, where $\mathcal{H}_j$ is an RKHS wrt $\langle \cdot, \cdot \rangle_{\alpha,j}$, for $1 \leq j \leq p$. Then the kernel APC problem (2.17) can be restated as*

$$\min_{\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}} \langle\!\langle \boldsymbol{\Phi}, \tilde{\mathbf{S}}^{(\alpha)} \boldsymbol{\Phi} \rangle\!\rangle_\alpha \qquad \textit{subject to} \qquad \langle\!\langle \boldsymbol{\Phi}, \boldsymbol{\Phi} \rangle\!\rangle_\alpha = 1, \tag{2.49}$$

*where $\tilde{\mathbf{S}}^{(\alpha)} : \boldsymbol{\mathcal{H}} \to \boldsymbol{\mathcal{H}}$ is defined by the component mapping*

$$(\tilde{\mathbf{S}}^{(\alpha)} \boldsymbol{\Phi})_i = \sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j + \phi_i, \tag{2.50}$$

*and* $\mathbf{S}_{ij}^{(\alpha)}$ *is the smoothing operator as defined in* (2.46)*. Moreover,* $\tilde{\mathbf{S}}^{(\alpha)}$ *is self-adjoint, positive, and bounded above by* $p$*.*

**Remark 4.** *Note that* (2.24a) *reduces to* (2.49) *when* $\alpha = 1$*. On the other hand, if we compare* (2.49) *and* (2.50) *with the population APC correspondences* (2.5) *and* (2.4)*, we see that* $\mathbf{S}_{ij}^{(\alpha)}$ *in the population kernel APC problem is an analogue of* $\mathbf{P}_{ij}$ *in the population APC problem, where* $L^2$*-orthogonal projection is replaced with smoothing. Roughly speaking,* $\mathbf{P}_{ij}$ *defined on finite-dimensional subspaces of* $L^2$*-spaces can be viewed as a special case of* $\mathbf{S}_{ij}$*: if we consider the* $L^2$*-spaces with orthogonal polynomial bases, then* $\mathbf{P}_{ij}$ *"smoothes" a function* $\phi_j$ *by taking only the leading terms in the basis expansion.*

By Theorem 2, $\mathbf{S}_{ij}^{(\alpha)}$ is compact. Although this does not imply compactness of $\tilde{\mathbf{S}}^{(\alpha)}$, one can readily verify the compactness of $\tilde{\mathbf{S}}^{(\alpha)} - \mathbf{I}$. Similar to the explanation for population APCs in Section 2.2.3, this means that $\tilde{\mathbf{S}}^{(\alpha)} - \mathbf{I}$ has an eigendecomposition with eigenvalues that can only accumulate at 0, which in turn implies that the eigenvalues of $\tilde{\mathbf{S}}^{(\alpha)}$ can only accumulate at $+1$. To this end, we see that the smallest kernel APC is given by the eigenvector corresponding to the smallest eigenvalue of $\tilde{\mathbf{S}}^{(\alpha)}$. Similarly, the $l^{th}$ smallest kernel APC is given by the eigenvector corresponding to the $l^{th}$ smallest eigenvalue of $\tilde{\mathbf{S}}^{(\alpha)}$ (where eigenvalues are repeated according to their multiplicity).

### 2.7.2 Power Algorithm for Kernel APCs

Applying the knowledge that vectors of kernel APC transformations are the eigenvectors of $\tilde{\mathbf{S}}^{(\alpha)}$ from a population standpoint, we execute the power algorithm on $\gamma\mathbf{I} - \tilde{\mathbf{S}}^{(\alpha)}$ to solve for the (smallest) kernel APC. The pseudocode is given below. Here $\gamma$ is taken to be $(p+1)/2$ since the spectrum of $\tilde{\mathbf{S}}^{(\alpha)}$ is bounded above by $p$, as claimed

in Theorem 3. Thus, solving for kernel APC reduces to iterative smoothing of each component $\phi_j$ against $X_i$, for $j \neq i$.

**Algorithm 1** Computation of kernel APCs

Let $\gamma = (p+1)/2$. Initialize $t = 0$, $\mathbf{\Phi}^{[0]} = (\phi_1^{[0]}, \phi_2^{[0]}, \ldots, \phi_p^{[0]})$.
**repeat**
  **for** $i = 1, \ldots, p$ **do**
    $\phi_i \leftarrow \gamma\phi_i^{[t]} - (\sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j^{[t]} + \phi_i^{[t]})$ ▷ Update steps
  **end for**
  Standardize with $c = (\sum \|\phi_i\|_{\alpha,i}^2)^{-1/2}$
  $(\phi_1^{[t+1]}, \phi_2^{[t+1]}, \ldots, \phi_p^{[t+1]}) \leftarrow (c\phi_1, c\phi_2, \ldots, c\phi_p)$
  $t \leftarrow t + 1$
**until** Var $\sum \phi_i^{[t]} + \alpha \sum \|\phi_i^{[t]}\|_{i,1}^2$ converges

To compute the $l^{th}$ smallest kernel APCs for $l > 1$, we just need to add a series of Gram-Schmidt steps

$$\phi_i \leftarrow \phi_i - \left( \sum_{j=1}^{p} \langle \phi_{\ell,j}, \phi_j^{[t]} \rangle_{\alpha,j} \right) \phi_{\ell,i}, \qquad 1 \leq \ell \leq l-1$$

following the update steps in Algorithm 1, to ensure that the orthogonality requirements (2.19) are satisfied. Here $\mathbf{\Phi}_\ell = (\phi_{\ell,1}, \cdots, \phi_{\ell,p})$, $1 \leq \ell \leq l-1$, stands for the $\ell^{th}$ smallest kernel APC that has been obtained beforehand.

The power algorithm is guaranteed to converge under mild conditions:

**Proposition 1.** *Suppose that the smallest eigenvalue of $\tilde{\mathbf{S}}^{(\alpha)}$ is of multiplicity one with corresponding unit eigenvector $\tilde{\mathbf{\Phi}}$. If the power algorithm is initialized with $\mathbf{\Phi}^{[0]}$ that has a nontrivial projection onto $\tilde{\mathbf{\Phi}}$, then the power algorithm converges.*

For implementation details see Appendix A.4.

**Remark 5.** All the results in this section still hold if we allow different penalty parameters $\alpha_j$ for $X_j$, $j = 1, \ldots, p$. In particular, Algorithm 1 can be easily modified to incorporate different values of $\alpha_j$.

## 2.8 Methodologies for Choosing Penalty Parameters

Any kernel calls implicitly for a multiplicative penalty parameter that controls the amount of regularization to balance bias and variance against each other. Methods that use multiple kernels will have as many penalty parameters as kernels. Choosing the penalty parameters in a given problem requires some principles for systematically selecting the values for these parameters. Such principles have been discussed at least as long as there have existed additive models (Hastie & Tibshirani, 1990), and APCs pose new problems only in so far as they use Rayleigh quotients as their optimization criteria rather than residual sums of squares or other regression loss functions as their minimization criteria. In this section, we discuss some possible ways to choose the penalty parameters $\alpha_1, \ldots, \alpha_p$ for estimating kernel APCs. An initial division of principles for penalty parameter selection is into a priori choice and data-driven choice.

### 2.8.1 A Priori Choice of Penalty Parameters

In order to make an informed a priori choice of a penalty parameter it must be translated into an interpretable form. The most common such form is in terms of a notion of "degrees of freedom" which can be heuristically rendered as "equivalent number of observations invested in estimating a transformation." To define degrees of freedom for kernelizing, note that in the power algorithm implementation in Section 2.7.2, the dependence of kernel APC on the tuning parameters $\alpha$ is through the smoothing operators $\mathbf{S}_{ij}^{(\alpha)}$ (defined in (2.46)). Empirically, for a penalty parameter $\alpha$ such a smoothing operation on regressor-response data $\{(x_i, y_i)\}_{i=1..n}$ is a linear operation $\mathbf{y} = (y_i)_{i=1..n} \mapsto \hat{\mathbf{y}} = (\hat{f}(x_i))_{i=1..n}$, $\mathbb{R}^n \to \mathbb{R}^n$, and can be represented by a

matrix operation $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where the $n \times n$ "smoother matrix" $\mathbf{S}$ is symmetric and non-negative definite, and all its eigenvalues are $\leq 1$. The matrix $\mathbf{S}$ depends on the penalty parameter $\alpha$, $\mathbf{S} = \mathbf{S}^{(\alpha)}$, and serves as the basis for defining notions of degrees of freedom. Several definitions exist, three of which are as follows (Buja et al., 1989):

- $df = \text{tr}(\mathbf{S}^2)$: This derives from the total variance in $\hat{\mathbf{y}}$, which under homoskedasticity is $\sum_i \text{Var}(\hat{y}_i) = \text{tr}(\mathbf{S}\mathbf{S}')\sigma^2$. Variance of fitted values is a measure of how much response variation has been invested in the fits.

- $df = \text{tr}(2\mathbf{S} - \mathbf{S}^2)$: This derives from the total residual variance in $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ under a homoskedasticity assumption: $\sum_i \text{Var}(r_i) = \text{tr}(\mathbf{I} - \mathbf{S} - \mathbf{S}' + \mathbf{S}\mathbf{S}')\sigma^2$. Variance of residuals, when substracted from $n\sigma^2$, is a measure of how much of the error variance has been lost to the fitted values.

- $df = \text{tr}(\mathbf{S})$: This derives from a Bayesian interpretation of kernelizing under a natural Bayes prior that results in $\mathbf{S}\sigma^2$ as the posterior covariance matrix of $\hat{\mathbf{y}}$. A frequentist derivation is obtained by generalizing Mallows' $C_p$ statistic which corrects the residual sum of squares with a term $2(df)\hat{\sigma}^2$ to make it unbiased for the predictive MSE; the appropriate generalization for smoothers is $df = \text{tr}(\mathbf{S})$.

Among these, the third is the most popular version. If $\mathbf{S}$ is a projection, all three definitions result in the same value, which is the projection dimension, but for kernels whose $\mathbf{S}$ contains eigenvalues strictly between 0 and 1 the three definitions are measures of different concepts. For general kernels the calculation of degrees of freedom for a ladder of penalty parameter values $\alpha$ may result in considerable computational expense, which is compounded by the fact that in practice for a prescribed degree of freedom several values of $\alpha$ need to be tried in a bisection search. Yet the translation of $\alpha$ to a degree of freedom may be the most natural device for deciding a priori on an approximate value of the penalty parameter. Selecting degrees of freedom sepa-

rately for each transformation $\phi_j$ is of course a heuristic for APCs, as it is for additive regression models, because what matters effectively is the total degrees of freedom in the additive function $\sum_{j=1}^{p} \hat{\phi}_j^{(n)}$. Summing up the individual degrees of freedom of $\hat{\phi}_j^{(n)}$ is only an approximation to the degrees of freedom of $\sum_{j=1}^{p} \hat{\phi}_j^{(n)}$ (Buja et al., 1989).

In practice one often decides on identical degrees of freedom *df* for all transforms $\hat{\phi}_j^{(n)}$ and chooses the sum $p \cdot df$ to be a fraction of $n$, such as $p \cdot df = n/10$.

### 2.8.2 Data-driven Choice of Penalty Parameters

The most popular data-driven method is based on cross-validation. A first question is what the criterion should be that is being cross-validated. We use as the relevant criterion the empirical, unpenalized sample eigenvalue:

$$\frac{\widehat{\mathrm{Var}}\,(\sum \hat{\phi}_j)}{\sum \widehat{\mathrm{Var}}\,(\hat{\phi}_j)}.$$

This is an estimate of $\lambda_1$ which, when small ($\ll 1$), suggests the existence of additive degeneracy in the data. Of course, the criterion that is actually being minimized in sample kernel APC is the penalized sample eigenvalue:

$$\hat{\lambda}_1 = \frac{\widehat{\mathrm{Var}}\,(\sum \hat{\phi}_j) + \sum \alpha_j \|\hat{\phi}_j\|_{j,1}^2}{\sum \widehat{\mathrm{Var}}\,(\hat{\phi}_j) + \sum \alpha_j \|\hat{\phi}_j\|_{j,1}^2}.$$

We treat this as a surrogate quantity that is not of substantive interest. (The distinction between quantity of interest and surrogate quantity is familiar from supervised classification where interest focuses on misclassification rates but minimization is carried out on surrogate loss functions such as logistic or exponential loss; accordingly it is misclassification rates that are used in cross-validation.)

To choose the penalty parameters in the simplest possible way, one often makes

them identical for all variables and then searches their common value $\alpha$ on a grid, minimizing the $k$-fold cross-validation criterion

$$\mathrm{CV}(\alpha) = \frac{1}{k}\sum_{i=1}^{k} \frac{\widehat{\mathrm{Var}}\left(\sum_{j=1}^{p}\hat{\phi}_{\{i\}j}\right)}{\sum_{j=1}^{p}\widehat{\mathrm{Var}}\left(\hat{\phi}_{\{i\}j}\right)}.$$

The variances $\widehat{\mathrm{Var}}$ are evaluated on the holdout sets while the transforms $\hat{\phi}_{\{i\}j}$ are estimated from the training sets.

Here, however, attention must be paid to the question of what "equal value of the penalty parameters" means. The issue is that the meaning of a penalty parameter $\alpha$ is very much scale dependent. For example, a standard Gaussian kernel $k(x, x') = \exp\{-\frac{1}{2}(x - x')^2\}$ is very different when a variable measured in miles is converted to a variable measured in feet. When all variables are continuous and come in different scales, one approach to equalizing the effect of scale on the penalties and kernels is to standardize all variables. Another approach is to calibrate all penalty parameters to produce the same degrees of freedom.

## 2.9 Methodology for Kernel APCs: Data Examples

In this section, we present the kernel APC methodology in terms of two data examples.

### 2.9.1 University Webpages

The major benefit of formulating APCs in the kernelizing framework is the flexibility of embedding the information contained in data objects in $p$ different $n \times n$ kernel matrices as opposed to an $n \times p$ feature matrix. Kernel matrices have an interpretation as similarity measures between pairs of data objects. It is therefore possible

to directly design similarity matrices (instead of features) for non-Euclidean data for use as kernels. Just as one extracts multiple features from data objects, one similarly extracts multiple similarity matrices to capture different topological information in data objects. Thus topological information between data objects captured by multiple kernels can be used to directly estimate APC transforms of non-quantitative data. APC finds associations between these kernels in terms of "implicit" redundancies. On data the APC variance is evaluated on the sum of "scorings" or "scalings" or "quantifications" (Section 2.2.1), and the penalties are obtained from the constructed kernel matrices. This methodology was not available at the time when the first article on APCs by Donnell et al. (1994) was written.

In this section, we consider data on university webpages from the "World Wide Knowledge Base" project at Carnegie Mellon University. This data set was preprocessed by Cardoso-Cachopo (2007) and previously studied in Guo et al. (2011) and Tan et al. (2015). It includes webpages from computer science departments at Cornell, University of Texas, University of Washington, and University of Wisconsin. In this analysis, we consider only the faculty webpages — resulting in a subset of $n = 374$ webpages and $d = 3901$ keywords that appear on these webpages. These webpages are data objects whose similarity in keywords form the raw ingredients for kernel APC analysis.

We now discuss how we constructed four similarity matrices to be used as kernels. Following Guo et al. (2011), first we reduced the number of keywords from 3901 to 100 by thresholding the entropy. Let $f_{ij}$ be the number of times the $j^{th}$ keyword appears in the $i^{th}$ webpage. (The entropy of the $j^{th}$ keyword is defined as $-\sum_{i=1}^{n} g_{ij} \log(g_{ij}) / \log(n)$, where $g_{ij} = f_{ij} / \sum_{i=1}^{n} f_{ij}$.) We then selected the 100 keywords with the largest entropy values and constructed an $n \times 100$ matrix $H$ whose $(i, j)$ element is $\log(1 + f_{ij})$. We further standardized each column to have zero mean

| group 1 | | group 2 | group 3 | | group 4 | |
|---|---|---|---|---|---|---|
| activ | student | address | acm | languag | advanc | receiv |
| area | teach | contact | algorithm | method | assist | scienc |
| book | work | cours | analysi | model | associ | softwar |
| build | year | depart | applic | network | center | state |
| california | | email | architectur | parallel | colleg | technolog |
| chair | | fall | base | problem | degre | univers |
| class | | fax | comput | process | director | |
| current | | hall | confer | program | educ | |
| faculti | | home | data | public | electr | |
| graduat | | inform | design | research | engin | |
| group | | link | develop | select | institut | |
| includ | | list | distribut | structur | intellig | |
| interest | | mail | gener | studi | laboratori | |
| introduct | | offic | high | system | mathemat | |
| paper | | page | ieee | techniqu | member | |
| project | | phone | implement | theori | number | |
| recent | | updat | investig | time | profession | |
| special | | web | journal | tool | professor | |

Table 2.1: Keywords in group 1 to group 4.

and unit variance. In order to obtain four different kernels, we applied the $k$-means algorithm to cluster the keywords in $H$ into $k = p = 4$ groups. Each group of keywords is represented as an $n \times m_j$ submatrix $H_j$, and we obtained the final $n \times n$ kernel matrix $K_j = H_j H_j^T / \mathrm{tr}(H_j H_j^T)$, where the normalization is to account for different group sizes. Thus the kernel matrix $K_j$ represents webpage-webpage similarities in terms of keywords in group $j$. Although a linear kernel is used here to construct $K_j$, any attempt at using the combined 100 keyword frequencies as features would hopelessly overfit the data given that $n = 374$. The approach based on kernels provides in this case four penalty parameters (one per kernel) to control overfitting, which we chose to be $\alpha_j = 0.0001$ $(j = 1, \ldots, 4)$ based on exploratory plots.

Table 2.1 shows the keywords in each group. Roughly, group 1 contains keywords related to teaching and current projects, group 2 contains keywords related to contact information, group 3 contains keywords related to research area, and group 4 contains

keywords related to biography of a faculty.

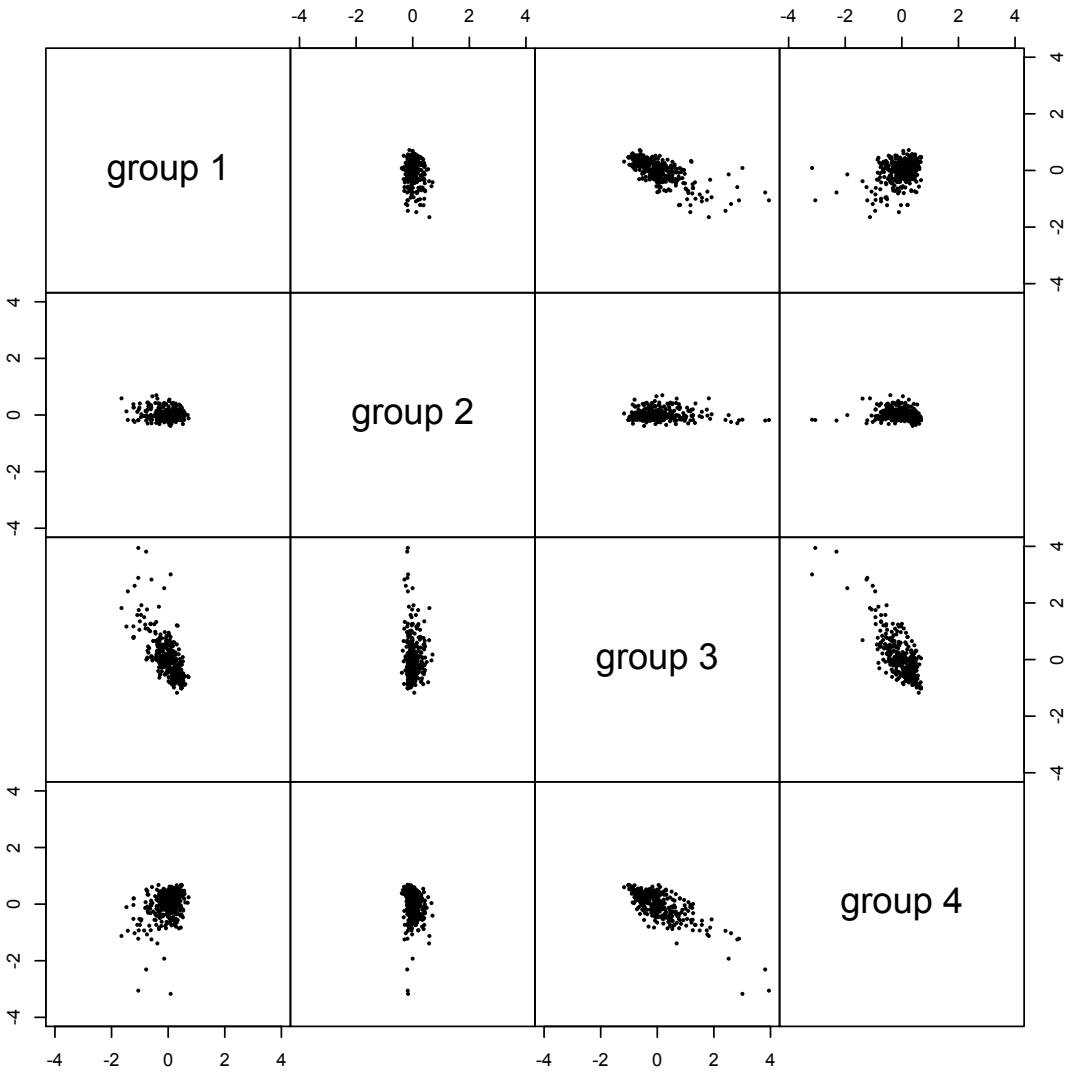


Figure 2.1: Pairwise scatterplot of the smallest kernel APC scores for the university webpages data. The eigenvalue for the APC is 0.0910.

From a kernel APC analysis using the kernel matrices $K_1, \ldots, K_4$ constructed above, we obtain score vectors $\hat{\phi}_1, \ldots, \hat{\phi}_4$. We can interpret the $n$-vector $\hat{\phi}_j$ as the kernel APC scores for individual webpages that reflect similarities based on the keywords in group $j$. Figure 2.1 shows the pairwise scatterplot of kernel APC scores between

different keyword groups. We see that $\hat{\phi}_3$ and $\hat{\phi}_4$ have strong negative correlation, $\hat{\phi}_1$ and $\hat{\phi}_3$ have moderately negative correlation, and $\hat{\phi}_1$ and $\hat{\phi}_4$ have weak positive correlation. The unpenalized sample eigenvalue $\hat{\lambda}_1 = (\widehat{\mathrm{Var}}\,(\sum \hat{\phi}_j)) \,/\, (\sum \widehat{\mathrm{Var}}\,(\hat{\phi}_j))$, which measures the strength of additive degeneracy, equals 0.0910, a value that is sufficiently close to zero to indicate considerable strength of additive association among the four kernels. (We form the ratio $\hat{\lambda}_1$ omitting the penalty terms; these are mere regularization devices for estimation and not of substantive interest.) The scores are centered to have zero mean and normalized to satisfy $\sum_{j=1}^{4} \widehat{\mathrm{Var}}\,(\hat{\phi}_j) = 1$. This standardization permits us to interpret $\widehat{\mathrm{Var}}\,(\hat{\phi}_j)$ as relative importance of group $j$ in the kernel APC solution. The variance of each group in the smallest kernel APC are: 0.1662 (group 1), 0.0305 (group 2), 0.5562 (group 3), 0.2471 (group 4). Ignoring group 2 which has the smallest weight, we see that, roughly, this means that

$$\hat{\phi}_1 + \hat{\phi}_3 + \hat{\phi}_4 \approx 0, \qquad \text{or, equivalently,} \qquad \hat{\phi}_4 \approx -\hat{\phi}_1 - \hat{\phi}_3.$$

If we plot $\hat{\phi}_4$ against $\hat{\phi}_1 + \hat{\phi}_3$, we obtain the scatterplot in Figure 2.2.

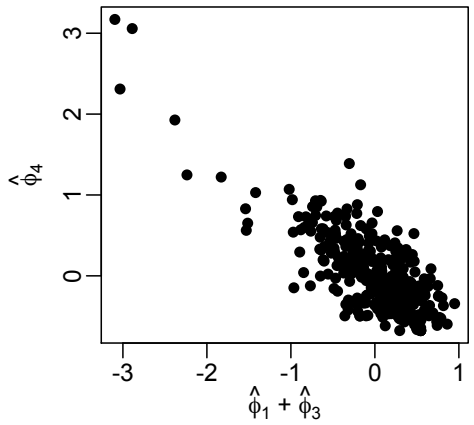


Figure 2.2: Plot of $\hat{\phi}_4$ against $\hat{\phi}_1 + \hat{\phi}_3$ in the smallest kernel APC for the university webpages data.

Recalling that the kernels were constructed to reflect similarity in terms of keywords related to (1) teaching and current projects, (2) contact information, (3) research area, and (4) biography, we obtain two results: contact information is related to neither of teaching and projects nor research, whereas biography is well predicted by teaching, projects and research. This is of course highly plausible for faculty webpages and academic biographies. — This example demonstrates the ability of kernel APCs to reveal associations among different topological representations encoded by multiple kernel matrices.

### 2.9.2 Air Pollution

In this section, we apply kernel APC analysis to a data set consisting of quantitative variables, where the purpose is to find nonlinear transformations that reflect additive redundancies among the variables. We analyze the $NO_2$ data that is publicly available on the StatLib data sets archive *http://lib.stat.cmu.edu/datasets/NO2.dat*. It contains a subsample of 500 observations from a data set collected by the Norwegian Public Roads Administration for studying the dependence of air pollution on traffic volume and meteorological condition. The response variable consists of hourly values of the log-concentration of $NO_2$ particles, measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. Because the posted data is only a subset of the original data, the middle chunk of observations is missing. To avoid artifacts, only the second half of the data (roughly November 2002 to May 2003) is used in the kernel APC analysis. Given below are descriptions for individual variables in the data:

| | |
|---|---|
| `NO2`: | hourly values of the logarithm of the concentration of $NO_2$ particles; |
| `Cars`: | logarithm of the number of cars per hour; |
| `TempAbove`: | temperature 2 meters above ground (degree C); |
| `Wind`: | wind speed (meters/second); |
| `TempDiff`: | temperature difference between 25 and 2 meters above ground (degree C); |
| `WindDir`: | wind direction (degrees between 0 and 360); |
| `HourOfDay`: | hour of day; |
| `DayNumber`: | day number from October 1, 2001. |

For each $j = 1, \ldots, p$, we use a Sobolev kernel corresponding to a cubic spline penalty $J_j(\phi_j) = \alpha_j \int (\phi_j''(x_j))^2 dx_j$. We first standardize all variables to unit variance and then choose the penalty parameters $\alpha_j$ to achieve "degrees of freedom" $= 4$ (for ways of selecting penalty parameters in terms of "degrees of freedom," see Section 2.8).

Figure 2.3 shows the transformations for each variable in the smallest kernel APC. As in Section 2.9.1, the transformed data points are centered to zero mean and normalized to satisfy $\sum \widehat{\mathrm{Var}}\,(\hat{\phi}_j) = 1$. The variables `Cars` and `HourOfDay` are the strongest variables with respective variances 0.51 and 0.304 under such a normalization. Holding other variables fixed, the approximate estimated constraint is $\hat{\phi}_2(\texttt{Cars}) + \hat{\phi}_7(\texttt{HourOfDay}) \approx 0$. Since $\hat{\phi}_2$ is monotone decreasing and the transformation of `HourOfDay` peaks around 4pm, we infer that the largest number of cars on the roads is found in the late afternoon, which is consistent with the daily experience of commuters.

In the second-smallest kernel APC, shown in Figure 2.4, the variables `TempAbove` and `DayNumber` play the dominant roles, and we have $\hat{\phi}_3(\texttt{TempAbove}) + \hat{\phi}_8(\texttt{DayNumber}) \approx 0$. Since $\hat{\phi}_3$ is monotone decreasing it follows that `TempAbove` decreases and then increases with respect to `DayNumber`. This relationship makes sense because our data span the period from November 2002 to May 2003, with the transition from fall and

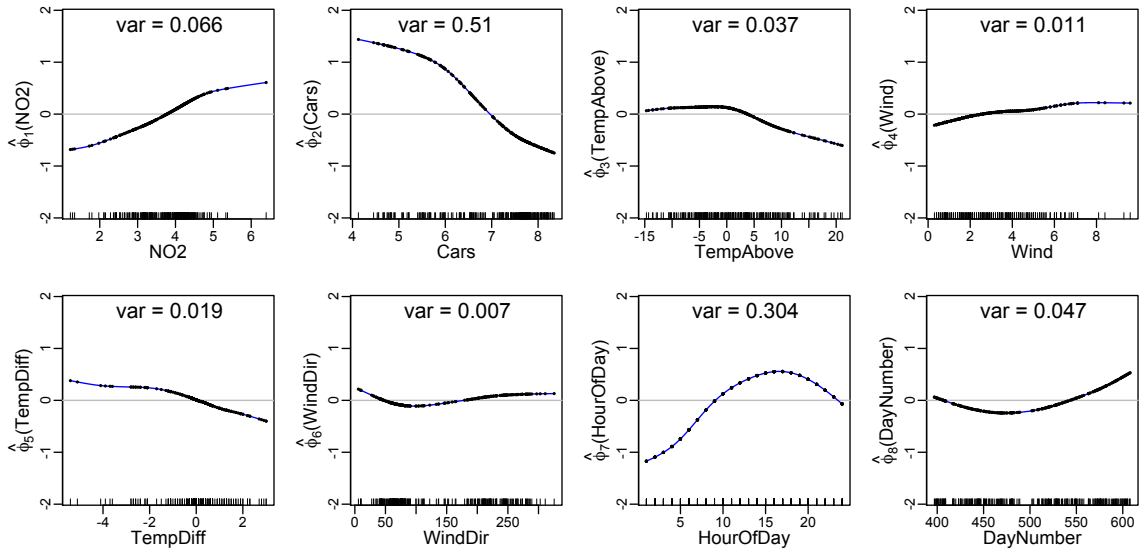


Figure 2.3: The smallest kernel APC transformations for the $NO_2$ data, using Sobolev kernel of order 2 for each variable. The eigenvalue for the APC is 0.0621. The black bars at the bottom of each panel indicate the location of data points for that variable.

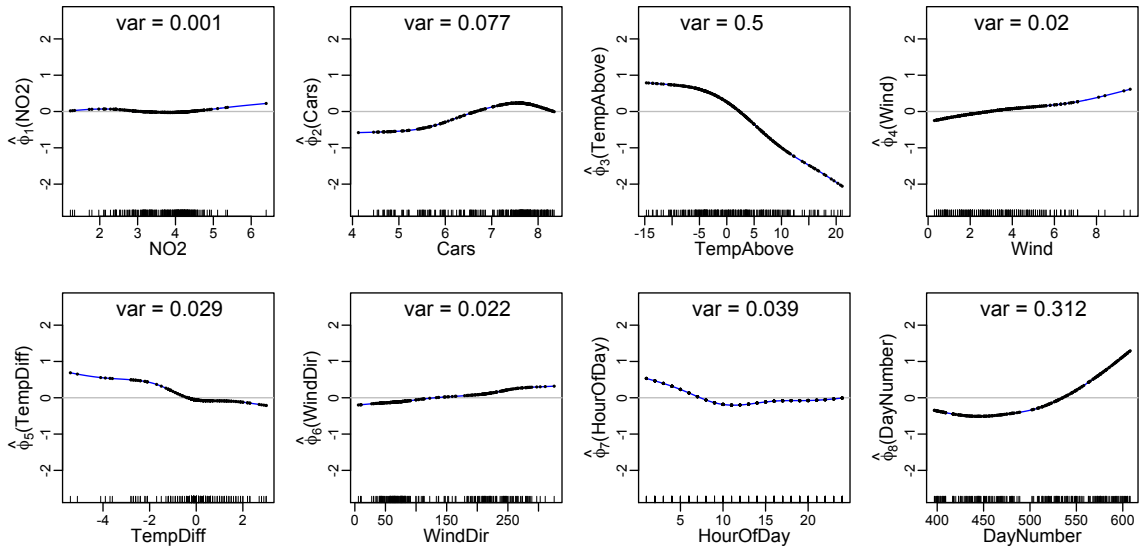


Figure 2.4: The second-smallest kernel APC transformations for the $NO_2$ data, using Sobolev kernel of order 2 for each variable. The eigenvalue for the APC is 0.0827. The black bars at the bottom of each panel indicate the location of data points for that variable.

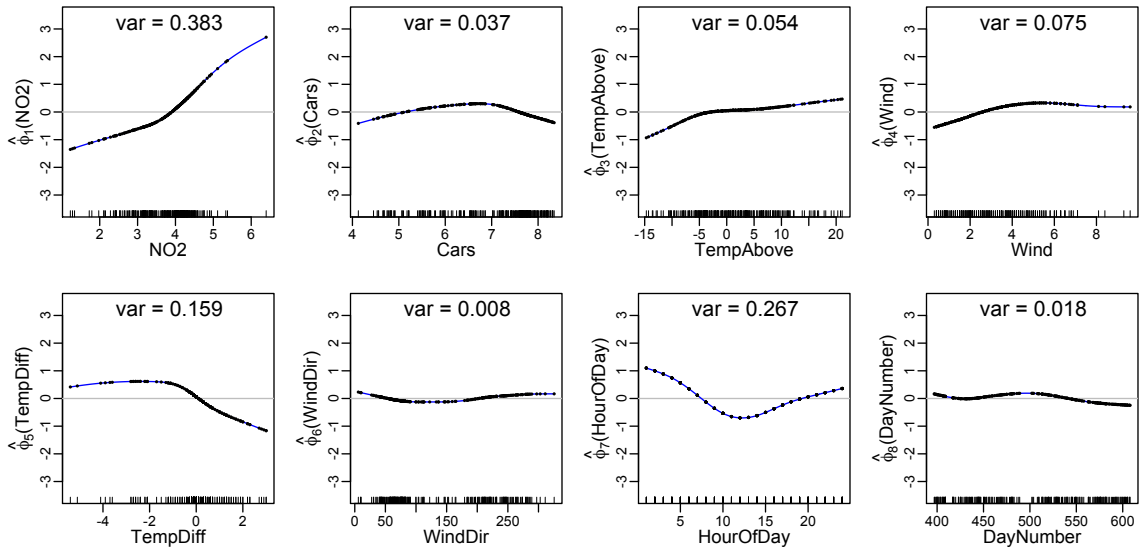


Figure 2.5: The third-smallest kernel APC transformations for the $NO_2$ data, using Sobolev kernel of order 2 for each variable. The eigenvalue for the APC is 0.189. The black bars at the bottom of each panel indicate the location of data points for that variable.

winter to early summer.

The response variable of interest in the original study, `NO2`, does not appear until the third-smallest kernel APC, shown in Figure 2.5. We have $\hat{\phi}_1(\texttt{NO2})+\hat{\phi}_5(\texttt{TempDiff})+\hat{\phi}_7(\texttt{HourOfDay}) \approx 0$. From the shape of $\hat{\phi}_7$ we see that the highest `NO2` occurs during lunch time, which makes sense as this is the time of greatest sun exposure. Note that surprisingly there is no interpretable association with `Cars` as its transformation has little variance and is not monotone (more cars should create more `NO2`). However, the strong association between `Cars` and `HourOfDay` in the smallest kernel APC creates an approximate non-identifiability between them, allowing `HourOfDay` to be a proxy for `Cars` in associations with other variables such as `NO2`. This explains the absence of association between `Cars` and `NO2`.

In summary, kernel APC analysis suggests a rich set of associations among the

variables. It also suggests that if an additive model had been fitted with NO2 as the response and all other variables as predictors, the estimated transforms of the predictors might suffer from interpretation problems due to the strong additive degeneracies discovered with the smallest and second-smallest kernel APCs.

## 2.10 Simulation

In this section, we evaluate the finite-sample performance of kernel APC on a simulated data for which the optimal transformations are known. We construct a simulated example consisting of four univariate random variables $X_1, \ldots, X_4$ with known population APC transformations $\phi_1(X_1), \ldots, \phi_4(X_4)$. This will be achieved by constructing them in such a way that the joint distribution of these transformations will be multivariate normal and highly collinear. The reason for this construction is that the extremal APCs of multivariate normal distributions are linear. (They also have APCs with non-extremal eigenvalues consisting of systems of Hermite polynomials; see Donnell et al. (1994).) This implies that if transformations $\phi_j(X_j)$ exist that result in a jointly multivariate normal distribution, they will constitute a population APC.

A simple procedure for simulating a situation with well-defined population APC is to first construct a multivariate normal distribution and transform its variables with the inverses of the desired transformations. APC estimation is then supposed to find approximations of these transformations from data simulated in this manner.

We start by constructing a multivariate normal distribution by using two independent variables $W_1, W_2 \sim \mathcal{N}(0, 1)$ to generate the underlying collinearity and four

independent variables $Z_1, Z_2, Z_3, Z_4 \sim \mathcal{N}(0, 0.1^2)$ to generate noise:

$$Y_1 = W_1 + Z_1, \qquad Y_2 = W_2 + Z_2, \qquad Y_3 = -W_1 - W_2 + Z_3, \qquad Y_4 = Z_4.$$

Thus the joint distribution features a collinearity of co-dimension 1 in the first three variables, and the fourth variable is independent of the rest. The correlation matrix of these four variables has a smallest eigenvalue of 0.007441113..., which will be the smallest population APC eigenvalue. The associated eigenvector is $(1/2, 1/2, 1/\sqrt{2}, 0)$, which indicates that the fourth transform will be zero, whereas the first three transforms will have variances 1/4, 1/4 and 1/2, respectively. The "observed" variables are constructed as marginal transformations $X_j = f_j(Y_j)$ using the following choices:

$$X_1 = \exp(Y_1), \qquad X_2 = -Y_2^{1/3}, \qquad X_3 = \exp(Y_3)/(1 + \exp(Y_3), \qquad X_4 = Y_4,$$

hence the APC transformations are

$$\phi_1^*(x) \sim \log(x), \qquad \phi_2^*(x) \sim -x^3, \qquad \phi_3^*(x) \sim \log(x/(1-x)), \qquad \phi_4^*(x) = 0.$$

As noted above the last transformation vanishes, and the other transformations are given only up to irrelevant additive constants as well as scales to achieve $\mathrm{Var}\,(\phi_1) = \mathrm{Var}\,(\phi_2) = 1/4$ and $\mathrm{Var}\,(\phi_3) = 1/2$.

Figure 2.6 shows the sample kernel APC for this data set ($n = 250$), with a common penalty parameter chosen by 5-fold cross-validation. As discussed at the end of Section 2.8.2, we standardized all variables to have unit variance before applying a standard Gaussian kernel $k(x, x') = \exp\{-\frac{1}{2}(x - x')^2\}$ for each variable $X_j$. The solid red line denotes the true transform $\phi_j^*$, while the dashed blue line denotes estimated transform $\hat{\phi}_j$. We see that for each variable, the two lines are almost indistinguishable,

though estimation performance worsens near the boundaries and on regions with few data points (location of data points are indicated by the black bars at the bottom of each plot). The transformed data points are centered to zero mean and normalized to $\sum_{j=1}^{4} \widehat{\mathrm{Var}}\left(\hat{\phi}_j\right) = 1$, so that $\widehat{\mathrm{Var}}\left(\hat{\phi}_j\right)$ indicates the relative importance of $\hat{\phi}_j$ in the estimated APCs. In fact, we see that $\widehat{\mathrm{Var}}\left(\hat{\phi}_j\right)$ is close to $\mathrm{Var}\left(Y_j\right)/[\sum_{i=1}^{4} \mathrm{Var}\left(Y_i\right)]$ in the data generating steps.

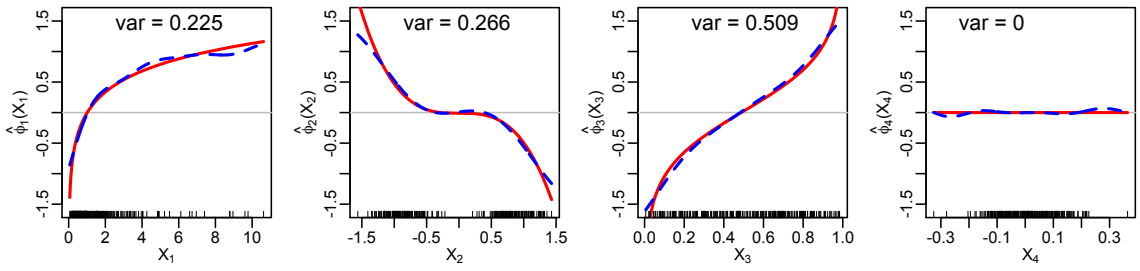


Figure 2.6: Plot of population APC transformations (——) and sample kernel APC transformations (– –). The eigenvalue for the sample kernel APC is 0.014. The black bars at the bottom of each panel indicate the location of data points for that variable.

## 2.11 Relation of APCs to Other Kernelized Multivariate Methods

### 2.11.1 Kernel PCA is NOT Kernel APC Analysis

Kernel principal component analysis (KPCA, Schölkopf et al. (1998), Schölkopf & Smola (2002)) is a well-known family of methods that begs the question of the relationship with kernel APC analysis. It can be shown (see, e.g., Appendix A.6) that, on a population level, the KPCA problem is equivalent to

$$\max_{\phi} \mathrm{Var}\left(\phi(X_1, \ldots, X_p)\right) \qquad \text{subject to} \qquad J(\phi) = 1. \tag{2.51}$$

In a (futile) attempt to reconstruct kernel APCs as a special case of KPCs, one would specialize $\phi$ to an additive functional form, $\phi(X_1, \ldots, X_p) = \sum \phi_j(X_j)$, and similarly for the penalty: $J(\phi) = \sum J_j(\phi_j)$. Thus the optimization problem (2.51) becomes

$$\max_{\phi_1,\ldots,\phi_p} \text{Var}\left(\sum \phi_j(X_j)\right) \qquad \text{subject to} \qquad \sum J_j(\phi_j) = 1. \tag{2.52}$$

Contrasting (2.52) with the kernel APC problem in (2.14), it becomes clear that additive KPCA and kernel APC analysis correspond to substantially different problems. Furthermore:

- As a maximization problem, (2.52) produces results that respond in an opaque way both to variance terms $\text{Var}(\phi_j)$ and to covariance terms $\text{Cov}(\phi_j, \phi_k)$. By comparison, kernel APCs are designed to respond solely to terms $\text{Cov}(\phi_j, \phi_k)$ and hence to association between variables alone.

- Converted to a minimization problem, (2.52) is meaningless because it is equivalent to *maximizing the penalty* $\sum J_j(\phi_j)$ subject to a constraint on the variance $\text{Var}\left(\sum \phi_j(X_j)\right)$. KPCA is intrinsically meaningful only for the upper end of the spectrum.

If the goal of PCA-related methods is to analyze associations among a set of variables, then kernel APC analysis represents a more limited yet more principled approach than KPCA. The limitations are due to APCs' focus on additivity, while a solid foundation for APCs is provided by the null comparison principle (Section 2.3). (We refer the reader to Appendix A.6 for further details on the comparison between KPCA and kernel APC analysis.)

### 2.11.2 Kernel CCA is a Special Case of Kernel APC Analysis

Although the focus on the lower end of the spectrum seems to have found little attention in the literature, the criterion we use for kernel APC can be related to existing proposals even if their focus is on the upper end of the spectrum. A special situation with precedent in the literature occurs for $p = 2$, in which case the kernel APC problem (2.14) reduces to the kernel canonical correlation analysis (CCA) problem discussed by Fukumizu et al. (2007). To see the equivalence, one may start with the simplified Rayleigh problem

$$\min/\max/\text{stationary}_{\phi_1,\phi_2} \ \frac{\text{Var}\,(\phi_1+\phi_2)+J_1(\phi_1)+J_2(\phi_2)}{\text{Var}\,(\phi_1)+\text{Var}\,(\phi_2)+J_1(\phi_1)+J_2(\phi_2)}. \tag{2.53}$$

It can be shown that stationary solutions satisfy

$$\text{Var}\,(\phi_1)+J_1(\phi_1) \ = \ \text{Var}\,(\phi_2)+J_2(\phi_2), \tag{2.54}$$

and it follows that the problem (2.53) is equivalent to

$$\min/\max/\text{stationary}_{\phi_1,\phi_2} \ \frac{\text{Cov}(\phi_1,\phi_2)}{\left(\text{Var}\,(\phi_1)+J_1(\phi_1)\right)^{1/2}\left(\text{Var}\,(\phi_2)+J_2(\phi_2)\right)^{1/2}},$$

where the normalization (2.54) can be enforced without loss of generality. This is recognized as a penalized form of CCA. It has been rediscovered several times over, in the machine learning literature by Bach & Jordan (2003), and earlier in the context of functional multivariate analysis by Leurgans et al. (1993).

Interesting is the work of Bach & Jordan (2003) which generalizes CCA to the case $p > 2$ but shows no interest in the results of such an analysis other than this becoming the building block in a method for independent components analysis (ICA), where the input variables $X_j$ are projections of multivariate data onto frames of orthogonal

unit vectors. Bach & Jordan (2003) correctly build up a finite-sample version of what amounts to APCs for $p > 2$ without a guiding principle other than the appearance of it being a "natural generalization". A population version and associated consistency theory is missing as their focus is on ICA and associated computational problems.

## 2.12 Concluding Remarks

APCs are a useful tool for exploring additive degeneracy in data. In this chapter, we propose the estimation of APCs using a regularization approach through kernelizing, and we establish the consistency of the resulting kernelized sample APCs. We also discuss computation of kernel APCs using power algorithm, and provide a theoretical justification for this.

It would be interesting to generalize our study of APCs in several directions. Due to the nonparametric nature of APC estimation, we have implicitly assumed that the sample size $n$ is large relative to the total number of variables $p$. It would be interesting to extend APCs to the high-dimensional setting where $p$ can be comparable to $n$. It would then be natural to impose additional structure such as sparsity in a flavor similar to the sparse additive models proposed by Ravikumar et al. (2009) in the regression framework. It would also be interesting to study the largest APCs and to examine whether it provides meaningful interpretation through dimensionality reduction as in conventional PCA.

Estimation of APCs is non-trivial due to its unsupervised learning nature. We have left open the problem of optimally and differentially select smoothing parameters for different variables within an APC and much less across different APCs, but this problem is unsolved even for additive regression, which is why such choices are usually made in terms of "degrees of freedom."

# Acknowledgements

We would like to acknowledge support for this project from the National Science Foundation under NSF grant DMS-1310795 to A.B. and NSF career grant DMS-1352060 to Z.M. as well as a grant from the Simons Foundation (SFARI) to A.B. We also thank Ming Yuan for valuable discussions.

# 3

# High-dimensional Robust Precision Matrix Estimation: Cellwise Corruption under $\epsilon$-Contamination*

## 3.1 Introduction

Covariance matrix estimation has long taken center stage in multivariate analysis (Anderson, 2003). The sample covariance estimator, which originates as the maximum likelihood estimator under a multivariate normal model, is optimal in many respects: It is unbiased, consistent, efficient under various distributional assumptions, and easily computable. Despite its positive traits, however, the sample covariance matrix is also highly non-robust when data are contaminated. Hence, various procedures in robust statistics have been derived to obtain a covariance matrix estimator that behaves well even in the presence of contaminated data (Huber, 1981; Hampel et al., 2011).

In other areas of multivariate analysis, the precision matrix $\mathbf{\Omega}^* := (\mathbf{\Sigma}^*)^{-1}$ is

*Joint work with Po-Ling Loh

of significant interest. Examples include computing Mahalanobis distances, linear discriminant analysis, and Gaussian graphical models. In the setting of graphical models, a random vector $\mathbf{X}$ is associated with an undirected graph $G = (V, E)$ that encodes conditional independence relations between components of $\mathbf{X}$ (Lauritzen, 1996). The vertex set $V$ contains $\{1, \ldots, p\}$, while the edge set $E$ consists of pairs $(i, j)$, where $(i, j) \in E$ if $X_i$ and $X_j$ are connected by an edge. For each non-edge $(i, j) \notin E$, the variables $X_i$ and $X_j$ are conditionally independent given all other variables. When $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$, pairwise conditional independence holds if and only if $\boldsymbol{\Omega}^*_{ij} = 0$. Thus, recovering the support of the precision matrix is equivalent to graphical model selection. The aforementioned observations have been used for network reconstruction in many scientific fields, including genetics and neuroscience (e.g., see Werhli et al. (2006); Smith et al. (2011) and the references cited therein). When the dimensionality $p$ is small compared to the number of samples $n$, a reasonable method for robust precision matrix estimation could consist of computing a robust estimate of the covariance matrix and then taking a matrix inverse.

With the recent deluge of high-dimensional data, however, a need has arisen to devise high-dimensional analogs of classical procedures that are both computable and possess rigorous theoretical guarantees. Although several methods, notably the graphical Lasso (GLasso) (Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) and the constrained $\ell_1$-minimization for inverse matrix estimation (CLIME) (Cai et al., 2011) estimator, have been proposed for high-dimensional precision matrix estimation, robust estimation of high-dimensional precision matrices has only recently emerged in the literature. The GLasso and CLIME estimators tend to perform poorly under contaminated data, since they take as input the sample covariance matrix that is sensitive to even a single outlier.

Popular classical robust covariance estimators are applicable in settings where less

than half the observation vectors are contaminated. Such an assumption is closely connected to the Tukey-Huber contamination model that underlies much of the existing robustness theory (Tukey, 1962; Huber, 1964). In the Tukey-Huber model, a mixture distribution with a dominant nominal component (such as a multivariate normal distribution) and a minority unspecified component are posited, and each observation vector is either completely clean or completely contaminated. Classical robust covariance estimators then involve downweighting contaminated observations in order to reduce their influence. When the dimension $p$ is large, however, the fraction of perfectly observed data vectors may be rather small: If all components of an observation vector had an independent chance of being contaminated, most observation vectors would be contaminated. Thus, downweighting an entire observation would waste the information contained in the clean components of the observation vector. This describes the setting of the *cellwise* contamination model, which was developed by Alqallaf et al. (2002). It generalizes the classical Tukey-Huber contamination model, which may be viewed as a case of *rowwise* contamination of the data matrix, and is fairly realistic for applications involving measurement error in DNA microarray analysis (Troyanskaya et al., 2001) or dropout measurements in sensor arrays (Swanson, 2000).

On the other hand, most existing approaches for robust covariance estimation focus on affine equivariance. These include the $M$-estimators (Maronna, 1976), Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) estimators (Rousseeuw, 1984, 1985), and Stahel-Donoho (SD) estimator (Stahel, 1981; Donoho, 1982). Although affine equivariance may be a desirable property under rowwise contamination, it is less appropriate in the setting of cellwise contamination, since linear combinations of observation vectors lead to a propagation of outliers (Alqallaf et al., 2009). In addition, the MVE, MCD, and SD estimators all require heavy

computational effort, rendering them impractical for high-dimensional datasets. To deal with cellwise contamination, Van Aelst (2016) proposed a modified SD estimator that adapts winsorization (Huber, 1981; Alqallaf et al., 2002) and a cellwise weighting scheme. Similar to the original SD estimator, however, computation is only feasible for small $p$. A recent approach by Agostinelli et al. (2015) is capable of dealing with both rowwise and cellwise outliers. The procedure consists of two steps: (1) flagging cellwise outliers as missing values; and (2) applying a rowwise robust method to the incomplete data. However, computation is again infeasible in high dimensions. Other recent proposals for robust high-dimensional covariance matrix estimation include those suggested by Chen et al. (2015) and Han et al. (2015), but both methods treat different contamination models and are not suitable to handle data with cellwise contamination: Han et al. (2015) study robust high-dimensional scatter matrix estimation when data are drawn from heavy-tailed distributions, and Chen et al. (2015) study a method based on "matrix depth" designed for handling rowwise contamination that is computationally intractable in high dimensions. However, note that our proposed estimators are computationally feasible.

In fact, relatively few approaches exist for robust high-dimensional precision matrix estimation under any form of contamination. One method is supplied by the TLasso estimator of Finegold & Drton (2011), which builds upon the GLasso and models the data as coming from the multivariate $t$-distribution, a long-tailed surrogate for the multivariate normal distribution. The "alternative multivariate $t$-distribution" is used to model a case where different coordinates of the distribution are obtained from the latent multivariate normal distribution using different weights. Although the TLasso demonstrates a higher degree of robustness than the GLasso under both rowwise and cellwise contamination in simulations, however, a theoretical analysis from the point of view of robust statistics has not been derived.

More recently, Oellerer & Croux (2014) and Tarr et al. (2015) propose a promising new method for high-dimensional precision matrix estimation, designed specifically for cellwise contamination. The method consists of combining a robust covariance estimator that may be computed efficiently with a suitable high-dimensional precision matrix estimation procedure. Similar plug-in estimators based on rank-based correlation matrix estimates were previously proposed by Liu et al. (2012) and Xue & Zou (2012) for model selection and parameter estimation in nonparanormal graphical models. However, a significant difference is that Liu et al. (2012) and Xue & Zou (2012) focus on establishing consistency *when the observations are drawn cleanly* from a nonparanormal model. Other follow-up work (Han & Liu, 2013, 2014; Fan et al., 2014, 2015; Wegkamp & Zhao, 2016) again focuses on establishing statistical consistency under transformational or heavy-tailed variants of the high-dimensional Gaussian model. In contrast, Oellerer & Croux (2014) and Tarr et al. (2015) study the behavior of robust estimators when a fraction of the data are contaminated, which is also the focus of this chapter. However, a rigorous high-dimensional analysis from the point of view of statistical consistency is absent from this line of work.

Our main contributions are to derive statistical error bounds in elementwise $\ell_\infty$-norm for robust precision matrix estimation procedures according to the proposals of Oellerer & Croux (2014) and Tarr et al. (2015). We study the setting of the cellwise $\epsilon$-contamination model, where at most an $\epsilon$ fraction of entries in the data matrix are corrupted by outliers. Our work thus fuses two threads of research involving classical robust statistics and high-dimensional estimation in a novel and rigorous manner. The bounds we derive match standard high-dimensional bounds for uncontaminated precision matrix estimation, up to a constant multiple of $\epsilon$. Furthermore, they are of a complementary nature to the theoretical results supplied by Oellerer & Croux (2014), since we are primarily concerned with robustness as measured from the viewpoint of

statistical consistency, rather than breakdown behavior.

More generally, our results reveal an interesting interplay between bounds for statistical error under $\epsilon$-contamination and classical measures of robustness such as the influence function (Hampel, 1974) and breakdown point (Donoho & Huber, 1983). Estimators with bounded influence have long been favored in classical robust statistics, as the rate of change in the statistical functional associated with the estimator is controlled when the nominal distribution is contaminated by an arbitrary point mass distribution. Our results show that a variety of bounded influence estimators, including Kendall's and Spearman's correlation coefficients, give rise to (inverse) covariance estimators with statistical error rates that depend linearly on the degree of contamination; the converse relationship may be seen to hold more generally as a result of our proof arguments. On the other hand, our discussion of the breakdown point of the precision matrix estimators, building upon the analysis of Oellerer & Croux (2014), emphasizes the significant differences between the notions of breakdown point and statistical consistency. Whereas our analysis shows that the robust CLIME and GLasso procedures have comparable behavior from the point of view of high-dimensional statistical consistency, the CLIME estimator has a substantially smaller breakdown point than the GLasso, due to its constrained feasibility region. Rather than advocating one measure of robustness over another, our discussion emphasizes the value of weighing different measures of robustness in selecting an appropriate estimator.

The remainder of this chapter is organized as follows: Section 3.2 furnishes the mathematical background for the cellwise contamination model and the robust covariance and precision matrix estimators to be considered in this chapter. Section 3.3 presents our main theoretical contributions, providing bounds on the statistical error of the covariance and precision matrix estimators under the cellwise contamination model, as well as concrete consequences in the presence of outliers and/or missing

data. Section 3.4 provides a discussion of the breakdown point for the robust GLasso and CLIME estimators. Section 3.5 contains simulation results that are used to validate the theoretical results of this chapter. We conclude with a discussion in Section 3.6, including some avenues for future research. The proof of main results in this chapter are relegated to Appendix B.

**Notation:** For a vector $\mathbf{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, we denote by $\|\mathbf{a}\|_1 = \sum_{i=1}^p |a_i|$ and $\|\mathbf{a}\|_2 = (\sum_{i=1}^p a_i^2)^{1/2}$ the $\ell_1$-norm and $\ell_2$-norm of $\mathbf{a}$, respectively. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we define the elementwise $\ell_1$-norm $\|\mathbf{A}\|_1 = \sum_{i=1}^p \sum_{j=1}^q |a_{ij}|$, the Frobenius norm $\|\mathbf{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^q a_{ij}^2)^{1/2}$, the elementwise $\ell_\infty$-norm $\|\mathbf{A}\|_\infty = \max_{1\le i\le p, 1\le j\le q} |a_{ij}|$, the spectral norm $\|\mathbf{A}\|_2 = \sup_{\|x\|\le 1} \|\mathbf{A}\mathbf{x}\|_2$, the matrix $\ell_1$-norm $\|\mathbf{A}\|_{L_1} = \max_{1\le j\le q} \sum_{i=1}^p |a_{ij}|$. We use $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots \ge \lambda_p(\mathbf{A})$ to denote the ordered eigenvalues of $\mathbf{A}$, and we write $\mathbf{A} \succ 0$ (respectively, $\mathbf{A} \succeq 0$) to indicate that $\mathbf{A}$ is positive definite (respectively, positive semidefinite). We write $\mathbf{I}$ for the identity matrix and $\mathbf{0}$ for the vector of all zeros (the respective dimension of which will be clear from context). The binary operation $\otimes$ denotes the tensor product.

## 3.2 Background and Problem Setup

We begin with a description of the cellwise contamination model, followed by a rigorous formulation of the robust covariance and precision matrix estimators to be studied in this chapter.

Following the notation of Alqallaf et al. (2002, 2009), we write the cellwise contamination model in the following form:

$$\mathbf{X}_k = (\mathbf{I} - \mathbf{B}_k)\mathbf{Y}_k + \mathbf{B}_k\mathbf{Z}_k, \qquad \forall k = 1, \ldots, n. \tag{3.1}$$

Here, we observe the contaminated random vector $\mathbf{X}_k \in \mathbb{R}^p$. The unobservable random vectors $\mathbf{Y}_k, \mathbf{Z}_k$, and $\mathbf{B}_k$ are independent, and $\mathbf{Y}_k \sim G$ (a nominal distribution) and $\mathbf{Z}_k \sim H^*$ (an unspecified outlier generating distribution). Furthermore, $\mathbf{B}_k = \text{diag}(B_{k1}, \ldots, B_{kp})$ is a diagonal matrix, where $B_{k1}, \ldots, B_{kp}$ are independent Bernoulli random variables with $P(B_{ki} = 1) = \epsilon_i$, for all $1 \leq i \leq p$.

When $\epsilon_1 = \cdots = \epsilon_p = \epsilon$, the probability of an observation vector having no contamination in any component is $(1-\epsilon)^p$, a quantity that decreases exponentially as the dimension increases. This probability goes below the critical value $1/2$ for $p \geq 14$ at $\epsilon = 0.05$, and for $p \geq 69$ at $\epsilon = 0.01$. Equation (3.1) is a special case of a more general model, where we allow other joint distributions for $B_{k1}, \ldots, B_{kp}$. For instance, if $B_{k1}, \ldots, B_{kp}$ were completely dependent (i.e., $P(B_{k1} = \cdots = B_{kp}) = 1$), we would obtain the rowwise contamination model. In that case, the probability of an observation vector being totally free of contamination would be $1 - \epsilon$, which is independent of the dimension. Alqallaf et al. (2009) also use the terms *fully independent contamination model (FICM)* and *fully dependent contamination model (FDCM)* to denote the cellwise and rowwise contamination settings, in order to distinguish the pattern of contamination across rows of the data matrix.

Throughout, we will work under the cellwise contamination model (3.1), and assume that $G$ is a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$. Our goal is to estimate the matrices $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Omega}^* = (\boldsymbol{\Sigma}^*)^{-1}$ from the (uncontaminated) normal component.

### 3.2.1 Covariance Matrix Estimation

When $\epsilon = 0$ (i.e., the data are uncontaminated), we may use the classical sample covariance matrix estimator $\tilde{\boldsymbol{\Sigma}}$, defined pairwise as

$$\tilde{\boldsymbol{\Sigma}}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j), \qquad \forall 1 \leq i, j \leq p,$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ki}$ and $\bar{X}_j = \frac{1}{n}\sum_{k=1}^{n} X_{kj}$. When $n \gg p$, the sample covariance is an efficient estimator for $\mathbf{\Sigma}^*$. However, when $\epsilon > 0$, the performance of $\tilde{\mathbf{\Sigma}}$ may be compromised depending on the properties of $H^*$: Under the cellwise contamination model, for $i \neq j$, we have

$$\begin{aligned}(\mathbf{\Sigma}_X^*)_{ij} &= (1-\epsilon_i)(1-\epsilon_j)\,(\mathbf{\Sigma}_Y^*)_{ij} + \epsilon_i\epsilon_j\,(\mathbf{\Sigma}_Z^*)_{ij} \\ &= (\mathbf{\Sigma}_Y^*)_{ij} - (\epsilon_i + \epsilon_j - \epsilon_i\epsilon_j)\,(\mathbf{\Sigma}_Y^*)_{ij} + \epsilon_i\epsilon_j\,(\mathbf{\Sigma}_Z^*)_{ij}\,.\end{aligned}$$

When no restrictions are placed on the covariance $\mathbf{\Sigma}_Z^*$ of the contaminating distribution, the elementwise deviations between $\mathbf{\Sigma}_X^*$ and $\mathbf{\Sigma}_Y^*$ (and consequently, also the sample covariance $\tilde{\mathbf{\Sigma}}_X := \tilde{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}_Y^*$) will in general behave arbitrary badly. Furthermore, note that even when $\mathbf{\Sigma}_Z^*$ is constrained to lie in a space where the deviations between $\mathbf{\Sigma}_X^*$ and $\mathbf{\Sigma}_Y^*$ are suitably bounded, we would require the contaminating distribution to have properties such as sub-Gaussian tails in order to ensure consistency of the sample covariance estimator on the order of $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$. When a procedure based on covariance estimation is used to estimate the precision matrix, the errors incurred during the covariance estimation step would propagate to the next step. For instance, this issue would arise in using the CLIME or GLasso estimator. In contrast, our theory for robust covariance estimators will not require any assumptions on either $\mathbf{\Sigma}_Z^*$ or the tail behavior of the contaminating distribution.

To deal with cellwise contamination in the high-dimensional setting, we therefore take the pairwise approach suggested by Oellerer & Croux (2014), where a robust covariance or correlation estimate is computed for each pair of variables. Early proposals of robust procedures are of this type (Bickel, 1964; Puri & Sen, 1971), where a coordinatewise approach is taken for robust estimation of location. In addition to having relatively low computational complexity, the pairwise approach is appealing because a high breakdown point of the pairwise estimators translates into a high

breakdown point of the overall covariance matrix. For $1 \leq i, j \leq p$, we write

$$\boldsymbol{\Sigma}^*_{ij} = \sigma_i \sigma_j \boldsymbol{\rho}_{ij}, \tag{3.2}$$

where $\sigma_i = [\mathrm{Var}\,(X_{ki})]^{1/2}$, $\sigma_j = [\mathrm{Var}\,(X_{kj})]^{1/2}$, and $\boldsymbol{\rho}_{ij} = \mathrm{Corr}\,(X_{ki}, X_{kj})$. We will take suitable robust estimators of $\hat{\sigma}_i$, $\hat{\sigma}_j$, and $\hat{\boldsymbol{\rho}}_{ij}$, to obtain the covariance matrix estimator $\hat{\boldsymbol{\Sigma}}$, with $(i, j)$ entry $\hat{\boldsymbol{\Sigma}}_{ij} = \hat{\sigma}_i \hat{\sigma}_j \hat{\boldsymbol{\rho}}_{ij}$.

To estimate $\sigma_i$, we consider the median absolute deviation from the median (MAD), a robust measure of scale. The MAD estimator was popularized by Hampel (1974), who attributes the concept to Gauss. It has a breakdown point of 50%. Let $X_{(1),i} \leq \cdots \leq X_{(n),i}$ denote the ordered values of $X_{1i}, \ldots, X_{ni}$. The sample median $\hat{m}_i$ and the sample MAD $\hat{d}_i$ are defined, respectively, as $\hat{m}_i = X_{(k^*),i}$ and $\hat{d}_i = W_{(k^*),i}$, where $W_{ki} = |X_{ki} - \hat{m}_i|$, for all $k = 1, \ldots, n$, and $k^* = \lceil n/2 \rceil$. Expressed another way,

$$\hat{d}_i = \underset{1 \leq k \leq n}{\mathrm{median}} \left( \left| X_{ki} - \underset{1 \leq \ell \leq n}{\mathrm{median}} (X_{\ell i}) \right| \right). \tag{3.3}$$

We then estimate $\sigma_i$ by $\hat{\sigma}_i = [\Phi^{-1}(0.75)]^{-1} \hat{d}_i$, where the constant $[\Phi^{-1}(0.75)]^{-1}$ is chosen in order to make the estimator consistent for $\sigma_i$ at normal distribution. The population-level median of a distribution with cdf $F$ is defined to be $m(F) := F^{-1}(0.5)$, where $F^{-1}(c) = \inf\{x : F(x) \geq c\}$, for $c \in [0, 1]$. Similarly, we may define the population-level MAD $d(F)$ to be the median of the distribution of $|X - m(F)|$, where $X$ has cdf $F$.

To estimate $\boldsymbol{\rho}_{ij}$, we consider the classical nonparametric correlation estimators, Kendall's tau and Spearman's rho:

**Kendall's tau** This statistic is given by

$$\boldsymbol{r}_{ij}^K = \frac{2}{n(n-1)} \sum_{k<\ell} \text{sign}(X_{ki} - X_{\ell i})\text{sign}(X_{kj} - X_{\ell j}), \tag{3.4}$$

where $\text{sign}(X) = 1$ if $X > 0$, $\text{sign}(X) = -1$ if $X < 0$, and $\text{sign}(0) = 0$.

**Spearman's rho** This statistic is given by

$$\boldsymbol{r}_{ij}^S = \frac{\sum_{k=1}^n [\text{rank}(X_{ki}) - (n+1)/2][\text{rank}(X_{kj}) - (n+1)/2]}{\sqrt{\sum_{k=1}^n [\text{rank}(X_{ki}) - (n+1)/2]^2 \sum_{k=1}^n [\text{rank}(X_{kj}) - (n+1)/2]^2}}, \tag{3.5}$$

where $\text{rank}(X_{ki})$ denotes the rank of $X_{ki}$ among $X_{1i}, \ldots, X_{ni}$.

The population versions of the estimators are given, respectively, by

$$\boldsymbol{\rho}_{ij}^K = E[\text{sign}(X_{1i} - X_{2i})\text{sign}(X_{1j} - X_{2j})], \tag{3.6a}$$

$$\boldsymbol{\rho}_{ij}^S = 3E[\text{sign}(X_{1i} - X_{2i})\text{sign}(X_{1j} - X_{3j})]. \tag{3.6b}$$

When $\epsilon_1 = \cdots = \epsilon_p = 0$, we have $\mathbf{X}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$; in this case, it is known that (Kendall, 1948; Kruskal, 1958)

$$\boldsymbol{\rho}_{ij} = \sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right) = 2\sin\left(\frac{\pi}{6}\boldsymbol{\rho}_{ij}^S\right).$$

Hence, for asymptotic consistency at normal distribution, our estimator for $\boldsymbol{\rho}_{ij}$ is the transformed version of Kendall's tau and Spearman's rho, given by $\sin(\frac{\pi}{2}\boldsymbol{r}_{ij}^K)$ and $2\sin(\frac{\pi}{6}\boldsymbol{r}_{ij}^S)$, respectively. We then define as $\hat{\boldsymbol{\Sigma}}$ our robust covariance matrix estimator, with

$$\hat{\boldsymbol{\Sigma}}_{ij}^K = \hat{\sigma}_i\hat{\sigma}_j \sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right), \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}}_{ij}^S = 2\hat{\sigma}_i\hat{\sigma}_j \sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right). \tag{3.7}$$

### 3.2.2 Precision Matrix Estimation

A long line of literature exists for precision matrix estimation in the high-dimensional setting. We will focus our attention on sparse precision matrix estimation; i.e., $\mathbf{\Omega}^*$ contains many zero entries. In this section, we review two techniques, the GLasso and CLIME, which produce a sparse precision matrix estimator based on optimizing a function of the sample covariance matrix. As proposed by Oellerer & Croux (2014) and Tarr et al. (2015), these methods may easily be modified to obtain robust versions, where the sample covariance matrix estimator is simply replaced by a robust covariance estimator $\hat{\mathbf{\Sigma}}$ as described in the previous section.

The graphical lasso (GLasso) estimator (Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) is defined as the maximizer of the following function:

$$\tilde{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succ 0}{\operatorname{argmin}} \left\{ \operatorname{tr}(\tilde{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \|\mathbf{\Omega}\|_1 \right\}.$$

Here, $\lambda > 0$ is a tuning parameter that controls the sparsity of the resulting precision matrix estimator.

In this chapter, we replace the sample covariance matrix $\tilde{\mathbf{\Sigma}}$ by the robust alternative $\hat{\mathbf{\Sigma}}$, and consider a variant where only the off-diagonal entries of the estimator are penalized:

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succ 0}{\operatorname{argmin}} \left\{ \operatorname{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \|\mathbf{\Omega}\|_{1,\text{off}} \right\}. \tag{3.8}$$

Note that although the program (3.8) is convex for any choice of $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$, several state-of-the-art algorithms for optimizing the GLasso require the matrix $\hat{\mathbf{\Sigma}}$ to be positive semidefinite (Friedman et al., 2008; Zhao et al., 2012; Hsieh et al., 2011). We will first derive statistical theory for the robust GLasso without a positive semidefinite projection step, and then discuss properties of the projected version in Section 3.4.

A popular alternative to the GLasso is the method of constrained $\ell_1$-minimization

for inverse matrix estimation (CLIME) proposed in Cai et al. (2011). The CLIME routine solves the following convex optimization problem by linear programming:

$$\tilde{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} \|\mathbf{\Omega}\|_1 \qquad \text{subject to} \qquad \|\tilde{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}\|_\infty \leq \lambda.$$

Note that here, no symmetry condition is imposed on $\mathbf{\Omega}$, and the solution is not symmetric in general. If a symmetric precision matrix estimate is desired, we may perform a post-symmetrization step on $\tilde{\mathbf{\Omega}} = (\tilde{\omega}^1_{ij})$ to obtain the symmetric matrix $\tilde{\mathbf{\Omega}}_{\text{sym}}$, defined by

$$\begin{aligned} &\tilde{\mathbf{\Omega}}_{\text{sym}} = (\tilde{\omega}_{ij}), \qquad \text{where} \\ &\tilde{\omega}_{ij} = \tilde{\omega}_{ji} = \tilde{\omega}^1_{ij} \mathbb{1}(|\tilde{\omega}^1_{ij}| \leq |\tilde{\omega}^1_{ji}|) + \tilde{\omega}^1_{ji} \mathbb{1}(|\tilde{\omega}^1_{ij}| > |\tilde{\omega}^1_{ji}|). \end{aligned} \tag{3.9}$$

In other words, between $\tilde{\omega}^1_{ij}$ and $\tilde{\omega}^1_{ji}$, we pick the entry with smaller magnitude. Similar to the GLasso case, we will robustify the CLIME estimator by solving

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} \|\mathbf{\Omega}\|_1 \qquad \text{subject to} \qquad \|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}\|_\infty \leq \lambda, \tag{3.10}$$

and then apply post-symmetrization (3.9) to obtain the robust CLIME estimator $\hat{\mathbf{\Omega}}_{\text{sym}}$.

We remark that the same estimators (3.8) and (3.10), based on plugging in a robust rank-based surrogate of the correlation matrix, also appeared in Liu et al. (2012) and Xue & Zou (2012). However, the focus of both papers was to derive consistency of the estimators under a nonparanormal model, rather than quantifying the effect of deviations from normality, which is the primary objective of the present chapter.

## 3.3 Main Results and Consequences

We now provide rigorous statements of the main results of this chapter. We first derive bounds for robust covariance matrix estimation, which are used to obtain bounds on the error incurred by the precision matrix estimator. Note, however, that the statistical error bounds presented in Section 3.3.1 are of independent interest; we believe they are the first bounds appearing in the literature that quantify the robustness of covariance matrix estimators under a cellwise contamination model.

### 3.3.1 Covariance Matrix Estimation

Throughout this section, we will assume that the standard deviations of the uncontaminated distributions are bounded as follows:

$$0 < \min_{1 \leq i \leq p} \sigma_i \leq \max_{1 \leq i \leq p} \sigma_i \leq M_\sigma. \tag{3.11}$$

We also define the expression

$$c(\sigma_i) = \frac{15}{64\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(1.1\sigma_i + 0.5)^2}{2\sigma_i^2}\right), \qquad \forall 1 \leq i \leq p. \tag{3.12}$$

Our first theorem provides a bound on the statistical error of the robust covariance estimator $\hat{\mathbf{\Sigma}}^K$ based on Kendall's tau correlations. Note that our result does *not* involve any assumptions on the contaminating distribution $H$. Thus, the distribution $H$ may contain point masses, and we do not require a probability density function of $H$ to even exist.

**Theorem 4.** *Under the cellwise contamination model* (3.1)*, suppose inequality* (3.11) *is satisfied, and* $\epsilon = \max_{1 \leq i \leq p} \epsilon_i \leq 0.02$*. Let* $C > \pi\sqrt{2}$ *and* $C' > \frac{1}{\Phi^{-1}(0.75) \min_{1 \leq i \leq p} c(\sigma_i)\sqrt{2}}$*,*

*and suppose*

$$\max\left\{C\sqrt{\frac{\log p}{n}}+26\pi\epsilon,\ C'\sqrt{\frac{\log p}{n}}+7.2M_\sigma\epsilon\right\}\leq 1, \tag{3.13}$$

*and* $\Phi^{-1}(0.75)C'\sqrt{\frac{\log p}{n}}<1$. *Then with probability at least*

$$1-2p^{-\left(\frac{C^2}{\pi^2}-2\right)}-6p^{-\{2[\Phi^{-1}(0.75)]^2C'^2\min_{1\leq i\leq p}c^2(\sigma_i)-1\}},$$

*the robust covariance estimator satisfies*

$$\begin{aligned}\left\|\hat{\mathbf{\Sigma}}^K-\mathbf{\Sigma}^*\right\|_\infty\leq&\left(C(M_\sigma^2+M_\sigma+1)+C'(2M_\sigma+1)\right)\sqrt{\frac{\log p}{n}}\\&+\left(97M_\sigma^2+89M_\sigma+82\right)\epsilon.\end{aligned} \tag{3.14}$$

The proof of Theorem 4 is provided in Section B.1.1.

**Remark 6.** *Theorem 4 clearly illustrates the effect of $\epsilon$-contamination on the estimation error of the covariance matrix estimator. Note that when $\epsilon=0$, we recover the minimax optimal rate for covariance matrix estimation in $\ell_\infty$-norm (Cai & Zhou, 2012); although the estimator $\hat{\mathbf{\Sigma}}^K$ is not equal to the sample covariance estimator in the uncontaminated case, the robust covariance estimator nonetheless converges to the true covariance matrix at the optimal rate. On the other hand, cellwise contamination introduces an extra term that is linear in $\epsilon$.*

*Another way to interpret the bound* (3.14) *is that if the level of contamination $\epsilon$ is bounded by a constant times $\sqrt{\frac{\log p}{n}}$, then the robust covariance estimator $\hat{\mathbf{\Omega}}^K$ will enjoy the same statistical error rate as the optimal covariance estimator in the uncontaminated case. As we will see in Theorems 6 and 7 below, the sample size requirements for precision matrix estimation are such that the condition $\epsilon\leq C\sqrt{\frac{\log p}{n}}$ still allows for a nonvanishing fraction of contamination. Furthermore, note that*

*although the restriction* $\epsilon \leq 0.02$ *may seem somewhat prohibitive, the proof of Theorem 4 reveals that the specific bound on* $\epsilon$ *is an artifact of the proof technique, and a more careful analysis would allow for a larger degree of contamination, at the expense of slightly looser constants in the covariance estimation bound* (3.14)*, as long as* $\epsilon$ *is bounded by some constant in* $[0, 1)$*.*

The following theorem is an analog of Theorem 4, derived for the robust covariance estimator $\hat{\mathbf{\Sigma}}^S$ based on Spearman's correlation coefficient. We assume that the ranks of variables between samples are distinct; note that this happens almost surely when the contaminating distribution has continuous density. The proof of Theorem 5 is provided in Section B.1.2.

**Theorem 5.** *Under the cellwise contamination model* (3.1)*, suppose the variable ranks are distinct. Also suppose inequality* (3.11) *is satisfied and* $\epsilon = \max_{1 \leq i \leq p} \epsilon_i \leq 0.01$*. Let* $C > 8\pi$ *and* $C' > \frac{1}{\Phi^{-1}(0.75) \min_{1 \leq i \leq p} c(\sigma_i)\sqrt{2}}$*, and suppose*

$$\max\left\{\frac{5C}{2}\sqrt{\frac{\log p}{n}} + 51\pi\epsilon,\ C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right\} \leq 1,$$

*and the sample size satisfies* $\Phi^{-1}(0.75)C'\sqrt{\frac{\log p}{n}} < 1$ *and* $n \geq \max\left\{15,\ \frac{16\pi^2}{C^2 \log p}\right\}$*. Then with probability at least*

$$1 - 2p^{-\left(\frac{C^2}{32\pi^2} - 2\right)} - 6p^{-\{2[\Phi^{-1}(0.75)]^2 C'^2 \min_{1 \leq i \leq p} c^2(\sigma_i) - 1\}},$$

*the robust covariance estimator satisfies*

$$\left\|\hat{\mathbf{\Sigma}}^S - \mathbf{\Sigma}^*\right\|_\infty \leq \left(\frac{5C}{2}(M_\sigma^2 + M_\sigma + 1) + C'(2M_\sigma + 1)\right)\sqrt{\frac{\log p}{n}} + \left(175M_\sigma^2 + 168M_\sigma + 161\right)\epsilon. \tag{3.15}$$

**Remark 7.** *The conclusion of Theorem 5 is very similar to that of Theorem 4, except for constants and an additional requirement on the size of* $n$*. However, note that when* $\frac{\log p}{n} = o(1)$*, implying the statistical consistency of the robust covariance estimator, the requirement* $n \geq \max\left\{15, \frac{16\pi^2}{C^2 \log p}\right\}$ *is essentially extraneous.*

Although the high-dimensional error bounds derived in Theorems 4 and 5 are substantially different from the canonical measures analyzed in the robust statistics literature, our bounds are somewhat related to the notion of the influence function of an estimator. The influence function (Hampel, 1974), defined at the population level, measures the infinitesimal change incurred by the statistical functional associated with an estimator when the underlying distribution is contaminated by a point mass. Thus, an estimator has a bounded influence function if the extent of the deviation in its functional representation due to contamination remains bounded, regardless of the location of the point mass. The error bounds (3.14) and (3.15) also reveal that the extent to which the error deviation between the robust covariance estimator and the true covariance grows is bounded by a constant depending only on $M_\sigma$. The two notions do not match precisely; for instance, our theorems allow contamination by an arbitrary distribution rather than simply a point mass, and we are comparing finite-sample deviations of an estimator from $\mathbf{\Sigma}^*$ rather than population-level deviations of a statistical functional under a contaminated distribution. However, note that by sending $n \to \infty$ in the finite-sample bounds and taking the contaminating distribution to be a point mass, we may conclude that the influence function of the robust covariance estimator is bounded when deviations are measured in the elementwise $\ell_\infty$-sense. Furthermore, the arguments in our proofs (cf. Lemmas 24 and 25 in Appendix B.4) may be used to derive the fact that the corresponding correlation estimators have a bounded influence function, the precise forms of which appear in Croux & Dehon (2010). The reverse implication, that a correlation estimator with

bounded influence (together with a bounded-influence scale estimator) gives rise to high-dimensional deviation bounds of the form in inequalities (3.14) and (3.15), is elaborated upon in Section 3.3.4 below.

Finally, note that although Theorems 4 and 5 have been derived under the assumption that the uncontaminated data follow a normal distribution, the same proof techniques may be applied to settings where the uncontaminated data are drawn from a different underlying distribution, as long as the uncontaminated distribution is suitably well-behaved. Since our primary goal is precision matrix estimation, we have focused only on the scenario where the uncontaminated data are drawn from a Gaussian distribution, in which case the structure of the precision matrix is of great interest in the statistical community.

### 3.3.2 Precision Matrix Estimation

Using the novel statistical error bounds derived in the previous section, we now provide statistical error bounds on the precision matrix estimators attained by plugging the robust covariance matrix estimates into the CLIME and GLasso. We provide explicit statements in the case of the covariance estimate based on Kendall's tau; analogous statements hold for Spearman's rho, assuming unique ranks.

We begin with the CLIME estimator. Consider the following uniformity class of matrices:

$$\mathcal{U}(q, s_0(p), M) = \left\{ \mathbf{\Omega} : \mathbf{\Omega} \succ 0, \|\mathbf{\Omega}\|_{L_1} \leq M, \max_{1 \leq i \leq p} \sum_{j=1}^{n} |\omega_{ij}|^q \leq s_0(p) \right\}, \tag{3.16}$$

for $0 \leq q < 1$, where $\mathbf{\Omega} := (\omega_{ij}) = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p)$. The following result provides an elementwise error bound on the estimation error between the CLIME output and the true precision matrix, provided the true precision matrix lies in the class (3.16)

defined above:

**Theorem 6.** *Under the cellwise contamination model* (3.1)*, suppose inequality* (3.11) *is satisfied, and* $\epsilon = \max_{1\le i\le p} \epsilon_i \le 0.02$*. Let* $C > \pi\sqrt{2}$ *and* $C' > \frac{1}{\Phi^{-1}(0.75)\min_{1\le i\le p} c(\sigma_i)\sqrt{2}}$*, and suppose inequality* (3.13) *also holds and* $\Phi^{-1}(0.75)C'\sqrt{\frac{\log p}{n}} < 1$*. If*

$$\lambda \ge M\left(C(M_\sigma^2 + M_\sigma + 1) + C'(2M_\sigma + 1)\right)\sqrt{\frac{\log p}{n}} + M\left(97M_\sigma^2 + 89M_\sigma + 82\right)\epsilon, \tag{3.17}$$

*then with probability at least*

$$1 - 2p^{-\left(\frac{C^2}{\pi^2}-2\right)} - 6p^{-\left\{2[\Phi^{-1}(0.75)]^2 C'^2 \min_{1\le i\le p} c^2(\sigma_i) - 1\right\}},$$

*the CLIME estimator* (3.10) *satisfies* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \le 4\|\mathbf{\Omega}^*\|_{L_1}\lambda$*.*

The proof of Theorem 6 is contained in Section B.1.3.

**Remark 8.** *Clearly, the optimal choice of* $\lambda$ *to minimize the estimation error bound in Theorem 6 is* $\lambda = C_1\sqrt{\frac{\log p}{n}} + C_2\epsilon$*, where* $C_1$ *and* $C_2$ *are the constant prefactors appearing on the right-hand side of inequality* (3.17)*. In this case,*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \le 4\|\mathbf{\Omega}^*\|_{L_1}\left(C_1\sqrt{\frac{\log p}{n}} + C_2\epsilon\right) \le 4M\left(C_1\sqrt{\frac{\log p}{n}} + C_2\epsilon\right).$$

For the GLasso, we focus on precision matrices satisfying the following assumption:

**Assumption 3** (Incoherence)**.** *There exists some* $0 < \alpha \le 1$ *such that*

$$\max_{e\in S^c} \|\mathbf{\Gamma}^*_{eS}(\mathbf{\Gamma}^*_{SS})^{-1}\|_{L_1} \le 1 - \alpha, \tag{3.18}$$

*where* $\mathbf{\Gamma}^* := \mathbf{\Sigma}^* \otimes \mathbf{\Sigma}^*$ *and* $S = \mathrm{supp}(\mathbf{\Omega}^*)$ *is the true edge set.*

We then have the following result, which is stated in terms of the population-level quantities $\kappa_{\Sigma^*} = \|\mathbf{\Sigma}^*\|_{L_1}$ and $\kappa_{\Gamma^*} = \|(\mathbf{\Gamma}^*_{SS})^{-1}\|_{L_1}$, as well as $k$, the maximum number

of nonzero elements in each row of $\mathbf{\Omega}^*$. The theorem also involves constants $C_0, C_1$, and $C_2$, which are independent of $\epsilon$ and the problem instances $n$, $p$, and $k$.

**Theorem 7.** *Under the cellwise contamination model* (3.1)*, suppose inequality* (3.11) *is satisfied, and* $\epsilon = \max_{1 \leq i \leq p} \epsilon_i \leq 0.02$*. Also suppose the sample size satisfies the scaling*

$$n \geq C_2 \tau \log p \cdot \left( \frac{1}{6(1+8/\alpha)k \max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa_{\Sigma^*}^3 \kappa_{\Gamma^*}^2\}} - C_0 \epsilon \right)^{-2}, \tag{3.19}$$

*and suppose Assumption 3 holds. Suppose* $\lambda = \frac{8}{\alpha}\left(C_0\epsilon + C_1\sqrt{\frac{\tau \log p}{n}}\right)$*. Then with probability at least* $1 - p^{2-\tau}$*, the GLasso estimator* (3.8) *satisfies* $\operatorname{supp}(\hat{\mathbf{\Omega}}) \subseteq \operatorname{supp}(\mathbf{\Omega}^*)$*, and*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leq 2\|(\mathbf{\Gamma}^*_{SS})^{-1}\|_{L_1}\left(1 + \frac{8}{\alpha}\right)\left(C_0\epsilon + C_1\sqrt{\frac{\tau \log p}{n}}\right).$$

The proof of Theorem 7 is contained in Section B.1.4. Theorem 7 implicitly assumes that $\epsilon \leq \frac{C}{k}$, so the expression in parentheses on the right-hand side of inequality (3.19) is positive.

**Remark 9.** *Comparing the results of Theorems 6 and 7, we see that as in the traditional uncontaminated setting, the GLasso delivers slightly stronger guarantees, at the expense of more stringent assumptions. In particular, the GLasso requires the sample size to scale as* $n \geq Ck^2 \log p$*, whereas the CLIME requires* $n \geq C'\|\mathbf{\Omega}^*\|_{L_1}^2 \log p$ *in order to achieve consistency. When the parameter* $M$ *defining the precision matrix class scales more slowly than* $k^2$*, the CLIME thus requires a weaker scaling. In addition, the GLasso result supposes Assumption 3, which posits an incoherence bound on submatrices of* $\mathbf{\Gamma}^*$*. On the other hand, Theorem 7 establishes that the* $\operatorname{supp}(\hat{\mathbf{\Omega}}) \subseteq \operatorname{supp}(\mathbf{\Omega}^*)$ *for the GLasso estimator, whereas Theorem 6 only guarantees consistency for the CLIME estimator in terms of* $\ell_\infty$*-norm. In the case of the CLIME estimator, however, the true support of* $\mathbf{\Omega}^*$ *may be obtained via thresholding, assuming the nonzero*

*elements of* $\mathbf{\Omega}^*$ *are of the order* $\Omega\left(\sqrt{\frac{\log p}{n}}\right)$.

Focusing on the level of contamination $\epsilon$ in relation to the problem dimensions, note that Theorems 6 and 7 both imply an $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right) + \mathcal{O}(\epsilon)$ error bound on the precision matrix estimator, under the corresponding assumptions. Hence, when $\epsilon \leq C\sqrt{\frac{\log p}{n}}$, the estimation error matches the error of the optimal precision matrix estimator in the uncontaminated case, up to a constant factor (Ren et al., 2015). Further note that when $\epsilon \leq C\sqrt{\frac{\log p}{n}}$, the condition $\epsilon = \mathcal{O}\left(\frac{1}{k}\right)$ required by the condition (3.19) in Theorem 7 clearly holds when the sample size satisfies $n \geq Ck^2 \log p$. Note that although the level of contamination tolerated by the estimator decreases as the level of sparsity increases, it is *not* required to decrease as $n$ and $p$ increase, as long as the ratio $\sqrt{\frac{\log p}{n}}$ remains fixed. Thus, the conclusions of Theorems 6 and 7 are truly high-dimensional. As in the case of the robust covariance matrix estimators, a nice feature is that when the data are uncontaminated ($\epsilon = 0$), the estimation error of the robust precision matrix estimator agrees with the optimal rate.

Lastly, note that since the inverse of the correlation matrix has the same support as the precision matrix, we could also estimate $\text{supp}(\mathbf{\Omega}^*)$ using the Kendall's or Spearman's correlation matrices $\hat{\boldsymbol{\rho}}^K, \hat{\boldsymbol{\rho}}^S$, defined by

$$\hat{\boldsymbol{\rho}}^K_{ij} = \sin\left(\frac{\pi}{2}\boldsymbol{r}^K_{ij}\right), \qquad \text{and} \qquad \hat{\boldsymbol{\rho}}^S_{ij} = 2\sin\left(\frac{\pi}{6}\boldsymbol{r}^S_{ij}\right), \tag{3.20}$$

respectively, as inputs to the CLIME (3.10) or GLasso (3.8). Indeed, Liu et al. (2012) and Xue & Zou (2012) proposed to plug in the correlation matrix estimators (3.20) into regularization routines for precision matrix estimation under the nonparanormal graphical model; in their case, the model under study is only identifiable up to centering and scaling, so a scale estimate is not necessary. In our setting, the same derivations as in Theorems 6 and 7, omitting the concentration bounds on the MAD

estimates of scale, would show convergence of $\hat{\boldsymbol{\rho}}^K$ and $\hat{\boldsymbol{\rho}}^S$ to the population correlation matrix $\boldsymbol{\rho}^*$ in $\ell_\infty$-norm, with the additional linear term in $\epsilon$. However, note that the conditions imposed for support recovery would need to hold for the correlation matrix $\boldsymbol{\rho}^*$, rather than for the precision matrix $\boldsymbol{\Omega}^*$. In particular, a minimum signal strength requirement on $\boldsymbol{\rho}^*$ is stronger than the same requirement imposed on $\boldsymbol{\Omega}^*$, since the latter can scale inversely with the standard deviations of individual variables in the joint distribution. We have therefore chosen to focus our attention in this chapter on the output of the CLIME and GLasso when applied to an estimate of the covariance instead of the correlation matrix.

### 3.3.3 Consequences for Robust Estimation

We now interpret the conclusions of our theorems in some concrete settings of interest.

**Constant fraction of outliers** We first briefly discuss the most basic setting of cellwise contamination, to emphasize the generality of our results. Following the model (3.1), suppose each entry of the data matrix $\mathbf{X}$ is contaminated independently with probability $\epsilon$. Furthermore, either all contaminated entries may be drawn independently from a fixed contaminating distribution, or the contaminated entries in each row may be drawn jointly from a fixed contaminating distribution. In each case, Theorems 4 and 5 provide elementwise error bounds on the robust covariance estimators, and Theorems 6 and 7 provide elementwise error bounds on the robust precision matrix estimators constructed from the CLIME and GLasso. The strength of the theorems lies in the fact that we do not make any side assumptions about the outlier distribution; it may be heavy-tailed and/or contain point masses. Hence, whereas statistics such as the sample covariance and sample correlation will have slower rates of convergence due to a constant fraction of outliers drawn from an ill-behaved dis-

tribution, their robust counterparts are agnostic to the outlier distribution.

It is also important to note that the bounds in the theorems of Sections 3.3.1 and 3.3.2 continue to hold when $\epsilon > C\sqrt{\frac{\log p}{n}}$. The difference is that in such scenarios, the statistical error will be of the order $\mathcal{O}(\epsilon)$ rather than $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$. However, the effect of an $\epsilon$ fraction of outliers nonetheless grows only linearly as a function of $\epsilon$. This emphasizes the robustness properties of the covariance and precision matrix estimators studied in this chapter.

**Missing data.** Turning to a somewhat different setting, note that missing data may also be seen as an instance of cellwise contamination. In this model, data are missing completely at random (MCAR), meaning that the probability of missingness is independent of the location of the unobserved entry of the data matrix (Little & Rubin, 1986). In other words, if we observe the matrix $\mathbf{X}^{\text{mis}}$ with missing entries, where the probability that an entry in column $i$ is missing is equal to $\epsilon_i$, we have

$$X_{ki}^{\text{mis}} = \begin{cases} Y_{ki}, & \text{with probability } 1-\epsilon_i, \\ \text{missing}, & \text{with probability } \epsilon_i, \end{cases} \tag{3.21}$$

where $\mathbf{Y}$ is the fully-observed matrix. Note that if we zero-fill the missing entries of $\mathbf{X}^{\text{mis}}$, the resulting matrix $\mathbf{X}$ exactly follows the cellwise contamination model (3.1), with $\mathbf{Z}_k = \mathbf{0}$ for all $k$. The following result is an immediate consequence of our theorems:

**Corollary 2.** *Suppose data are drawn from the missing data model* (3.21)*, and the matrix* $\mathbf{X}$ *is the zero-filled data matrix. Let* $\epsilon = \max_{1\le i\le p}\epsilon_i$*. Under the same conditions as in Theorem 6, we have*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \le 4\|\mathbf{\Omega}^*\|_{L_1}\lambda,$$

*for the robust CLIME estimator constructed from* $\mathbf{X}$*. Under the same conditions as in Theorem 7, we have* $\mathrm{supp}(\hat{\mathbf{\Omega}}) \subseteq \mathrm{supp}(\mathbf{\Omega}^*)$ *and*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leq 2\|(\mathbf{\Gamma}^*_{SS})^{-1}\|_{L_1} \left(1 + \frac{8}{\alpha}\right) \left(C_0\epsilon + C_1\sqrt{\frac{\tau \log p}{n}}\right),$$

*for the robust GLasso estimator constructed from* $\mathbf{X}$*.*

Note that the conclusion of Corollary 2 does not require the matrix $\mathbf{X}$ to be zero-filled for missing values; in fact, we could fill the missing entries with samples generated according to any distribution (as long as the distribution remains the same across rows). This is because the missing entries are treated as outliers. Of course, our bounds should only be interpreted up to constant factors, and filling missing entries in a strategic way, e.g., filling entries in column $i$ with the mean $E(X_{ki})$, could lead to smaller estimation error in practice.

**Rowwise contamination.** Although we have thus far assumed that data are contaminated according to a cellwise mechanism, we now show that the same results apply for rowwise contamination, as well. Recall that each row in the data matrix for the rowwise contamination model with contamination level $\epsilon$ is given by

$$\mathbf{X}_k = (1 - B_k)\mathbf{Y}_k + B_k\mathbf{Z}_k, \qquad \forall 1 \leq k \leq n, \tag{3.22}$$

where $\mathbf{Y}_k$ is the uncontaminated row vector, $\mathbf{Z}_k$ is the contamination vector, and $B_k \sim \mathrm{Bernoulli}(\epsilon)$.

Although model (3.22) differs from model (3.1), a simple inspection of the proofs of Theorems 6 and 7 shows that only Lemma 13 needs to be modified. Furthermore,

equation (B.14) simply needs to be replaced by the equation

$$(X_{ki}, X_{kj}) \overset{\text{i.i.d.}}{\sim} F_{ij} = (1-\epsilon)\Phi_{\boldsymbol{\mu}_{\{i,j\}}, \boldsymbol{\Sigma}_{\{i,j\}}} + \epsilon H_{ij}, \qquad \forall 1 \leq k \leq n, \tag{3.23}$$

in the proof of Lemma 13. Equation (3.23) comes from the fact that the pair is either drawn jointly from a normal distribution with probability $1-\epsilon$, or from the contaminating distribution with probability $\epsilon$. Then the remainder of the argument follows as before, implying that the same conclusion of Lemma 13 applies. (We could obtain a smaller prefactor for $\epsilon$ in the bound (B.1), since $2\epsilon$ is replaced by $\epsilon$, but we are not concerned about optimizing constants here.) We therefore arrive at the following result:

**Corollary 3.** *Under the rowwise contamination model* (3.22)*, the same conclusions as in Corollary 2 hold for the CLIME and GLasso estimators constructed from* $\mathbf{X}$*.*

We emphasize that the rowwise contamination model (3.22) is *not* in general a special case of the cellwise contamination model (3.1); rather, the proof techniques for analyzing the cellwise model may be used to handle the rowwise model, as well.

### 3.3.4 Extensions

In fact, our proofs reveal that the key inequalities required in establishing our theorems are the following error bounds on the entrywise correlation and scale estimators:

$$\max_{1 \leq i,j \leq p} |\hat{\boldsymbol{\rho}}_{ij} - \boldsymbol{\rho}_{ij}| \leq C_1 \sqrt{\frac{\log p}{n}} + C_2 \epsilon, \text{ and}$$
$$\max_{1 \leq i \leq p} |\hat{\sigma}_i - \sigma_i| \leq C_1' \sqrt{\frac{\log p}{n}} + C_2' \epsilon.$$

The $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ terms arise from fast concentration of the estimators $\hat{\boldsymbol{\rho}}$ and $\hat{\sigma}$ to their means $E(\hat{\boldsymbol{\rho}})$ and $E(\hat{\sigma})$, respectively (via a Hoeffding inequality + union bound

argument), whereas the $\mathcal{O}(\epsilon)$ terms arise from bounding the deviations $|E(\hat{\boldsymbol{\rho}}) - \boldsymbol{\rho})|$ and $|E(\hat{\sigma}) - \sigma|$ under an $\epsilon$-contamination model. This is essentially a bounded-influence property of the robust correlation and scale estimators used to define the robust precision matrix. We summarize these ides in the following meta-theorem:

**Theorem 8** (Meta-Theorem)**.** *Suppose a robust covariance estimator is defined elementwise according to* $\hat{\boldsymbol{\Sigma}}_{ij} = \hat{\sigma}_i \hat{\sigma}_j \hat{\boldsymbol{\rho}}_{ij}$*. Also suppose:*

*(i) The correlation and scale estimators satisfy the deviation bounds*

$$\max_{1 \le i,j \le p} |\hat{\boldsymbol{\rho}}_{ij} - E(\hat{\boldsymbol{\rho}}_{ij})| \le C_1 \sqrt{\frac{\log p}{n}}, \text{ and} \tag{3.24a}$$

$$\max_{1 \le i \le p} |\hat{\sigma}_i - E(\hat{\sigma}_i)| \le C_1' \sqrt{\frac{\log p}{n}}. \tag{3.24b}$$

*(ii) The correlation and scale estimators satisfy the bounded-influence inequalities*

$$\max_{1 \le i,j \le p} |E(\hat{\boldsymbol{\rho}}_{ij}) - \boldsymbol{\rho}_{ij}| \le C_2 \epsilon, \text{ and} \tag{3.25a}$$

$$\max_{1 \le i \le p} |E(\hat{\sigma}_i) - \sigma_i| \le C_2' \epsilon, \tag{3.25b}$$

*when samples are drawn i.i.d. from an $\epsilon$-contaminated Gaussian distribution. Then the GLasso and CLIME estimators based on $\hat{\boldsymbol{\Sigma}}$ yield precision matrix estimators satisfying the error bound*

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_\infty \le C \sqrt{\frac{\log p}{n}} + C' \epsilon.$$

**Remark 10.** *When $\hat{\boldsymbol{\rho}}_{ij}$ is the Kendall's tau correlation and $\sigma_i$ is the MAD estimator, inequalities* (3.24a) *and* (3.25a) *are essentially established in Lemmas 13 and 24, whereas inequalities* (3.24b) *and* (3.25b) *are derived in Lemmas 14 and 22. Simi-*

*larly, inequalities* (3.24a) *and* (3.25a) *are derived for Spearman's rho correlation in Lemmas 15 and 25.*

The framework of Theorem 8 enables us to extend our analysis to other natural robust candidates for $\hat{\mathbf{\Sigma}}$, composed of entrywise correlation and scale estimates. To illustrate this point, we mention several examples below:

- **Quadrant correlation estimator.** The quadrant correlation estimator is defined by

$$r_{ij}^{Q} = \frac{1}{n}\sum_{k=1}^{n} \operatorname{sign}\left(X_{ki} - \underset{1\le\ell\le n}{\operatorname{median}} X_{\ell i}\right) \operatorname{sign}\left(X_{kj} - \underset{1\le\ell\le n}{\operatorname{median}} X_{\ell j}\right),$$

  and is also known to have bounded influence (Shevlyakov & Vilchevski, 2002). One can show that the quadrant correlation estimator also satisfies the inequalities (3.24a) and (3.25a) appearing in Theorem 8; the derivations are similar to those employed for Kendall's tau and Spearman's rho correlation, so we do not provide the details here.

- **Gnanadesikan-Kettenring estimator.** Tarr et al. (2015) and Oellerer & Croux (2014) also propose to use the following estimator for pairwise covariances: Noting that

$$\operatorname{Cov}(X,Y) = \frac{1}{4\alpha\beta}\left[\operatorname{Var}\left(\alpha X + \beta Y\right) - \operatorname{Var}\left(\alpha X - \beta Y\right)\right],$$

  the proposal is to replace the variance estimator by a robust variance estimator (e.g., the square of the MAD estimator). The drawback of this estimator in comparison to the covariance estimators based on Kendall's tau and Spearman's rho is that the covariance estimator has a maximal breakdown point of 25% under cellwise contamination, since the argument in the variance involves a

sum of variables, and any robust variance estimator has a maximal breakdown point of 50%. However, from the point of view of statistical consistency, the Gnanadesikan-Kettenring covariance estimator may be seen to perform equally well. Indeed, consider the covariance estimator

$$\frac{1}{4}\left(\hat{\sigma}^2_{(i,j),+} - \hat{\sigma}^2_{(i,j),-}\right), \tag{3.26}$$

where $\hat{\sigma}_{(i,j),+}$ is the (rescaled) MAD statistic computed from $\{X_{ki} + X_{kj} : 1 \leq k \leq n\}$, and $\hat{\sigma}_{(i,j),-}$ is analogously defined to be the MAD statistic computed from $\{X_{ki} - X_{kj} : 1 \leq k \leq n\}$. Then our derivations showing the consistency of the MAD estimator (cf. Lemmas 22 and 23, with minor modifications) show that

$$\max_{1\leq i,j\leq p} |\hat{\sigma}_{(i,j),+} - \sigma_{(i,j),+}| \leq C_1\sqrt{\frac{\log p}{n}} + C_2\epsilon, \qquad \text{and}$$
$$\max_{1\leq i,j\leq p} |\hat{\sigma}_{(i,j),-} - \sigma_{(i,j),-}| \leq C_1\sqrt{\frac{\log p}{n}} + C_2\epsilon,$$

for data from the cellwise contamination model, where $\sigma_{(i,j),+}$ and $\sigma_{(i,j),-}$ are the population-level standard deviations of the distributions of $X_{ki} + X_{kj}$ and $X_{ki} - X_{kj}$, respectively. Thus,

$$\max_{1\leq i,j\leq p} |\hat{\sigma}^2_{(i,j),+} - \sigma^2_{(i,j),+}|, \ \max_{1\leq i,j\leq p} |\hat{\sigma}^2_{(i,j),-} - \sigma^2_{(i,j),-}| \leq C'\sqrt{\frac{\log p}{n}} + C''\epsilon,$$

as well, from which we may conclude that the pairwise covariance estimator (3.26) deviates from the true covariance $\mathrm{Cov}(X_{ki}, X_{kj})$ by the same margin.

- $Q_n$ **estimator.** Finally, consider the $Q_n$ scale estimator (Rousseeuw & Croux, 1993), defined by

$$Q_n = c\{|X_k - X_\ell| : k < \ell\}_{(k^*)},$$

where $c$ is a constant factor and $k^* = \lceil \binom{n}{2}/4 \rceil$. The $Q_n$ estimator is also known to have a bounded influence property for real-valued data. Since the $Q_n$ estimator is also based on quantiles, essentially the same types of arguments used to derive MAD concentration (cf. Appendix B.3) may be used to establish the desired bounds (3.24b) and (3.25b) appearing in Theorem 8.

## 3.4 Breakdown Point

We now turn to a brief discussion of the breakdown point of the estimators studied in this chapter. As discussed in Donoho & Huber (1983) and Hampel et al. (2011), breakdown analysis concerns the *global* behavior of a procedure, under large departures from an assumed situation. On the other hand, the theoretical analysis of statistical consistency and efficiency are related to notions of infinitesimal robustness, and quantifies the *local* behavior of a procedure at or near the assumed situation. Donoho & Huber (1983) draw an analogy between the fields of material science and statistics, where the notions of stiffness (resistance of a material to displacements caused by a small load) and breaking strength (the amount of load required to make the material fracture) parallel those of the influence function and the breakdown point. Ideally, a procedure should perform well both locally and globally; optimizing either measure alone is unwise. Our key result of this section shows that although the GLasso and CLIME estimators both enjoy roughly the same statistical rate of estimation, the CLIME does *not* perform as well as the GLasso when the breakdown point is used to quantify the degree of robustness.

Our analysis of the GLasso estimator closely follows that of Oellerer & Croux (2014); however, since the specific precision matrix estimators analyzed in this chapter differ slightly, we include the full argument for the sake of completeness. We define

the finite-sample breakdown point of the precision matrix estimator under cellwise contamination to be

$$\epsilon_n(\hat{\mathbf{\Omega}}, \mathbf{X}) := \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{\mathbf{X}^m} D(\hat{\mathbf{\Omega}}(\mathbf{X}), \hat{\mathbf{\Omega}}(\mathbf{X}^m)) = \infty \right\}, \tag{3.27}$$

where

$$D(\mathbf{A}, \mathbf{B}) := \max \left\{ |\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p^{-1}(\mathbf{A}) - \lambda_p^{-1}(\mathbf{B})| \right\},$$

and $\mathbf{X}^m$ is a data matrix obtained from $\mathbf{X}$ by replacing at most $m$ entries in each column by arbitrary elements. We also define the explosion finite sample breakdown point of a covariance matrix estimator as follows:

$$\epsilon_n^+(\mathbf{S}, \mathbf{X}) := \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{\mathbf{X}^m} |\lambda_1(\mathbf{S}(\mathbf{X})) - \lambda_1(\mathbf{S}(\mathbf{X}^m))| = \infty \right\} \tag{3.28}$$

(cf. Maronna & Zamar (2002)). Note that the explosion breakdown point only accounts for maximum eigenvalues, whereas the overall covariance matrix estimator breaks down under explosion or *implosion* (i.e., arbitrarily small minimum eigenvalues). Also, the breakdown point under cellwise contamination is less than or equal to the breakdown point under rowwise contamination.

We will consider the breakdown behavior of a slightly tweaked version of the GLasso presented earlier. Consider the matrix

$$\check{\mathbf{\Sigma}}(\mathbf{X}) := \operatorname*{argmin}_{\mathbf{M} \succeq 0} \|\hat{\mathbf{\Sigma}} - \mathbf{M}\|_\infty, \tag{3.29}$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}(\mathbf{X})$ is the robust covariance matrix estimator constructed from the data matrix $\mathbf{X}$. Let

$$\check{\mathbf{\Omega}}(\mathbf{X}) := \operatorname*{argmin}_{\mathbf{\Omega} \succ 0} \left\{ \operatorname{tr}(\check{\mathbf{\Sigma}}\mathbf{\Omega}) - \log\det(\mathbf{\Omega}) + \lambda \|\mathbf{\Omega}\|_{1,\text{off}} \right\} \tag{3.30}$$

be the corresponding GLasso estimator. Note that from a computational standpoint, the projection step (3.29) is important so that fast solvers for the GLasso program (3.30) may be applied (Friedman et al., 2008). Furthermore, the projection step (3.29) is convex, and the additional computational time is negligible compared to the computation required for running the GLasso. We have the following result, proved in Section B.1.5:

**Theorem 9.** *Consider the positive semidefinite version of the robust GLasso estimator* (3.30)*. Under the same conditions as in Theorem 7, we have* $\mathrm{supp}(\check{\mathbf{\Omega}}) \subseteq \mathrm{supp}(\mathbf{\Omega}^*)$ *and*

$$\|\check{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leq 2\|(\mathbf{\Gamma}^*_{SS})^{-1}\|_{L_1} \left(1 + \frac{8}{\alpha}\right) \left(C_0'\epsilon + C_1'\sqrt{\frac{\tau \log p}{n}}\right). \tag{3.31}$$

*Furthermore, for any data matrix* $\mathbf{X} \in \mathbb{R}^{n \times p}$*, the breakdown point satisfies* $\epsilon_n(\check{\mathbf{\Omega}}, \mathbf{X}) = 50\%$*.*

**Remark 11.** *Note that Theorem 9 guarantees that the robust GLasso estimator* $\check{\mathbf{\Omega}}$ *obtained from a semidefinite projection of the robust covariance estimator shares the same level of statistical consistency achieved by the robust GLasso estimator* $\hat{\mathbf{\Omega}}$*. In addition, the precision matrix estimator* $\check{\mathbf{\Omega}}$ *has a breakdown point of 50%. Although other authors also suggest projecting the robust covariance estimator onto the positive semidefinite cone before applying the GLasso (Oellerer & Croux, 2014; Tarr et al., 2015), they advocate a projection in terms of the Frobenius norm rather than the* $\ell_\infty$*-norm in the optimization program* (3.29)*. As can be seen in the proof of Theorem 9, minimizing the elementwise* $\ell_\infty$*-norm is much more natural from the point of view of statistical consistency, since it guarantees that the* $\ell_\infty$*-error between the precision matrix estimate and the true precision matrix grows by at most a factor of two.*

We now show that although the CLIME is as robust as the GLasso in terms of statistical consistency under the cellwise contamination model, it has much poorer

breakdown behavior. Consider the CLIME estimator based on corrupted data:

$$\min \|\boldsymbol{\Omega}\|_1 \quad \text{s.t.} \quad \|\hat{\boldsymbol{\Sigma}}(\mathbf{X}^m)\boldsymbol{\Omega} - I\|_\infty \leq \lambda, \tag{3.32}$$

where $\hat{\boldsymbol{\Sigma}}(\mathbf{X}^m)$ is the robust covariance estimator based on a data matrix with at most $m$ arbitrarily corrupted entries per column. Since the CLIME estimator arises as the solution to a constrained linear program, the solution is undefined (infinite) when the problem is infeasible. Indeed, we will show in the following theorem that such a case may arise even by corrupting at most *one* entry in each column of the data matrix.

**Theorem 10.** *In the case when $p = 2$, there exists $\mathbf{X} \in \mathbb{R}^{n\times 2}$ such that $\epsilon_n(\hat{\boldsymbol{\Omega}}, \mathbf{X}) = \frac{1}{n}$, where $\hat{\boldsymbol{\Omega}}$ denotes the CLIME estimator.*

The proof of Theorem 10, supplied in Section B.1.6, provides the construction of a data matrix $\mathbf{X} \in \mathbb{R}^{n\times 2}$ where the CLIME estimator becomes infeasible after perturbing a single entry in each column. This is in stark contrast to the result in Theorem 9, which establishes that the breakdown point of the robust GLasso estimator is 50%, for *any* data matrix $\mathbf{X}$.

**Remark 12.** *Although Theorem 10 is stated for the case $p = 2$, the argument used to prove the theorem is readily generalizable to higher dimensions, as well, in which case we would also have a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfying $\epsilon_n(\boldsymbol{\Omega}, \mathbf{X}) = \frac{1}{n}$. For instance, we could construct an $n \times p$ matrix $\mathbf{X}^1$ such that $\boldsymbol{\Sigma}(\mathbf{X}^1)$ is a block matrix with upper-left block equal to the matrix constructed in the proof of Theorem 10, lower-left block equal to the identity, and off-diagonal blocks equal to zero.*

The conclusion of Theorem 10 underscores the fact that consistency and breakdown point under cellwise contamination are in some sense orthogonal measures of robustness. As demonstrated in the previous section, the robust CLIME and GLasso both enjoy good rates of statistical consistency when the contamination fraction $\epsilon$

is sufficiently small relative to the problem parameters. On the other hand, the results of this section show that the CLIME is extremely non-robust in terms of its breakdown point. Similarly, procedures such as the Gnanadesikan-Kettenring estimator (3.26) may be shown to be statistically consistent under cellwise contamination (cf. Section 3.3.4), but as discussed in Oellerer & Croux (2014), the breakdown point of the covariance estimator $\hat{\mathbf{\Sigma}}$ is at most 25%, which leads to error propagation in $\hat{\mathbf{\Omega}}$.

Finally, we note that the notion of breakdown point that we consider in equation (3.27) is defined with respect to a finite sample, without recourse to probability distributions. Other notions of breakdown point, defined with respect to an $\epsilon$-contaminated distribution, have also been studied in the literature (Hampel et al., 2011). For some alternative measures of breakdown robustness, the CLIME estimator may have a more controlled breakdown behavior, but we have not explored them here.

## 3.5 Simulation

In this section, we perform simulation studies to examine the performance of the two robust covariance matrix estimators introduced in Section 3.2, and also the robust precision matrix estimators obtained using the GLasso. We will refer to the two type of estimators as `Kendall` and `Spearman`, respectively.

For comparison, we also compute the following robust covariance matrix estimators, which are similarly plugged into the GLasso to obtain robust precision matrix estimators:

- `SpearmanU`: The pairwise covariance matrix estimator proposed in Oellerer & Croux (2014), where the MAD estimator is combined with Spearman's rho

(without transformation):

$$\hat{\mathbf{\Sigma}}_{ij} = \hat{\sigma}_i \hat{\sigma}_j \boldsymbol{r}_{ij}^S, \qquad \text{where} \qquad \hat{\sigma}_i = [\Phi^{-1}(0.75)]^{-1} \hat{d}_i.$$

- `OGK`: The OGK estimator proposed in Maronna & Zamar (2002), with scale estimator $Q_n$.

- `NPD`: The pairwise covariance matrix estimator considered in Tarr et al. (2015), where

$$\tilde{\mathbf{\Sigma}}_{ij} = \frac{1}{4} \left( \hat{\sigma}^2_{(i,j),+} - \hat{\sigma}^2_{(i,j),-} \right).$$

  Here, $\hat{\sigma}_{(i,j),+}$ is the $Q_n$ statistic computed from $\{X_{ki} + X_{kj} : 1 \leq k \leq n\}$ and $\hat{\sigma}_{(i,j),-}$ is the $Q_n$ statistic computed from $\{X_{ki} - X_{kj} : 1 \leq k \leq n\}$. An NPD projection is applied to $\tilde{\mathbf{\Sigma}}$ to obtain the final positive semidefinite covariance matrix estimator.

Further details for the orthogonalized Gnanedesikan-Kettenring (OGK) and nearest positive definite (NPD) procedures may be found in Maronna & Zamar (2002) and Higham (2002), respectively. The nonrobust GLasso, which takes the sample covariance matrix estimator as an input (`SampleCov`), as well as the inverse sample covariance matrix estimator (`InvCov`), applicable in the case $p < n$, are used as points of reference.

An implementation of the GLasso that allows the diagonal entries of the precision matrix estimator to be unpenalized is provided in the widely used `glasso` package. In this chapter, however, we use the GLasso implementation from the `QUIC` package (Hsieh et al., 2011), since it does not require the input covariance matrix to be positive semidefinite, and speeds up substantially over `glasso`. We select the tuning parameter $\lambda$ in GLasso by cross-validation: We first split the data into $K$ groups, or folds, of nearly equal size. For a given $\lambda$ and $1 \leq k \leq K$, we take the $k^{th}$

fold as the test set, and compute the precision matrix estimate $\hat{\mathbf{\Omega}}_\lambda^{(-k)}$ based on the remaining $K-1$ folds. We then compute the negative log-likelihood on the test set, $L^{(k)}(\lambda) = -\log\det\hat{\mathbf{\Omega}}_\lambda^{(-k)} + \operatorname{tr}\left(\hat{\mathbf{\Sigma}}^{(k)}\hat{\mathbf{\Omega}}_\lambda^{(-k)}\right)$, where $\hat{\mathbf{\Sigma}}^{(k)}$ is the robust covariance estimate obtained from the test set. This is done over a logarithmically spaced grid of 15 values between $\lambda_{\max} = \max_{i\neq j}|\hat{\mathbf{\Sigma}}_{ij}|$ and $\lambda_{\min} = 0.01\lambda_{\max}$, where $\hat{\mathbf{\Sigma}}$ is the robust covariance estimate computed from the whole data set. The value of $\lambda$ that minimizes $\frac{1}{K}\sum_{k=1}^{K} L^{(k)}(\lambda)$ is selected as the final tuning parameter.

**Simulation settings** We consider the following sampling schemes, covering different structures of the precision matrix $\mathbf{\Omega}^* \in \mathbb{R}^{p\times p}$:

- Banded: $\mathbf{\Omega}^*_{ij} = 0.6^{|i-j|}$.
- Sparse: $\mathbf{\Omega}^* = \mathbf{B} + \delta\mathbf{I}_p$, where $b_{ii} = 0$ and $b_{ij} = b_{ji}$, with $P(b_{ij} = 0.5) = 0.1$ and $P(b_{ij} = 0) = 0.9$, for $i \neq j$. The parameter $\delta$ is chosen such that the condition number of $\mathbf{\Omega}^*$ equals $p$. The matrix is then standardized to have unit diagonals.
- Dense: $\mathbf{\Omega}^*_{ii} = 1$ and $\mathbf{\Omega}^*_{ij} = 0.5$, for $i \neq j$.
- Diagonal: $\mathbf{\Omega}^* = \mathbf{I}_p$.

For each sampling scheme and dimension $p \in \{120, 400\}$, we generate $B = 100$ samples of size $n = 200$ from the multivariate normal distribution $N(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$. We then add 5% or 10% of rowwise or cellwise contamination to the data, where the outliers are sampled independently from $N(10, 0.2)$. We also simulate model deviation by generating all observations from either the multivariate $t$-distribution, $t_3(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$, or the alternative $t$-distribution, $t_3^*(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$, each with three degrees of freedom. Recall that $\mathbf{X} \sim t_\nu(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$, where $t_\nu(\mathbf{0}, \mathbf{\Omega}^*)^{-1})$ denotes the multivariate $t$-distribution with $\nu$ degrees of freedom, if $\mathbf{X} = \mathbf{Y}/\sqrt{\tau}$, where $\mathbf{Y} \sim N(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$ and $\tau \sim \Gamma(\nu/2, \nu/2)$. The alternative $t$-distribution, denoted by $t_\nu^*$, is proposed in

Finegold & Drton (2011) as a generalization of the multivariate $t$-distribution. We say that $\mathbf{X} \sim t_\nu^*(\mathbf{0}, (\mathbf{\Omega}^*)^{-1})$ if $X_i = Y_i/\sqrt{\tau_i}$, for all $1 \leq i \leq p$, where the divisors $\tau_i \sim \Gamma(\nu/2, \nu/2)$ are independent. In this case, the heaviness of the tails are different for different components of $\mathbf{X}$.

**Performance measures** We assess the performance of the covariance and precision matrix estimators via the deviations $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$ and $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty$, respectively. We also consider the false positive (FP) and false negative (FN) rates:

$$\text{FP} = \frac{|\{(i,j) : \hat{\mathbf{\Omega}}_{ij} \neq 0, \mathbf{\Omega}^*_{ij} = 0\}|}{|\{(i,j) : \mathbf{\Omega}^*_{ij} = 0\}|}, \quad \text{and} \quad \text{FN} = \frac{|\{(i,j) : \hat{\mathbf{\Omega}}_{ij} = 0, \mathbf{\Omega}^*_{ij} \neq 0\}|}{|\{(i,j) : \mathbf{\Omega}^*_{ij} \neq 0\}|}.$$

FP gives the proportion of zero elements in the true precision matrix that are incorrectly estimated to be nonzero, while FN gives the proportion of nonzero elements in the true precision matrix that are incorrectly estimated to be zero. Note that if $\mathbf{\Omega}^*$ has no zero entries, as in the case of the banded and dense structures, the quantity FP is undefined.

**Results** Tables 3.1 and 3.2 show the results for $n = 200$ and $p = 120$. We summarize the salient points below:

- When the dataset is clean, `SampleCov` performs best in terms of both covariance and precision matrix estimation, across all sampling schemes. Note that even though the data are uncontaminated, `InvCov` performs poorly, due to the fact that the sample covariance matrix has low precision when $p > n/2$.

- In the case of rowwise contamination, the nonrobust `SampleCov` has the largest estimation error for the covariance matrix, as expected. Curiously, the precision matrix estimation error based on `SampleCov` is the lowest among all estimators. We do not have good explanation for this, but the tuning parameter selected

for `SampleCov` by cross-validation tends to be smaller (as can be seen from its relatively low FN). `NPD`, `Kendall`, `Spearman`, and `SpearmanU` have similar performance in terms of both covariance and precision matrix estimation. In all sampling schemes, `OGK` outperforms these four estimators for covariance estimation, but not consistently so for precision matrix estimation.

- For covariance and precision matrix estimation under cellwise contamination, the `Kendall`, `Spearman`, and `SpearmanU` estimators perform the best. `NPD` performs the worst among all cellwise robust covariance matrix estimators. Nonetheless, `NPD` still beats `OGK`, which is designed to work well under rowwise contamination, and also beats the nonrobust `SampleCov`.

- When the data are generated from the multivariate $t$-distribution or alternative $t$-distribution, we again see that `Kendall`, `Spearman`, and `SpearmanU` behave similarly and outperform all other estimators, across all sampling schemes.

- When $\mathbf{\Omega}^*$ is either sparse or diagonal, FP is low for all estimators except `InvCov`, under all contamination mechanisms.

- Except for `InvCov`, FN is high when $\mathbf{\Omega}^*$ is banded or dense, under all contamination mechanisms. This is expected because GLasso implicitly assumes the underlying $\mathbf{\Omega}^*$ to be sparse, which is not true in these cases. When $\mathbf{\Omega}^*$ is sparse, the FN for `Kendall`, `Spearman`, and `SpearmanU` are relatively low compared to the other estimators.

Tables 3.3 and 3.4 show the results for $n = 200$ and $p = 400$. Since $p > n$, the inverse sample covariance matrix cannot be computed, hence is excluded from the analysis. Overall, we obtain conclusions similar to those obtained in the first set of simulations:

- When the data are clean, `SampleCov` perform best in terms of estimation error, across all sampling schemes. Immediately following are `OGK` and `NPD`, and

then `Kendall`, `Spearman`, and `SpearmanU` (the last three have nearly the same performance).

- Under rowwise contamination, `SampleCov` has the worst covariance estimation error, but also the best precision estimation error, across all sampling schemes. `OGK` performs best in terms of covariance estimation, but not precision estimation. `NPD`, `Kendall`, `Spearman`, and `SpearmanU` have similar performance in nearly all cases. When $\mathbf{\Omega}^*$ is diagonal and the contamination fraction is 10%, `Kendall` turns out to have high precision estimation error, possibly because the selected tuning parameter in GLasso is too small (as can be seen by the high FP).

- In terms of estimation error under cellwise contamination, `OGK` performs nearly as badly as `SampleCov`. `Kendall`, `Spearman`, and `SpearmanU` perform equally well, while `NPD` is slightly worse off.

- When the data are generated from the multivariate $t$-distribution or alternative $t$-distribution, `SampleCov` performs badly. `Kendall`, `Spearman`, and `SpearmanU` perform similarly and outperform `OGK` and `NPD`, across all sampling schemes.

- In general, under all contamination mechanisms, when $\mathbf{\Omega}^*$ is either sparse or diagonal, FP is low for all estimators. On the other hand, when $\mathbf{\Omega}^*$ is banded or dense, FN is high, as expected. When $\mathbf{\Omega}^*$ is sparse, FN is not as low as desired.

In summary, `SampleCov` performs best for clean data. Under rowwise contamination, `OGK` yields the best results in terms of covariance estimation; under cellwise contamination, `Kendall`, `Spearman`, and `SpearmanU` equally share the best performance, while `NPD` is slightly worse off. `Kendall`, `Spearman`, and `SpearmanU` also perform very well when the data are generated from a multivariate $t$-distribution or

the alternative $t$-distribution, although these latter cases are not covered by our theory. Empirical results of a similar flavor were obtained in Liu et al. (2012), although their paper does not provide theoretical guarantees for the behavior of the estimators under contaminated data.

| | | clean | | | | 5% rowwise | | | | 10% rowwise | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN |
| Banded | SampleCov | 1.11 | 0.30 | | 0.85 | 5.91 | 0.31 | | 0.60 | 10.44 | 0.31 | | 0.61 |
| | OGK | 1.20 | 0.32 | | 0.88 | 1.98 | 0.37 | | 0.90 | 2.91 | 0.41 | | 0.91 |
| | NPD | 1.26 | 0.35 | | 0.96 | 2.24 | 0.37 | | 0.72 | 3.39 | 0.39 | | 0.71 |
| | Kendall | 1.73 | 0.33 | | 0.87 | 2.50 | 0.32 | | 0.63 | 3.37 | 0.31 | | 0.63 |
| | Spearman | 1.73 | 0.33 | | 0.87 | 2.50 | 0.33 | | 0.64 | 3.37 | 0.33 | | 0.64 |
| | SpearmanU | 1.73 | 0.34 | | 0.88 | 2.50 | 0.34 | | 0.64 | 3.37 | 0.34 | | 0.63 |
| | InvCov | 1.11 | 1.68 | | 0.00 | 5.91 | 1.83 | | 0.00 | 10.44 | 2.09 | | 0.00 |
| Sparse | SampleCov | 0.70 | 0.34 | 0.19 | 0.11 | 5.57 | 0.35 | 0.36 | 0.30 | 10.09 | 0.32 | 0.36 | 0.32 |
| | OGK | 0.79 | 0.39 | 0.18 | 0.15 | 1.62 | 0.51 | 0.18 | 0.20 | 2.39 | 0.59 | 0.17 | 0.24 |
| | NPD | 0.82 | 0.47 | 0.09 | 0.32 | 1.63 | 0.55 | 0.21 | 0.66 | 2.58 | 0.61 | 0.20 | 0.76 |
| | Kendall | 1.15 | 0.43 | 0.17 | 0.16 | 1.63 | 0.41 | 0.32 | 0.37 | 2.36 | 0.40 | 0.32 | 0.41 |
| | Spearman | 1.15 | 0.43 | 0.17 | 0.16 | 1.64 | 0.43 | 0.32 | 0.37 | 2.38 | 0.43 | 0.31 | 0.42 |
| | SpearmanU | 1.15 | 0.45 | 0.17 | 0.15 | 1.65 | 0.45 | 0.33 | 0.36 | 2.37 | 0.46 | 0.31 | 0.41 |
| | InvCov | 0.70 | 2.83 | 1.00 | 0.00 | 5.57 | 3.14 | 1.00 | 0.00 | 10.09 | 3.54 | 1.00 | 0.00 |
| Dense | SampleCov | 0.60 | 0.60 | | 0.99 | 5.54 | 0.61 | | 0.75 | 10.05 | 0.60 | | 0.75 |
| | OGK | 0.63 | 0.61 | | 0.99 | 1.18 | 0.68 | | 0.99 | 1.88 | 0.74 | | 0.99 |
| | NPD | 0.67 | 0.62 | | 0.99 | 1.23 | 0.65 | | 0.82 | 1.89 | 0.69 | | 0.79 |
| | Kendall | 1.00 | 0.66 | | 0.99 | 1.37 | 0.64 | | 0.79 | 1.91 | 0.64 | | 0.78 |
| | Spearman | 1.00 | 0.66 | | 0.99 | 1.37 | 0.64 | | 0.79 | 1.91 | 0.64 | | 0.77 |
| | SpearmanU | 0.99 | 0.66 | | 0.99 | 1.37 | 0.64 | | 0.78 | 1.91 | 0.65 | | 0.77 |
| | InvCov | 0.60 | 2.63 | | 0.00 | 5.54 | 1.28 | | 0.00 | 10.05 | 1.48 | | 0.00 |
| Diagonal | SampleCov | 0.30 | 0.31 | 0.00 | 0.00 | 5.31 | 0.26 | 0.24 | 0.00 | 9.84 | 0.28 | 0.24 | 0.00 |
| | OGK | 0.32 | 0.33 | 0.00 | 0.00 | 0.55 | 0.35 | 0.00 | 0.00 | 0.80 | 0.44 | 0.00 | 0.00 |
| | NPD | 0.33 | 0.35 | 0.00 | 0.00 | 0.63 | 0.31 | 0.18 | 0.00 | 0.98 | 0.39 | 0.21 | 0.00 |
| | Kendall | 0.51 | 0.62 | 0.00 | 0.00 | 0.68 | 0.51 | 0.20 | 0.00 | 0.96 | 0.46 | 0.21 | 0.00 |
| | Spearman | 0.51 | 0.62 | 0.00 | 0.00 | 0.68 | 0.52 | 0.21 | 0.00 | 0.96 | 0.47 | 0.22 | 0.00 |
| | SpearmanU | 0.51 | 0.62 | 0.00 | 0.00 | 0.68 | 0.52 | 0.21 | 0.00 | 0.96 | 0.45 | 0.23 | 0.00 |
| | InvCov | 0.30 | 2.81 | 1.00 | 0.00 | 5.31 | 3.19 | 1.00 | 0.00 | 9.84 | 3.60 | 1.00 | 0.00 |

Table 3.1: Simulation results for seven estimators and four sampling schemes, when $n = 200$ and $p = 120$. Performance is measured by $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$ for covariance matrix estimation (Cov), $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty$ for precision matrix estimation (Prec), and false positive rate (FP) and false negative rate (FN) for support recovery of the true precision matrix. The results are averaged over 100 replications.

| | | 5% cellwise | | | | 10% cellwise | | | | multivariate $t$ | | | | alternative $t$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN |
| Banded | SampleCov | 8.33 | 0.51 | | 0.97 | 13.09 | 0.54 | | 0.99 | 18.31 | 0.49 | | 0.87 | 57.72 | 0.57 | | 0.93 |
| | OGK | 8.10 | 0.51 | | 0.95 | 13.15 | 0.54 | | 0.99 | 3.85 | 0.43 | | 0.92 | 12.15 | 0.53 | | 0.92 |
| | NPD | 2.78 | 0.41 | | 0.95 | 4.70 | 0.46 | | 0.96 | 4.06 | 0.44 | | 0.96 | 4.53 | 0.46 | | 0.96 |
| | Kendall | 2.43 | 0.40 | | 0.92 | 3.67 | 0.45 | | 0.92 | 3.32 | 0.41 | | 0.90 | 3.60 | 0.42 | | 0.90 |
| | Spearman | 2.43 | 0.41 | | 0.92 | 3.67 | 0.45 | | 0.92 | 3.32 | 0.41 | | 0.91 | 3.60 | 0.42 | | 0.90 |
| | SpearmanU | 2.43 | 0.41 | | 0.93 | 3.67 | 0.45 | | 0.93 | 3.32 | 0.42 | | 0.91 | 3.60 | 0.43 | | 0.90 |
| | InvCov | 8.33 | 0.41 | | 0.00 | 13.09 | 0.46 | | 0.00 | 18.31 | 1.26 | | 0.00 | 57.72 | 0.53 | | 0.00 |
| Sparse | SampleCov | 8.39 | 0.90 | 0.05 | 0.81 | 13.25 | 0.93 | 0.01 | 0.91 | 11.47 | 0.77 | 0.14 | 0.43 | 32.95 | 0.94 | 0.12 | 0.44 |
| | OGK | 8.18 | 0.90 | 0.06 | 0.77 | 13.71 | 0.94 | 0.01 | 0.90 | 3.38 | 0.65 | 0.16 | 0.23 | 8.67 | 0.86 | 0.16 | 0.34 |
| | NPD | 2.15 | 0.61 | 0.06 | 0.45 | 4.04 | 0.73 | 0.05 | 0.59 | 3.17 | 0.69 | 0.08 | 0.45 | 3.31 | 0.71 | 0.07 | 0.49 |
| | Kendall | 1.58 | 0.61 | 0.16 | 0.30 | 2.44 | 0.72 | 0.13 | 0.46 | 2.34 | 0.58 | 0.15 | 0.25 | 2.32 | 0.62 | 0.16 | 0.22 |
| | Spearman | 1.58 | 0.62 | 0.15 | 0.30 | 2.44 | 0.73 | 0.13 | 0.46 | 2.34 | 0.59 | 0.15 | 0.25 | 2.32 | 0.62 | 0.15 | 0.23 |
| | SpearmanU | 1.58 | 0.63 | 0.16 | 0.30 | 2.44 | 0.73 | 0.13 | 0.46 | 2.34 | 0.60 | 0.15 | 0.25 | 2.32 | 0.63 | 0.16 | 0.22 |
| | InvCov | 8.39 | 0.77 | 1.00 | 0.00 | 13.25 | 0.85 | 1.00 | 0.00 | 11.47 | 2.10 | 1.00 | 0.00 | 32.95 | 0.87 | 1.00 | 0.00 |
| Dense | SampleCov | 8.39 | 0.90 | | 0.99 | 13.25 | 0.93 | | 0.99 | 10.06 | 0.88 | | 0.98 | 31.24 | 0.95 | | 0.99 |
| | OGK | 8.02 | 0.90 | | 0.99 | 13.14 | 0.93 | | 0.99 | 2.14 | 0.76 | | 0.99 | 6.82 | 0.89 | | 0.99 |
| | NPD | 1.51 | 0.71 | | 0.99 | 2.64 | 0.78 | | 0.99 | 2.21 | 0.76 | | 0.99 | 2.50 | 0.78 | | 0.99 |
| | Kendall | 1.36 | 0.70 | | 0.99 | 2.00 | 0.75 | | 0.99 | 1.84 | 0.74 | | 0.99 | 2.08 | 0.75 | | 0.99 |
| | Spearman | 1.36 | 0.70 | | 0.99 | 2.00 | 0.75 | | 0.99 | 1.84 | 0.74 | | 0.99 | 2.08 | 0.75 | | 0.99 |
| | SpearmanU | 1.36 | 0.70 | | 0.99 | 2.00 | 0.75 | | 0.99 | 1.84 | 0.74 | | 0.99 | 2.08 | 0.75 | | 0.99 |
| | InvCov | 8.39 | 0.78 | | 0.00 | 13.25 | 0.85 | | 0.00 | 10.06 | 1.88 | | 0.00 | 31.24 | 0.88 | | 0.00 |
| Diagonal | SampleCov | 8.44 | 0.89 | 0.00 | 0.00 | 13.37 | 0.93 | 0.00 | 0.00 | 5.07 | 0.77 | 0.01 | 0.00 | 15.41 | 0.90 | 0.00 | 0.00 |
| | OGK | 7.89 | 0.89 | 0.00 | 0.00 | 13.15 | 0.93 | 0.00 | 0.00 | 1.07 | 0.51 | 0.00 | 0.00 | 3.44 | 0.77 | 0.00 | 0.00 |
| | NPD | 0.76 | 0.43 | 0.00 | 0.00 | 1.37 | 0.58 | 0.00 | 0.00 | 1.11 | 0.52 | 0.00 | 0.00 | 1.25 | 0.55 | 0.00 | 0.00 |
| | Kendall | 0.70 | 0.44 | 0.00 | 0.00 | 1.00 | 0.50 | 0.00 | 0.00 | 0.93 | 0.48 | 0.00 | 0.00 | 1.02 | 0.50 | 0.00 | 0.00 |
| | Spearman | 0.70 | 0.44 | 0.00 | 0.00 | 1.00 | 0.50 | 0.00 | 0.00 | 0.93 | 0.48 | 0.00 | 0.00 | 1.02 | 0.50 | 0.00 | 0.00 |
| | SpearmanU | 0.70 | 0.44 | 0.00 | 0.00 | 1.00 | 0.50 | 0.00 | 0.00 | 0.93 | 0.48 | 0.00 | 0.00 | 1.02 | 0.50 | 0.00 | 0.00 |
| | InvCov | 8.44 | 0.76 | 1.00 | 0.00 | 13.37 | 0.85 | 1.00 | 0.00 | 5.07 | 2.12 | 1.00 | 0.00 | 15.41 | 0.92 | 1.00 | 0.00 |

Table 3.2: Simulation results for seven estimators and four sampling schemes, when $n = 200$ and $p = 120$. Performance is measured by $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$ for covariance matrix estimation (Cov), $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty$ for precision matrix estimation (Prec), and false positive rate (FP) and false negative rate (FN) for support recovery of the true precision matrix. The results are averaged over 100 replications.

| | | clean | | | | 5% rowwise | | | | 10% rowwise | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN |
| Banded | SampleCov | 1.24 | 0.33 | | 0.96 | 5.98 | 0.34 | | 0.85 | 10.34 | 0.35 | | 0.86 |
| | OGK | 1.38 | 0.34 | | 0.96 | 2.20 | 0.38 | | 0.95 | 3.10 | 0.41 | | 0.95 |
| | NPD | 1.64 | 0.38 | | 0.99 | 2.75 | 0.40 | | 0.89 | 3.95 | 0.42 | | 0.89 |
| | Kendall | 2.07 | 0.37 | | 0.97 | 2.76 | 0.34 | | 0.85 | 3.73 | 0.35 | | 0.86 |
| | Spearman | 2.07 | 0.37 | | 0.97 | 2.76 | 0.35 | | 0.86 | 3.73 | 0.35 | | 0.86 |
| | SpearmanU | 2.07 | 0.37 | | 0.97 | 2.76 | 0.35 | | 0.86 | 3.73 | 0.35 | | 0.86 |
| Sparse | SampleCov | 0.81 | 0.44 | 0.09 | 0.56 | 5.61 | 0.43 | 0.14 | 0.73 | 9.93 | 0.40 | 0.14 | 0.74 |
| | OGK | 0.96 | 0.45 | 0.09 | 0.59 | 1.86 | 0.53 | 0.09 | 0.62 | 2.87 | 0.61 | 0.10 | 0.62 |
| | NPD | 1.11 | 0.59 | 0.03 | 0.79 | 2.14 | 0.63 | 0.08 | 0.93 | 3.61 | 0.68 | 0.08 | 0.95 |
| | Kendall | 1.35 | 0.50 | 0.09 | 0.60 | 1.76 | 0.48 | 0.12 | 0.77 | 2.71 | 0.47 | 0.12 | 0.79 |
| | Spearman | 1.35 | 0.50 | 0.08 | 0.60 | 1.77 | 0.49 | 0.12 | 0.77 | 2.72 | 0.49 | 0.12 | 0.79 |
| | SpearmanU | 1.35 | 0.51 | 0.09 | 0.60 | 1.78 | 0.51 | 0.13 | 0.77 | 2.72 | 0.51 | 0.12 | 0.79 |
| Dense | SampleCov | 0.69 | 0.62 | | 1.00 | 5.53 | 0.62 | | 0.91 | 9.90 | 0.60 | | 0.91 |
| | OGK | 0.78 | 0.64 | | 1.00 | 1.29 | 0.69 | | 1.00 | 1.92 | 0.74 | | 1.00 |
| | NPD | 0.89 | 0.65 | | 1.00 | 1.54 | 0.68 | | 0.93 | 2.24 | 0.72 | | 0.91 |
| | Kendall | 1.17 | 0.68 | | 1.00 | 1.54 | 0.65 | | 0.92 | 2.12 | 0.70 | | 0.91 |
| | Spearman | 1.17 | 0.68 | | 1.00 | 1.54 | 0.65 | | 0.92 | 2.12 | 0.65 | | 0.91 |
| | SpearmanU | 1.17 | 0.68 | | 1.00 | 1.54 | 0.66 | | 0.92 | 2.12 | 0.65 | | 0.91 |
| Diagonal | SampleCov | 0.34 | 0.37 | 0.00 | 0.00 | 5.28 | 0.26 | 0.09 | 0.00 | 9.64 | 0.32 | 0.09 | 0.00 |
| | OGK | 0.38 | 0.38 | 0.00 | 0.00 | 0.58 | 0.36 | 0.00 | 0.00 | 0.78 | 0.44 | 0.00 | 0.00 |
| | NPD | 0.45 | 0.32 | 0.00 | 0.00 | 0.78 | 0.37 | 0.07 | 0.00 | 1.15 | 0.44 | 0.09 | 0.00 |
| | Kendall | 0.59 | 0.72 | 0.00 | 0.00 | 0.78 | 0.60 | 0.08 | 0.00 | 1.07 | 4.83 | 0.33 | 0.00 |
| | Spearman | 0.59 | 0.72 | 0.00 | 0.00 | 0.78 | 0.60 | 0.08 | 0.00 | 1.07 | 0.57 | 0.08 | 0.00 |
| | SpearmanU | 0.59 | 0.72 | 0.00 | 0.00 | 0.78 | 0.59 | 0.08 | 0.00 | 1.07 | 0.56 | 0.09 | 0.00 |

Table 3.3: Simulation results for six estimators and four sampling schemes, when $n = 200$ and $p = 400$. Performance is measured by $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$ for covariance matrix estimation (Cov), $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty$ for precision matrix estimation (Prec), and false positive rate (FP) and false negative rate (FN) for support recovery of the true precision matrix. The results are averaged over 100 replications.

| | | 5% cellwise | | | | 10% cellwise | | | | multivariate $t$ | | | | alternative $t$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN | Cov | Prec | FP | FN |
| Banded | `SampleCov` | 8.90 | 0.48 | | 0.69 | 13.70 | 0.46 | | 0.44 | 22.41 | 0.45 | | 0.87 | 137.82 | 0.57 | | 0.86 |
| | `OGK` | 8.79 | 0.48 | | 0.66 | 13.89 | 0.46 | | 0.39 | 3.97 | 0.44 | | 0.95 | 18.41 | 0.51 | | 0.53 |
| | `NPD` | 4.04 | 0.45 | | 0.98 | 7.03 | 0.45 | | 0.78 | 5.03 | 0.46 | | 0.94 | 5.83 | 0.48 | | 0.97 |
| | `Kendall` | 2.89 | 0.42 | | 0.96 | 4.11 | 0.46 | | 0.98 | 3.69 | 0.42 | | 0.96 | 3.99 | 0.43 | | 0.97 |
| | `Spearman` | 2.89 | 0.42 | | 0.96 | 4.11 | 0.46 | | 0.98 | 3.69 | 0.42 | | 0.96 | 3.99 | 0.43 | | 0.97 |
| | `SpearmanU` | 2.89 | 0.42 | | 0.96 | 4.11 | 0.46 | | 0.97 | 3.69 | 0.42 | | 0.96 | 3.99 | 0.44 | | 0.97 |
| Sparse | `SampleCov` | 8.98 | 0.91 | 0.01 | 0.96 | 13.82 | 0.85 | 0.52 | 0.45 | 13.53 | 0.79 | 0.05 | 0.80 | 79.44 | 0.96 | 0.04 | 0.85 |
| | `OGK` | 8.83 | 0.91 | 0.02 | 0.94 | 14.48 | 0.88 | 0.57 | 0.40 | 3.83 | 0.66 | 0.10 | 0.62 | 12.48 | 0.90 | 0.07 | 0.77 |
| | `NPD` | 3.10 | 0.72 | 0.03 | 0.83 | 6.15 | 0.82 | 0.02 | 0.87 | 4.40 | 0.76 | 0.03 | 0.84 | 4.67 | 0.78 | 0.03 | 0.86 |
| | `Kendall` | 1.80 | 0.64 | 0.06 | 0.74 | 2.94 | 0.74 | 0.05 | 0.82 | 2.61 | 0.63 | 0.07 | 0.69 | 2.71 | 0.66 | 0.07 | 0.67 |
| | `Spearman` | 1.80 | 0.65 | 0.06 | 0.74 | 2.94 | 0.74 | 0.05 | 0.82 | 2.61 | 0.64 | 0.07 | 0.69 | 2.71 | 0.66 | 0.07 | 0.67 |
| | `SpearmanU` | 1.80 | 0.65 | 0.07 | 0.73 | 2.94 | 0.75 | 0.05 | 0.82 | 2.61 | 0.64 | 0.07 | 0.68 | 2.71 | 0.66 | 0.07 | 0.67 |
| Dense | `SampleCov` | 8.96 | 0.90 | | 0.96 | 13.81 | 0.85 | | 0.46 | 12.64 | 0.88 | | 0.99 | 79.01 | 0.98 | | 1.00 |
| | `OGK` | 8.62 | 0.90 | | 0.93 | 13.64 | 0.85 | | 0.38 | 2.24 | 0.76 | | 1.00 | 10.33 | 0.92 | | 1.00 |
| | `NPD` | 2.35 | 0.77 | | 1.00 | 4.28 | 0.84 | | 1.00 | 2.82 | 0.79 | | 1.00 | 3.22 | 0.81 | | 1.00 |
| | `Kendall` | 1.64 | 0.72 | | 1.00 | 2.29 | 0.77 | | 1.00 | 2.12 | 0.75 | | 1.00 | 2.25 | 0.76 | | 1.00 |
| | `Spearman` | 1.64 | 0.72 | | 1.00 | 2.29 | 0.77 | | 1.00 | 2.12 | 0.75 | | 1.00 | 2.25 | 0.76 | | 1.00 |
| | `SpearmanU` | 1.64 | 0.72 | | 1.00 | 2.29 | 0.77 | | 1.00 | 2.12 | 0.75 | | 1.00 | 2.25 | 0.76 | | 1.00 |
| Diagonal | `SampleCov` | 9.03 | 0.90 | 0.00 | 0.00 | 13.93 | 0.87 | 0.47 | 0.00 | 6.33 | 0.77 | 0.01 | 0.00 | 39.73 | 0.95 | 0.00 | 0.00 |
| | `OGK` | 8.60 | 0.90 | 0.00 | 0.00 | 13.74 | 0.87 | 0.54 | 0.00 | 1.11 | 0.52 | 0.00 | 0.00 | 5.17 | 0.84 | 0.00 | 0.00 |
| | `NPD` | 1.20 | 0.54 | 0.00 | 0.00 | 2.19 | 0.69 | 0.00 | 0.00 | 1.42 | 0.58 | 0.00 | 0.00 | 1.62 | 0.62 | 0.00 | 0.00 |
| | `Kendall` | 0.81 | 0.52 | 0.00 | 0.00 | 1.15 | 0.54 | 0.00 | 0.00 | 1.06 | 0.52 | 0.00 | 0.00 | 1.14 | 0.54 | 0.00 | 0.00 |
| | `Spearman` | 0.81 | 0.52 | 0.00 | 0.00 | 1.15 | 0.54 | 0.00 | 0.00 | 1.06 | 0.52 | 0.00 | 0.00 | 1.14 | 0.54 | 0.00 | 0.00 |
| | `SpearmanU` | 0.81 | 0.52 | 0.00 | 0.00 | 1.15 | 0.54 | 0.00 | 0.00 | 1.06 | 0.52 | 0.00 | 0.00 | 1.14 | 0.54 | 0.00 | 0.00 |

Table 3.4: Simulation results for six estimators and four sampling schemes, when $n = 200$ and $p = 400$. Performance is measured by $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$ for covariance matrix estimation (Cov), $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty$ for precision matrix estimation (Prec), and false positive rate (FP) and false negative rate (FN) for support recovery of the true precision matrix. The results are averaged over 100 replications.

## 3.6 Discussion

We have derived statistical error bounds for high-dimensional robust precision matrix estimators, when data are drawn from a multivariate normal distribution and then observed subject to cellwise contamination. We show that in such settings, the precision matrix estimators that are obtained by plugging in pairwise robust covariance estimators to the GLasso or CLIME routine, as suggested by Oellerer & Croux (2014) and Tarr et al. (2015), have error bounds that match standard high-dimensional bounds for uncontaminated precision matrix estimation, up to an additive factor involving a constant multiple of the contamination fraction $\epsilon$. Our results for precision matrix estimators are derived via estimation error bounds for robust covariance matrix estimators, which have similar deviation properties.

The results of this chapter naturally suggest several venues for future work. In particular, it would be interesting to relate the nonasymptotic statistical error bounds to the behavior of the sensitivity curve of the robust covariance estimator, which is the finite-sample analog of the influence function. We have also left open the question of calculating the breakdown point for the CLIME estimator with respect to more general data matrices, as well as the breakdown behavior of CLIME and GLasso under different notions of breakdown point. Although our results imply the superiority of the GLasso over the CLIME estimator from the perspective of the finite-sample breakdown point, this may only be part of the story.

Lastly, it would be interesting to generalize our study to other classes of distributions. In one direction, it would be possible to study contaminated versions of other distributions besides the multivariate Gaussian, for which the precision matrix encodes information about the underlying graphical model (e.g., Ising models on trees). A harder question to tackle would be the problem of robust graphical model estimation in settings where the structure of the graph is not encoded in the preci-

sion matrix alone. Finally, one could consider robust estimation of scatter matrices, when the uncontaminated data are drawn from an elliptical distribution. In that case, the proposed Kendall's tau and Spearman's rho correlation coefficients would still be Fisher consistent upon taking the respective sine transformations, so similar error bounds should hold. As demonstrated in our simulation results, the pairwise covariance estimators based on Kendall's tau and Spearman's rho perform reasonably well when data are generated from either the multivariate $t$-distribution or the alternative $t$-distribution. This motivates studying the convergence rates of the same covariance matrix estimators under heavy-tailed or elliptical distributions.

The problem of estimating high-dimensional covariance matrices under various structural assumptions has also been widely studied. Various families of structured covariance matrices have been introduced, including bandable matrices (Cai et al., 2010), Toeplitz matrices (Cai et al., 2013), and sparse matrices (Bickel & Levina, 2008; Cai & Zhou, 2012). The proposed covariance matrix estimators involve regularizing the sample covariance matrix in accordance to structural assumptions. It would be interesting to study robust versions of these structured covariance matrix estimators under a model such as cellwise contamination. Besides graphical models, covariance matrix estimation is also useful for statistical methods such as linear discriminant analysis and principal component analysis. Several high-dimensional procedures have been proposed with proven theoretical guarantees when data are uncontaminated (Cai & Liu, 2011; Vu et al., 2013), and it would be interesting to study robust adaptations of these procedures, as well.

# 4

# Optimal Estimation of A Quadratic Functional under the Gaussian Two-Sequence Model

## 4.1 Introduction

The problem of estimating the quadratic functional $\int f^2$ occupies an important position in nonparametric statistical inference literature. In the density estimation setting where one observes an i.i.d. sample from a distribution with density function $f$, Bickel & Ritov (1988) was the first to show that there is an interesting phase transition where the minimax rate of convergence for estimating $\int f^2$ under mean squared error is the usual parametric rate when the Hölder smoothness parameter of the density function is greater than 1/4, and is otherwise slower than the parametric rate. Giné & Nickl (2008) constructed an adaptive estimator of $\int f^2$ in the density estimation setting. Donoho & Nussbaum (1990) developed a minimax theory for estimating quadratic functionals of periodic functions in the nonparametric regression model.

Quadratic functional estimation has been particularly well studied in the Gaussian

*Joint work with T. Tony Cai

sequence model:

$$Y_i = \theta_i + \sigma_n z_i, \qquad i = 1, 2, \ldots, \tag{4.1}$$

where $z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. The model (4.1) is equivalent to the white noise with drift model and can be used to approximate other nonparametric function estimation models. Estimating the quadratic functional $Q(\theta) = \sum \theta_i^2$ under (4.1) is the analog of estimating $\int f^2$ in the density estimation or nonparametric regression model. Fan (1991) and Efromovich & Low (1996) developed a minimax theory for estimating $Q(\theta) = \sum \theta_i^2$ over quadratically convex parameter spaces such as hyperrectangles and Sobolev balls. Cai & Low (2005, 2006b) further extended this theory to minimax and adaptive estimation over parameter spaces that are not necessarily quadratically convex. It is shown that the problem exhibits different phase transition phenomena in such a setting. A more recent paper by Collier et al. (2015) gave a non-asymptotic analysis of estimation of the quadratic functional over ellipsoids and classes of sparse vectors. The focus so far has been on the one-sequence case.

There are close connections between the problem of quadratic functional estimation and that of signal detection under (4.1). Specifically, for a mean vector $\theta$, we say that there is a signal at location $i$ if $\theta_i \neq 0$. The problem of signal detection is then to distinguish between $\theta = 0$ and $\theta \neq 0$. Since $Q(\theta) = 0$ if and only if $\theta = 0$, it is not surprising that estimators of $Q(\theta)$ can be used to construct procedures that are effective for detecting signals. See, for instance, Cai & Low (2005) and the references therein. The results on estimating the quadratic functional $Q(\theta)$ also have important implications on hypothesis testing and construction of confidence balls. See, for example, Li (1989), Dümbgen (1998), Lepski & Spokoiny (1999), Ingster & Suslina (2003), Baraud (2004), Genovese & Wasserman (2005), and Cai & Low (2006a,b).

In this chapter, we consider the estimation of the quadratic functional

$$Q(\mu, \theta) = \frac{1}{n} \sum_{i=1}^{n} \mu_i^2 \theta_i^2 \tag{4.2}$$

under the Gaussian two-sequence model,

$$X_i = \mu_i + \sigma z_i', \quad Y_i = \theta_i + \sigma z_i, \qquad i = 1, \ldots, n, \tag{4.3}$$

where $z_1', \ldots, z_n', \ z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\sigma$ is the noise level. The goal is to optimally estimate $Q(\mu, \theta)$ based on the observed data $(X_i, Y_i)$, $i = 1, ..., n$. Strictly speaking, $Q(\mu, \theta)$ is a quartic functional, but we will refer to it as a quadratic functional in the two-sequence case, as it is quadratic in $\mu$ given $\theta$, and vice versa. We are particularly interested in the case where both mean vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\theta = (\theta_1, \ldots, \theta_n)$ are sparse.

In addition to being of significant theoretical interest in its own right, this estimation problem is also motivated by the problem of simultaneous signal detection in integrative genomics, where it is of interest to test whether there are single nucleotide polymorphisms (SNPs) that are simultaneously associated with multiple human traits or disorders (Consortium, 2011; Cotsapas et al., 2011; Sivakumaran et al., 2011; Rankinen et al., 2015; Li et al., 2015). More specifically, let $X_i$ be the Z-score of the association between trait 1 and the $i^{th}$ SNP, and let $Y_i$ be the Z-score of the association between trait 2 and the $i^{th}$ SNP, for $i = 1, \ldots, n$. When the SNPs are chosen from different linkage equilibrium blocks, then it is approximately true that the $X_i$'s are independent, as are the $Y_i$'s. Moreover, when $X_i$ and $Y_i$ are calculated in independent datasets, then for each $i$, $X_i$ is independent of $Y_i$. In a simplified statistical framework, the simultaneous signal detection problem can then be studied under the Gaussian two-sequence model (4.3), where the goal is to detect the

presence of location $i$ with $\mu_i\theta_i \neq 0$. Equivalently, we want to distinguish between $\mu \star \theta = 0$ and $\mu \star \theta \neq 0$, where $\mu \star \theta = (\mu_1\theta_1, \ldots, \mu_n\theta_n)$ is the coordinate-wise product of $\mu$ and $\theta$. Of particular interest is the setting where the proportion of signals is small, and the signal strengths are relatively weak. This is indeed the setting in the genomics context, as only a small number of SNPs are expected to be associated with both traits. Moreover, the association, if it exists, is weak. Since $Q(\mu, \theta) = 0$ if and only if $\mu \star \theta = 0$, one might expect a connection similar to that in the single Gaussian sequence model to exist between the estimation problem and the simultaneous signal detection problem. More discussions on the application of quadratic functional estimators to the problem of simultaneous signal detection are given in Section 4.4.

In this chapter, we focus on studying the estimation of $Q(\mu, \theta)$. We propose optimal estimators of $Q(\mu, \theta)$ over a family of parameter spaces to be introduced, and establish the minimax rates of convergence. It is shown that the optimal rate exhibits interesting phase transitions in this family. Along with the establishment of the minimax rates of convergence, we explain the intuition behind the construction of the optimal estimators.

The rest of the chapter is organized as follows: Section 4.2 considers estimation of the functional $Q(\mu, \theta)$ and establishes the minimax rates of convergence. Section 4.3 complements our theoretical study with some simulation results. We conclude the chapter with a discussion in Section 4.4. Additional results not included in this chapter as well as the proofs of main results are relegated to Appendix C.

## 4.2 Optimal Estimation of $Q(\mu, \theta)$

In this section, we consider the estimation of the quadratic functional $Q(\mu, \theta) = \frac{1}{n}\sum_{i=1}^n \mu_i^2\theta_i^2$ of two sparse normal mean vectors $\mu = (\mu_1, \ldots, \mu_n)$ and $\theta = (\theta_1, \ldots, \theta_n)$

under the Gaussian two-sequence model (4.3). An additional constraint is imposed on the number of coordinates that are simultaneously nonzero for both mean vectors. The noise level $\sigma$ in model (4.3) is assumed to be known. Estimation of the noise level, $\sigma$, is relatively easy under the sparse sequence model (4.3) and will be discussed in Section 4.3.

We begin by introducing some notation that will be used throughout this chapter. Given a vector $\theta = (\theta_1, \ldots, \theta_n)$, we denote by $\|\theta\|_0 = \mathrm{Card}(\{i : \theta_i \neq 0\})$ the $\ell_0$-quasi-norm of $\theta$, $\|\theta\|_2 = \sqrt{\sum_{i=1}^n \theta_i^2}$ its $\ell_2$-norm, and $\|\theta\|_\infty = \max_{1\leq i\leq n} |\theta_i|$ its $\ell_\infty$-norm. For any real numbers $a$ and $b$, we set $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$ and $a_+ = a \vee 0$. Throughout, the notation $a_n \asymp b_n$ means that there exists some numerical constants $c$ and $C$ such that $c \leq \frac{a_n}{b_n} \leq C$ when $n$ is large. By "numerical constants" we usually mean constants that might depend on the characteristics of the problem but whose specific values are of little interest to us. The precise values of the numerical constants $c$ and $C$ may also vary from line to line.

Adopting an asymptotic framework where the vector size $n$ is the driving variable, we parameterize the signal strength, sparsity, and simultaneous sparsity of $\mu$ and $\theta$ as functions of $n$. Specifically, we consider the family of parameter spaces

$$\Omega(\beta, \epsilon, b) = \{(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}^n : \|\mu\|_0 \leq k_n, \|\mu\|_\infty \leq s_n, \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n, \\ \|\mu \star \theta\|_0 \leq q_n\}, \tag{4.4}$$

indexed by three parameters $\beta, \epsilon$, and $b$. We have the sparsity parameterization

$$k_n = n^\beta, \qquad 0 < \beta < \frac{1}{2}, \tag{4.5}$$

the simultaneous sparsity parameterization

$$q_n = n^{\epsilon}, \qquad 0 < \epsilon \leq \beta, \tag{4.6}$$

and the signal strength parametrization

$$s_n = n^{b}, \qquad b \in \mathbb{R}. \tag{4.7}$$

In principle, $\beta$ can take any value between 0 and 1. We are primarily interested in the estimation problem for the range $0 < \beta < \frac{1}{2}$, as it is well-known that this corresponds to the case of rare signals (Donoho & Jin, 2004).

Our goal is to derive the minimax rate of convergence for $Q(\mu, \theta)$ over $\Omega(\beta, \epsilon, b)$:

$$R^*(n, \Omega(\beta, \epsilon, b)) = \inf_{\widehat{Q}} \sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q} - Q(\mu, \theta))^2.$$

We will show that $R^*(n, \Omega(\beta, \epsilon, b))$ satisfies

$$R^*(n, \Omega(\beta, \epsilon, b)) \asymp \gamma_n(\beta, \epsilon, b), \tag{4.8}$$

where $\gamma_n(\beta, \epsilon, b)$ is a function of $n$ indexed by $\beta, \epsilon$ and $b$. There are two main tasks in establishing the minimax rate of convergence. For each triple $(\beta, \epsilon, b)$ satisfying $0 < \epsilon \leq \beta < \frac{1}{2}$ and $b \in \mathbb{R}$, we construct an estimator $\widehat{Q}^*$ that satisfies

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}^* - Q(\mu, \theta))^2 \leq C\gamma_n(\beta, \epsilon, b),$$

and show that $R^*(n, \Omega(\beta, \epsilon, b)) \geq c\gamma_n(\beta, \epsilon, b)$, where $C$ and $c$ are numerical constants that depend only on $\beta, \epsilon, b$, and $\sigma$. Combining these upper and lower bounds yields the minimax rate of convergence (4.8). In this case, we say that the estimator $\widehat{Q}^*$

attains the minimax rate of convergence over the parameter space $\Omega(\beta, \epsilon, b)$.

Interestingly, the estimation problem exhibits different phase transitions for the minimax rate $\gamma_n(\beta, \epsilon, b)$ in three regimes: the *sparse* regime where $0 < \epsilon < \frac{\beta}{2}$, the *moderately dense* regime where $\frac{\beta}{2} \leq \epsilon \leq \frac{3\beta}{4}$, and the *strongly dense* regime where $\frac{3\beta}{4} < \epsilon \leq \beta$. Collectively, we call $\frac{\beta}{2} \leq \epsilon \leq \beta$ the *dense* regime. In the sparse regime, the simultaneous signal is sparse in the sense that $q_n \ll \sqrt{k_n}$, while in the dense regime, the simultaneous signal is dense in the sense that $q_n \gg \sqrt{k_n}$. This is analogous to the terminology used in the one-sequence model, where the signal is called sparse if $0 < \beta < \frac{1}{2}$ ($k_n \ll \sqrt{n}$), and dense if $\frac{1}{2} \leq \beta \leq 1$ ($k_n \gg \sqrt{n}$). The key distinction is that, in the two-sequence case, sparseness or denseness is used to describe the relationship between simultaneous sparsity $q_n$ and sparsity $k_n$, as opposed to between $k_n$ and the vector size $n$. We remark that our use of the terminology is not superficial — a detailed analysis of lower and upper bounds for the estimation problem does reveal an intimate connection to the corresponding regimes in the one-sequence case. In particular, when the signal is moderately strong, the hardness of the two-sequence estimation problem is essentially characterized by an underlying one-sequence problem that displays different behavior in the sparse and the dense regimes. On the other hand, we construct optimal estimators for $Q(\mu, \theta)$, borrowing intuition from optimal estimators for $Q(\theta)$ in respective regimes.

Intuitively, when $b$ is very small (i.e., signal is very weak), we are better off estimating $Q(\mu, \theta)$ by

$$\widehat{Q}_0 = 0, \tag{4.9}$$

since any attempt to estimate $Q(\mu, \theta)$ will incur a greater estimation risk. On the other hand, when $b$ is sufficiently large (i.e., signal is strong), it is desirable to estimate $Q(\mu, \theta)$ based on the observed data $(X_i, Y_i)$, $i = 1, \ldots, n$. With a slight abuse of terminology, we say that the signal is weak if it corresponds to the region where $\widehat{Q}_0$ is

optimal, and we say that the signal is strong otherwise. In Sections 4.2.1 and 4.2.2, we construct two estimators of $Q(\mu, \theta)$ that respectively attain the minimax rates of convergence over the sparse and dense regimes when the signal is sufficiently large.

It is possible to generalize our parametrization to the case where $\mu$ and $\theta$ have different levels of both sparsity and signal strengths. This amounts to estimating $Q(\mu, \theta)$ over the parameter space

$$\Omega(\alpha, \beta, \epsilon, a, b) = \{(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}^n : \|\mu\|_0 \leq j_n, \|\mu\|_\infty \leq r_n, \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n, \\ \|\mu \star \theta\|_0 \leq q_n\}, \tag{4.10}$$

where $j_n = n^\alpha, k_n = n^\beta, q_n = n^\epsilon$ with $0 < \epsilon \leq \alpha \wedge \beta < \frac{1}{2}$, and $r_n = n^a, s_n = n^b$ with $a, b \in \mathbb{R}$. In this section, however, we will focus on the simplest case where $j_n = k_n = n^\beta$ and $r_n = s_n = n^b$, since the technical analysis is similar to that for the more general case (4.10), but less tedious. We did derive the minimax rates of convergence for the case where $j_n = k_n = n^\beta$ but $r_n$ and $s_n$ are allowed to differ. As the phase transitions for the minimax rates of convergence in this case are much more sophisticated, but also are less easily digestible, we opt to defer its presentation to Appendix C. The analysis for the general case (4.10) where no equality constraint is imposed on either the sparsity or signal strength of $\mu$ and $\theta$ follows similarly, provided that the magnitude of the simultaneous sparsity $\epsilon$ is compared to $\alpha$ if $a \geq b$, and to $\beta$ if $b \geq a$, for the determination of sparse and dense regimes.

### 4.2.1 Estimation in the Sparse Regime

We begin with the estimation of $Q(\mu, \theta) = \frac{1}{n} \sum \mu_i^2 \theta_i^2$ over the parameter space $\Omega(\beta, \epsilon, b)$ in the sparse regime, where $q_n$ is calibrated as in expression (4.6) with $0 < \epsilon < \frac{\beta}{2}$.

To construct an optimal estimator for $Q(\mu, \theta)$, we base our intuition on the estimation of the quadratic functional $Q(\theta) = \frac{1}{n}\sum \theta_i^2$, in the case where we only have one sequence of observations $Y_i$, $i = 1, \ldots, n$, from model (4.3). Consider the family of parameter spaces indexed by $k_n = n^\beta, 0 < \beta < 1$ and $s_n = n^b, b \in \mathbb{R}$:

$$\Theta(\beta, b) = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n\}. \tag{4.11}$$

It can be shown that for $0 < \beta < \frac{1}{2}$, the minimax rate of convergence for $Q(\theta)$ over $\Theta(\beta, b)$ satisfies

$$R^*(n, \Theta(\beta, b)) := \inf_{\widehat{Q}} \sup_{\theta \in \Theta(\beta,b)} E_\theta(\widehat{Q} - Q(\theta))^2 \asymp \gamma_n(\beta, b), \tag{4.12}$$

where

$$\gamma_n(\beta, b) = \begin{cases} n^{2\beta+4b-2} & \text{if } b \leq 0, \\ n^{2\beta-2}(\log n)^2 & \text{if } 0 < b \leq \frac{\beta}{2}, \\ n^{\beta+2b-2} & \text{if } b > \frac{\beta}{2}. \end{cases} \tag{4.13}$$

When $0 < \beta < \frac{1}{2}$, we have $k_n \ll \sqrt{n}$. Thus, we anticipate only very few coordinates of $\theta$ to be nonzero. If, in addition, $b < 0$, then the signal is both rare and weak, and one can do no better than simply estimating $Q(\theta)$ by $\widehat{Q}_0 = 0$. Nonetheless, when $b > 0$, the signal is rare but sufficiently strong, and the estimator

$$\widehat{Q}_1 = \frac{1}{n}\sum_{i=1}^n [(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0], \qquad \text{where } \theta_0 := E(Z^2 - \sigma^2\tau_n)_+, Z \sim N(0, \sigma^2), \tag{4.14}$$

that performs coordinate-wise thresholding on $Y_i^2$ with choice of tuning parameter $\tau_n = 2\log n$ is optimal. Each term $\theta_i^2$ is estimated independently by $(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0$, since the sparsity pattern is unstructured. The estimator (4.14) involves a thresholding step, $(Y_i^2 - \sigma^2\tau_n)_+$, for denoising, and a de-bias step by subtracting $\theta_0$ from the

thresholded term so that we estimate the zero coordinates of $\theta$ unbiasedly. This is important because the proportion of zero entries in this case is relatively large, and a biased estimator for these coordinates will unnecessarily inflates the estimation risk.

The results on the estimation of one-sequence quadratic functional over classes of sparse vectors in (4.11)-(4.14) (and that over classes of dense vectors in (4.19)-(4.20)) are new, though we were made aware of the appearance of similar results in the concurrent work of Collier et al. (2015). The focus and main contribution of this chapter is on the estimation of the quadratic functional $Q(\mu, \theta)$ in the two-sequence case.

We now return to the sparse regime in the two-sequence setting, where $0 < \epsilon < \frac{\beta}{2}$ and $0 < \beta < \frac{1}{2}$. In this case, $k_n \ll \sqrt{n}$, so the signal of individual sequences is rare. Moreover, the simultaneous sparsity $q_n \ll \sqrt{k_n}$ implies that we rarely have signals occurring simultaneously at the same coordinate of each sequence. This means that if we know for sure that $\mu_i$ is nonzero, it is unclear if $\theta_i$ is nonzero unless $|\theta_i|$ is large enough (and vice versa). Such an intuition motivates the estimator

$$\widehat{Q}_2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i^2 - \sigma^2\tau_n)_+ - \mu_0][(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0], \tag{4.15}$$

where $\mu_0 = \theta_0 := E(Z^2 - \sigma^2\tau_n)_+$ with the threshold level $\tau_n = \log n$, where $Z \sim N(0, \sigma^2)$. The construction of $\widehat{Q}_2$ is a straightforward extension of the construction of $\widehat{Q}_1$: each term $\mu_i^2\theta_i^2$ is estimated independently by the product $[(X_i^2 - \sigma^2\tau_n)_+ - \mu_0][(Y_i^2 - \sigma^2\tau_n)_+ - \theta_0]$. Since $q_n \ll \sqrt{k_n}$, following our previous argument, thresholding $X_i^2$ and $Y_i^2$ *independently* at a common threshold level is natural.

We now present a theorem on the upper bound of the mean squared error of $\widehat{Q}_2$.

**Theorem 11** (Sparse Regime: Upper Bound)**.** *For $b > 0$, the estimator $\widehat{Q}_2$, as in*

(4.15) *with* $\tau_n = \log n$, *satisfies*

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2 - Q(\mu,\theta))^2 \leq C\Big[n^{2\epsilon+4b-2}(\log n)^2 + n^{\epsilon+6b-2}\Big]. \tag{4.16}$$

Straightforward calculation shows that for the estimator $\widehat{Q}_0 = 0$,

$$\begin{aligned}\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_0 - Q(\mu,\theta))^2 &= \sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} \left(\frac{1}{n}\sum_{i=1}^{n}\mu_i^2\theta_i^2\right)^2 \\ &= q_n^2 s_n^8 n^{-2} = n^{2\epsilon+8b-2}, \end{aligned}\tag{4.17}$$

for $0 < \epsilon \leq \beta < \frac{1}{2}$ and $b \in \mathbb{R}$. We now show that the combination of $\widehat{Q}_0$ (when $b < 0$) and $\widehat{Q}_2$ (when $b \geq 0$) is optimal, by providing a matching lower bound.

**Theorem 12** (Sparse Regime: Lower Bound)**.** *Let* $0 < \epsilon < \frac{\beta}{2}$ *and* $0 < \beta < \frac{1}{2}$. *Then*

$$R^*(n, \Omega(\beta,\epsilon,b)) \geq c\gamma_n(\beta,\epsilon,b),$$

*where*

$$\gamma_n(\beta,\epsilon,b) = \begin{cases} n^{2\epsilon+8b-2} & \text{if } b \leq 0, \\ n^{2\epsilon+4b-2}(\log n)^2 & \text{if } 0 < b \leq \frac{\epsilon}{2}, \\ n^{\epsilon+6b-2} & \text{if } b > \frac{\epsilon}{2}. \end{cases} \tag{4.18}$$

Crucial to the derivation of the lower bound is the Constrained Risk Inequality (CRI) given in Brown & Low (1996). To apply CRI, it suffices to construct two priors supported on $\Omega(\beta,\epsilon,b)$ that have small chi-square distance but a large difference in the expected values of the resulting quadratic functionals. The cases $b \leq \frac{\epsilon}{2}$ and $b > \frac{\epsilon}{2}$ correspond to choices of distinct pairs of priors. For $b > \frac{\epsilon}{2}$, the CRI boils down to the standard technique of inscribing a hardest hyperrectangle, with the Bayes risk for a simple prior supported on the hyperrectangle being a lower bound for the minimax risk. Nevertheless, the case $b \leq \frac{\epsilon}{2}$ requires the use of a rich collection of hyperrect-

angles and a mixture prior which mixes over the vertices of the hyperrectangles in this collection. Mixing increases the difficulty of the Bayes estimation problem and is needed here to attain a sharp lower bound.

**Remark 13.** Combining (4.16), (4.17) and (4.18), we see that when $0 < \epsilon < \frac{\beta}{2}$ and $0 < \beta < \frac{1}{2}$, $\widehat{Q}_2$ attains the optimal rate of convergence over $\Omega(\beta, \epsilon, b)$ when $b > 0$. On the other hand, $\widehat{Q}_0$ attains the optimal rate of convergence over $\Omega(\beta, \epsilon, b)$ when $b \leq 0$.

**Remark 14.** So far, we have implicitly assumed that $\beta$ is fixed and we characterize each regime by the relative magnitude of $\epsilon$ to $\beta$. It is possible to turn this view the other way around, to assume that $\epsilon$ is fixed and to characterize each regime by the relative magnitude of $\beta$ to $\epsilon$. We then see from (4.18) that within the sparse regime where $0 < 2\epsilon < \beta < \frac{1}{2}$, the minimax rate of convergence $\gamma_n(\beta, \epsilon, b)$ for a fixed $\epsilon$ does not involve $\beta$. Such a lack of dependency on $\beta$ is also highlighted in the two plot panels in the bottom row of Figure 4.1.

### 4.2.2 Estimation in the Dense Regime

We now consider estimating $Q(\mu, \theta)$ in the dense regime, where $q_n$ is calibrated as in expression (4.6) with $\frac{\beta}{2} \leq \epsilon \leq \beta$. The dense regime is subdivided into two cases: the moderately dense case with $\frac{\beta}{2} \leq \epsilon \leq \frac{3\beta}{4}$ and the strongly dense case with $\frac{3\beta}{4} < \epsilon \leq \beta$.

In the dense regime, the estimator $\widehat{Q}_2$ defined in (4.15) is suboptimal, as the thresholding step in both $X_i^2$ and $Y_i^2$ ends up thresholding too many coordinates when the signal is weak. Note that the simultaneous sparsity $q_n \gg \sqrt{k_n}$ suggests that for each coordinate $i$ with $\mu_i \neq 0$, it is more often the case that $\theta_i \neq 0$ (compared to when $q_n \ll \sqrt{k_n}$), and vice versa. Therefore, it is no longer reasonable to perform thresholding on $X_i^2$ and $Y_i^2$ independently. The additional knowledge of relatively

high proportion of simultaneous nonzero entries suggests that whenever we observe a large value of $X_i^2$ (an implication of $\mu_i \neq 0$), then even if $Y_i^2$ is small, we should still estimate $\mu_i^2\theta_i^2$ rather than setting it equals zero. The same reasoning applies to the case where $X_i^2$ is small but $Y_i^2$ is large.

To construct an optimal estimator in the dense regime, we again borrow some intuition from the estimation of the quadratic functional $Q(\theta) = \frac{1}{n}\sum \theta_i^2$ in the one-sequence case. We consider the family of parameter spaces given in (4.11), but for $\frac{1}{2} \leq \beta < 1$. The minimax rate of convergence once again satisfies (4.12), but with

$$\gamma_n(\beta, b) = \begin{cases} n^{2\beta+4b-2} & \text{if } b \leq \frac{1-2\beta}{4}, \\ n^{-1} & \text{if } \frac{1-2\beta}{4} < b \leq \frac{1-\beta}{2}, \\ n^{\beta+2b-2} & \text{if } b > \frac{1-\beta}{2}. \end{cases} \tag{4.19}$$

When $\frac{1}{2} \leq \beta < 1$, we have $k_n \gg \sqrt{n}$, meaning that $\theta$ contains a relatively large number of non-zero coordinates compared to the case when $0 < \beta < \frac{1}{2}$. The characterization of weak and strong signal is no longer $b < 0$ versus $b \geq 0$ as in the case of $0 < \beta < \frac{1}{2}$, but $b \leq \frac{1-2\beta}{4}$ versus $b > \frac{1-2\beta}{4}$. That is, given the same signal strength $b$, the relatively large number of nonzero coordinates of $\theta$ when $k_n \gg \sqrt{n}$ collectively represents a stronger signal as compared to the case when $k_n \ll \sqrt{n}$. Thus, the threshold of "strong" signal as encoded by $b$ is lowered when $k_n \gg \sqrt{n}$. It is not surprising that for the range of weak signal $b \leq \frac{1-2\beta}{4}$, the estimator $\widehat{Q}_0 = 0$ is optimal. On the other hand, when $b > \frac{1-2\beta}{4}$, the optimal estimator for $Q(\theta)$ is the unbiased estimator

$$\widehat{Q}_3 = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - \sigma^2). \tag{4.20}$$

An optimal estimator is often one that strikes an appropriate balance between bias and variance in its mean squared error. The estimators $\widehat{Q}_0$ and $\widehat{Q}_3$ represent two extremes in terms of bias-variance tradeoff. We see that the $\widehat{Q}_0$ that is optimal for

exceedingly weak signal has zero variance, while the $\widehat{Q}_3$ that is optimal for sufficiently strong signal has zero bias. Due to the denseness of nonzero coordinates when $k_n \gg \sqrt{n}$, one could not afford to introduce bias to the estimator in the hope of achieving smaller variance. Without additional information about the sparsity structure, the unbiased estimator $\widehat{Q}_3$ is necessary for optimal estimation of $Q(\theta)$.

We now return to the two-sequence setting for the estimation of $Q(\mu, \theta)$, for the case $\frac{\beta}{2} \leq \epsilon \leq \beta$ and $0 < \beta < \frac{1}{2}$. Although the signal for individual sequences is sparse ($k_n \ll \sqrt{n}$), the simultaneous signal is dense in the sense that $q_n \gg \sqrt{k_n}$. The intuition garnered from the one-sequence case motivates the estimator

$$\widehat{Q}_4 = \frac{1}{n} \sum_{i=1}^{n} \big[ (X_i^2 - \sigma^2)(Y_i^2 - \sigma^2) \mathbb{1}(X_i^2 \vee Y_i^2 > \sigma^2 \tau_n) - \eta \big], \tag{4.21}$$

where

$$\eta = E[(Z_1^2 - \sigma^2)(Z_2^2 - \sigma^2) \mathbb{1}(Z_1^2 \vee Z_2^2 > \sigma^2 \tau_n)], \qquad Z_1, Z_2 \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

From $\widehat{Q}_4$, we see that each term $\mu_i^2 \theta_i^2$ is estimated unbiasedly (modulo $\eta$) by $(X_i^2 - \sigma^2)(Y_i^2 - \sigma^2)$ whenever at least one of $X_i^2$ and $Y_i^2$ is sufficiently large. This is in accordance with our previous argument that estimation should be done whenever we have at least one large value of $X_i^2$ or $Y_i^2$. The threshold $\tau_n$ is a tuning parameter whose value is yet to be determined during the analysis of the mean squared error of $\widehat{Q}_4$, though it turns out that $\tau_n = c \log n$ for any $c \geq 4$ attains the optimal rate of convergence. The subtraction of $\eta$ from $(X_i^2 - \sigma^2)(Y_i^2 - \sigma^2) \mathbb{1}(X_i^2 \vee Y_i^2 > \sigma^2 \tau_n)$ is needed because the majority of coordinates $i$ has $\mu_i = \theta_i = 0$. A biased estimator for these coordinates unavoidably inflates the estimation risk. The naive unbiased estimator

$$\frac{1}{n} \sum_{i=1}^{n} (X_i^2 - \sigma^2)(Y_i^2 - \sigma^2)$$

does not seem to perform well when $0 < \beta < \frac{1}{2}$ due to the rarity of nonzero coordinates in individual sequences. A thresholding step $\mathbb{1}(X_i^2 \vee Y_i^2 > \sigma^2\tau_n)$ is needed to guard against estimating entries with $\mu_i = \theta_i = 0$ with noise.

Note that $\widehat{Q}_2$ defined in (4.15) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}[(X_i^2 - \sigma^2\tau_n)\mathbb{1}(X_i^2 > \sigma^2\tau_n) - \mu_0][(Y_i^2 - \sigma^2\tau_n)\mathbb{1}(Y_i^2 > \sigma^2\tau_n) - \theta_0].$$

Comparing this expression with $\widehat{Q}_4$, we see that when both $X_i^2$ and $Y_i^2$ are large, the term $\mu_i^2\theta_i^2$ is roughly estimated as $(X_i^2 - \sigma^2\tau_n)(Y_i^2 - \sigma^2\tau_n)$. Moreover, $(X_i^2 - \sigma^2\tau_n)(Y_i^2 - \sigma^2\tau_n)$ is a biased estimator of $\mu_i^2\theta_i^2$ when $\tau_n > 1$.

We present an upper bound on the mean squared error of $\widehat{Q}_4$ in the following theorem.

**Theorem 13** (Dense Regime: Upper Bound)**.** *For $b > 0$, the estimator $\widehat{Q}_4$, as in* (4.21) *with $\tau_n = 4\log n$, satisfies*

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_4 - Q(\mu,\theta))^2 \leq C\max\left\{n^{2\epsilon-2}(\log n)^4, n^{\epsilon+6b-2}, n^{\beta+4b-2}\right\}. \quad (4.22)$$

We now provide a matching lower bound to complement the upper bound in the dense regime.

**Theorem 14** (Dense Regime: Lower Bound)**.** *Let $\frac{\beta}{2} \leq \epsilon \leq \beta$ and $0 < \beta < \frac{1}{2}$. Then*

$$R^*(n, \Omega(\beta,\epsilon,b)) \geq c\gamma_n(\beta,\epsilon,b),$$

*where*

$$\gamma_n(\beta,\epsilon,b)=\begin{cases} n^{2\epsilon+8b-2} & \text{if } b\le 0,\\ n^{2\epsilon-2}(\log n)^4 & \text{if } 0<b\le \frac{2\epsilon-\beta}{4},\\ n^{\beta+4b-2} & \text{if } \frac{2\epsilon-\beta}{4}<b\le \frac{\beta-\epsilon}{2},\\ n^{\epsilon+6b-2} & \text{if } b>\frac{\beta-\epsilon}{2}, \end{cases} \tag{4.23}$$

*when* $\frac{\beta}{2}\le\epsilon\le\frac{3\beta}{4}$, *and*

$$\gamma_n(\beta,\epsilon,b)=\begin{cases} n^{2\epsilon+8b-2} & \text{if } b\le 0,\\ n^{2\epsilon-2}(\log n)^4 & \text{if } 0<b\le \frac{\epsilon}{6},\\ n^{\epsilon+6b-2} & \text{if } b>\frac{\epsilon}{6}, \end{cases} \tag{4.24}$$

*when* $\frac{3\beta}{4}<\epsilon\le\beta$.

The minimax rates of convergence display different phase transitions within the two subdivisions of the dense regime. In the moderately dense regime where $\frac{\beta}{2}\le\epsilon\le\frac{3\beta}{4}$, there are phase transitions at $b=\frac{2\epsilon-\beta}{4}$ and $b=\frac{\beta-\epsilon}{2}$, given in (4.23). Note that $\frac{2\epsilon-\beta}{4}\le\frac{\beta-\epsilon}{2}$ if and only if $\epsilon\le\frac{3\beta}{4}$. In the strongly dense regime where $\epsilon>\frac{3\beta}{4}$, the phase $\frac{2\epsilon-\beta}{4}<b\le\frac{\beta-\epsilon}{2}$ is non-existent, and we only have one intermediate phase, $0<b\le\frac{\epsilon}{6}$, given in (4.24).

We establish the lower bound by constructing least favorable priors and applying CRI. Except for the rate $n^{\epsilon+6b-2}$, which is obtained through the inscription of a hardest hyperrectangle, all other cases require some forms of mixing over the vertices of a rich collection of hyperrectangles.

**Remark 15.** Combining (4.17), (4.22), (4.23), and (4.24), we see that for the parameter space $\Omega(\beta,\epsilon,b)$ with $\frac{\beta}{2}\le\epsilon\le\beta<\frac{1}{2}$, $\widehat{Q}_4$ attains the minimax rate of convergence when $b>0$. On the other hand, $\widehat{Q}_0=0$ attains the minimax rate of convergence when $b\le 0$.

**Remark 16.** Following Remark 14, we see that similar to the sparse regime, the

minimax rate of convergence $\gamma_n(\beta, \epsilon, b)$ for a fixed $\epsilon$ does not involve $\beta$ in the strongly dense regime where $\epsilon \leq \beta < \frac{4\epsilon}{3}$. In contrast, $\gamma_n(\beta, \epsilon, b)$ for a fixed $\epsilon$ depends explicitly on $\beta$ in the moderately dense regime where $\frac{4\epsilon}{3} \leq \beta \leq 2\epsilon$. The dependency or lack of dependency of $\gamma_n(\beta, \epsilon, b)$ on $\beta$ within each regime is also illustrated in the two plot panels at the bottom of Figure 4.1.

Interestingly, in the two-sequence case, the regions $\{b : b \leq 0\}$ and $\{b : b > 0\}$ appear to constitute the regions of weak signal and strong signal, respectively, regardless of the level of simultaneous sparsity. This is in contrast to the one-sequence case where the dividing line is $b = 0$ when $k_n \ll \sqrt{n}$, and $b = \frac{1-2\beta}{4}$ when $k_n \gg \sqrt{n}$. We caution that this apparent "reconciliation" in the two-sequence case is simply because the signal strengths are taken to be the same for both sequences $\mu$ and $\theta$ in the simplified results presented above.

**Remark 17.** When the signal strengths $r_n = n^a$ and $s_n = n^b$ of $\mu$ and $\theta$ are allowed to differ, it turns out that $\{(a, b) : a \wedge b \leq 0\}$ characterizes the region of weak signal when $q_n \ll \sqrt{k_n}$, while $\{(a, b) : a \vee b \leq 0\} \cup \{(a, b) : a \wedge b \leq \frac{\beta-2\epsilon}{4}\}$ comprises the region of weak signal when $q_n \gg \sqrt{k_n}$. We refer the readers to Appendix C for more details.

### 4.2.3 Phase Transitions in the Minimax Rates of Convergence

We see from Sections 4.2.1 and 4.2.2 that within each regime, the minimax rates of convergence exhibit several phase transitions. In addition, each transition is governed by a change in the relative magnitudes of the sparsity parameter $\beta$, the simultaneous sparsity parameter $\epsilon$, and the signal strength parameter $b$. In fact, it is the way phase transitions occur within each regime that characterizes the regime itself. Furthermore,

the phase transitions actually display "continuity" across the boundaries of different regimes.

To depict what we meant graphically, first note that from Sections 4.2.1 and 4.2.2, the minimax rates of convergence

$$\gamma_n(\beta, \epsilon, b) \asymp n^{r(\beta,\epsilon,b)}, \tag{4.25}$$

modulo a factor involving $\log n$ when applicable. In Figure 4.1, we plot the rate exponent $r(\beta, \epsilon, b)$ against $b$ for the sparse, moderately dense, and strongly dense regimes.

Specifically, in the top row of Figure 4.1, we fix $\beta = 0.45$ and plot $r(\beta, \epsilon, b)$ against $b$ for a range of $\epsilon$ values in $(0, \beta)$. The top left panel of Figure 4.1 provides a continuum view of $r(\beta, \epsilon, b)$, as $\epsilon$ increases from 0 to $\beta$. Each piecewise straight line corresponds to an $\epsilon$ value in the considered range. To highlight the discrepancy among the three regimes, we color the sparse regime ($0 < \epsilon < \frac{\beta}{2}$) in red, the moderately dense regime ($\frac{\beta}{2} \leq \epsilon \leq \frac{3\beta}{4}$) in green, and the strongly dense regime ($\frac{3\beta}{4} < \epsilon \leq \beta$) in blue. We see that the three regimes have somewhat different behaviors for small positive values of $b$. In particular, the sparse regime and the strongly dense regime experience two transitions (three different slopes), while the moderately dense regime experiences three transitions (four different slopes). Note that the difference in the number of transitions is restored at the intersection of the blue region and the red region. Thus, the phase transition is in some sense "continuous" across the regime boundaries — the piecewise straight lines corresponding to $r(\beta, \epsilon, b)$'s exhibit smooth transition as $\epsilon$ increases from 0 to $\beta$. The top right panel of Figure 4.1 provides a static view for each regime. We plot $r(\beta, \epsilon, b)$ against $b$ for three values of $\epsilon$ corresponding to three different regimes: $\epsilon = 0.12$ (sparse regime), $\epsilon = 0.28$ (moderately dense regime), and $\epsilon = 0.4$ (strongly dense regime). The knots on each dashed line indicate the transition

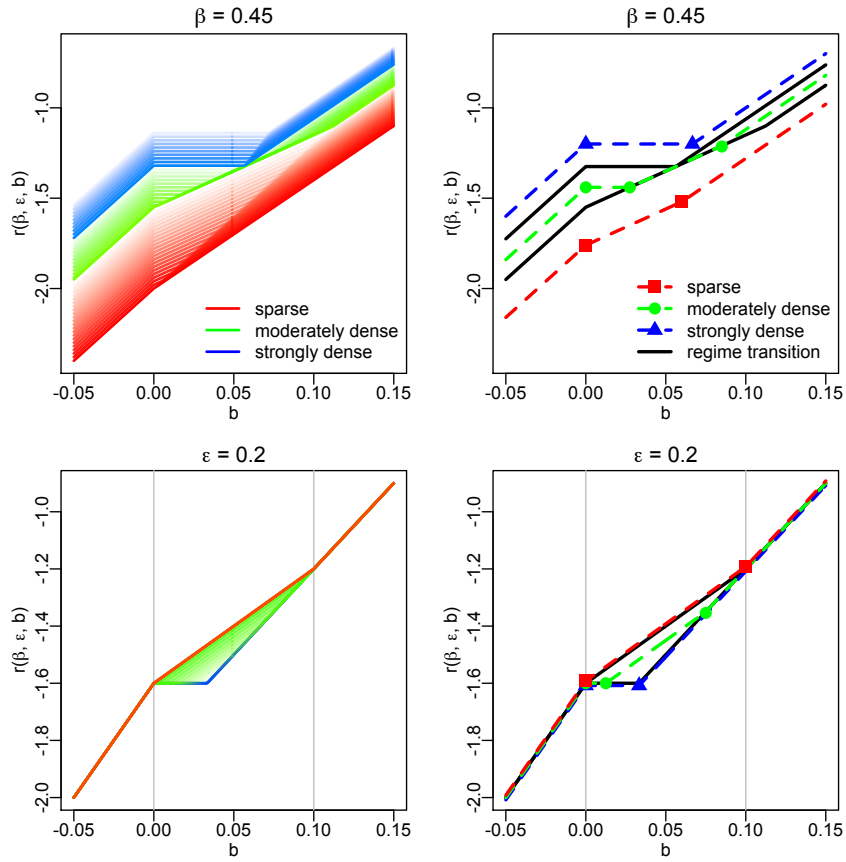


Figure 4.1: Plot of the rate exponent $r(\beta, \epsilon, b)$ against the signal strength $b$. In the sparse regime (—), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon + 4b - 2, \epsilon + 6b - 2$. In the moderately dense regime (—), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon - 2, \beta + 4b - 2, \epsilon + 6b - 2$. In the strongly dense regime (—), $r(\beta, \epsilon, b)$ changes in the order $2\epsilon + 8b - 2, 2\epsilon - 2, \epsilon + 6b - 2$. Top row, left panel: a continuum view of $r(\beta, \epsilon, b)$ as $\epsilon$ increases from 0 to $\beta = 0.45$ (color changes from red to blue). Top row, right panel: a static view of each regime: sparse ($\epsilon = 0.12$), moderately dense ($\epsilon = 0.28$), and strongly dense ($\epsilon = 0.4$). Transition points are indicated by the knots on the dashed lines. Bottom row, left panel: a continuum view of $r(\beta, \epsilon, b)$ as $\beta$ increases from $\epsilon = 0.2$ to 0.5 (color changes from blue to red). Grey vertical lines indicate $b = 0$ and $b = \frac{\epsilon}{2}$. Bottom row, right panel: a static view of each regime: strongly dense ($\beta = 0.25$), moderately dense ($\beta = 0.35$), and sparse ($\beta = 0.45$).

points for the slope of the line.

On the other hand, in the bottom row of Figure 4.1, we fix $\epsilon = 0.2$ and plot $r(\beta, \epsilon, b)$ against $b$ for a range of $\beta$ values in $(\epsilon, 0.5)$. The bottom left panel of Figure 4.1 provides a continuum view of $r(\beta, \epsilon, b)$, as $\beta$ increases from $\epsilon$ to 0.5. Again, the the strongly dense regime ($\epsilon \leq \beta < \frac{4\epsilon}{3}$) is colored in blue, the moderately dense regime ($\frac{4\epsilon}{3} \leq \beta \leq 2\epsilon$) in green, and the sparse regime ($\beta > 2\epsilon$) in red, with each piecewise straight line corresponding to a $\beta$ value in the considered range. The two grey vertical lines indicate the locations $b = 0$ and $b = \frac{\epsilon}{2}$. Note that all the red lines overlap (so do all the blue lines), indicating that $r(\beta, \epsilon, b)$ for a fixed $\epsilon$ is independent of $\beta$ in the sparse regime and the strongly dense regime. In the moderately dense regime, $r(\beta, \epsilon, b)$ only depends on $\beta$ when $0 < b < \frac{\epsilon}{2}$. The bottom right panel of Figure 4.1 provides a static view for each regime. We plot $r(\beta, \epsilon, b)$ against $b$ for three values of $\beta$: $\beta = 0.25$ (strongly dense regime), $\beta = 0.35$ (moderately dense regime), and $\beta = 0.45$ (sparse regime). Due to the overlap of all lines in the range $b \leq 0$ and $b > \frac{\epsilon}{2}$, we shift the dashed lines corresponding to $\beta = 0.45$ and $\beta = 0.25$ (in red and in blue, respectively) slightly to aid distinguishing the changes of $r(\beta, \epsilon, b)$ in different regimes.

## 4.3 Simulation

In this section, we report on simulation studies to compare the performance of the three estimators $\widehat{Q}_0 = 0, \widehat{Q}_2$ as in (4.15), and $\widehat{Q}_4$ as in (4.21), under different scenarios. We computed the mean squared error (MSE) of the three estimators to show that our simulation results are compatible with the theoretical results given in Section 4.2.

So far, we have assumed that the noise level $\sigma$ is known. In practice, $\sigma$ is typically unknown and needs to be estimated. Under the sparse setting of the present chapter,

$\sigma$ is easily estimable. Let $M \in \mathbb{R}^{2n}$ have $M_{2i-1} = X_i$ and $M_{2i} = Y_i$ for $i = 1, \ldots, n$. A simple robust estimator of the noise level $\sigma$ can be obtained from the median absolute deviation (MAD) of the combined sample:

$$\hat{\sigma} = \frac{\text{median}_j |M_j - \text{median}_k(M_k)|}{0.6745}.$$

Such an estimator has been used in Donoho & Johnstone (1994) for wavelet estimation.

We considered simulation studies over a range of sample size $n$, sparsity $k_n = n^\beta$, simultaneous sparsity $q_n = n^\epsilon$, and signal strength $s_n = n^b$. More specifically, we took $n \in \{10^3, 10^4, \ldots, 10^7\}$, $\beta = 0.45$ for individual sequences, $b \in \{-0.1, 0.15, 0.2\}$, and three values of simultaneous sparsity, one for each regime: $\epsilon = 0.02$ (sparse regime), $\epsilon = 0.3$ (moderately dense regime) and $\epsilon = 0.44$ (strongly dense regime). For each $(n, \beta, \epsilon, b)$, we generated data from the Gaussian two-sequence model (4.3) with $\mu, \theta \in \{0, \pm n^b\}^n$, $\|\mu\|_0 = \|\theta\|_0 = [n^\beta]$, and $\|\mu \star \theta\|_0 = [n^\epsilon]$, where $[\cdot]$ denotes rounding to the nearest integer. Figure 4.2 is the plot of the MSE (averaged over 200 replications) of the three estimators against sample size in the log-log scale, for each combination of simultaneous sparsity and signal strength.

The theoretical results in Section 4.2 indicate that for $\widehat{Q} = \widehat{Q}_0, \widehat{Q}_2$, or $\widehat{Q}_4$,

$$\sup_{(\mu,\theta) \in \Omega(\beta,\epsilon,b)} E(\widehat{Q} - Q(\mu, \theta))^2 \asymp n^{r(\beta,\epsilon,b)}$$

for some rate exponent $r(\beta, \epsilon, b)$ (modulo a logarithmic factor when applicable). Thus, it is not surprising that the results in Figure 4.2 (mostly) exhibit a linear pattern. When the signal is weak with $b = -0.1$ (see the first row of Figure 4.2), we see that $\widehat{Q}_0$ (wide-dashed line) and $\widehat{Q}_4$ (dotted line) have the lowest mean squared error. Note that we expect $\widehat{Q}_0$ to be optimal when the signal is weak. We observe that $\widehat{Q}_4$

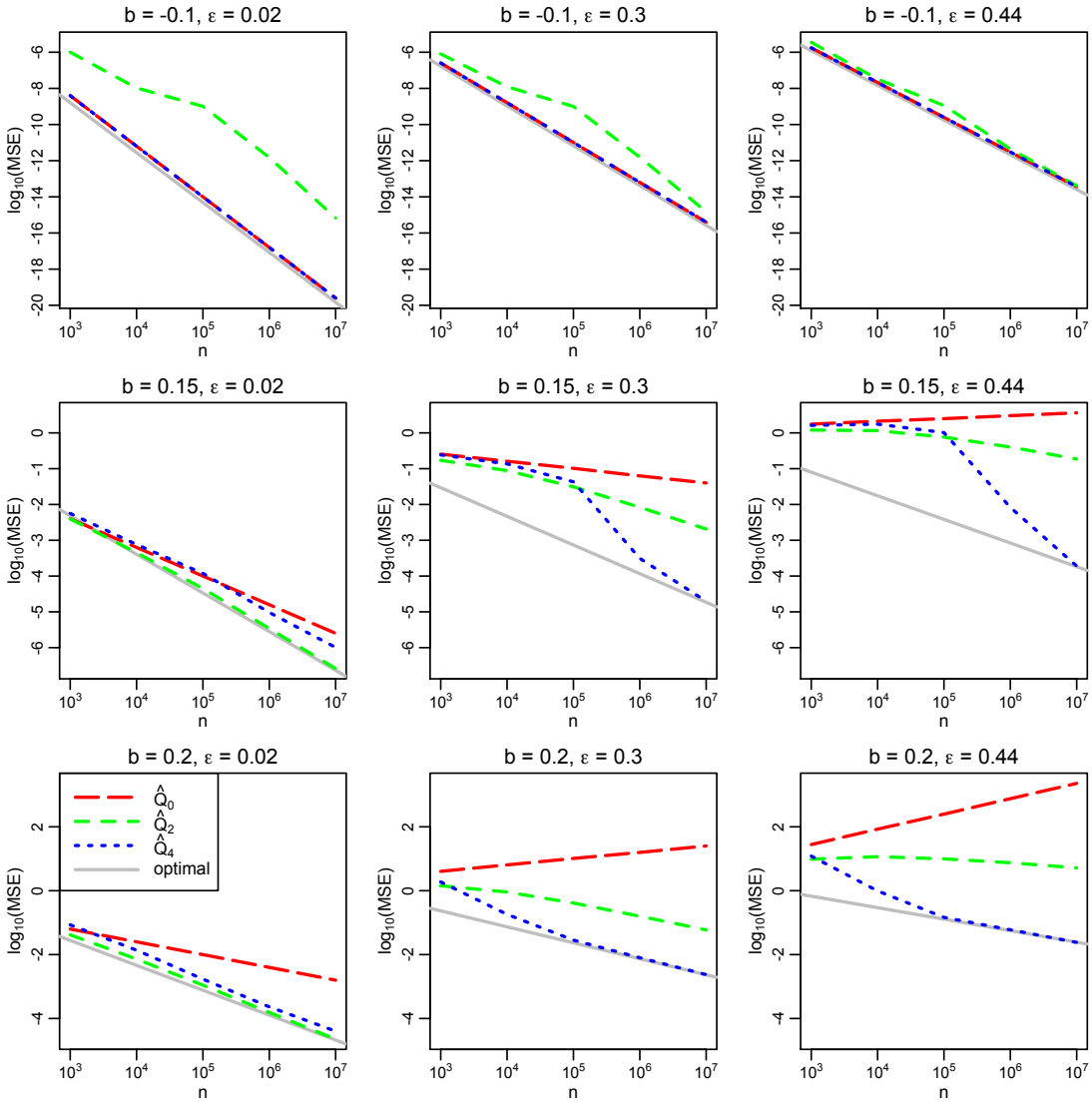


Figure 4.2: Plot of MSE for the estimators $\widehat{Q}_0$, $\widehat{Q}_2$, and $\widehat{Q}_4$ over different sample sizes $n \in \{10^3, \ldots, 10^7\}$, in the log-log scale. Fixing $\beta = 0.45$, the columns are ordered from left to right as $\epsilon = 0.02$ (sparse regime), $\epsilon = 0.3$ (moderately dense regime), and $\epsilon = 0.44$ (strongly dense regime). The rows are ordered from top to bottom in increasing signal strength: $b \in \{-0.1, 0.15, 0.2\}$. Solid line has a slope equal to that of the optimal rate exponent $r(\beta, \epsilon, b)$.

is nearly as good as $\widehat{Q}_0$ from Figure 4.2. This is because when the signal is weak, the thresholding step $\mathbb{1}(X_i^2 \vee Y_i^2 \geq \sigma^2\tau_n)$ thresholds both noise and weak signals, and the de-bias term $\eta$ is extremely small when $n$ is moderately large, resulting in $\widehat{Q}_4 \approx \widehat{Q}_0 = 0$. As the signal becomes sufficiently strong ($b \in \{0.15, 0.2\}$), $\widehat{Q}_2$ starts to dominate in the sparse regime ($\epsilon = 0.02$) while $\widehat{Q}_4$ dominates in the moderately dense and strongly dense regimes ($\epsilon \in \{0.3, 0.44\}$). When the signal is sufficiently large ($b \in \{0.15, 0.2\}$), $\widehat{Q}_0$ is clearly suboptimal. In particular, in the case where signal is both dense and strong ($b = 0.2, \epsilon \in \{0.3, 0.44\}$), the MSE of $\widehat{Q}_0$ diverges to infinity, as indicated by the positive slope of the wide-dashed line. Note also that as either $\epsilon$ or $b$ increases, MSE increases, as can be seen by the flattening or reversing of slopes towards the right end or bottom of the plot panel. This is compatible with the fact that $r(\beta, \epsilon, b)$ increases with respect to both $\epsilon$ and $b$.

For each combination $(\beta, \epsilon, b)$, the solid line has a slope equal to the optimal rate exponent $r(\beta, \epsilon, b)$, and an intercept deliberately selected so that it lies close to the line corresponding to the optimal estimator. We see from Figure 4.2 that for all combinations of $(b, \epsilon)$ except $b = 0.15, \epsilon \in \{0.3, 0.44\}$, the slope of the solid line aligns well with that of the optimal estimator, confirming the validity of our theoretical results. We conjecture that in the case $b = 0.15, \epsilon \in \{0.3, 0.44\}$, the worst case rate of the optimal estimator $\widehat{Q}_4$ in $\Omega(\beta, \epsilon, b)$ is not attained at the configuration of location and magnitude of nonzero entries in $\mu, \theta$ considered in the simulation. This can be seen from the fact that $\widehat{Q}_4$ has a steeper slope than the optimal one (i.e., faster rate of convergence) for sufficiently large $n$.

## 4.4 Discussion

In this chapter, we discuss the estimation of the quadratic functional $Q(\mu, \theta) = \frac{1}{n}\sum \mu_i^2\theta_i^2$ over a family of parameter spaces where $\mu$ and $\theta$ are constrained in terms of the magnitude, sparsity, and simultaneous sparsity. Similar to the one-sequence estimation problem, we show that the minimax rates of convergence display different phase transitions over the sparse regime and the dense regime. Different from the one-sequence estimation problem, in the two-sequence case, the dense regime can be further subdivided into the moderately dense regime and the strongly dense regime. Despite the similarity in terminology, we emphasize that denseness and sparseness refer to the relationship between simultaneous sparsity and individual sparsity in the two-sequence problem, rather than that between sparsity and vector size as in the one-sequence problem. The construction of the optimal estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ are inspired by their one-sequence correspondence in respective regimes, with appropriate modification that accounts for the structure of the two-sequence problem.

Our study of the two-sequence estimation problem can be generalized in several aspects. In Appendix C, we show that the optimal rates of convergence for estimation of $Q(\mu, \theta)$ continue to subsume the aforementioned regimes, when $\mu$ and $\theta$ are allowed unequal signal strengths. Moreover, the optimal rates are attained by the same estimators in respective regimes. Nonetheless, the distinction between the sparse and dense regimes is more apparent in this setting. In the sparse regime, estimation is only desirable when the signal strengths of both sequences are sufficiently strong. In contrast, in the dense regime, estimation is desirable whenever at least one sequence admits a sufficiently strong signal (and the signal strength of the other sequence is not too weak). Throughout this chapter, we assume that the sequences $\{X_i : 1 \le i \le n\}$ and $\{Y_i : 1 \le i \le n\}$ have a common noise level $\sigma$. Our analysis can be easily extended to the case where $\sigma_X \neq \sigma_Y$, by appropriately replacing the threshold levels

in the proposed estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ with ones that involve $\sigma_X$ or $\sigma_Y$. Such a modification yields estimators which attain minimax rates of convergence that are identical to that given in this chapter. When $\sigma_X$ and $\sigma_Y$ are unknown, we can use MAD to estimate the noise level of each sequence and plug in to the modified estimators.

The focus of this chapter is on minimax rates of convergence for the estimation of $Q(\mu, \theta)$. Adaptive estimation of $Q(\mu, \theta)$ is an interesting but technically challenging problem. Cai & Low (2005) introduced a block thresholding estimator for adaptive estimation of the quadratic functional in the one-sequence setting. It would be interesting to explore whether a similar idea could be used for adaptively estimating the quadratic functional in the two-sequence setting. In this chapter, we consider the estimation of $Q(\mu, \theta)$ over the parameter space defined in (4.4), where signal strengths are incorporated through the $\ell_\infty$-norm. For future work, it would also be interesting to study the behavior of the estimation problem under an $\ell_p$-norm constraint on the signal strengths, where $p \in (0, \infty)$.

A problem that is closely related to the estimation of the quadratic functional $Q(\mu, \theta)$ is the simultaneous signal detection problem, where the goal is to distinguish between $\mu \star \theta = 0$ and $\mu \star \theta \neq 0$. In the single Gaussian sequence setting where one observes $Y_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$, it is of interest to test $\theta = 0$ against $\theta \neq 0$, and there are two natural approaches: the sum of squares type test statistic $\sum Y_i^2$ and the max-type test statistic $\max |Y_i|$. Simultaneous signal detection generalizes the one-sequence testing problem and arises frequently in the context of integrative genomics. In genetics, for instance, it is often of interest to identify polymorphisms that are associated with multiple related conditions (Rankinen et al., 2015; Li et al., 2015). The problem of simultaneous signal detection has been studied by Zhao et al. (2012) under a mixture model framework, and a max-type statistic, $\max(|X_i| \wedge |Y_i|)$,

is proposed for detecting sparse simultaneous signals. On the other hand, in this chapter we study the estimation of quadratic functional under the sequence model framework. The proposed estimators $\widehat{Q}_2$ and $\widehat{Q}_4$ can be applied to the simultaneous signal detection problem as well. Similar to the problem of quadratic functional estimation, it turns out that the simultaneous signal detection problem behaves differently over two regimes. In the dense regime, a signal is detectable provided the signal strength of at least one of the sequences is sufficiently strong and the signal strength of the other sequence is not too weak. In contrast, in the sparse regime, a signal is only detectable when both sequences admit sufficiently strong signals. A crude analysis shows that the test procedures based on the statistics $\widehat{Q}_2$ and $\widehat{Q}_4$ are effective in detecting simultaneous signals over the respective detectable regions. A complete analysis of the optimality and adaptivity of such a test procedure is an interesting but challenging problem which we leave for future work.

# A

# Supplement for Chapter 2

This chapter contains supporting materials for Chapter 2. We present in Section A.1 the proofs for technical results given in Section 2.5. Proofs of the consistency results stated in Section 2.6 are given in Section A.2, whereas proofs related to the power algorithm of Section 2.7 are given in Section A.3. Section A.4 contains implementation details for the power algorithm, while Section A.5 contains an alternative linear algebra method for computing sample kernel APCs. A comparison of kernel APC with kernel PCA is given in Section A.6.

## A.1 Proofs for Section 2.5

### Proof of Lemma 1

*Proof.* Since $\mathcal{H}^0$ is finite-dimensional and the covariance matrix of a basis of $\mathcal{H}^0$ is of full-rank, $\mathrm{Var}\,(\cdot)$ induces a norm on $\mathcal{H}^0$, thereby turns it into a Hilbert space. It is easy to check that any finite-dimensional Hilbert space is also an RKHS, so $\mathcal{H}^0$ is an RKHS with respect to $\mathrm{Var}\,(\cdot)$. To simplify notation, in below we will write $\|\phi\|_0^2$ for $\mathrm{Var}\,(\phi)$. We now show boundedness of evaluation functionals on $\mathcal{H}$ wrt $\|\phi\|_\alpha^2 = \|\phi\|_0^2 + \alpha\|\phi\|_1^2$, where $\alpha > 0$. Let $\phi = \phi^1 + \phi^0$ be uniquely decomposed into $\phi^0 \in \mathcal{H}^0$ and

$\phi^1 \in \mathcal{H}^1$, and note $\|\phi^1\|_1^2 = \|\phi\|_1^2$ because $\phi^0$ is in the null space of $\|\cdot\|_1$. To express statements such as $\|\phi^1\|_0^2 \leq c\|\phi^1\|_1^2$ for some constant $c$ not depending on $\phi^1$, we use the simplifying notation $\|\phi^1\|_0^2 \lesssim \|\phi^1\|_1^2$. Under the assumption $E(k^1(X,X)) < \infty$, we do have $\|\phi^1\|_0^2 = \mathrm{Var}\,(\phi^1(X)) \leq E(\phi^1(X)^2) = E(\langle \phi^1, k_x^1\rangle_1^2) \leq E(\|\phi^1\|_1^2\|k_x^1\|_1^2) = \|\phi^1\|_1^2 E(k^1(X,X)) \lesssim \|\phi^1\|_1^2$. Facts such as $\|\phi+\psi\|^2 \leq 2(\|\phi\|^2+\|\psi\|^2)$ can be expressed as $\|\phi+\psi\|^2 \lesssim \|\phi\|^2+\|\psi\|^2$. In the following derivation, explanations in parens describe the action needed to step to the next line:

$$
\begin{array}{rcccll}
|\phi(x)|^2 & \lesssim & |\phi^0(x)|^2 & + & |\phi^1(x)|^2 & \text{(apply RKHS assumptions)} \\
& \lesssim & \|\phi^0\|_0^2 & + & \|\phi^1\|_1^2 & \text{(use } \phi^0 = \phi - \phi^1, \quad \|\phi^1\|_1^2 = \|\phi\|_1^2\text{)} \\
& \lesssim & (\|\phi\|_0^2 + \|\phi^1\|_0^2) & + & \|\phi\|_1^2 & \text{(use } \|\phi^1\|_0^2 \lesssim \|\phi^1\|_1^2\text{)} \\
& \lesssim & (\|\phi\|_0^2 + \|\phi^1\|_1^2) & + & \|\phi\|_1^2 & \text{(use } \|\phi^1\|_1^2 = \|\phi\|_1^2\text{)} \\
& \lesssim & (\|\phi\|_0^2 + \|\phi\|_1^2) & + & \|\phi\|_1^2 & \text{(use } \alpha > 0\text{)} \\
& \lesssim & \|\phi\|_0^2 + \alpha\|\phi\|_1^2 & & & \\
& = & \|\phi\|_\alpha^2 & & &
\end{array}
$$

We show next completeness of $\mathcal{H}$ wrt $\|\phi\|_\alpha^2$: Assume the sequence $\phi^{0(n)} + \phi^{1(n)}$ is Cauchy, i.e., $\|(\phi^{0(n)}+\phi^{1(n)}) - (\phi^{0(m)}+\phi^{1(m)})\|_\alpha^2 \to 0$ as $m, n \to \infty$. We then note:

$$\|(\phi^{0(n)}+\phi^{1(n)})-(\phi^{0(m)}+\phi^{1(m)})\|_\alpha^2 \;=\; \|(\phi^{0(n)}+\phi^{1(n)})-(\phi^{0(m)}+\phi^{1(m)})\|_0^2 + \alpha\|\phi^{1(n)}-\phi^{1(m)}\|_1^2$$

It follows that both terms on the right hand side converge to zero as $m, n \to \infty$. Convergence of the term $\|\phi^{1(n)} - \phi^{1(m)}\|_1^2$ implies that the sequence $\phi^{1(n)}$ is Cauchy in $\mathcal{H}^1$ wrt $\|\cdot\|_1$. By assumption $\mathcal{H}^1$ is RKHS, hence complete, granting that the sequence has a limit $\phi^{1(\infty)}$.

To address the existence of a limit for the sequence $\phi^{0(n)}$, we start by using the fact $\|\phi^1\|_0 \lesssim \|\phi^1\|_1$, which implies that $\|\phi^{1(n)} - \phi^{1(m)}\|_0$ also converges to zero. We use

next the following bound:

$$\begin{aligned}\|(\phi^{0(n)}+\phi^{1(n)}) - (\phi^{0(m)}+\phi^{1(m)})\|_0 &= \|(\phi^{0(n)}-\phi^{0(m)})+(\phi^{1(n)}-\phi^{1(m)})\|_0 \\ &\geq \big|\, \|\phi^{0(n)}-\phi^{0(m)}\|_0 - \|\phi^{1(n)}-\phi^{1(m)})\|_0 \,\big|\end{aligned}$$

The left hand term on the first line converges to zero by assumption, and we just showed that the term $\|\phi^{1(n)}-\phi^{1(m)})\|_0$ also converges to zero, implying together that $\|\phi^{0(n)}-\phi^{0(m)}\|_0$ must converge to zero as well. Hence the sequence $\phi^{0(n)}$ is Cauchy in $\mathcal{H}^0$ under $\|\cdot\|_0$ and has a limit $\phi^{0(\infty)}$ since $\mathcal{H}^0$ is an RKHS under $\|\cdot\|_0$.

We still need to show that the sequences $\phi^{0(n)}$ and $\phi^{1(n)}$ converge to their limits in the norm $\|\cdot\|_\alpha$, but this follows from $\|\phi^{0(n)} - \phi^{0(\infty)}\|_0 = \|\phi^{0(n)} - \phi^{0(\infty)}\|_\alpha$ and $\|\phi^{1(n)}-\phi^{1(\infty)}\|_\alpha \lesssim \|\phi^{1(n)}-\phi^{1(\infty)}\|_1$. It is finally proven that $\phi^{0(n)}+\phi^{1(n)} \to \phi^{0(\infty)}+\phi^{1(\infty)}$ in the norm $\|\cdot\|_\alpha$. □

## Proof of Lemma 3

*Proof.* Isometry follows from $\|\tilde{\phi}_1\|_1 = \|\phi^1 - T_0(\phi^1)\|_1 = \|\phi^1\|_1$ because $T_0(\phi^1) \in \mathcal{H}^0$. Boundedness of evaluation functionals is seen as follows, abbreviating $\phi^0 = T_0(\phi^1)$:

$$\begin{aligned}|\tilde{\phi}_1(x)|^2 &\lesssim |\phi^1(x)|^2 + |\phi^0(x)|^2 && \text{(apply RKHS assumptions)} \\ &\lesssim \|\phi^1\|_1^2 + \|\phi^0\|_0^2 && \text{(use continuity of } T_0: \ \|\phi^0\|_0 \lesssim \|\phi^1\|_1, \ ) \\ &\lesssim \|\phi^1\|_1^2 + \|\phi^1\|_1^2 \\ &\lesssim \|\phi^1\|_1^2 \\ &= \|\tilde{\phi}_1\|_1^2\end{aligned}$$

We now check that $E(\tilde{k}^1(X,X)) < \infty$. Since $\mathcal{H}^0$ is a finite-dimensional RKHS with respect to $\mathrm{Var}(\cdot)$, it is easy to check that its reproducing kernel $k^0$ satisfies

$E(k^0(X,X)) < \infty$. On the other hand,

$$\begin{aligned}
|\tilde{\phi}^1(x)| &\leq |\phi^1(x)| + |\phi^0(x)| && \text{(by reproducing property)} \\
&= |\langle \phi^1, k_x^1 \rangle_1| + |\langle \phi^0, k_x^0 \rangle_0| && \text{(by Cauchy-Schwarz Inequality)} \\
&\leq \|\phi^1\|_1 \|k_x^1\|_1 + \|\phi^0\|_0 \|k_x^0\|_0 && \text{(use continuity of } T_0 : \ \|\phi^0\|_0 \lesssim \|\phi^1\|_1) \\
&\lesssim \|\phi^1\|_1 \|k_x^1\|_1 + \|\phi^1\|_1 \|k_x^0\|_0 && \text{(use } \|\phi^1\|_1 = \|\tilde{\phi}\|_1) \\
&= \|\tilde{\phi}^1\|_1 \|k_x^1\|_1 + \|\tilde{\phi}^1\|_1 \|k_x^0\|_0
\end{aligned}$$

Plugging in $\tilde{\phi}^1 = \tilde{k}_x^1$, we obtain $\|\tilde{k}_x^1\|_1^2 = \tilde{k}_x^1(x) \lesssim \|\tilde{k}_x^1\|_1 \|k_x^1\|_1 + \|\tilde{k}_x^1\|_1 \|k_x^0\|_0$, and therefore, $\|\tilde{k}_x^1\|_1 \lesssim \|k_x^1\|_1 + \|k_x^0\|_0$. It follows that $E(\tilde{k}^1(X,X)) \lesssim E(k^1(X,X)) + E(k^0(X,X)) < \infty$. □

## Proof of Lemma 4

*Proof.* If $\mathcal{H}^1$ is an RKHS complement granted by the assumptions of Lemma 1, let $T_0$ be the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}^0$ restricted to $\mathcal{H}^1$, where orthogonality is wrt $\|\cdot\|_\alpha^2 = \mathrm{Var}(\cdot) + \alpha\|\cdot\|_1^2$. Then $T_0 : \mathcal{H}^1 \to \mathcal{H}^0$ is bounded, and the associated RKHS complement $\tilde{\mathcal{H}}^1 = \{\phi^1 - T_0(\phi^1) \,|\, \phi^1 \in \mathcal{H}^1\}$ granted by Lemma 3 is the orthogonal complement of $\mathcal{H}^0$ wrt $\|\cdot\|_\alpha$. That is, $\tilde{\mathcal{H}}^1 = \{\phi \in \mathcal{H} : \langle \phi, \phi^0 \rangle_\alpha = 0 \ \forall \phi^0 \in \mathcal{H}^0\}$. Finally, $\tilde{\mathcal{H}}^1$ does not depend on $\|\cdot\|_1$ because $\langle \phi, \phi^0 \rangle_\alpha = \mathrm{Cov}(\phi, \phi^0)$ due to $\langle \phi, \phi^0 \rangle_1 = 0$ for all $\phi^0 \in \mathcal{H}^0$. □

## Proof of Lemma 5

*Proof.* We first decompose $\mathcal{H}$ as $\mathcal{H} = \mathcal{H}^0 \oplus \tilde{\mathcal{H}}^1$, where $\tilde{\mathcal{H}}^1$ is the canonical complement of $\mathcal{H}^0$ wrt $\langle \cdot, \cdot \rangle_\alpha$. Let $k, k^0, \tilde{k}^\alpha$ denote the reproducing kernel of $\mathcal{H}, \mathcal{H}^0, \tilde{\mathcal{H}}^1$, respectively, wrt $\langle \cdot, \cdot \rangle_\alpha$, and let $\tilde{k}^1$ denote the reproducing kernel of $\tilde{\mathcal{H}}^1$ wrt $\langle \cdot, \cdot \rangle_1$. Then $k = k^0 + \tilde{k}^\alpha$. That $\mathcal{H}^0$ is a finite-dimensional RKHS wrt $\mathrm{Cov}(\cdot, \cdot)$ implies that $E(k^0(X,X)) < \infty$,

so it suffices to show that $E(\tilde{k}^\alpha(X,X)) < \infty$. For this, we need to connect $\tilde{k}^\alpha$ with $\tilde{k}^1$. By Lemma 3, $E(\tilde{k}^1(X,X)) < \infty$. This in turn implies that the covariance operator $\mathbf{C} : \tilde{\mathcal{H}}^1 \to \tilde{\mathcal{H}}^1$ given by $\langle \phi, \mathbf{C}\psi \rangle_1 = \mathrm{Cov}(\phi, \psi)$ exists and is bounded (Fukumizu et al., 2007). Given $\phi \in \tilde{\mathcal{H}}^1$, we then have

$$\begin{aligned}\phi(x) = \langle \phi, \tilde{k}^1_x \rangle_1 = \langle \phi, \tilde{k}^\alpha_x \rangle_\alpha &= \mathrm{Cov}(\phi, \tilde{k}^\alpha_x) + \alpha \langle \phi, \tilde{k}^\alpha_x \rangle_1 \\ &= \langle \phi, \mathbf{C}\tilde{k}^\alpha_x \rangle_1 + \alpha \langle \phi, \tilde{k}^\alpha_x \rangle_1 = \langle \phi, (\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1}) \tilde{k}^\alpha_x \rangle_1.\end{aligned}$$

It follows that $\langle \phi, \tilde{k}^\alpha_x \rangle_\alpha = \langle \phi, (\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1}) \tilde{k}^\alpha_x \rangle_1 \ \forall \phi \in \tilde{\mathcal{H}}^1$, and $\tilde{k}^1_x = (\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1}) \tilde{k}^\alpha_x$. Therefore,

$$\begin{aligned}\tilde{k}^\alpha(x,x) = \langle \tilde{k}^\alpha_x, \tilde{k}^\alpha_x \rangle_\alpha = \langle \tilde{k}^\alpha_x, (\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1}) \tilde{k}^\alpha_x \rangle_1 &= \langle \tilde{k}^\alpha_x, \tilde{k}^1_x \rangle_1 \\ = \langle (\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1})^{-1} \tilde{k}^1_x, \tilde{k}^1_x \rangle_1 &\leq \|(\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1})^{-1}\| \tilde{k}^1(x,x).\end{aligned}$$

Since $\|(\mathbf{C} + \alpha \mathbf{Id}_{\tilde{\mathcal{H}}^1})^{-1}\| \leq \alpha^{-1} < \infty$, we obtain $E(\tilde{k}^\alpha(X,X)) < \infty$. □

## A.2 Consistency Proof of Section 2.6

In this section, we give the consistency proof for sample kernel APCs. We begin by presenting in Section A.2.1 some basic properties of the operators $\hat{\mathbf{C}}^{(n)}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj})$ and $\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj})$, which forms the building blocks for the proof of three key lemmas in Section A.2.2.

### A.2.1 Proofs of Supporting Lemmas

We consider some lemmas that will be directly useful for establishing the proofs in Section A.2.2. The following lemma corresponds to Lemma 5 in Fukumizu et al.

(2007), and bounds the Hilbert-Schmidt norm of the difference between the empirical cross-covariance operator and the (population) cross-covariance operator.

**Lemma 10.** *The cross-covariance operator* $\mathbf{C}_{ij}$ *is Hilbert-Schmidt, and*

$$E\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\|_{\mathrm{HS}} = O(n^{-1/2}),$$

*where* $\|\cdot\|_{\mathrm{HS}}$ *denotes the Hilbert-Schmidt norm of a Hilbert-Schmidt operator.*

Corollary 4 is an immediate consequence of Lemma 10.

**Corollary 4.** *The cross-covariance operator* $\mathbf{C}_{ij}$ *satisfies*

$$P(\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\| > \epsilon) \leq d\epsilon^{-1}n^{-1/2},$$

*where* $d$ *is some constant that does not depend on* $n$*.*

*Proof.* Since the operator norm of an operator is dominated by its Hilbert-Schmidt norm, it follows from Lemma 10 that

$$P(\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\| > \epsilon) \leq P(\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\|_{\mathrm{HS}} > \epsilon) \leq \epsilon^{-1}E\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\|_{\mathrm{HS}} \leq d\epsilon^{-1}n^{-1/2}.$$

□

## Proof of Lemma 6

Since $\mathbf{C}_{jj} \succeq 0$, it is easy to see that $\mathbf{C}_{jj}+\alpha(\mathbf{Id}_{\mathcal{H}_j}-\mathbf{C}_{jj}) = (1-\alpha)\mathbf{C}_{jj}+\alpha\mathbf{Id}_{\mathcal{H}_j} \succeq \alpha\mathbf{Id}_{\mathcal{H}_j}$ for $0 < \alpha \leq 1$. On the other hand, by Corollary 4, there exist constants $d_j$ not depending on $n$ such that

$$P(\|\hat{\mathbf{C}}_{jj}^{(n)} - \mathbf{C}_{jj}\| > \epsilon) \leq d_j\epsilon^{-1}n^{-1/2}, \qquad 1 \leq j \leq p. \tag{A.1}$$

Since $\alpha_n \to 0$, for sufficiently large values of $n$, we have

$$\begin{aligned}\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) &= \left(\hat{\mathbf{C}}_{jj}^{(n)} - \mathbf{C}_{jj}\right) + \left(\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj})\right) \\ &\succeq \left(\hat{\mathbf{C}}_{jj}^{(n)} - \mathbf{C}_{jj}\right) + \alpha_n \mathbf{Id}_{\mathcal{H}_j} \\ &= \left(\hat{\mathbf{C}}_{jj}^{(n)} - \mathbf{C}_{jj} + \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j}\right) + \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j},\end{aligned}$$

and it follows that

$$\begin{aligned}P\left(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \succeq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j}\right) &\geq P\left(\hat{\mathbf{C}}_{jj}^{(n)} - \mathbf{C}_{jj} + \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j} \succeq \mathbf{0}\right) \\ &= P\left(\mathbf{C}_{jj} - \hat{\mathbf{C}}_{jj}^{(n)} \preceq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j}\right) \\ &\geq P\left(\|\mathbf{C}_{jj} - \hat{\mathbf{C}}_{jj}^{(n)}\| \leq \frac{\alpha_n}{2}\right) \\ &\geq 1 - 2d_j\alpha_n^{-1}n^{-1/2},\end{aligned}$$

where the last inequality is due to (A.1). Applying a union bound, we obtain

$$\begin{aligned}&P\left(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \succeq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j} \text{ for } 1 \leq j \leq p\right) \\ &= 1 - P\left(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \preceq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j} \text{ for some } j\right) \\ &\geq 1 - \sum_{j=1}^{p} P\left(\hat{\mathbf{C}}_{jj}^{(n)} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \preceq \frac{\alpha_n}{2}\mathbf{Id}_{\mathcal{H}_j}\right) \\ &\geq 1 - \delta,\end{aligned}$$

where $\delta = 2(\sum_{j=1}^{p} d_j)\alpha_n^{-1}n^{-1/2}$.

**Lemma 11.** *Suppose that Assumption 2 hold, and $\alpha_n \to 0$. Then, for sufficiently*

*large values of* $n$,

$$\inf_{\|\mathbf{\Phi}\|_\star=1} \left\{ \langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star \right\} \geq \alpha_n.$$

*Proof.* By Lemma 6,

$$\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}) \succeq \alpha_n \mathbf{Id}_{\mathcal{H}_j}, \qquad 1 \leq j \leq p,$$

for sufficiently large values of $n$. Hence,

$$\begin{aligned}
&\inf_{\|\mathbf{\Phi}\|_\star=1} \left\{ \langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star \right\} \\
&= \inf_{\sum_{j=1}^p \|\phi_j\|_{\star,j}^2=1} \left( \sum_{j=1}^p \langle \phi_j, (\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))\phi_j \rangle_{\star,j} \right) \\
&= \min_{1\leq j\leq p} \inf_{\|\phi_j\|_{\star,j}^2=1} \langle \phi_j, (\mathbf{C}_{jj} + \alpha_n(\mathbf{Id}_{\mathcal{H}_j} - \mathbf{C}_{jj}))\phi_j \rangle_{\star,j} \\
&\geq \alpha_n.
\end{aligned}$$

□

**Lemma 12.** *Suppose that Assumption 2 hold, and* $\alpha_n \to 0$. *Then, for sufficiently large values of* $n$,

$$P\left( \sup_{\mathbf{\Phi}\in\boldsymbol{\mathcal{H}}} \left| \frac{\langle \mathbf{\Phi}, \mathrm{diag}(\hat{\mathbf{C}}^{(n)})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star}{\langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star} - 1 \right| > \epsilon \right) \leq d\epsilon^{-1}\alpha_n^{-1}n^{-1/2},$$

*where* $d$ *is a constant not depending on* $n$.

*Proof.*

$$\sup_{\mathbf{\Phi}\in\boldsymbol{\mathcal{H}}} \left| \frac{\langle \mathbf{\Phi}, \mathrm{diag}(\hat{\mathbf{C}}^{(n)})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star}{\langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi} \rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi} \rangle_\star} - 1 \right| = \sup_{\|\mathbf{\Phi}\|_\star=1} \left| \frac{\langle \mathbf{\Phi}, (\mathrm{diag}(\hat{\mathbf{C}}^{(n)}) - \mathrm{diag}(\mathbf{C}))\mathbf{\Phi} \rangle_\star}{\langle \mathbf{\Phi}, (\mathrm{diag}(\mathbf{C}) + \mathbf{J}^{(n)})\mathbf{\Phi} \rangle_\star} \right|$$

$$
\begin{aligned}
&\leq \sup_{\|\mathbf{\Phi}\|_\star=1} \frac{|\langle \mathbf{\Phi}, (\mathrm{diag}(\hat{\mathbf{C}}^{(n)}) - \mathrm{diag}(\mathbf{C}))\mathbf{\Phi}\rangle_\star|}{\alpha_n} \\
&\leq \max_{1\leq j\leq p} \sup_{\|\phi_j\|^2_{\star,j}=1} \frac{|\langle \phi_j, (\hat{\mathbf{C}}^{(n)}_{jj} - \mathbf{C}_{jj})\phi_j\rangle_{\star,j}|}{\alpha_n} \\
&= \max_{1\leq j\leq p} \frac{\|\hat{\mathbf{C}}^{(n)}_{jj} - \mathbf{C}_{jj}\|}{\alpha_n},
\end{aligned}
$$

where the first inequality is due to Lemma 11. By Corollary 4, there exist constants $d_j$ not depending on $n$ such that

$$
P(\|\hat{\mathbf{C}}^{(n)}_{jj} - \mathbf{C}_{jj}\| > \epsilon) \leq d_j \epsilon^{-1} n^{-1/2}, \qquad 1 \leq j \leq p.
$$

Applying a union bound, it follows that with probability at most $(\sum_{j=1}^p d_j)\epsilon^{-1}\alpha_n^{-1}n^{-1/2}$,

$$
\max_{1\leq j\leq p} \frac{\|\hat{\mathbf{C}}^{(n)}_{jj} - \mathbf{C}_{jj}\|}{\alpha_n} \geq \sup_{\mathbf{\Phi}\in\mathcal{H}} \left| \frac{\langle \mathbf{\Phi}, \mathrm{diag}(\hat{\mathbf{C}}^{(n)})\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star}{\langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star} - 1 \right| > \epsilon.
$$

□

### A.2.2 Proofs of Main Lemmas

We are now ready to prove the three key lemmas given in Section 2.6.

## Proof of Lemma 7

*Proof.* By Lemma 12,

$$
\begin{aligned}
\hat{R}_{\alpha_n}(\mathbf{\Phi}) &= \frac{\langle \mathbf{\Phi}, \hat{\mathbf{C}}^{(n)}\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star}{\langle \mathbf{\Phi}, \mathrm{diag}(\hat{\mathbf{C}}^{(n)})\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star} \\
&= \frac{\langle \mathbf{\Phi}, \hat{\mathbf{C}}^{(n)}\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star}{\langle \mathbf{\Phi}, \mathrm{diag}(\mathbf{C})\mathbf{\Phi}\rangle_\star + \langle \mathbf{\Phi}, \mathbf{J}^{(n)}\mathbf{\Phi}\rangle_\star}(1 + O_p(\alpha_n^{-1}n^{-1/2})),
\end{aligned}
$$

where $O_p(\alpha_n^{-1}n^{-1/2})$ is a quantity that, when divided by $\alpha_n^{-1}n^{-1/2}$, is bounded in probability uniformly over all $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}$.

On the other hand, by Corollary 4, there exists a constant $d$ not depending on $n$ such that

$$P(\|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\| > \epsilon) \leq d\epsilon^{-1}n^{-1/2}, \qquad 1 \leq i, j \leq p.$$

Combined with Lemma 11, we obtain

$$\begin{aligned}
&\sup_{\boldsymbol{\Phi}\in\boldsymbol{\mathcal{H}}} \left| \frac{\langle \boldsymbol{\Phi}, \hat{\mathbf{C}}^{(n)}\boldsymbol{\Phi}\rangle_\star + \langle \boldsymbol{\Phi}, \mathbf{J}^{(n)}\boldsymbol{\Phi}\rangle_\star}{\langle \boldsymbol{\Phi}, \mathrm{diag}(\mathbf{C})\boldsymbol{\Phi}\rangle_\star + \langle \boldsymbol{\Phi}, \mathbf{J}^{(n)}\boldsymbol{\Phi}\rangle_\star} - R_{\alpha_n}(\boldsymbol{\Phi}) \right| \\
&= \sup_{\boldsymbol{\Phi}\in\boldsymbol{\mathcal{H}}} \left| \frac{\langle \boldsymbol{\Phi}, (\hat{\mathbf{C}}^{(n)} - \mathbf{C})\boldsymbol{\Phi}\rangle_\star}{\langle \boldsymbol{\Phi}, \mathrm{diag}(\mathbf{C})\boldsymbol{\Phi}\rangle_\star + \langle \boldsymbol{\Phi}, \mathbf{J}^{(n)}\boldsymbol{\Phi}\rangle_\star} \right| \\
&\leq \sup_{\|\boldsymbol{\Phi}\|_\star=1} \frac{|\langle \boldsymbol{\Phi}, (\hat{\mathbf{C}}^{(n)} - \mathbf{C})\boldsymbol{\Phi}\rangle_\star|}{\alpha_n} \\
&= \frac{\|\hat{\mathbf{C}}^{(n)} - \mathbf{C}\|}{\alpha_n} \\
&\leq \frac{p^2 \max_{1\leq i,j\leq p} \|\hat{\mathbf{C}}_{ij}^{(n)} - \mathbf{C}_{ij}\|}{\alpha_n} \\
&= \frac{p^2 O_p(n^{-1/2})}{\alpha_n} \\
&= O_p(\alpha_n^{-1}n^{-1/2}).
\end{aligned}$$

Under condition (2.30), $\alpha_n^{-1}n^{-1/2} \to 0$, so we conclude that

$$\hat{R}_{\alpha_n}(\boldsymbol{\Phi}) = (R_{\alpha_n}(\boldsymbol{\Phi}) + O_p(\alpha_n^{-1}n^{-1/2}))(1 + O_p(\alpha_n^{-1}n^{-1/2})) = R_{\alpha_n}(\boldsymbol{\Phi}) + o_p(1),$$

where $o_p(1)$ is a quantity to converges to 0 in probability uniformly over all $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}$. The proof is complete. □

We now turn to the proof of Lemma 8.

## Proof of Lemma 8

*Proof.* To simplify notation, we write $\|\phi\|^2$ for $\mathrm{Var}\,(\phi)$. The probability measure for which variance is taken should be clear from context.

Given $\epsilon \in (0,1)$ and $\mathbf{\Psi} \in \boldsymbol{\mathcal{H}}$, let $\alpha(\epsilon) = (\epsilon \sum \|\psi_j\|^2)/(2\sum \|\psi_j\|_{j,1}^2)$. Then

$$R_{\alpha(\epsilon)}(\mathbf{\Psi}) = \frac{\|\sum \psi_j\|^2 + \alpha(\epsilon)\sum \|\psi_j\|_{j,1}^2}{\sum \|\psi_j\|^2 + \alpha(\epsilon)\sum \|\psi_j\|_{j,1}^2} \leq \frac{\|\sum \psi_j\|^2 + \alpha(\epsilon)\sum \|\psi_j\|_{j,1}^2}{\sum \|\psi_j\|^2} \leq \frac{\|\sum \psi_j\|^2}{\sum \|\psi_j\|^2} + \frac{\epsilon}{2}.$$

Hence, to establish (2.34), it suffices to show that there exists $\mathbf{\Psi} \in \boldsymbol{\mathcal{H}}$ such that

$$\frac{\|\sum \psi_j\|^2}{\sum \|\psi_j\|^2} < \lambda_1 + \frac{\epsilon}{2}. \tag{A.2}$$

To this end, let $\delta = \epsilon/(6\lambda_1/\sqrt{p} + 4\sqrt{\lambda_1} + 3\epsilon/\sqrt{p} + 2) \in (0,1)$. Under the assumption that $\mathcal{H}_j$ is dense in $L^2(\mathcal{X}_j, dP_j)$, there exists $\psi_j \in \mathcal{H}_j$ such that $\|\psi_j - \phi_j^*\| < \delta/p$, for $j = 1, \ldots, p$. It follows that

$$\begin{aligned}\left|\left\|\sum \psi_j\right\| - \left\|\sum \phi_j^*\right\|\right| &< \left\|\sum \psi_j - \sum \phi_j^*\right\| \\ &= \left\|\sum (\psi_j - \phi_j^*)\right\| \leq \sum \|\psi_j - \phi_j^*\| \leq p \cdot \frac{\delta}{p} = \delta. \end{aligned} \tag{A.3}$$

To establish (A.2) for such a choice of $\mathbf{\Psi}$, we want to find an upper bound for $\|\sum \psi_j\|^2$ and a lower bound for $\sum \|\psi_j\|^2$. By definition, the population APC $\mathbf{\Phi}^*$ satisfies $\sum \|\phi_j^*\|^2 = 1$. Hence,

$$\begin{aligned}\left|\sum \|\psi_j\|^2 - 1\right| &= \left|\sum \|\psi_j\|^2 - \sum \|\phi_j^*\|^2\right| \\ &\leq \sum \left|\|\psi_j\|^2 - \|\phi_j^*\|^2\right| \\ &= \sum \left|\|\psi_j\| - \|\phi_j^*\|\right| \cdot \left|\|\psi_j\| + \|\phi_j^*\|\right| \\ &= \sum \frac{\delta}{p}\left(2\|\phi_j^*\| + \frac{\delta}{p}\right)\end{aligned}$$

$$= \frac{2\delta}{p} \sum \|\phi_j^*\| + \frac{\delta^2}{p}$$
$$\leq \frac{2\delta}{\sqrt{p}} + \frac{\delta^2}{p} \leq \frac{3\delta}{\sqrt{p}}, \tag{A.4}$$

where the second to the last inequality follows from the fact that $\sum \|\phi_j^*\| \leq (p \sum \|\phi_j^*\|^2)^{1/2}$ $= \sqrt{p}$, and the last inequality follows from $0 < \delta < 1$.

On the other hand, given that the population APC $\mathbf{\Phi}^*$ satisfies $\|\sum \phi_j^*\|^2 = \lambda_1$, we have

$$\begin{aligned}\left| \left\| \sum \psi_j \right\|^2 - \lambda_1 \right| &= \left| \left\| \sum \psi_j \right\|^2 - \left\| \sum \phi_j^* \right\|^2 \right| \\ &= \left| \left\| \sum \psi_j \right\| - \left\| \sum \phi_j^* \right\| \right| \cdot \left| \left\| \sum \psi_j \right\| + \left\| \sum \phi_j^* \right\| \right| \\ &\leq \delta \left( 2 \left\| \sum \phi_j^* \right\| + \delta \right) = \delta(2\sqrt{\lambda_1} + \delta) \leq 2\sqrt{\lambda_1}\delta + \delta, \end{aligned} \tag{A.5}$$

where the first inequality is due to (A.3).

Note that $0 < 3\delta/\sqrt{p} < 1$. Combining (A.4) and (A.5), we obtain

$$\frac{\|\sum \psi_j\|^2}{\sum \|\psi_j\|^2} \leq \frac{\lambda_1 + 2\sqrt{\lambda_1}\delta + \delta}{1 - \frac{3\delta}{\sqrt{p}}} = \lambda_1 + \frac{\frac{3\lambda_1}{\sqrt{p}}\delta + 2\sqrt{\lambda_1}\delta + \delta}{1 - \frac{3\delta}{\sqrt{p}}} \leq \lambda_1 + \frac{\epsilon}{2},$$

where the last inequality follows from the definition of $\delta$. This completes the proof. □

## Proof of Lemma 9

*Proof.* Based on the discussion in Section 2.2.3, the operator $\mathbf{P} : \mathbf{H}^* \to \mathbf{H}^*$ can be expressed as follows:

$$\mathbf{P} = \sum_{\nu=1}^{\infty} \lambda_\nu \langle \cdot, \mathbf{\Phi}_\nu \rangle_P \mathbf{\Phi}_\nu,$$

where $\{\lambda_\nu\}$ is the set of eigenvalues with $+1$ as the only possible accumulation point, and $\{\mathbf{\Phi}_\nu\}$ is the corresponding eigenfunctions so that $\{\mathbf{\Phi}_\nu\}$ forms a complete or-

thonormal basis system of $\mathbf{H}^*$.

Let $\lambda_1$ and $\lambda_2$ denote the smallest and the second smallest eigenvalue of $\mathbf{P}$, respectively. Under Assumptions 1(a)−(c), $\lambda_1 < 1$ is not an accumulation point and it has multiplicity one, so $\lambda_1 < \lambda_2$. It follows that the smallest population APC $\mathbf{\Phi}^* = \mathbf{\Phi}_1$. By definition, $\|\mathbf{\Phi}^*\|_P^2 = \sum \mathrm{Var}\,(\phi_j^*) = 1$.

Let $\mathbf{\Phi}_N^{(n)} = \mathbf{\Phi}^{(n)}/\|\mathbf{\Phi}^{(n)}\|_P$, and let $\delta_n = \langle \mathbf{\Phi}_N^{(n)}, \mathbf{\Phi}^* \rangle_P$. Then

$$\begin{aligned}
\langle \mathbf{\Phi}_N^{(n)}, \mathbf{P}\mathbf{\Phi}_N^{(n)} \rangle_P &= \sum_{\nu=1}^{\infty} \lambda_\nu \langle \mathbf{\Phi}_N^{(n)}, \mathbf{\Phi}_\nu \rangle_P^2 \\
&\geq \lambda_1 \langle \mathbf{\Phi}_N^{(n)}, \mathbf{\Phi}^* \rangle_P^2 + \lambda_2 \sum_{\nu=2}^{\infty} \langle \mathbf{\Phi}_N^{(n)}, \mathbf{\Phi}_\nu \rangle_P^2 \\
&= \lambda_1 \delta_n^2 + \lambda_2 (1 - \delta_n^2) \\
&\geq \lambda_1. \qquad \text{(A.6)}
\end{aligned}$$

Hence,

$$R_0(\mathbf{\Phi}^{(n)}) = \frac{\mathrm{Var}\,(\sum \phi_j^{(n)})}{\sum \mathrm{Var}\,(\phi_j^{(n)})} = \frac{\langle \mathbf{\Phi}^{(n)}, \mathbf{P}\mathbf{\Phi}^{(n)} \rangle_P}{\|\mathbf{\Phi}^{(n)}\|_P^2} = \langle \mathbf{\Phi}_N^{(n)}, \mathbf{P}\mathbf{\Phi}_N^{(n)} \rangle_P \geq \lambda_1. \qquad \text{(A.7)}$$

By assumption, $R_0(\mathbf{\Phi}^{(n)}) \to \lambda_1$, so all the inequalities in (A.6) becomes equalities, and we conclude that

$$\delta_n^2 = \langle \mathbf{\Phi}_N^{(n)}, \mathbf{\Phi}^* \rangle_P^2 = \frac{\langle \mathbf{\Phi}^{(n)}, \mathbf{\Phi}^* \rangle_P^2}{\|\mathbf{\Phi}^{(n)}\|_P^2 \|\mathbf{\Phi}^*\|_P^2} = \frac{\Big(\sum \mathrm{Cov}(\phi_j^{(n)}, \phi_j^*)\Big)^2}{\Big(\sum \mathrm{Var}\,(\phi_j^{(n)})\Big)\Big(\sum \mathrm{Var}\,(\phi_j^*)\Big)} \to 1. \qquad \text{(A.8)}$$

□

## A.3 Proofs for Section 2.7

This section contains proofs for theorems in Section 2.7.

## Proof of Theorem 2

*Proof.* To see that $\mathbf{S}_{ij}^{(\alpha)}$ is well-defined, we need to show the existence and uniqueness of solution to the regularized population regression problem.

First, note that for a given $\phi_j \in \mathcal{H}_j$, the operator $\mathrm{Cov}(\phi_j(X_j), \cdot(X_i)) : \mathcal{H}_i \to \mathbb{R}$ is a bounded linear functional on $\mathcal{H}_i$. By the Riesz Representation Theorem, there exists a unique $h \in \mathcal{H}_i$ such that $\mathrm{Cov}(\phi_j(X_j), f(X_i)) = \langle h, f \rangle_{\alpha,i}$ for all $f \in \mathcal{H}_i$. It then follows that

$$\begin{aligned}
&\operatorname*{argmin}_{f \in \mathcal{H}_i} \left\{ \mathrm{Var}\left(\phi_j(X_j) - f(X_i)\right) + \alpha \|f\|_{i,1}^2 \right\} \\
&= \operatorname*{argmin}_{f \in \mathcal{H}_i} \left\{ -2\mathrm{Cov}(\phi_j(X_j), f(X_i)) + \mathrm{Var}\left(f(X_i)\right) + \alpha \|f\|_{i,1}^2 \right\} \\
&= \operatorname*{argmin}_{f \in \mathcal{H}_i} \left\{ -2\langle h, f \rangle_{\alpha,i} + \|f\|_{\alpha,i}^2 \right\} \\
&= h.
\end{aligned}$$

That is, we have $\mathbf{S}_{ij}^{(\alpha)} \phi_j = h$, where $h$ is unique and satisfies $\mathrm{Cov}(\phi_j(X_j), f(X_i)) = \langle h, f \rangle_{\alpha,i}$ for all $f \in \mathcal{H}_i$. Equivalently,

$$\mathrm{Cov}(\phi_i(X_i), \phi_j(X_j)) = \langle \phi_i, \mathbf{S}_{ij}^{(\alpha)} \phi_j \rangle_{\alpha_i}, \qquad \forall \phi_i \in \mathcal{H}_i, \phi_j \in \mathcal{H}_j.$$

Thus, we see that $\mathbf{S}_{ij}^{(\alpha)}$ is the cross-covariance operator from $\mathcal{H}_j$ to $\mathcal{H}_i$. It follows that $\mathbf{S}_{ij}^{(\alpha)}$ is Hilbert-Schmidt (Fukumizu et al., 2007), hence compact.

To show (2.48), recall that Riesz Representation Theorem says that if $\ell$ is a bounded linear functional on $\mathcal{H}$ with representer $h_\ell \in \mathcal{H}$, i.e. $\ell(f) = \langle h_\ell, f \rangle_\alpha$ for all $f \in \mathcal{H}$, then $\|\ell\| = \|h_\ell\|_\alpha$. In the case that $\ell(f) = \mathrm{Cov}(\phi_j(X_j), f(X_i)) = \langle h, f \rangle_{\alpha,i}$

for all $f \in \mathcal{H}_i$, we have

$$\begin{aligned}\|\mathbf{S}_{ij}^{(\alpha)}\phi_j\|_{\alpha,i} = \|h\|_{\alpha,i} = \|\mathrm{Cov}(\phi_j(X_j), \cdot(X_i))\| &= \sup_{\|f\|_{\alpha,i}\leq 1} |\mathrm{Cov}(\phi_j(X_j), f(X_i))| \\ &\leq \sup_{\|f\|_{\alpha,i}\leq 1} \left(\mathrm{Var}\left(\phi_j(X_j)\right)\mathrm{Var}\left(f(X_i)\right)\right)^{1/2} \\ &\leq \sup_{\|f\|_{\alpha,i}\leq 1} \left(\mathrm{Var}\left(\phi_j(X_j)\right)\right)^{1/2}\|f\|_{\alpha,i} = \left(\mathrm{Var}\left(\phi_j(X_j)\right)\right)^{1/2} \leq \|\phi_j\|_{\alpha,j}.\end{aligned}$$

□

## Proof of Theorem 3

*Proof.* First, note that we can rewrite the optimization criterion in the population kernel APC problem as

$$\begin{aligned}&\mathrm{Var}\left(\sum_i \phi_i(X_i)\right) + \alpha\sum_i \|\phi_i\|_{i,1}^2 \\ &= \sum_i \mathrm{Var}\left(\phi_i(X_i)\right) + \alpha\sum_i \|\phi_i\|_{i,1}^2 + \sum_i\sum_{j\neq i}\mathrm{Cov}(\phi_i(X_i), \phi_j(X_j)) \\ &= \sum_i \|\phi_i\|_{\alpha,i}^2 + \sum_i\sum_{j\neq i}\langle\phi_i, \mathbf{S}_{ij}^{(\alpha)}\phi_j\rangle_{\alpha,i} \\ &= \sum_i \left\langle \phi_i, \sum_{j\neq i}\mathbf{S}_{ij}^{(\alpha)}\phi_j + \phi_i\right\rangle_{\alpha,i} \\ &= \langle\mathbf{\Phi}, \tilde{\mathbf{S}}^{(\alpha)}\mathbf{\Phi}\rangle_\alpha \qquad \geq 0.\end{aligned}$$

Hence, $\tilde{\mathbf{S}}^{(\alpha)}$ is positive. That the constraint $\sum \mathrm{Var}\,\phi_i(X_i) + \alpha\sum\|\phi_i\|_{i,1}^2 = \langle\mathbf{\Phi}, \mathbf{\Phi}\rangle_\alpha$ follows by definition.

To see that $\tilde{\mathbf{S}}^{(\alpha)}$ is self-adjoint, we need to show that $\langle\mathbf{\Phi}, \tilde{\mathbf{S}}^{(\alpha)}\mathbf{\Psi}\rangle_\alpha = \langle\tilde{\mathbf{S}}^{(\alpha)}\mathbf{\Phi}, \mathbf{\Psi}\rangle_\alpha$. Since

$$\mathrm{Cov}(\phi_i(X_i), \psi_j(X_j)) = \langle\phi_i, \mathbf{S}_{ij}^{(\alpha)}\psi_j\rangle_{\alpha,i} = \langle\mathbf{S}_{ji}^{(\alpha)}\phi_i, \psi_j\rangle_{\alpha,j}, \qquad \forall\phi_i \in \mathcal{H}_i, \psi_j \in \mathcal{H}_j,$$

we see that $(\mathbf{S}_{ij}^{(\alpha)})^* = \mathbf{S}_{ji}^{(\alpha)}$. It follows that

$$\begin{aligned}
\langle \mathbf{\Phi}, \tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Psi} \rangle_\alpha &= \sum_i \langle \phi_i, (\tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Psi})_i \rangle_{\alpha,i} = \sum_i \left\langle \phi_i, \sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \psi_j + \psi_i \right\rangle_{\alpha,i} \\
&= \sum_i \sum_{j \neq i} \langle \phi_i, \mathbf{S}_{ij}^{(\alpha)} \psi_j \rangle_{\alpha,i} + \sum_i \langle \phi_i, \psi_i \rangle_{\alpha,i} \\
&= \sum_j \sum_{i \neq j} \langle \mathbf{S}_{ji}^{(\alpha)} \phi_i, \psi_j \rangle_{\alpha,j} + \sum_j \langle \phi_j, \psi_j \rangle_{\alpha,j} \\
&= \sum_j \left\langle \sum_{i \neq j} \mathbf{S}_{ji}^{(\alpha)} \phi_i + \phi_j, \psi_j \right\rangle_{\alpha,j} = \sum_j \langle (\tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Phi})_j, \psi_j \rangle_{\alpha,j} \\
&= \langle \tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Phi}, \mathbf{\Psi} \rangle_\alpha,
\end{aligned}$$

so $\tilde{\mathbf{S}}^{(\alpha)}$ is self-adjoint.

To check that $\tilde{\mathbf{S}}^{(\alpha)}$ is bounded above by $p$, by (2.48), we have $\|\mathbf{S}_{ij}^{(\alpha)} \phi_j\|_{\alpha,i} \leq \|\phi_j\|_{\alpha,j}$. Therefore,

$$\begin{aligned}
\|\tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Phi}\|_\alpha^2 &= \sum_{i=1}^p \left\| \sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j + \phi_i \right\|_{\alpha,i}^2 \\
&\leq \sum_{i=1}^p \left( \sum_{j \neq i} \|\mathbf{S}_{ij}^{(\alpha)} \phi_j\|_{\alpha,i} + \|\phi_i\|_{\alpha,i} \right)^2 && \text{(use } \|\mathbf{S}_{ij}^{(\alpha)} \phi_j\|_{\alpha,i} \leq \|\phi_j\|_{\alpha,j}\text{)} \\
&\leq \sum_{i=1}^p \left( \sum_{j=1}^p \|\phi_j\|_{\alpha,j} \right)^2 && \text{(use } (\textstyle\sum_{j=1}^p a_j)^2 \leq p \sum_{j=1}^p a_j^2\text{)} \\
&\leq p \cdot p \sum_{j=1}^p \|\phi_j\|_{\alpha,j}^2 \\
&= p^2 \|\mathbf{\Phi}\|_\alpha^2,
\end{aligned}$$

so $\|\tilde{\mathbf{S}}^{(\alpha)}\| = \sup_{\|\mathbf{\Phi}\|_\alpha = 1} \|\tilde{\mathbf{S}}^{(\alpha)} \mathbf{\Phi}\|_\alpha \leq p$. □

## Proof of Proposition 1

*Proof.* Let $\mathbf{M} = \gamma \mathbf{Id}_{\boldsymbol{\mathcal{H}}} - \tilde{\mathbf{S}}^{(\alpha)}$, where $\gamma = (p+1)/2$. Then $\tilde{\boldsymbol{\Phi}}$ is the unit eigenfunction corresponding to the largest eigenvalue $\lambda$ of $\mathbf{M}$, and it is assumed that $\lambda$ has multiplicity one. By assumption, the power algorithm is initialized with $\boldsymbol{\Phi}^{[0]}$ that satisfies

$$\boldsymbol{\Phi}^{[0]} = a_0 \tilde{\boldsymbol{\Phi}} + \boldsymbol{\Psi}^{[0]}, \qquad \text{where } \boldsymbol{\Psi}^{[0]} \perp \tilde{\boldsymbol{\Phi}} \text{ and } a_0 > 0.$$

Let

$$\boldsymbol{\Phi}^{[t+1]} = \frac{\mathbf{M}\boldsymbol{\Phi}^{[t]}}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha},$$

and suppose that

$$\boldsymbol{\Phi}^{[t]} = a_t \tilde{\boldsymbol{\Phi}} + \boldsymbol{\Psi}^{[t]}, \qquad \text{where } \boldsymbol{\Psi}^{[t]} \perp \tilde{\boldsymbol{\Phi}}.$$

Then

$$\boldsymbol{\Phi}^{[t+1]} = \frac{\mathbf{M}\boldsymbol{\Phi}^{[t]}}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha} = \frac{\mathbf{M}(a_t\tilde{\boldsymbol{\Phi}} + \boldsymbol{\Psi}^{[t]})}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha} = \frac{a_t\lambda}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha}\tilde{\boldsymbol{\Phi}} + \frac{\mathbf{M}\boldsymbol{\Psi}^{[t]}}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha}.$$

Matching the coefficients, we see that

$$a_{t+1} = \frac{a_t\lambda}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha}, \qquad \boldsymbol{\Psi}^{[t+1]} = \frac{\mathbf{M}\boldsymbol{\Psi}^{[t]}}{\|\mathbf{M}\boldsymbol{\Phi}^{[t]}\|_\alpha}, \tag{A.9}$$

and it follows that $a_0 > 0$ implies $a_t > 0$ for all $t \in \mathbb{N}$. Now note that for $\boldsymbol{\Psi} \perp \tilde{\boldsymbol{\Phi}}$,

$$\|\mathbf{M}\boldsymbol{\Psi}\|_\alpha \le r\|\boldsymbol{\Psi}\|_\alpha, \qquad \text{where } r < \lambda, \tag{A.10}$$

so by (A.9) and (A.10),

$$\frac{\|\boldsymbol{\Psi}^{[t+1]}\|_\alpha}{a_{t+1}} = \frac{\|\mathbf{M}\boldsymbol{\Psi}^{[t]}\|_\alpha}{a_t\lambda} \le \left(\frac{r}{\lambda}\right)\frac{\|\boldsymbol{\Psi}^{[t]}\|_\alpha}{a_t},$$

which in turn implies

$$\frac{\|\mathbf{\Psi}^{[t]}\|_\alpha}{a_t} \leq \left(\frac{r}{\lambda}\right)^t \frac{\|\mathbf{\Psi}^{[0]}\|_\alpha}{a_0} \to 0 \text{ as } t \to \infty. \tag{A.11}$$

To show that $\mathbf{\Phi}^{[t]} \to \tilde{\mathbf{\Phi}}$, note that $\|\mathbf{\Phi}^{[t]}\|_\alpha = 1$ implies that

$$\|a_t\tilde{\mathbf{\Phi}} + \mathbf{\Psi}^{[t]}\|_\alpha = 1 \Leftrightarrow a_t^2 + \|\mathbf{\Psi}^{[t]}\|_\alpha^2 = 1 \Leftrightarrow 1 + \frac{\|\mathbf{\Psi}^{[t]}\|_\alpha^2}{a_t^2} = \frac{1}{a_t^2}.$$

From (A.11), we conclude that $a_t^2 \to 1$ and $\|\mathbf{\Psi}^{[t]}\|_\alpha^2 \to 0$, hence

$$\|\mathbf{\Phi}^{[t]} - \tilde{\mathbf{\Phi}}\|_\alpha^2 = (1 - a_t)^2 + \|\mathbf{\Psi}^{[t]}\|_\alpha^2 \to 0.$$

□

## A.4 Implementation Details of the Power Algorithm

We justified the use of a smoothing-based power algorithm in computing population kernel APCs in Section 2.7. In this section, we give a detailed description of its empirical implementation.

### A.4.1 The Representer Theorem for Kernel APCs

We first need to resolve the issue that the spaces $\mathcal{H}_j$ in $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$ in the sample kernel APC problem

$$\min_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} \widehat{\text{Var}}\,(\sum_{j=1}^p \phi_j) \; + \; \sum_{j=1}^p \alpha_j \|\phi_j\|_{1,j}^2 \quad \text{subject to} \quad \sum_{j=1}^p \widehat{\text{Var}}\,(\phi_j) \; + \; \sum_{j=1}^p \alpha_j \|\phi_j\|_{1,j}^2 = 1 \tag{A.12}$$

is (almost always) infinite-dimensional, which can pose challenges computationally. As will be shown, the beauty of the RKHS framework for APCs estimation is that for suitable RKHSs $\boldsymbol{\mathcal{H}}$, the solution to (A.12) always lie in a finite-dimensional subspace of $\boldsymbol{\mathcal{H}}$ and thus can be computed in closed form.

To begin, consider the smoothing splines problem

$$\min_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \alpha\|f\|_1^2\right\}, \tag{A.13}$$

where $\mathcal{H}$ is an RKHS with semi-norm $\|\cdot\|_1$. To be more concrete, we suppose that $\mathcal{H}$ is associated with reproducing kernel $k$ and inner product $\langle\cdot,\cdot\rangle_k$. Moreover,

$$\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1, \tag{A.14}$$

where $\mathcal{H}^0$ is a finite-dimensional linear subspace of $\mathcal{H}$ with basis $\{q_1,\ldots,q_m\}$, $m = \dim(\mathcal{H}^0) < n$, and $\mathcal{H}^1$ is the orthogonal complement of $\mathcal{H}^0$. With the decomposition (A.14), the reproducing kernel $k$ can also be uniquely decomposed into $k = k^0 + k^1$, where $k^0(x,\cdot) = \mathbf{P}^0 k(x,\cdot)$, $k^1(x,\cdot) = \mathbf{P}^1 k(x,\cdot)$, and $\mathbf{P}^0$ and $\mathbf{P}^1$ denote the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}^0$ and $\mathcal{H}^1$, respectively. One can check that $\mathcal{H}^0$ and $\mathcal{H}^1$ are RKHSs with reproducing kernels $k^0$ and $k^1$, respectively (Aronszajn, 1950). Denote the respective RKHS inner products on $\mathcal{H}^0$ and $\mathcal{H}^1$ by $\langle\cdot,\cdot\rangle_0$ and $\langle\cdot,\cdot\rangle_1$. Then the inner product $\langle f,g\rangle_k$ on $\mathcal{H}$ bears the decomposition

$$\langle f,g\rangle_k = \langle f^0,g^0\rangle_0 + \langle f^1,g^1\rangle_1, \qquad f,g\in\mathcal{H}, \tag{A.15}$$

with $f = f^0 + f^1$, $g = g^0 + g^1$, and $f^0, g^0 \in \mathcal{H}^0$, $f^1, g^1 \in \mathcal{H}^1$, the decomposition is again unique. In this case, we define the penalty term

$$\|f\|_1^2 := \|\mathbf{P}^1 f\|_k^2 = \|f^1\|_k^2 = \|f^1\|_1^2, \qquad f\in\mathcal{H},$$

and it goes without saying that $\mathcal{H}^0$ is the null space of the semi-norm $\|\cdot\|_1$, so the functions in $\mathcal{H}^0$ are not penalized in (A.13). RKHSs constructed as above are covered by the general spaces discussed in Section 2.5.1, and the Sobolev space is an example of such RKHSs.

It is known (Wahba, 1990) that the solution $\hat{f}$ of (A.13) must lie in a finite-dimensional subspace of $\mathcal{H}$. Specifically, write $\hat{f} = \hat{f}^0 + \hat{f}^1$ with $\hat{f}^0 \in \mathcal{H}^0, \hat{f}^1 \in \mathcal{H}^1$, then

$$\hat{f}^1 \in \text{span}\{k^1(x_i, \cdot) : 1 \leq i \leq n\}.$$

In essence, this means that to solve (A.13), only the representers of the evaluation functionals (projected to $\mathcal{H}^1$) at the locations of the observed data matters. This is known as the *Representer Theorem for smoothing splines*. A more general version of this Representer Theorem, adapted to the case of kernel APCs, states that for any probability measure $P_j(dx_j)$, not necessarily an empirical measure, only the representers of the evaluation functionals at the locations that belong to the support of $P_j$ matters.

**Theorem 15** (Representer Theorem for Kernel APCs)**.** *Let $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$, where $\mathcal{H}_j = \mathcal{H}_j^0 \oplus \mathcal{H}_j^1$ is an RKHS with reproducing kernel $k_j = k_j^0 + k_j^1$. Then, the solution to the kernel APC problem* (A.12)*, if exists, is taken on the subspace $\boldsymbol{\mathcal{H}}_P := \mathcal{H}_{P_1} \times \mathcal{H}_{P_2} \times \cdots \times \mathcal{H}_{P_p}$, where*

$$\mathcal{H}_{P_j} := \mathcal{H}_j^0 \oplus \overline{\text{span}\{k_j^1(x, \cdot) - \mu_j^1 : x \in \text{supp}(P_j)\}},$$

*and $\mu_j^1$ is the mean element of $\mathcal{H}_j^1$ with respect to the marginal probability measure $P_j$.*

*Proof.* Let $\mu_j$ be the mean element of $\mathcal{H}_j$ with respect to $P_j$:

$$\langle \phi_j, \mu_j \rangle_{k_j} = E(\langle \phi_j, k_{X_j} \rangle_{k_j}) = E(\phi_j(X_j)) \qquad \forall \phi_j \in \mathcal{H}_j,$$

and suppose that $\mu_j = \mu_j^0 + \mu_j^1$, and $\mu_j^0 \in \mathcal{H}_j^0$, $\mu_j^1 \in \mathcal{H}_j^1$. For $\psi_j \in \mathcal{H}_j$, we have

$$\psi_j \perp \mathcal{H}_{P_j} \Rightarrow \psi_j(x) - E(\psi_j) = 0, \qquad \text{for } x \in \text{supp}(P_j).$$

This is because $\psi_j(x) - E(\psi_j) = \langle \psi_j, k_j(x,\cdot) - \mu_j \rangle_{k_j} = \langle \psi_j, k_j^0(x,\cdot) - \mu_j^0 \rangle_{k_j} + \langle \psi_j, k_j^1(x,\cdot) - \mu_j^1 \rangle_{k_j}$ and $k_j^0(x,\cdot) - \mu_j^0 \in \mathcal{H}_j^0$. So given $\psi_j \perp \mathcal{H}_{P_j}$ and $\phi_j \in \mathcal{H}_{P_j}$, we have $\text{Var}\,(\psi_j) = 0$ and $\|\phi_j + \psi_j\|_{j,1}^2 = \|\phi_j\|_{j,1}^2 + \|\psi_j\|_{j,1}^2$, which implies that

$$\text{Var} \sum_{j=1}^{p} (\phi_j + \psi_j) = \text{Var}\,(\sum_{j=1}^{p} \phi_j), \qquad \sum_{j=1}^{p} \alpha_j \|\phi_j + \psi_j\|_{j,1}^2 \geq \sum_{j=1}^{p} \alpha_j \|\phi_j\|_{j,1}^2,$$

and the inequality is strict when $\psi_i \not\equiv 0$ for some $1 \leq i \leq p$.

Now, suppose on the contrary that $(\phi_1^* + \psi_1^*, \ldots, \phi_p^* + \psi_p^*)$ is the optimal solution of the kernel APC problem, where $\phi_j^* \in \mathcal{H}_{P_j}, \psi_j^* \perp \mathcal{H}_{P_j}$ and $\psi_i^* \not\equiv 0$ for some $1 \leq i \leq p$. Let

$$\delta = \sum_{j=1}^{p} \alpha_j \|\phi_j^* + \psi_j^*\|_{j,1}^2,$$

then $(\phi_1^* + \psi_1^*, \ldots, \phi_p^* + \psi_p^*)$ is also an optimal solution of the following optimization problem:

$$\min_{\mathbf{\Phi} \in \boldsymbol{\mathcal{H}}} \text{Var}\,(\sum_{j=1}^{p} \phi_j) + \sum_{j=1}^{p} \alpha_j \|\phi_j\|_{j,1}^2 \qquad \text{subject to} \qquad \sum_{j=1}^{p} \text{Var}\,(\phi_j) = 1 - \delta. \quad \text{(A.16)}$$

But as argued before we have $\text{Var} \sum(\phi_j^* + \psi_j^*) = \text{Var}\,(\sum \phi_j^*)$ and $\sum \alpha_j \|\phi_j^* + \psi_j^*\|_{j,1}^2 > \sum \alpha_j \|\phi_j^*\|_{j,1}^2$. Also, subject to the constraint that $\sum \text{Var}\,(\phi_j^* + \psi_j^*) = 1 - \delta$, we have $\sum \text{Var}\,(\phi_j^*) = 1 - \delta$. This gives the desired contradiction since in this case $(\phi_1^*, \ldots, \phi_p^*)$ is a better solution of (A.16) comparing to the optimal solution $(\phi_1^* + \psi_1^*, \ldots, \phi_p^* + \psi_p^*)$. Therefore, we must have $\psi_j^* \equiv 0$ for $1 \leq j \leq p$. This completes the proof. □

Note that in the case where $P_j$ denotes the empirical probability measure with

only finitely many values $\{x_{1j}, \ldots, x_{nj}\}$ in its support, Theorem 15 specializes to the finite-sample version of the Representer Theorem for kernel APCs:

**Corollary 5.** *Let $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$, where $\mathcal{H}_j = \mathcal{H}_j^0 \oplus \mathcal{H}_j^1$ is an RKHS with reproducing kernel $k_j = k_j^0 + k_j^1$. Given data $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $1 \leq i \leq n$, the solution to the sample kernel APC problem* (2.20)*, if exists, is taken on the finite-dimensional subspace $\boldsymbol{\mathcal{H}}_n := \mathcal{H}_{n,1} \times \cdots \times \mathcal{H}_{n,p}$, where*

$$\mathcal{H}_{n,j} := \mathcal{H}_j^0 \oplus \operatorname{span}\left\{ k_j^1(x_{ij}, \cdot) - \frac{1}{n}\sum_{a=1}^{n} k_j^1(x_{aj}, \cdot) : 1 \leq i \leq n \right\}.$$

One can similarly show that other higher-order sample kernel APCs, if exists, also lie in the finite-dimensional subspace $\boldsymbol{\mathcal{H}}_n$.

### A.4.2 Smoothing in RKHSs with Null Spaces

To implement the power algorithm presented in Algorithm 1, it follows from Corollary 5 that it suffices to work with the coefficients of the basis of $\mathcal{H}_{n,i}$. Specifically, let

$$\phi_i = \sum_{\ell=1}^{n} \beta_{\ell i} f_{\ell i} + \sum_{\ell=1}^{m_i} \beta_{n+\ell,i} q_{\ell i},$$

where $f_{\ell i} = k_i^1(x_{\ell i}, \cdot) - \frac{1}{n}\sum_{a=1}^{n} k_i^1(x_{ai}, \cdot)$ for $1 \leq \ell \leq n$ and $\{q_{\ell i}\}_{\ell=1}^{m_i}$ forms a basis for $\mathcal{H}_i^0$. Then the update steps $\phi_i \leftarrow \gamma\phi_i^{[t]} - (\sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j^{[t]} + \phi_i^{[t]})$ in Algorithm 1 becomes

$$\begin{aligned} \beta_{\ell i} &\leftarrow (\gamma - 1)\beta_{\ell i}^{[t]} - c_{\ell i}, && 1 \leq \ell \leq n, \\ \beta_{n+\ell,i} &\leftarrow (\gamma - 1)\beta_{n+\ell,i}^{[t]} - d_{\ell i}, && 1 \leq \ell \leq m_i, \end{aligned} \tag{A.17}$$

where $\{c_{\ell i}\}_{\ell=1}^{n}$ and $\{d_{\ell i}\}_{\ell=1}^{m_i}$ are two sets of coefficients obtained from the smoothing step $\sum_{j \neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j^{[t]}$, to be derived shortly.

Let $\boldsymbol{\beta}_i = (\beta_{1i}, \ldots, \beta_{ni}) \in \mathbb{R}^n$, and let $\mathbf{G}_i$ be the $n \times n$ centered kernel matrix

associated with $k_i^1$, with $(j, \ell)$ entry

$$
\begin{aligned}
(\mathbf{G}_i)_{j\ell} &= \langle f_{ji}, f_{\ell i} \rangle_{i,1} \qquad (A.18)\\
&= k_i^1(x_{ji}, x_{\ell i}) - \frac{1}{n}\sum_{b=1}^{n} k_i^1(x_{ji}, x_{bi}) - \frac{1}{n}\sum_{a=1}^{n} k_i^1(x_{ai}, x_{\ell i}) + \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n} k_i^1(x_{ai}, x_{bi}).
\end{aligned}
$$

Then the normalizing constant $c$ in Algorithm 1 can be obtained upon computation of the variance of the transformed data points $\{\phi_i(x_{\ell i})\}_{\ell=1}^n$ and the penalty term $\|\phi_i\|_{i,1}^2 = \boldsymbol{\beta}_i^T \mathbf{G}_i \boldsymbol{\beta}_i$, for $1 \le i \le p$.

We now consider the smoothing step $\sum_{j\neq i} \mathbf{S}_{ij}^{(\alpha)} \phi_j$, which by linearity of smoothing is empirically the regularized least squares regression of $\sum_{j\neq i} \phi_j(X_j)$ on $X_i$. This amounts to solving the following optimization problem:

$$
\min_{f \in \mathcal{H}_i} \left\{ \widehat{\mathrm{Var}} \left( \sum_{j\neq i} \phi_j(X_j) - f(X_i) \right) + \alpha_i \|f\|_{i,1}^2 \right\}, \qquad (A.19)
$$

where $\widehat{\mathrm{Var}}\left(\sum_{j\neq i} \phi_j(X_j) - f(X_i)\right)$ evaluates to

$$
\frac{1}{n}\sum_{\ell=1}^{n} \left[ \sum_{j\neq i} \left( \phi_j(x_{\ell j}) - \frac{1}{n}\sum_{b=1}^{n} \phi_j(x_{bj}) \right) - \left( f(x_{\ell i}) - \frac{1}{n}\sum_{a=1}^{n} f(x_{ai}) \right) \right]^2.
$$

We see that (A.19) is essentially the smoothing splines problem (A.13) (modulo centering), hence it is not surprising that its solution lies in $\mathcal{H}_{n,i}$ as well.

Following Wahba (1990) (page 11-12), let the closed form solution of (A.19) be

$$
f = \sum_{\ell=1}^{n} c_{\ell i} f_{\ell i} + \sum_{\ell=1}^{m_i} d_{\ell i} q_{\ell i}.
$$

Then, (A.19) can be restated as

$$
\min_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^{m_i}} \left\{ \frac{1}{n} \|\mathbf{y} - (\mathbf{G}_i\mathbf{c} + \mathbf{Q}_i\mathbf{d})\|^2 + \alpha_i \mathbf{c}^T \mathbf{G}_i \mathbf{c} \right\}, \qquad (A.20)
$$

where $\mathbf{c}^T = (c_{1i}, \ldots, c_{ni})$, $\mathbf{d}^T = (d_{1i}, \ldots, d_{m_i i})$, $\mathbf{y}^T = (y_1, \ldots, y_n)$ with $y_\ell = \sum_{j \neq i}(\phi_j(x_{\ell j}) - \frac{1}{n}\sum_{b=1}^{n} \phi_j(x_{bj}))$ for $1 \leq \ell \leq n$, $\mathbf{G}_i$ is as given in (A.18), and $\mathbf{Q}_i$ is the column-centered version of

$$\tilde{\mathbf{Q}}_i = \begin{pmatrix} q_{1i}(x_{1i}) & \cdots & q_{m_i i}(x_{1i}) \\ \vdots & \vdots & \vdots \\ q_{1i}(x_{ni}) & \cdots & q_{m_i i}(x_{ni}) \end{pmatrix}.$$

It then follows that the solution of (A.20) is

$$\mathbf{d} = (\mathbf{Q}_i^T \mathbf{M}_i^{-1} \mathbf{Q}_i)^{-1} \mathbf{Q}_i^T \mathbf{M}_i^{-1} \mathbf{y}, \qquad \mathbf{c} = \mathbf{M}_i^{-1}(\mathbf{y} - \mathbf{Q}_i \mathbf{d}),$$

where $\mathbf{M}_i = \mathbf{G}_i + n\alpha_i \mathbf{I}$, $\mathbf{I}$ being the $n \times n$ identity matrix. Plugging $\mathbf{c}$ and $\mathbf{d}$ into (A.17) completes the update steps.

## A.5 A Direct Approach for Computing Kernel APCs

In this section, we give a direct approach for computing kernel APCs.

From Corollary 5, we know that the solution $\hat{\mathbf{\Phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)$ of the sample kernel APC problem (2.20) lies in the finite-dimensional function space $\boldsymbol{\mathcal{H}}_n = \mathcal{H}_{n,1} \times \cdots \times \mathcal{H}_{n,p}$. In the following, we derive the resulting linear algebra problem in terms of the coefficients with respect to the basis of $\mathcal{H}_{n,j}$'s. We will focus on the case where there are no null spaces, i.e. $\mathcal{H}_j = \mathcal{H}_j^1$ and $k_j = k_j^1$, for $1 \leq j \leq p$. The case with null spaces requires the use of the additional basis $\{q_{1j}, \ldots, q_{m_j j}\}$ for $\mathcal{H}_j^0$, $1 \leq j \leq p$, which is tractable but with slightly more tedious derivation. We recommend the use of power algorithm described in Section 2.7 when dealing with cases involving null spaces. The power algorithm is computationally more attractive than the direct linear algebra approach given below, when the interest is only in extracting a few eigenfunctions.

For each $1 \leq j \leq p$, we express $\phi_j \in \mathcal{H}_{n,j}$ as $\phi_j = \sum_{i=1}^n \beta_{ij} f_{ij}$, where

$$f_{ij}(\cdot) := k_j(x_{ij}, \cdot) - \frac{1}{n}\sum_{a=1}^n k_j(x_{aj}, \cdot), \qquad 1 \leq i \leq n.$$

Then

$$\begin{aligned}\sum_{j=1}^p \phi_j &= \sum_{j=1}^p \sum_{i=1}^n \beta_{ij} f_{ij} = \sum_{j=1}^p \boldsymbol{\beta}_j^T \mathbf{f}_j \\ &\qquad \text{where } \boldsymbol{\beta}_j^T = (\beta_{1j}, \ldots, \beta_{nj}),\ \mathbf{f}_j^T = (f_{1j}, \ldots, f_{nj}) \\ &= \boldsymbol{\beta}^T \mathbf{F} \qquad \text{where } \boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T),\ \mathbf{F}^T = (\mathbf{f}_1^T, \ldots, \mathbf{f}_p^T).\end{aligned}$$

The penalty term associated with $\phi_j$ evaluates to

$$\|\phi_j\|_{k_j}^2 = \left\langle \sum_{i=1}^n \beta_{ij} f_{ij}, \sum_{\ell=1}^n \beta_{\ell j} f_{\ell j} \right\rangle_{k_j} = \sum_{i=1}^n \sum_{\ell=1}^n \beta_{ij}\beta_{\ell j} \langle f_{ij}, f_{\ell j}\rangle_{k_j} = \boldsymbol{\beta}_j^T \mathbf{G}_j \boldsymbol{\beta}_j,$$

where $\mathbf{G}_j$ is the centered kernel matrix associated with $k_j$, with $(i, \ell)$ entry

$$\begin{aligned}(\mathbf{G}_j)_{i\ell} &= \langle f_{ij}, f_{\ell j}\rangle_{k_j} \\ &= k_j(x_{ij}, x_{\ell j}) - \frac{1}{n}\sum_{b=1}^n k_j(x_{ij}, x_{bj}) - \frac{1}{n}\sum_{a=1}^n k_j(x_{aj}, x_{\ell j}) + \frac{1}{n^2}\sum_{a=1}^n \sum_{b=1}^n k_j(x_{aj}, x_{bj}).\end{aligned}$$

Therefore, we can rewrite the penalty term as

$$\sum_{j=1}^p \alpha_j \|\phi_j\|_{k_j}^2 = \sum_{j=1}^p \alpha_j \boldsymbol{\beta}_j^T \mathbf{G}_j \boldsymbol{\beta}_j.$$

The variance term in the sample kernel APC criterion evaluates to

$$\widehat{\mathrm{Var}}\,(\sum_{j=1}^p \phi_j) = \widehat{\mathrm{Var}}\,(\boldsymbol{\beta}^T \mathbf{F}) = \frac{1}{n}\boldsymbol{\beta}^T \mathbf{G}\mathbf{G}^T \boldsymbol{\beta},$$

where $\mathbf{G}^T = (\mathbf{G}_1, \cdots, \mathbf{G}_p)$. Meanwhile, the variance term in the sample kernel APC constraint is

$$\sum_{j=1}^{p} \widehat{\mathrm{Var}}\,(\phi_j) = \sum_{j=1}^{p} \widehat{\mathrm{Var}}\,(\boldsymbol{\beta}_j^T \mathbf{f}_j) = \frac{1}{n} \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \mathbf{G}_j^2 \boldsymbol{\beta}_j.$$

Hence, the optimization problem (2.20), expressed in linear algebra notation, becomes

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{pn}}} \quad & \frac{1}{n} \boldsymbol{\beta}^T \mathbf{G}\mathbf{G}^T \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathrm{diag}(\alpha_1 \mathbf{G}_1, \ldots, \alpha_p \mathbf{G}_p) \boldsymbol{\beta} \\ \text{subject to} \quad & \frac{1}{n} \boldsymbol{\beta}^T \mathrm{diag}(\mathbf{G}_1^2, \ldots, \mathbf{G}_p^2) \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathrm{diag}(\alpha_1 \mathbf{G}_1, \ldots, \alpha_p \mathbf{G}_p) \boldsymbol{\beta} = 1. \end{aligned} \tag{A.21}$$

Equivalently, we want to solve the following generalized eigenvalue problem:

$$\begin{aligned} &\begin{pmatrix} \mathbf{G}_1^2 + n\alpha_1 \mathbf{G}_1 & \mathbf{G}_1 \mathbf{G}_2 & \cdots & \mathbf{G}_1 \mathbf{G}_p \\ \mathbf{G}_2 \mathbf{G}_1 & \mathbf{G}_2^2 + n\alpha_2 \mathbf{G}_2 & \cdots & \mathbf{G}_2 \mathbf{G}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_p \mathbf{G}_1 & \mathbf{G}_p \mathbf{G}_2 & \cdots & \mathbf{G}_p^2 + n\alpha_p \mathbf{G}_p \end{pmatrix} \boldsymbol{\beta} \\ &= \lambda \begin{pmatrix} \mathbf{G}_1^2 + n\alpha_1 \mathbf{G}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2^2 + n\alpha_2 \mathbf{G}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_p^2 + n\alpha_p \mathbf{G}_p \end{pmatrix} \boldsymbol{\beta}. \end{aligned} \tag{A.22}$$

Following Bach & Jordan (2003), we can approximate the diagonal blocks $\mathbf{G}_j^2 + n\alpha_j \mathbf{G}_j$ in (A.22) by $(\mathbf{G}_j + \frac{n\alpha_j}{2}\mathbf{I})^2$. Letting $\boldsymbol{\gamma}_j = (\mathbf{G}_j + \frac{n\alpha_j}{2}\mathbf{I})\boldsymbol{\beta}_j$ allows the reformulation of the generalized eigenproblem above as an eigenproblem, in which case we just need

to perform eigendecomposition on

$$\mathbf{R} = \begin{pmatrix} \mathbf{I} & \mathbf{R}_1^T\mathbf{R}_2 & \cdots & \mathbf{R}_1^T\mathbf{R}_p \\ \mathbf{R}_2^T\mathbf{R}_1 & \mathbf{I} & \cdots & \mathbf{R}_2^T\mathbf{R}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_p^T\mathbf{R}_1 & \mathbf{R}_p^T\mathbf{R}_2 & \cdots & \mathbf{I} \end{pmatrix},$$

where $\mathbf{R}_j = \mathbf{G}_j(\mathbf{G}_j + \frac{n\alpha_j}{2}\mathbf{I})^{-1}$ and $\mathbf{I}$ is the $n \times n$ identity matrix, to get its eigenvector $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ (corresponding to the smallest eigenvalue). The desired (approximate) solution of (A.21) can then be obtained as $\hat{\boldsymbol{\beta}}_j = (\mathbf{G}_j + \frac{n\alpha_j}{2}\mathbf{I})^{-1}\hat{\boldsymbol{\gamma}}_j$, while the (mean-centered) estimated transform evaluated at the data points is

$$\hat{\boldsymbol{\phi}}_j = \mathbf{G}_j\hat{\boldsymbol{\beta}}_j = \mathbf{G}_j\Big(\mathbf{G}_j + \frac{n\alpha_j}{2}\mathbf{I}\Big)^{-1}\hat{\boldsymbol{\gamma}}_j.$$

The second-smallest and subsequent higher order sample kernel APCs can be obtained similarly by extracting the eigenvector corresponding to the second-smallest and subsequent smallest eigenvalue of $\mathbf{R}$.

We remark that the linear algebra problem (A.21) is often numerically ill-conditioned due to low-rankness of $\mathbf{G}_j$, so one has to make adjustment in order to solve for APCs. This, however, introduces undesirable arbitrariness to the resulting optimization problem.

## A.6 A Comparison of Kernel APC with Kernel PCA

Kernel PCA (KPCA) provides a nonlinear generalization of standard PCA though kernelizing. By the use of the kernel trick, KPCA enables one to perform PCA in a

high-dimensional feature space (usually taken to be an RKHS) that is related to the original input space by some nonlinear mapping (Schölkopf et al., 1998; Schölkopf & Smola, 2002).

To compare kernel APCs with KPCs, we first set up some notations. Let $(\mathcal{X}_1, \mathcal{B}_{\mathcal{X}_1})$, ..., $(\mathcal{X}_p, \mathcal{B}_{\mathcal{X}_p})$ be measurable spaces, and consider a random vector $\mathbf{X} = (X_1, \ldots, X_p)$ taking values in $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ with distribution $P_{1:p}$. Let $\mathcal{H}$ be the RKHS associated with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ consists of real-valued functions with common domain $\mathcal{X}$ and has an inner product $\langle \cdot, \cdot \rangle_k$. Consider the mapping $\mathcal{X} \to \mathcal{H}$, $\mathbf{X} \mapsto k_{\mathbf{X}}(\cdot) := k(\mathbf{X}, \cdot)$. To perform PCA in $\mathcal{H}$, we solve

$$\max_{\phi \in \mathcal{H}} \mathrm{Var}\left(\langle \phi, k_{\mathbf{X}} \rangle_k\right) \qquad \text{subject to} \qquad \|\phi\|_k^2 = 1. \tag{A.23}$$

By the reproducing property, $\langle \phi, k_{\mathbf{X}} \rangle_k = \phi(\mathbf{X}) = \phi(X_1, \ldots, X_p)$, so (A.23) is equivalent to

$$\max_{\phi} \mathrm{Var}\left(\phi(X_1, \ldots, X_p)\right) \qquad \text{subject to} \qquad \|\phi\|_k^2 = 1.$$

To compare KPCA and kernel APC, we consider using an additive kernel in the KPCA problem. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is additive if it can be written as a sum of the kernel function of each dimension:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{p} k_j(x_j, x_j').$$

Then each $\phi \in \mathcal{H}$ has a decomposition $\phi(\mathbf{x}) = \sum \phi_j(x_j)$, where $\phi_j \in \mathcal{H}_j$ and $\mathcal{H}_j$ is the RKHS associated with $k_j$, endowed with an inner product $\langle \cdot, \cdot \rangle_{k_j}$. Hence, the KPCA problem with an additive kernel reduces to

$$\max_{\phi^1 \in \mathcal{H}_1, \ldots, \phi_p \in \mathcal{H}_p} \mathrm{Var}\left(\sum_{j=1}^{p} \phi_j\right) \qquad \text{subject to} \qquad \sum_{j=1}^{p} \|\phi_j\|_{k_j}^2 = 1. \tag{A.24}$$

Using the notation in Section 2.5.5, we can rewrite (A.24) in terms of quadratic forms in $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$:

$$\max_{\boldsymbol{\Phi} \in \boldsymbol{\mathcal{H}}} \langle \boldsymbol{\Phi}, \mathbf{C}\boldsymbol{\Phi} \rangle_k \qquad \text{subject to} \qquad \langle \boldsymbol{\Phi}, \boldsymbol{\Phi} \rangle_k = 1, \tag{A.25}$$

where $\boldsymbol{\Phi} = (\phi^1, \ldots, \phi_p)$, $\mathbf{C} = (\mathbf{C}_{ij})_{i,j}$, and $\mathbf{C}_{ij}$ is the cross-covariance operator of $(X_i, X_j)$. Contrast (A.25) with the population kernel APC problem in (2.24b), we see that KPCA and kernel APC are substantially different. Even if we try to match the objective function and compare (A.25) with (2.24a) instead, KPCA is still different from APC since (A.25) is an eigenproblem in $\boldsymbol{\mathcal{H}}$ whereas (2.24a) is a generalized eigenproblem in $\boldsymbol{\mathcal{H}}$.

Another distinctive difference between kernel APC and KPCA is that kernel APC focuses on minimization for concurvity detection, whereas KPCA focuses on maximization for dimension reduction and minimization does not even make sense. To see this, note that solving the following minimization version of the KPCA problem

$$\min_{\phi_1 \in \mathcal{H}_1, \ldots, \phi_p \in \mathcal{H}_p} \operatorname{Var}\left(\sum_{j=1}^{p} \phi_j\right) \qquad \text{subject to} \qquad \sum_{j=1}^{p} \|\phi_j\|_{k_j}^2 = 1 \tag{A.26}$$

is equivalent to solving

$$\max_{\phi_1 \in \mathcal{H}_1, \ldots, \phi_p \in \mathcal{H}_p} \sum_{j=1}^{p} \|\phi_j\|_{k_j}^2 \qquad \text{subject to} \qquad \operatorname{Var}\left(\sum_{j=1}^{p} \phi_j\right) = 1.$$

Since the penalties $\|\phi_j\|_{k_j}^2$ are usually considered as a measure of regularity (i.e., "smoothness") of a function $\phi_j$, it makes no sense that one is interested in obtaining transformations that have maximum "wiggliness".

While one might argue that (A.26) still yields a solution with small $\operatorname{Var}\left(\sum \phi_j\right)$ that could potentially be interesting, an issue concerns computation arises: the empirical

solution of the minimization version of KPCA (with or without additive structure) does not necessarily lie in a finite-dimensional subspace of $\mathcal{H}$ (i.e., there is no Representer Theorem such as Theorem 15 for kernel APC), renders the use of RKHS unappealing from a computational standpoint. To see this, consider the following unconstrained minimization version of KPCA:

$$\min_{\phi \in \mathcal{H}} \frac{\mathrm{Var}\left(\phi(\mathbf{X})\right)}{\|\phi\|_k^2}. \tag{A.27}$$

Let $\mu_{\mathbf{X}} \in \mathcal{H}$ be the mean element that satisfies $\langle \phi, \mu_{\mathbf{X}} \rangle_k = E(\phi(\mathbf{X}))$ for all $\phi \in \mathcal{H}$. For any function $\phi + \psi$, $\phi \in \mathcal{H}_{P_{\mathbf{X}}} := \mathrm{span}\{k(\mathbf{x}, \cdot) - \mu_{\mathbf{X}} : \mathbf{x} \in \mathrm{supp}(P_{\mathbf{X}})\}$ and $\psi \perp \mathcal{H}_{P_{\mathbf{X}}}, \psi \not\equiv 0$, we have $\mathrm{Var}\left(\psi(\mathbf{X})\right) = 0$ since $\psi(\mathbf{x}) - E(\psi(\mathbf{X})) = \langle \psi, k(\mathbf{x}, \cdot) - \mu_{\mathbf{X}} \rangle_k = 0$ for $\mathbf{x} \in \mathrm{supp}(P_{\mathbf{X}})$. Hence,

$$\frac{\mathrm{Var}\left(\phi(\mathbf{X}) + \psi(\mathbf{x})\right)}{\|\phi + \psi\|_k^2} = \frac{\mathrm{Var}\left(\phi(\mathbf{X})\right)}{\|\phi\|_k^2 + \|\psi\|_k^2} < \frac{\mathrm{Var}\left(\phi(\mathbf{X})\right)}{\|\phi\|_k^2}.$$

So the minimum in (A.27), if attained, is not in $\mathcal{H}_{P_{\mathbf{X}}}$.

# B

# Supplement for Chapter 3

This chapter contains the proofs of main theorems presented in Chapter 3. Section B.1 contains an outline of the proofs of the main theorems, whereas Sections B.2-B.4 contain proofs of the more technical supporting lemmas.

## B.1 Proofs for Main Results in Section 3.3

### B.1.1 Proof of Theorem 4

The proof is based on Lemma 13, which gives an error bound for the pairwise terms $\sin(\frac{\pi}{2}\boldsymbol{r}_{ij}^K)$, and Lemma 14, which gives an error bound for the scale estimates $\hat{\sigma}_i$. Note that we require the bound $\epsilon \leq 0.02$ on the level of contamination in Lemma 13, but the requirement could be relaxed with a more refined proof technique. The proofs of Lemmas 13 and 14 are provided in Sections B.2.1 and B.2.2.

**Lemma 13.** *Under model* (3.1)*, let* $\epsilon = \max_{1\leq i\leq p}\epsilon_i \leq 0.02$*. For any constant* $C > \pi\sqrt{2}$*, we have*

$$\max_{1\leq i,j\leq p}\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| \leq C\sqrt{\frac{\log p}{n}} + 26\pi\epsilon, \tag{B.1}$$

*with probability at least* $1 - 2p^{-(C^2/\pi^2-2)}$*.*

**Lemma 14.** *Under model* (3.1)*, suppose* $0 < \min_{1\le i\le p}\sigma_i \le \max_{1\le i\le p}\sigma_i \le M_\sigma$*, and the maximum contamination error satisfies* $\epsilon = \max_{1\le i\le p}\epsilon_i \le \frac{1}{16}$*. Let* $c(\sigma_i)$ *be defined as in equation* (3.12)*, and suppose* $C' > \frac{1}{\Phi^{-1}(0.75)\min_{1\le i\le p} c(\sigma_i)\sqrt{2}}$ *and* $\Phi^{-1}(0.75)C'\sqrt{\frac{\log p}{n}} < 1$*. Then with probability at least* $1 - 6p^{-\{2[\Phi^{-1}(0.75)]^2 C'^2 \min_{1\le i\le p} c^2(\sigma_i)-1\}}$*, we have*

$$\max_{1\le i\le p}|\hat{\sigma}_i - \sigma_i| \le C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon.$$

By the triangle inequality, we may decompose $|\hat{\sigma}_i\hat{\sigma}_j \sin(\frac{\pi}{2}\boldsymbol{r}_{ij}^K) - \boldsymbol{\Sigma}_{ij}^*|$ as follows:

$$\begin{aligned}
&\left|\hat{\sigma}_i\hat{\sigma}_j \sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \sigma_i\sigma_j\boldsymbol{\rho}_{ij}\right| \\
&\le |\hat{\sigma}_i - \sigma_i||\hat{\sigma}_j - \sigma_j|\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + \left|\sigma_i \sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right)\right||\hat{\sigma}_j - \sigma_j| \\
&\qquad + |\hat{\sigma}_i\sigma_j|\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + |\hat{\sigma}_j\rho_{ij}||\hat{\sigma}_i - \sigma_i| \\
&\overset{(i)}{\le} |\hat{\sigma}_i - \sigma_i||\hat{\sigma}_j - \sigma_j|\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + \sigma_i|\hat{\sigma}_j - \sigma_j| \\
&\qquad + |\hat{\sigma}_i\sigma_j|\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + \hat{\sigma}_j|\hat{\sigma}_i - \sigma_i| \\
&\le |\hat{\sigma}_i - \sigma_i||\hat{\sigma}_j - \sigma_j|\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + \sigma_i|\hat{\sigma}_j - \sigma_j| \\
&\qquad + (|\hat{\sigma}_i - \sigma_i| + \sigma_i)\sigma_j\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| + (|\hat{\sigma}_j - \sigma_j| + \sigma_j)|\hat{\sigma}_i - \sigma_i|,
\end{aligned}$$

where $(i)$ uses the facts that $|\sin(x)| \le 1$ for all $x$, and $|\boldsymbol{\rho}_{ij}| \le 1$, since it is a correlation coefficient. Using Lemmas 13 and 14 and the assumption (3.13), we obtain the overall bound

$$\left(C\sqrt{\frac{\log p}{n}} + 26\pi\epsilon\right)\left(C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right)^2 + M_\sigma\left(C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right) + \left(M_\sigma + C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right)\left\{\left(C\sqrt{\frac{\log p}{n}} + 26\pi\epsilon\right)M_\sigma + \left(C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right)\right\}$$

$$\leq (M_\sigma(M_\sigma+1)+1)\left(C\sqrt{\frac{\log p}{n}}+26\pi\epsilon\right)+(2M_\sigma+1)\left(C'\sqrt{\frac{\log p}{n}}+7.2M_\sigma\epsilon\right),$$

implying inequality (3.14).

### B.1.2 Proof of Theorem 5

The proof is based on Lemma 15, which gives an error bound for $2\sin(\frac{\pi}{6}\boldsymbol{r}_{ij}^S)$, and Lemma 14, which gives an error bound for $\hat{\sigma}_i$. Note that we require the bound $\epsilon \leq 0.01$ on the level of contamination in Lemma 15, but the requirement could again be relaxed with a more refined proof technique. The proof of Lemma 15 is contained in Section B.2.3.

**Lemma 15.** *Under model* (3.1)*, let* $\epsilon = \max_{1\leq i\leq p}\epsilon_i \leq 0.01$*. Suppose* $C > 8\pi$ *and the sample size satisfies* $n \geq \max\left\{15,\ \frac{16\pi^2}{C^2\log p}\right\}$*. Then*

$$\max_{1\leq i,j\leq p}\left|2\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right)-\boldsymbol{\rho}_{ij}^S\right| \leq \frac{5C}{2}\sqrt{\frac{\log p}{n}}+51\pi\epsilon, \tag{B.2}$$

*with probability at least* $1-2p^{-\left\{\frac{C^2}{32\pi^2}-2\right\}}$.

Using a similar decomposition as in the proof of Theorem 4, we have

$$\begin{aligned}
&\left|2\hat{\sigma}_i\hat{\sigma}_j\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right)-\sigma_i\sigma_j\boldsymbol{\rho}_{ij}\right| \\
&\leq |\hat{\sigma}_i-\sigma_i||\hat{\sigma}_j-\sigma_j|\left|2\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right)-\boldsymbol{\rho}_{ij}\right|+\sigma_i|\hat{\sigma}_j-\sigma_j| \\
&\qquad +(|\hat{\sigma}_i-\sigma_i|+\sigma_i)\sigma_j\left|2\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right)-\boldsymbol{\rho}_{ij}\right|+(|\hat{\sigma}_j-\sigma_j|+\sigma_j)|\hat{\sigma}_i-\sigma_i|.
\end{aligned}$$

Using Lemmas 14 and 15, we then obtain the overall upper bound

$$\left(\frac{5C}{2}\sqrt{\frac{\log p}{n}}+51\pi\epsilon\right)\left(C'\sqrt{\frac{\log p}{n}}+7.2M_\sigma\epsilon\right)^2+M_\sigma\left(C'\sqrt{\frac{\log p}{n}}+7.2M_\sigma\epsilon\right)+$$

$$\left(M_\sigma + C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right)\left\{M_\sigma\left(\frac{5C}{2}\sqrt{\frac{\log p}{n}} + 51\pi\epsilon\right) + \left(C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right)\right\}$$
$$\leq (M_\sigma(M_\sigma+1)+1)\left(\frac{5C}{2}\sqrt{\frac{\log p}{n}} + 51\pi\epsilon\right) + (2M_\sigma+1)\left(C'\sqrt{\frac{\log p}{n}} + 7.2M_\sigma\epsilon\right),$$

which is easily simplified to obtain the prescribed bound.

### B.1.3 Proof of Theorem 6

Clearly, it suffices to prove the elementwise deviation bound for the unsymmetrized matrix $\hat{\mathbf{\Omega}}$. A version of the following result appears in Cai et al. (2011); the proof is provided in Section B.2.4 for completeness.

**Lemma 16.** *Suppose $\mathbf{\Omega}^* \in \mathcal{U}(q, s_0(p), M)$. If $\hat{\mathbf{\Omega}}$ is the output of the CLIME estimator* (3.10) *with $\lambda \geq M\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty$, then $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leq 4\|\mathbf{\Omega}^*\|_{L_1}\lambda$.*

Combining Lemma 16 with Theorem 4, we obtain the desired result.

### B.1.4 Proof of Theorem 7

Our proof is based on the following result:

**Lemma 17** (Theorem 1 in Ravikumar et al. (2011))**.** *Suppose $\mathbf{\Omega}^*$ satisfies the incoherence condition* (3.18)*, and that for all $1 \leq i, j \leq p$, the tail condition*

$$P\left(|\hat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}^*_{ij}| \geq \delta\right) \leq \frac{1}{f(n,\delta)}, \qquad \forall \delta > 0, \tag{B.3}$$

*holds, for some function $f$ that is monotonically increasing in $n$. Also suppose*

$$n > \bar{n}_f\left(\frac{1}{6(1+8/\alpha)k\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}},\ p^\tau\right),$$

*where* $\bar{n}_f(\delta; r) = \text{argmax}\{n : f(n, \delta) \leq r\}$ *and* $\bar{\delta}_f(n; r) := \text{argmax}\{\delta : f(n, \delta) \leq r\}$. *Then with probability at least* $1 - p^{2-\tau}$, *for the choice* $\lambda = \frac{8}{\alpha}\bar{\delta}_f(n, p^\tau)$, *the GLasso estimator satisfies* $\text{supp}(\hat{\mathbf{\Omega}}) \subseteq \text{supp}(\mathbf{\Omega}^*)$ *and*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leq 2\kappa_{\Gamma^*}\left(1 + \frac{8}{\alpha}\right)\bar{\delta}_f(n, p^\tau).$$

Inspecting the proofs of the technical lemmas employed in proving Theorem 4, inequality (B.3) holds with the function $f(n, \delta) = c_1 \exp(c_2 n(\delta - c_0\epsilon)^2)$, defined for $\delta > c_0\epsilon$, where $c_0, c_1$, and $c_2$ are appropriately chosen constants. An easy calculation shows that $\bar{\delta}_f(n, r) = c_0\epsilon + \sqrt{\frac{1}{c_2 n}\log\left(\frac{r}{c_1}\right)}$, so $\bar{\delta}_f(n, p^\tau) = c_0\epsilon + C_1\sqrt{\frac{\tau \log p}{n}}$. Similarly, we may easily verify that $\bar{n}_f(\delta, p^\tau) = C_2\frac{\tau \log p}{(\delta - c_0\epsilon)^2}$. Lemma 17 then implies the desired conclusions.

### B.1.5 Proof of Theorem 9

Note that $\check{\mathbf{\Sigma}}$ is the projection of the robust covariance estimator $\hat{\mathbf{\Sigma}}$ onto the positive semidefinite cone, where the distance is measured in the elementwise $\ell_\infty$-norm. Furthermore, note that $\|\check{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\|_\infty \leq \|\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}\|_\infty$, since $\mathbf{\Sigma}^* \succeq 0$. Hence,

$$\|\check{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \|\check{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\|_\infty + \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq 2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty. \tag{B.4}$$

This implies that the bound (B.3) in Lemma 17 holds with $\hat{\mathbf{\Sigma}}$ replaced by $\check{\mathbf{\Sigma}}$, and $f(n, \delta)$ replaced by $f(n, \delta/2)$. Proceeding as in the proof of Theorem 7, we arrive at the bound (3.31).

Turning to the derivation of the breakdown point, note that by Theorem 1 of Oellerer & Croux (2014), we have

$$\epsilon_n(\check{\mathbf{\Omega}}(\mathbf{X}), \mathbf{X}) \geq \epsilon_n^+(\check{\mathbf{\Sigma}}(\mathbf{X}), \mathbf{X}). \tag{B.5}$$

Consider the estimator $\check{\mathbf{\Sigma}}(\mathbf{X}^m)$, based on corrupted data. We have

$$\|\check{\mathbf{\Sigma}}(\mathbf{X}^m) - \mathbf{\Sigma}^*\|_\infty \leq 2\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m) - \mathbf{\Sigma}^*\|_\infty \leq 2\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m)\|_\infty + 2\|\mathbf{\Sigma}^*\|_\infty, \tag{B.6}$$

where the first inequality follows from the bound (B.4), and the second inequality comes from the triangle inequality. Furthermore, note that since $\check{\mathbf{\Sigma}}(\mathbf{X}^m) \succeq 0$ by construction,

$$\begin{aligned}\lambda_1(\check{\mathbf{\Sigma}}(\mathbf{X}^m)) = \|\check{\mathbf{\Sigma}}(\mathbf{X}^m)\|_2 &\leq \|\check{\mathbf{\Sigma}}(\mathbf{X}^m) - \mathbf{\Sigma}^*\|_2 + \|\mathbf{\Sigma}^*\|_2 \\ &\leq p\|\check{\mathbf{\Sigma}}(\mathbf{X}^m) - \mathbf{\Sigma}^*\|_\infty + \|\mathbf{\Sigma}^*\|_2, \end{aligned} \tag{B.7}$$

where we have used the bound $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq p\|\mathbf{A}\|_\infty$, for all $\mathbf{A} \in \mathbb{R}^{p\times p}$, in the last inequality. Combining inequalities (B.6) and (B.7), we then obtain

$$\lambda_1(\check{\mathbf{\Sigma}}(\mathbf{X}^m)) \leq 2p\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m)\|_\infty + 2p\|\mathbf{\Sigma}^*\|_\infty + \|\mathbf{\Sigma}^*\|_2,$$

so

$$\begin{aligned}&\left|\lambda_1(\check{\mathbf{\Sigma}}(\mathbf{X}^m)) - \lambda_1(\check{\mathbf{\Sigma}}(\mathbf{X}))\right| \\ &\quad\leq \lambda_1(\check{\mathbf{\Sigma}}(\mathbf{X})) + \left(2p\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m)\|_\infty + 2p\|\mathbf{\Sigma}^*\|_\infty + \|\mathbf{\Sigma}^*\|_2\right). \end{aligned} \tag{B.8}$$

Finally, since the correlation estimators are bounded in magnitude by 1, we have

$$\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m)\|_\infty \leq \max_{1\leq i,j\leq p} \hat{\sigma}_i(\mathbf{X}^m)\hat{\sigma}_j(\mathbf{X}^m), \tag{B.9}$$

where $\{\hat{\sigma}_i(\mathbf{X}^m)\}_{1\leq i\leq p}$ are the robust scale estimators based on $\mathbf{X}^m$, given by the MAD estimators calculated from the corresponding columns. Furthermore, the breakdown point of the MAD is 50% (Huber, 1981), so the quantity on the right-hand side of

inequality (B.9) is finite when $\frac{m}{n} < 50\%$. Then by inequality (B.8) and the definition of the explosion breakdown point, we conclude that $\epsilon_n^+(\check{\mathbf{\Sigma}}(\mathbf{X}), \mathbf{X}) \geq 50\%$. By inequality (B.5), we therefore have $\epsilon_n(\check{\mathbf{\Omega}}(\mathbf{X}), \mathbf{X}) \geq 50\%$, as well.

We now establish that $\epsilon_n(\check{\mathbf{\Omega}}(\mathbf{X}), \mathbf{X}) = 50\%$. Note that if we are allowed to corrupt more than 50% of the entries in each column of the data matrix, the columnwise MAD estimates may be made arbitrarily small (say, smaller than some value $a$); indeed, we may simply replace more than half of the entries in each column by values in $(0, a)$. Consequently, the overall covariance estimator $\hat{\mathbf{\Sigma}}(\mathbf{X}^m)$ will have all entries bounded in magnitude by $[\Phi^{-1}(0.75)]^{-2}a^2$. We claim that the diagonal elements of $\check{\mathbf{\Sigma}}(\mathbf{X}^m)$ must therefore be bounded in magnitude by $2[\Phi^{-1}(0.75)]^{-2}a^2$. Indeed, note that the matrix $\text{diag}(\hat{\mathbf{\Sigma}}(\mathbf{X}^m))$ is feasible for the projection (3.29). Hence, we must have

$$\|\hat{\mathbf{\Sigma}}(\mathbf{X}^m) - \check{\mathbf{\Sigma}}(\mathbf{X}^m)\|_\infty \leq \|\hat{\mathbf{\Sigma}}(\mathbf{X}^m) - \text{diag}(\hat{\mathbf{\Sigma}}(\mathbf{X}^m))\|_\infty \leq [\Phi^{-1}(0.75)]^{-2}a^2,$$

implying in particular that

$$\begin{aligned}\|\text{diag}(\check{\mathbf{\Sigma}}(\mathbf{X}^m))\|_\infty &\leq \|\text{diag}(\hat{\mathbf{\Sigma}}(\mathbf{X}^m))\|_\infty + \|\text{diag}(\hat{\mathbf{\Sigma}}(\mathbf{X}^m)) - \text{diag}(\check{\mathbf{\Sigma}}(\mathbf{X}^m))\|_\infty \\ &\leq \frac{2a^2}{[\Phi^{-1}(0.75)]^2},\end{aligned}$$

as claimed. Now note that the first-order optimality condition for the GLasso is given by

$$\check{\mathbf{\Sigma}}(\mathbf{X}^m) - \left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1} + \lambda \cdot \text{sign}\{\check{\mathbf{\Omega}}(\mathbf{X}^m) - \text{diag}(\check{\mathbf{\Omega}}(\mathbf{X}^m))\} = 0,$$

where the sign function is computed entrywise, omitting the diagonal elements of $\check{\mathbf{\Omega}}(\mathbf{X}^m)$. In particular, this implies that the $\text{diag}(\check{\mathbf{\Sigma}}(\mathbf{X}^m)) = \text{diag}\left\{\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right\}$, so the diagonal elements of $\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}$ are bounded in magnitude by $2[\Phi^{-1}(0.75)]^{-2}a^2$

as well. Hence,

$$\lambda_p\left(\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right) = \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^T\left(\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right)\mathbf{v} \leq \min_{1\leq j\leq p} \mathbf{e}_j^T\left(\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right)\mathbf{e}_j$$
$$\leq \left\|\text{diag}\left\{\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right\}\right\|_\infty \leq 2[\Phi^{-1}(0.75)]^{-2}a^2,$$

where the $\mathbf{e}_j$'s are the canonical basis vectors, and we have used the variational representation of eigenvalues of a Hermitian matrix to show that the minimum eigenvalue is bounded by the minimum diagonal entry. This allows us to conclude that

$$1 = \lambda_p\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right) \leq \lambda_1\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)\lambda_p\left(\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)^{-1}\right)$$
$$\leq \lambda_1\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right)\cdot\frac{2a^2}{[\Phi^{-1}(0.75)]^2},$$

where we have used the inequality $\lambda_p(\mathbf{A}\boldsymbol{B}) \leq \lambda_1(\mathbf{A})\lambda_p(\boldsymbol{B})$, for $\mathbf{A}, \boldsymbol{B} \succeq 0$, in the first inequality (Zhang, 2011). Hence, $\lambda_1\left(\check{\mathbf{\Omega}}(\mathbf{X}^m)\right) \geq \frac{[\Phi^{-1}(0.75)]^2}{2a^2}$. However, we may choose $a$ to be arbitrarily close to 0, implying that the maximum eigenvalue of $\check{\mathbf{\Omega}}(\mathbf{X}^m)$ may be made arbitrarily large, and the estimator breaks down. This concludes the proof.

### B.1.6 Proof of Theorem 10

Clearly, $\epsilon_n(\hat{\mathbf{\Omega}}, \mathbf{X}) \geq \frac{1}{n}$ for any $\mathbf{X}$, by the definition of the breakdown point. To show equality, we now provide a data matrix $X$ and a corrupted data matrix $\mathbf{X}^1$, where $\mathbf{X}^1$ differs from $\mathbf{X}$ in at most one element per column, and the CLIME problem is

feasible for $\hat{\mathbf{\Sigma}}(\mathbf{X})$ but infeasible for $\hat{\mathbf{\Sigma}}(\mathbf{X}^1)$. Let

$$\mathbf{X}^1 = \begin{pmatrix} a_1 & -a_1 \\ a_2 & -a_2 \\ \vdots & \vdots \\ a_n & -a_n \end{pmatrix},$$

where the $a_k$'s are all distinct. Note that the columns of $\mathbf{X}^1$ are perfectly negatively correlated; hence, the correlation matrix (computed from either Kendall's tau or Spearman's rho, for instance) is $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. Furthermore, we have $\hat{\sigma}_1 = \hat{\sigma}_2 := \hat{\sigma}$, since the data in the two columns are negatives of each other. It follows that $\hat{\mathbf{\Sigma}}(\mathbf{X}^1) = \hat{\sigma}^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. Clearly, the problem

$$\beta_1 : \left\| \hat{\mathbf{\Sigma}}(\mathbf{X}^1)\beta_1 - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|_\infty \leq \lambda$$

is infeasible for $\lambda < \frac{1}{2}$. Hence, the CLIME estimator based on $\hat{\mathbf{\Sigma}}(\mathbf{X}^1)$ is infeasible.

On the other hand, we may construct an initial data matrix $\mathbf{X}$ such that the CLIME program based on $\hat{\mathbf{\Sigma}}(\mathbf{X})$ is feasible, simply by altering the last row of $\mathbf{X}^1$. Suppose we change the last row of $\mathbf{X}^1$ to $(a_n, a_n)$. Then the columns are no longer perfectly negatively correlated, and it is easy to check that the correlation matrix of $\mathbf{X}$ will take the form $\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$, for some $|a| < 1$. Denoting the corresponding

estimates of scale as $\hat{\sigma}_1$ and $\hat{\sigma}_2$, we then have

$$\hat{\mathbf{\Sigma}}(\mathbf{X}) = \begin{pmatrix} \hat{\sigma}_1^2 & a\hat{\sigma}_1\hat{\sigma}_2 \\ a\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{pmatrix}.$$

Note that $\det\{\hat{\mathbf{\Sigma}}(\mathbf{X})\} = \hat{\sigma}_1^2\hat{\sigma}_2^2(1-a^2) > 0$. It follows that $\hat{\mathbf{\Sigma}}(\mathbf{X})$ is invertible. In particular, the matrix $\left(\hat{\mathbf{\Sigma}}(\mathbf{X})\right)^{-1}$ is always a feasible point for the CLIME program based on $\hat{\mathbf{\Sigma}}(\mathbf{X})$.

Hence, we conclude that the CLIME program breaks down when even one corruption per column is allowed. It follows that $\epsilon_n(\hat{\mathbf{\Omega}}, \mathbf{X}) = \frac{1}{n}$ for the constructed value of $\mathbf{X}$.

## B.2 Supporting proofs for Section 3.3

In this section, we provide the proofs of the technical lemmas used to establish the theorems in Section 3.3.

### B.2.1 Proof of Lemma 13

When $i = j$, we have

$$\begin{aligned} \boldsymbol{r}_{ii}^K &= \frac{2}{n(n-1)} \sum_{k<\ell} \operatorname{sign}^2(X_{ki} - X_{\ell i}) \\ &= \frac{2}{n(n-1)} \sum_{k<\ell} (1 - \mathbb{1}(X_{ki} = X_{\ell i})) \\ &= 1 - \frac{2}{n(n-1)} \sum_{k<\ell} \mathbb{1}(X_{ki} = X_{\ell i}). \end{aligned}$$

Hence,

$$\begin{aligned}\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ii}^K\right)-\boldsymbol{\rho}_{ii}\right| &= \left|\sin\left(\frac{\pi}{2}-\frac{\pi}{n(n-1)}\sum_{k<\ell}\mathbb{1}(X_{ki}=X_{\ell i})\right)-1\right| \\ &= \left|\cos\left(\frac{\pi}{n(n-1)}\sum_{k<\ell}\mathbb{1}(X_{ki}=X_{\ell i})\right)-\cos(0)\right| \\ &\leq \frac{\pi}{2}q_i,\end{aligned}$$

where

$$q_i = \frac{2}{n(n-1)}\sum_{k<\ell}\mathbb{1}(X_{ki}=X_{\ell i})$$

is a $U$-statistic, and the last inequality follows from the fact that $\cos(x)$ is 1-Lipschitz. By Hoeffding's inequality for $U$-statistics, we have

$$P\left(\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ii}^K\right)-\boldsymbol{\rho}_{ii}\right|\geq t\right)\leq P\left(q_i\geq\frac{2t}{\pi}\right)\leq\exp\left(-\frac{4nt^2}{\pi^2}\right). \tag{B.10}$$

Now, consider the case where $i\neq j$. Note that

$$\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right)-\boldsymbol{\rho}_{ij}\right|\leq\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right)-\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right)\right|+\left|\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right)-\boldsymbol{\rho}_{ij}\right|, \tag{B.11}$$

where $\boldsymbol{\rho}_{ij}^K = E(\boldsymbol{r}_{ij}^K)$ and the expectation is with respect to the distribution under model (3.1). Since $\boldsymbol{r}_{ij}^K$ is a $U$-statistic with kernel bounded between $-1$ and 1, Hoeffding's inequality and the fact that $\sin(x)$ is 1-Lipschitz implies that the first term on the right-hand side of inequality (B.11) satisfies

$$P\left(\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right)-\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right)\right|\geq t\right)\leq P\left(|\boldsymbol{r}_{ij}^K-\boldsymbol{\rho}_{ij}^K|\geq\frac{2}{\pi}t\right)\leq 2\exp\left(-\frac{nt^2}{\pi^2}\right). \tag{B.12}$$

Combining inequalities (B.10) and (B.12) and taking $t = C\sqrt{\frac{\log p}{n}}$, we conclude that

with probability at least $1 - 2p^{-(C^2/\pi^2-2)}$,

$$\max_{1\le i\le p}\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ii}^K\right)-\boldsymbol{\rho}_{ii}\right|\le C\sqrt{\frac{\log p}{n}},\qquad\text{and}\tag{B.13a}$$

$$\max_{i\neq j}\left|\sin\left(\frac{\pi}{2}\boldsymbol{r}_{ij}^K\right)-\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right)\right|\le C\sqrt{\frac{\log p}{n}}.\tag{B.13b}$$

For the second term on the right-hand side of equation (B.11), we have under model (3.1) that for any pair $i \neq j$,

$$(X_{ki}, X_{kj}) \overset{\text{i.i.d.}}{\sim} F_{ij} = (1-\gamma_{ij})\Phi_{\boldsymbol{\mu}_{\{i,j\}},\boldsymbol{\Sigma}_{\{i,j\}}} + \gamma_{ij}H_{ij},\qquad \forall 1\le k\le n,\tag{B.14}$$

where $\Phi_{\boldsymbol{\mu}_{\{i,j\}},\boldsymbol{\Sigma}_{\{i,j\}}} = N(\boldsymbol{\mu}_{\{i,j\}}, \boldsymbol{\Sigma}_{\{i,j\}})$ is the marginal distribution of $(Y_{ki}, Y_{kj})$, $H_{ij}$ is a mixture of the distributions of $Y_{ki}, Y_{kj}, Z_{ki}$, and $Z_{kj}$, and $1-\gamma_{ij} = (1-\epsilon_i)(1-\epsilon_j)$.

By Lemma 24, we have $\boldsymbol{\rho}_{ij}^K = \frac{2}{\pi}\sin^{-1}\boldsymbol{\rho}_{ij} + R_{ij}$, where $|R_{ij}| \le 12\gamma_{ij} + 17\gamma_{ij}^2$. Setting $R'_{ij} = \frac{\pi}{2}R_{ij}$, we then have

$$\begin{aligned}
\left|\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right)-\boldsymbol{\rho}_{ij}\right| &= \left|\sin\left(\sin^{-1}(\boldsymbol{\rho}_{ij})+R'_{ij}\right)-\boldsymbol{\rho}_{ij}\right| \\
&= \left|\sin(\sin^{-1}(\boldsymbol{\rho}_{ij}))\cos(R'_{ij})+\cos(\sin^{-1}(\boldsymbol{\rho}_{ij}))\sin(R'_{ij})-\boldsymbol{\rho}_{ij}\right| \\
&= \left|\boldsymbol{\rho}_{ij}\cos(R'_{ij})+\sqrt{1-\boldsymbol{\rho}_{ij}^2}\sin(R'_{ij})-\boldsymbol{\rho}_{ij}\right| \\
&\le \left|\boldsymbol{\rho}_{ij}\left(1-\cos(R'_{ij})\right)\right| + \left|\sqrt{1-\boldsymbol{\rho}_{ij}^2}\sin(R'_{ij})\right| \\
&\le \left[1-\cos(R'_{ij})\right] + \left|\sin(R'_{ij})\right|.
\end{aligned}$$

Note that $\gamma_{ij} = \epsilon_i + \epsilon_j - \epsilon_i\epsilon_j \le 2\epsilon$, so

$$|R'_{ij}| \le \frac{\pi}{2}(12\gamma_{ij}+17\gamma_{ij}^2) \le \frac{\pi}{2}\left(12\cdot 2\epsilon + 17(2\epsilon)^2\right) = 12\pi\epsilon + 34\pi\epsilon^2.$$

In particular, this bound is less than 1 when $\epsilon \le 0.02$. Then using the fact that

$|\sin(x) - x| \leq \frac{|x|^3}{3!}$ and $|1 - \cos(x)| \leq \frac{x^2}{2!}$ for $|x| \leq 1$, we conclude that

$$\begin{aligned}\max_{1\leq i,j\leq p}\left|\sin\left(\frac{\pi}{2}\boldsymbol{\rho}_{ij}^K\right) - \boldsymbol{\rho}_{ij}\right| &\leq \max_{1\leq i,j\leq p}\left[|R'_{ij}| + \frac{(R'_{ij})^2}{2} + \frac{|R'_{ij}|^3}{6}\right] \\ &\leq 2\max_{1\leq i,j\leq p}|R'_{ij}| \leq 26\pi\epsilon. \end{aligned} \tag{B.15}$$

Combining inequalities (B.13) and (B.15) then proves the desired result.

### B.2.2 Proof of Lemma 14

Under model (3.1), we have the marginal distributions

$$X_{ki} \overset{\text{i.i.d.}}{\sim} F_i = (1-\epsilon_i)\Phi_{\mu_i,\sigma_i} + \epsilon_i H_i, \quad \forall 1 \leq k \leq n,$$

for each $1 \leq i \leq p$, where $\Phi_{\mu_i,\sigma_i} = N(\mu_i, \sigma_i^2)$ is the marginal distribution of $Y_{ki}$ and $H_i$ is the marginal distribution of $Z_{ki}$.

Let $d(F_i)$ and $d(\Phi_{\mu_i,\sigma_i})$ denote the population MADs corresponding to $F_i$ and $\Phi_{\mu_i,\sigma_i}$, respectively. Since $\hat{\sigma}_i = [\Phi^{-1}(0.75)]^{-1}\hat{d}_i$ and $\sigma_i = [\Phi^{-1}(0.75)]^{-1}d(\Phi_{\mu_i,\sigma_i})$, with $\hat{d}_i$ defined as in equation (3.3), it suffices to bound the term $|\hat{d}_i - d(\Phi_{\mu_i,\sigma_i})|$, which we decompose as follows:

$$|\hat{d}_i - d(\Phi_{\mu_i,\sigma_i})| \leq |\hat{d}_i - d(F_i)| + |d(F_i) - d(\Phi_{\mu_i,\sigma_i})|.$$

By Lemma 23, for $0 < t < 1$,

$$\begin{aligned} P\big(\max_{1\leq i\leq p}|\hat{d}_i - d(F_i)| > t\big) &\leq \sum_{i=1}^{p} P\big(|\hat{d}_i - d(F_i)| > t\big) \\ &\leq 6p\max_{1\leq i\leq p}\left\{\exp(-2nc^2(\sigma_i)t^2)\right\} \\ &= 6p\exp\left(-2n\min_{1\leq i\leq p}c^2(\sigma_i)t^2\right). \end{aligned}$$

Let $t = \Phi^{-1}(0.75)C'\sqrt{\frac{\log p}{n}} < 1$. With probability at least

$$1 - 6p^{-\{2[\Phi^{-1}(0.75)]^2 C'^2 \min_{1\le i\le p} c^2(\sigma_i)-1\}},$$

we then have

$$\max_{1\le i\le p} |\hat{d}_i - d(F_i)| \le \Phi^{-1}(0.75)C'\sqrt{\frac{\log(p)}{n}}.$$

On the other hand, by Lemma 22, we have

$$\max_{1\le i\le p} |d(F_i) - d(\Phi_{\mu_i,\sigma_i})| \le 4.8 \max_{1\le i\le p} \sigma_i\epsilon_i \le 4.8M_\sigma\epsilon.$$

Thus, with probability at least $1 - 6p^{-\{2[\Phi^{-1}(0.75)]^2 C'^2 \min_{1\le i\le p} c^2(\sigma_i)-1\}}$,

$$\max_{1\le i\le p} |\hat{d}_i - d(\Phi_{\mu_i,\sigma_i})| \le \Phi^{-1}(0.75)C'\sqrt{\frac{\log(p)}{n}} + 4.8M_\sigma\epsilon.$$

It follows that with the same probability,

$$\max_{1\le i\le p} |\hat{\sigma}_i - \sigma_i| = [\Phi^{-1}(0.75)]^{-1} \max_{1\le i\le p} |\hat{d}_i - d(\Phi_{\mu_i,\sigma_i})| \le C'\sqrt{\frac{\log(p)}{n}} + 7.2M_\sigma\epsilon.$$

### B.2.3 Proof of Lemma 15

When $i = j$, we have $2\sin(\frac{\pi}{6}r_{ii}^S) = \rho_{ii} = 1$; hence, we only need to consider the case when $i \ne j$. First, note that

$$\begin{aligned}\left|2\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right) - \boldsymbol{\rho}_{ij}\right| \le 2\left|\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right) - \sin\left(\frac{\pi}{6}E(\boldsymbol{r}_{ij}^S)\right)\right| \\ + \left|2\sin\left(\frac{\pi}{6}E(\boldsymbol{r}_{ij}^S)\right) - \boldsymbol{\rho}_{ij}\right|, \qquad \text{(B.16)}\end{aligned}$$

where the expectation is taken with respect to the distribution under model (3.1). By Lemma 26, we have $\boldsymbol{r}_{ij}^S = \frac{n-2}{n+1}U_{ij} + \frac{3}{n+1}\boldsymbol{r}_{ij}^K$, where $U_{ij}$ is a $U$-statistic with kernel bounded between $-3$ and 3, and $\boldsymbol{r}_{ij}^K$ is the Kendall's tau correlation. Using the fact that $\sin(x)$ is 1-Lipschitz, we then have

$$
\begin{aligned}
&P\left(2\left|\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right) - \sin\left(\frac{\pi}{6}E(\boldsymbol{r}_{ij}^S)\right)\right| \geq t\right) \\
&\leq P\left(|\boldsymbol{r}_{ij}^S - E(\boldsymbol{r}_{ij}^S)| \geq \frac{3t}{\pi}\right) \\
&= P\left(\left|\frac{n-2}{n+1}(U_{ij} - E(U_{ij})) + \frac{3}{n+1}(\boldsymbol{r}_{ij}^K - \boldsymbol{\rho}_{ij}^K)\right| \geq \frac{3t}{\pi}\right) \\
&\leq P\left(|U_{ij} - E(U_{ij}))| + \frac{6}{n+1} \geq \frac{3t}{\pi}\right) \\
&\leq P\left(|U_{ij} - E(U_{ij})| \geq \frac{3t}{2\pi}\right),
\end{aligned}
$$

where the last inequality follows from the choice $t = C\sqrt{\frac{\log p}{n}}$ and the fact that $\frac{6}{n+1} \leq \frac{3t}{2\pi}$ when $n \geq \frac{16\pi^2}{C^2 \log p}$. Furthermore, Hoeffding's inequality implies

$$
P\left(|U_{ij} - E(U_{ij})| \geq \frac{3t}{2\pi}\right) \leq 2\exp\left(-2\left\lfloor\frac{n}{3}\right\rfloor\left(\frac{3t}{2\pi}\right)^2\frac{1}{6^2}\right) \leq 2\exp\left(-\frac{nt^2}{32\pi^2}\right).
$$

Plugging in $t = C\sqrt{\frac{\log p}{n}}$ and using a union bound, we then have

$$
\begin{aligned}
&P\left(\max_{1\leq i,j\leq p} 2\left|\sin\left(\frac{\pi}{6}\boldsymbol{r}_{ij}^S\right) - \sin\left(\frac{\pi}{6}E(\boldsymbol{r}_{ij}^S)\right)\right| \geq C\sqrt{\frac{\log p}{n}}\right) \\
&\leq 2p^2\exp\left(-\frac{C^2\log p}{32\pi^2}\right) = 2p^{-\left\{\frac{C^2}{32\pi^2}-2\right\}}. \qquad \text{(B.17)}
\end{aligned}
$$

For the second term on the right-hand side of equation (B.11), we have under model (3.1) that for any pair $i \neq j$,

$$
(X_{ki}, X_{kj}) \overset{\text{i.i.d.}}{\sim} F_{ij} = (1-\gamma_{ij})\Phi_{\boldsymbol{\mu}_{\{i,j\}},\boldsymbol{\Sigma}_{\{i,j\}}} + \gamma_{ij}H_{ij}, \qquad \forall 1 \leq k \leq n,
$$

where $\Phi_{\boldsymbol{\mu}_{\{i,j\}},\boldsymbol{\Sigma}_{\{i,j\}}} = N(\boldsymbol{\mu}_{\{i,j\}}, \boldsymbol{\Sigma}_{\{i,j\}})$ is the marginal distribution of $(Y_{ki}, Y_{kj})$, $H_{ij}$ is a mixture of the distributions of $Y_{ki}, Y_{kj}, Z_{ki}$, and $Z_{kj}$, and $1 - \gamma_{ij} = (1 - \epsilon_i)(1 - \epsilon_j)$.

By Lemma 25, we have $E(\boldsymbol{r}_{ij}^S) = \frac{6}{\pi} \sin^{-1}\left(\frac{\rho_{ij}}{2}\right) + R_{ij}$, where $|R_{ij}| \leq 48\gamma_{ij} + 129\gamma_{ij}^2 + 88\gamma_{ij}^3 + \frac{12}{n+1}$. Setting $R'_{ij} = \frac{\pi}{6} R_{ij}$, we then have

$$
\begin{aligned}
&\left| 2\sin\left(\frac{\pi}{6} E(\boldsymbol{r}_{ij}^S)\right) - \boldsymbol{\rho}_{ij} \right| \\
&= \left| 2\sin\left(\sin^{-1}(\boldsymbol{\rho}_{ij}/2) + R'_{ij}\right) - \boldsymbol{\rho}_{ij} \right| \\
&= \left| 2\sin(\sin^{-1}(\boldsymbol{\rho}_{ij}/2))\cos(R'_{ij}) + 2\cos(\sin^{-1}(\boldsymbol{\rho}_{ij}/2))\sin(R'_{ij}) - \boldsymbol{\rho}_{ij} \right| \\
&= \left| \boldsymbol{\rho}_{ij}\cos(R'_{ij}) + 2\sqrt{1 - \boldsymbol{\rho}_{ij}^2/4} \cdot \sin(R'_{ij}) - \boldsymbol{\rho}_{ij} \right| \\
&\leq \left| \boldsymbol{\rho}_{ij}\left(1 - \cos(R'_{ij})\right) \right| + 2\left| \sqrt{1 - \boldsymbol{\rho}_{ij}^2/4} \cdot \sin(R'_{ij}) \right| \\
&\leq \left[1 - \cos(R'_{ij})\right] + 2\left|\sin(R'_{ij})\right|.
\end{aligned}
$$

Note that $\gamma_{ij} = \epsilon_i + \epsilon_j - \epsilon_i\epsilon_j \leq 2\epsilon$, so

$$
\begin{aligned}
|R'_{ij}| &\leq \frac{\pi}{6}\left(48\gamma_{ij} + 129\gamma_{ij}^2 + 88\gamma_{ij}^3 + \frac{12}{n+1}\right) \\
&\leq \frac{\pi}{6}\left(48 \cdot 2\epsilon + 129(2\epsilon)^2 + 88(2\epsilon)^3 + \frac{12}{n+1}\right) \\
&\leq 16\pi\epsilon + 86\pi\epsilon^2 + 118\pi\epsilon^3 + \frac{2\pi}{n+1}.
\end{aligned}
$$

In particular, this bound is less than 1 when $\epsilon \leq 0.01$ and $n \geq 15$. Then using the fact that $|\sin(x) - x| \leq \frac{|x|^3}{3!}$ and $|\cos(x) - 1| \leq \frac{x^2}{2!}$ for $|x| \leq 1$, we conclude that

$$
\begin{aligned}
\max_{1\leq i,j\leq p} \left| 2\sin\left(\frac{\pi}{6} E(\boldsymbol{r}_{ij}^S)\right) - \boldsymbol{\rho}_{ij} \right| &\leq \max_{1\leq i,j\leq p} \left[ 2|R'_{ij}| + \frac{(R'_{ij})^2}{2} + \frac{|R'_{ij}|^3}{3} \right] \\
&\leq 3 \max_{1\leq i,j\leq p} |R'_{ij}| \\
&\leq 48\pi\epsilon + 258\pi\epsilon^2 + 354\pi\epsilon^3 + \frac{6\pi}{n+1}
\end{aligned}
$$

$$\leq 51\pi\epsilon + \frac{3C}{2}\sqrt{\frac{\log p}{n}},$$

where the final inequality uses the assumption $n \geq \frac{16\pi^2}{C^2 \log p}$ once more. Combining this bound with inequality (B.17) implies the desired result.

### B.2.4 Proof of Lemma 16

We have

$$\|\mathbf{I} - \hat{\mathbf{\Sigma}}\mathbf{\Omega}^*\|_\infty = \|(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{\Omega}^*\|_\infty \leq \|\mathbf{\Omega}^*\|_{L_1}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \lambda, \tag{B.18}$$

the first inequality is due to $\|\mathbf{AB}\|_\infty \leq \|\mathbf{A}\|_\infty\|\mathbf{B}\|_{L_1}$, and the second inequality follows by assumption. Then

$$\|\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty \leq \|\hat{\mathbf{\Sigma}}\hat{\mathbf{\Omega}} - \mathbf{I}\|_\infty + \|\mathbf{I} - \hat{\mathbf{\Sigma}}\mathbf{\Omega}^*\|_\infty \leq 2\lambda.$$

For $1 \leq i \leq p$, let $\mathbf{e}_i$ be the canonical vector with 1 in the $i^{th}$ coordinate and 0 in all other coordinates, and let $\hat{\boldsymbol{\beta}}_i$ be the solution of the following convex optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{\beta}\|_1 \qquad \text{subject to} \qquad \|\hat{\mathbf{\Sigma}}\boldsymbol{\beta} - \mathbf{e}_i\|_\infty \leq \lambda.$$

Note that $\hat{\mathbf{\Omega}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p)$ (cf. Lemma 1 in Cai et al. (2011)). It follows that $\|\hat{\boldsymbol{\beta}}_i\|_1 \leq \|\mathbf{\Omega}^*\|_{L_1}$, for $1 \leq i \leq p$, so $\|\hat{\mathbf{\Omega}}\|_{L_1} \leq \|\mathbf{\Omega}^*\|_{L_1}$. Hence,

$$\begin{aligned}
\|\mathbf{\Sigma}^*(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty &\leq \|\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty + \|(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty \\
&\leq 2\lambda + \|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{L_1}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \\
&\leq 2\lambda + \|\hat{\mathbf{\Omega}}\|_{L_1}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty + \|\mathbf{\Omega}^*\|_{L_1}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \\
&\leq 4\lambda.
\end{aligned}$$

Finally,

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty = \|\mathbf{\Omega}^*\mathbf{\Sigma}^*(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty \leq \|\mathbf{\Omega}^*\|_{L_1}\|\mathbf{\Sigma}^*(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*)\|_\infty \leq 4\|\mathbf{\Omega}^*\|_{L_1}\lambda.$$

## B.3 Lemmas for MAD concentration

In this section, we prove several lemmas that are needed in deriving consistency of the MAD estimator. We begin with some results concerning the concentration of sample medians from an arbitrary distribution. A version of Lemmas 19 and 20 is also contained in Serfling & Mazumder (2009).

**Lemma 18.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F$, and let $\hat{m}$ be the sample median. If $\hat{m} < c$, then $|\{X_i : X_i \leq c\}| \geq \frac{n}{2}$. If $\hat{m} > c$, then $|\{X_i : X_i \leq c\}| \leq \frac{n}{2}$.*

*Proof.* This result follows easily from the definition of the sample median. □

**Lemma 19.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution $F$. Let $m$ be the population median and let $\hat{m}$ be the sample median. Then*

$$P\left(|\hat{m} - m| > \frac{t}{2}\right) \leq 2\exp(-2nb^2(t)),$$

*where $b(t) = \min\left\{F(m + \frac{t}{2}) - \frac{1}{2}, \frac{1}{2} - F(m - \frac{t}{2})\right\}$.*

*Proof.* By Lemma 18,

$$\begin{aligned} P\left(\hat{m} > m + \frac{t}{2}\right) &\leq P\left(\left|\left\{X_i : X_i \leq m + \frac{t}{2}\right\}\right| \leq \frac{n}{2}\right) \\ &= P\left(\sum_{i=1}^{n} \mathbb{1}\left\{X_i \leq m + \frac{t}{2}\right\} \leq \frac{n}{2}\right) \\ &= P\left(\sum_{i=1}^{n}(Y_i - EY_i) \leq \frac{n}{2} - np_1\right) \end{aligned}$$

$$= \exp\left[ -2n\Big(p_1 - \frac{1}{2}\Big)^2\right], \tag{B.19}$$

where $Y_i = \mathbb{1}\left\{X_i \leq m + \frac{t}{2}\right\}$ and $p_1 = F(m + \frac{t}{2})$, and the last inequality follows from Hoeffding's inequality. Similarly, we have

$$\begin{aligned} P\Big(\hat{m} < m - \frac{t}{2}\Big) &\leq P\Big(\Big|\Big\{X_i : X_i \leq m - \frac{t}{2}\Big\}\Big| \geq \frac{n}{2}\Big) \\ &= P\Big(\sum_{i=1}^{n} \mathbb{1}(X_i \leq m - \frac{t}{2}) \geq \frac{n}{2}\Big) \\ &= P\Big(\sum_{i=1}^{n} (Z_i - EZ_i) \geq \frac{n}{2} - np_2\Big) \\ &\leq \exp\left[ -2n\Big(p_2 - \frac{1}{2}\Big)^2\right], \end{aligned} \tag{B.20}$$

where $Z_i = \mathbb{1}\left\{X_i \leq m - \frac{t}{2}\right\}$ and $p_2 = F(m - \frac{t}{2})$. Combining expressions (B.19) and (B.20), we then obtain

$$P\Big(|\hat{m} - m| > \frac{t}{2}\Big) \leq \exp\left[ -2n\Big(p_1 - \frac{1}{2}\Big)^2\right] + \exp\left[ -2n\Big(p_2 - \frac{1}{2}\Big)^2\right] \leq 2\exp(-2nb^2(t)).$$

□

**Lemma 20.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F$. Let $m$ and $d$ denote the population median and MAD, respectively, and let $\hat{m}$ and $\hat{d}$ denote the sample median and MAD. Let $G$ be the distribution of $|X_i - m|$. Then*

$$P(|\hat{d} - d| > t) \leq 6\exp(-2na^2(t)), \tag{B.21}$$

*where*

$$a(t) = \min\left\{ F\left(m + \frac{t}{2}\right) - \frac{1}{2},\ \frac{1}{2} - F\left(m - \frac{t}{2}\right),\ G\left(d + \frac{t}{2}\right) - \frac{1}{2},\ \frac{1}{2} - G\left(d - \frac{t}{2}\right) \right\}.$$

*Proof.* Let $W_i = |X_i - \hat{m}|$. By the definition of the sample MAD, Lemma 18 gives

$$\begin{aligned}
P(\hat{d} > d + t) &\leq P\bigg(|\{W_i : W_i \leq d + t\}| \leq \frac{n}{2}\bigg) \\
&= P\bigg(|\{X_i : |X_i - \hat{m}| \leq d + t\}| \leq \frac{n}{2}\bigg) \\
&\leq P\bigg(|\{X_i : |X_i - \hat{m}| \leq d + t\}| \leq \frac{n}{2}, \text{ and } |\hat{m} - m| \leq \frac{t}{2}\bigg) \\
&\quad + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg) \\
&\leq P\bigg(\bigg|\bigg\{X_i : |X_i - m| \leq d + \frac{t}{2}\bigg\}\bigg| \leq \frac{n}{2}\bigg) + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg) \\
&= P\bigg(\sum_{i=1}^{n} \mathbb{1}\bigg\{|X_i - m| \leq d + \frac{t}{2}\bigg\} \leq \frac{n}{2}\bigg) + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg) \\
&= P\bigg(\sum_{i=1}^{n} (Y_i - EY_i) \leq \frac{n}{2} - np_3\bigg) + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg),
\end{aligned}$$

where $Y_i = \mathbb{1}\left\{|X_i - m| \leq d + \frac{t}{2}\right\}$ and $p_3 = G(d + \frac{t}{2})$. Then by Hoeffding's inequality and Lemma 19, the last quantity is bounded by

$$\exp\bigg[-2n\Big(p_3 - \frac{1}{2}\Big)^2\bigg] + 2\exp(-2nb^2(t)). \tag{B.22}$$

Similarly,

$$\begin{aligned}
P(\hat{d} < d - t) &\leq P\bigg(|\{W_i : W_i \leq d - t\}| \geq \frac{n}{2}\bigg) \\
&= P\bigg(|\{X_i : |X_i - \hat{m}| \leq d - t\}| \geq \frac{n}{2}\bigg) \\
&\leq P\bigg(|\{X_i : |X_i - \hat{m}| \leq d - t\}| \geq \frac{n}{2}, \text{ and } |\hat{m} - m| \leq \frac{t}{2}\bigg) \\
&\quad + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg) \\
&\leq P\bigg(\bigg|\bigg\{X_i : |X_i - m| \leq d - \frac{t}{2}\bigg\}\bigg| \geq \frac{n}{2}\bigg) + P\bigg(|\hat{m} - m| > \frac{t}{2}\bigg)
\end{aligned}$$

$$= P\Big(\sum_{i=1}^{n} \mathbb{1}\left\{|X_i - m| \le d - \frac{t}{2}\right\} \ge \frac{n}{2}\Big) + P\Big(|\hat{m} - m| > \frac{t}{2}\Big)$$
$$= P\Big(\sum_{i=1}^{n} (Z_i - EZ_i) \ge \frac{n}{2} - np_4\Big) + P\Big(|\hat{m} - m| > \frac{t}{2}\Big),$$

where $Z_i = \mathbb{1}\left\{|X_i - m| \le d - \frac{t}{2}\right\}$ and $p_4 = G(d - \frac{t}{2})$. By Hoeffding's inequality and Lemma 19, the last quantity is upper-bounded by

$$\exp\Big[-2n\Big(p_4 - \frac{1}{2}\Big)^2\Big] + 2\exp(-2nb^2(t)). \tag{B.23}$$

Combining expressions (B.22) and (B.23) then yields

$$P(|\hat{d} - d| > t) \le 4\exp(-2nb^2(t)) + \exp\Big[-2n\Big(p_3 - \frac{1}{2}\Big)^2\Big] + \exp\Big[-2n\Big(p_4 - \frac{1}{2}\Big)^2\Big]$$
$$\le 6\exp(-2na^2(t)).$$

□

Next, we prove two population-level lemmas for the $\epsilon$-contamination model. As remarked in the introduction, we use the notation $F^{-1}(c) = \inf\{x : F(x) \ge c\}$, which is defined even if the cdf $F$ is not surjective on the interval $[0, 1]$. Note that Lemmas 21 and 22 do not impose any conditions on the contaminating distribution $H$.

**Lemma 21.** *Let $F = (1-\epsilon)\Phi_{\mu,\sigma} + \epsilon H$, where $\Phi_{\mu,\sigma}$ denotes the $N(\mu, \sigma^2)$ distribution and $H$ is an arbitrary distribution. Let $\Phi := \Phi_{0,1}$ be the standard normal cdf and suppose that $0 \le \epsilon < 1$. Then*

$$\mu + \Phi^{-1}\Big(\frac{c-\epsilon}{1-\epsilon}\Big)\sigma = \Phi^{-1}_{\mu,\sigma}\left(\frac{c-\epsilon}{1-\epsilon}\right) \le F^{-1}(c) \le \Phi^{-1}_{\mu,\sigma}\left(\frac{c}{1-\epsilon}\right) = \mu + \Phi^{-1}\Big(\frac{c}{1-\epsilon}\Big)\sigma. \tag{B.24}$$

*Proof.* Let $F = (1-\epsilon)\Phi_{\mu,\sigma} + \epsilon H$. Then

$$\begin{aligned} F\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c}{1-\epsilon}\Big)\Big) &= (1-\epsilon)\Phi_{\mu,\sigma}\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c}{1-\epsilon}\Big)\Big) + \epsilon H\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c}{1-\epsilon}\Big)\Big) \\ &\geq (1-\epsilon)\cdot\frac{c}{1-\epsilon} = c, \end{aligned} \tag{B.25}$$

where by a slight abuse of notation, we use $F$ and $H$ to denote the cdfs of the corresponding distributions. In addition,

$$\begin{aligned} 1 - F\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c-\epsilon}{1-\epsilon}\Big)\Big) &= (1-\epsilon)\Big[1 - \Phi_{\mu,\sigma}\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c-\epsilon}{1-\epsilon}\Big)\Big)\Big] + \epsilon\Big[1 - H\Big(\Phi^{-1}_{\mu,\sigma}\Big(\frac{c-\epsilon}{1-\epsilon}\Big)\Big)\Big] \\ &\geq (1-\epsilon)\left(1 - \frac{c-\epsilon}{1-\epsilon}\right) = 1 - c. \end{aligned} \tag{B.26}$$

Combining equations (B.25) and (B.26), and using the facts that $F$ is monotonically increasing, we then obtain the desired bound (B.24). Note that the outer equalities hold since $\Phi^{-1}_{\mu,\sigma}(x) = \mu + \Phi^{-1}(x)\sigma$. □

**Lemma 22.** *Let $F = (1-\epsilon)\Phi_{\mu,\sigma} + \epsilon H$, where $\Phi_{\mu,\sigma}$ denotes the $N(\mu, \sigma^2)$ distribution and $H$ is an arbitrary distribution. Suppose $0 \leq \epsilon \leq \frac{1}{16}$. Let $d(F)$ and $d(\Phi_{\mu,\sigma})$ denote the population MADs corresponding to $F$ and $\Phi_{\mu,\sigma}$, respectively. Then*

$$|d(F) - d(\Phi_{\mu,\sigma})| \leq 4.8\sigma\epsilon.$$

*Proof.* By an abuse of notation, we also use $F$ to denote the cdf of the contaminated distribution. Then $F^{-1}$ is the quantile function. Note in particular that the following statements hold, where $X \sim F$, as an easy consequence of the definition of $F^{-1}$:

(i) $d(F) \leq a$ if $P(|X - F^{-1}(0.5)| \leq a) \geq 0.5$,

(ii) $d(F) > a$ if $P(|X - F^{-1}(0.5)| \leq a) < 0.5$.

Furthermore, we may write

$$\begin{aligned} P(|X - F^{-1}(0.5)| \leq a) &\geq (1-\epsilon) \cdot P(|Z - F^{-1}(0.5)| \leq a) \\ &= (1-\epsilon) \left\{ \Phi_{\mu,\sigma} \left( F^{-1}(0.5) + a \right) - \Phi_{\mu,\sigma} \left( F^{-1}(0.5) - a \right) \right\}, \end{aligned}$$

where $Z \sim N(\mu, \sigma^2)$. By Lemma 21, the last expression is further lower-bounded by

$$(1-\epsilon) \left\{ \Phi_{\mu,\sigma} \left( \Phi_{\mu,\sigma}^{-1} \left( \frac{0.5-\epsilon}{1-\epsilon} \right) + a \right) - \Phi_{\mu,\sigma} \left( \Phi_{\mu,\sigma}^{-1} \left( \frac{0.5}{1-\epsilon} \right) - a \right) \right\}.$$

We will take

$$a = \Phi_{\mu,\sigma}^{-1} \left( \frac{0.75}{1-\epsilon} \right) - \Phi_{\mu,\sigma}^{-1} \left( \frac{0.5-\epsilon}{1-\epsilon} \right) = \Phi_{\mu,\sigma}^{-1} \left( \frac{0.5}{1-\epsilon} \right) - \Phi_{\mu,\sigma}^{-1} \left( \frac{0.25-\epsilon}{1-\epsilon} \right),$$

where the second inequality comes from the fact that $\Phi_{\mu,\sigma}^{-1}(b) = -\Phi_{\mu,\sigma}^{-1}(1-b)$. Then the lower bound becomes

$$(1-\epsilon) \left( \frac{0.75}{1-\epsilon} - \frac{0.25-\epsilon}{1-\epsilon} \right) \geq 0.5.$$

Putting the bounds together, we have

$$P(|X - F^{-1}(0.5)| \leq a) \geq 0.5,$$

so by the implication (i) above, it follows that

$$d(F) \leq \Phi_{\mu,\sigma}^{-1} \left( \frac{0.75}{1-\epsilon} \right) - \Phi_{\mu,\sigma}^{-1} \left( \frac{0.5-\epsilon}{1-\epsilon} \right). \tag{B.27}$$

Similarly, we may derive a lower bound on $d(F)$ by writing

$$P(|X - F^{-1}(0.5)| > a) \geq (1 - \epsilon) \cdot P(|Z - F^{-1}(0.5)| > a),$$

where $Z \sim N(\mu, \sigma^2)$. Furthermore,

$$\begin{aligned} P(|Z - F^{-1}(0.5)| \leq a) &= \Phi_{\mu,\sigma}\left(F^{-1}(0.5) + a\right) - \Phi_{\mu,\sigma}\left(F^{-1}(0.5) - a\right) \\ &\leq \Phi_{\mu,\sigma}\left(\Phi_{\mu,\sigma}^{-1}\left(\frac{0.5}{1-\epsilon}\right) + a\right) - \Phi_{\mu,\sigma}\left(\Phi_{\mu,\sigma}^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right) - a\right), \end{aligned}$$

using Lemma 21. Taking

$$a = \Phi_{\mu,\sigma}^{-1}\left(\frac{0.75 - 2\epsilon}{1 - 2\epsilon}\right) - \Phi_{\mu,\sigma}^{-1}\left(\frac{0.5}{1-\epsilon}\right) = \Phi_{\mu,\sigma}^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right) - \Phi_{\mu,\sigma}^{-1}\left(\frac{0.25}{1-2\epsilon}\right),$$

we then have the bound

$$P(|Z - F^{-1}(0.5)| \leq a) \leq \frac{0.75 - 2\epsilon}{1 - 2\epsilon} - \frac{0.25}{1 - 2\epsilon} = \frac{0.5 - 2\epsilon}{1 - 2\epsilon},$$

implying that

$$P(|X - F^{-1}(0.5)| > a) \geq (1 - \epsilon) \cdot \left(1 - \frac{0.5 - 2\epsilon}{1 - 2\epsilon}\right) > 0.5.$$

It follows that

$$P(|X - F^{-1}(0.5)| \leq a) < 0.5,$$

so by implication (ii) above,

$$d(F) > \Phi_{\mu,\sigma}^{-1}\left(\frac{0.75 - 2\epsilon}{1 - 2\epsilon}\right) - \Phi_{\mu,\sigma}^{-1}\left(\frac{0.5}{1-\epsilon}\right). \tag{B.28}$$

Using the fact that $d(\Phi_{\mu,\sigma}) = \Phi_{\mu,\sigma}^{-1}(0.75)$ and $\Phi_{\mu,\sigma}^{-1}(0.5) = 0$, inequality (B.27)

implies that

$$\begin{aligned} d(F) - d(\Phi_{\mu,\sigma}) &\leq \left\{ \Phi_{\mu,\sigma}^{-1}\left(\frac{0.75}{1-\epsilon}\right) - \Phi_{\mu,\sigma}^{-1}(0.75) \right\} + \left\{ \Phi_{\mu,\sigma}^{-1}(0.5) - \Phi_{\mu,\sigma}^{-1}\left(\frac{0.5-\epsilon}{1-\epsilon}\right) \right\} \\ &\leq 3.6\sigma \left\{ \left(\frac{0.75}{1-\epsilon} - 0.75\right) + \left(0.5 - \frac{0.5-\epsilon}{1-\epsilon}\right) \right\} \\ &= 3.6\sigma \cdot \frac{1.25\epsilon}{1-\epsilon} \\ &\leq 4.8\sigma\epsilon, \end{aligned}$$

where the second inequality comes from Lemma 28 and the observation $\Phi_{\mu,\sigma}^{-1}(x) = \mu + \sigma\Phi_{0,1}^{-1}(x)$, along with the assumption $\epsilon \leq \frac{1}{16}$. Similarly, inequality (B.28) implies that

$$\begin{aligned} d(F) - d(\Phi_{\mu,\sigma}) &\geq \left\{ \Phi_{\mu,\sigma}^{-1}\left(\frac{0.75-2\epsilon}{1-2\epsilon}\right) - \Phi_{\mu,\sigma}^{-1}(0.75) \right\} + \left\{ \Phi_{\mu,\sigma}^{-1}(0.5) - \Phi_{\mu,\sigma}^{-1}\left(\frac{0.5}{1-\epsilon}\right) \right\} \\ &\geq -3.6\sigma \left\{ \left(0.75 - \frac{0.75-2\epsilon}{1-2\epsilon}\right) + \left(\frac{0.5}{1-\epsilon} - 0.5\right) \right\} \\ &= -3.6\sigma \left( \frac{0.5\epsilon}{1-2\epsilon} + \frac{0.5\epsilon}{1-\epsilon} \right) \\ &\geq -3.98\sigma\epsilon. \end{aligned}$$

Thus, we have the desired result. □

We conclude with the main lemma of this section, which establishes the consistency of the sample MAD to its population-level version.

**Lemma 23.** *Let $X_1, \ldots, X_n$ be a random sample from $F = (1-\epsilon)\Phi_{\mu,\sigma} + \epsilon H$, where $0 \leq \epsilon \leq \frac{1}{16}$, $\Phi_{\mu,\sigma}$ denotes the $N(\mu, \sigma^2)$ distribution, and $H$ is an arbitrary distribution. Let $d := d(F)$ be the population MAD corresponding to $F$, and let $\hat{d}$ be the sample MAD. Then for $0 < t < 1$, we have*

$$P(|\hat{d} - d| > t) \leq 6\exp(-2nc^2(\sigma)t^2), \tag{B.29}$$

*where* $c(\sigma) = \frac{15}{64\sqrt{2\pi}\sigma} \exp\left(-\frac{(1.1\sigma+0.5)^2}{2\sigma^2}\right)$.

*Proof.* By Lemma 20, it suffices to show that

$$a(t) \geq c(\sigma)t,$$

for the $\epsilon$-contaminated distribution, with $a(t)$ as defined in the lemma. With an abuse of notation, let $F, \Phi_{\mu,\sigma}$, and $H$ denote the cdfs of the respective distributions. Let

$$G(c) = P(|X_i - m| \leq c),$$

where $m$ denotes the median of the contaminated distribution. Note that by the definition of the median, we have $F(m) \geq \frac{1}{2}$ and $G(d) \geq \frac{1}{2}$. Define

$$\begin{aligned}
b_1 &= F\Big(m + \frac{t}{2}\Big) - \frac{1}{2} \geq F\Big(m + \frac{t}{2}\Big) - F(m), \\
b_2 &= \frac{1}{2} - F\Big(m - \frac{t}{2}\Big) \geq F\left(m - \frac{t}{4}\right) - F\Big(m - \frac{t}{2}\Big), \\
b_3 &= G\Big(d + \frac{t}{2}\Big) - \frac{1}{2} \geq G\Big(d + \frac{t}{2}\Big) - G(d), \qquad \text{and} \\
b_4 &= \frac{1}{2} - G\Big(d - \frac{t}{2}\Big) \geq G\left(d - \frac{t}{4}\right) - G\Big(d - \frac{t}{2}\Big),
\end{aligned}$$

where we have used the fact that $F\left(m - \frac{t}{4}\right) < \frac{1}{2}$ and $G\left(d - \frac{t}{4}\right) < \frac{1}{2}$ in the second and fourth inequalities. Then $a(t) = \min\{b_1, b_2, b_3, b_4\}$.

Note that

$$\begin{aligned}
b_1 &\geq (1-\epsilon)\left(\Phi_{\mu,\sigma}\left(m + \frac{t}{2}\right) - \Phi_{\mu,\sigma}(m)\right) + \epsilon\left(H\left(m + \frac{t}{2}\right) - H(m)\right) \\
&\geq (1-\epsilon)\left(\Phi_{\mu,\sigma}\left(m + \frac{t}{2}\right) - \Phi_{\mu,\sigma}(m)\right).
\end{aligned}$$

Similarly, we can check that

$$
\begin{aligned}
b_2 &\geq (1-\epsilon)\left(\Phi_{\mu,\sigma}\left(m-\frac{t}{4}\right)-\Phi_{\mu,\sigma}\left(m-\frac{t}{2}\right)\right),\\
b_3 &\geq (1-\epsilon)\left(G_\Phi\left(d+\frac{t}{2}\right)-G_\Phi(d)\right), \qquad \text{and}\\
b_4 &\geq (1-\epsilon)\left(G_\Phi\left(d-\frac{t}{4}\right)-G_\Phi\left(d+\frac{t}{2}\right)\right),
\end{aligned}
$$

where $G_\Phi(c) := \Phi_{\mu,\sigma}(m+c)-\Phi_{\mu,\sigma}(m-c)$. By the mean value theorem, we have $c_1, c_2, c_3$, and $c_4$ such that

$$
\begin{aligned}
b_1 &\geq (1-\epsilon)\Phi'_{\mu,\sigma}(c_1)\frac{t}{2}, && m \leq c_1 \leq m+\frac{t}{2},\\
b_2 &\geq (1-\epsilon)\Phi'_{\mu,\sigma}(c_2)\frac{t}{4}, && m-\frac{t}{2} \leq c_2 \leq m-\frac{t}{4},\\
b_3 &\geq (1-\epsilon)G'_\Phi(c_3)\frac{t}{2}\\
&= (1-\epsilon)\left(\Phi'_{\mu,\sigma}(m+c_3)+\Phi'_{\mu,\sigma}(m-c_3)\right)\frac{t}{2}, && d \leq c_3 \leq d+\frac{t}{2},\\
b_4 &\geq (1-\epsilon)G'_\Phi(c_4)\frac{t}{4}\\
&= (1-\epsilon)\left(\Phi'_{\mu,\sigma}(m+c_4)+\Phi'_{\mu,\sigma}(m-c_4)\right)\frac{t}{4}, && d-\frac{t}{2} \leq c_4 \leq d-\frac{t}{4}.
\end{aligned}
$$

Note in particular that

$$
c_1,\ c_2,\ m+c_3,\ m-c_3,\ m+c_4,\ m-c_4 \in \left[m-d-\frac{t}{2},\ m+d+\frac{t}{2}\right].
$$

Let $d(\Phi_{\mu,\sigma}) = \Phi^{-1}(0.75)\sigma$ be the MAD estimator corresponding to $\Phi_{\mu,\sigma}$. By Lemma 21, for $0 \leq \epsilon \leq \frac{1}{16}$, the median $m = F^{-1}(0.5)$ satisfies

$$
\mu+\Phi^{-1}\left(\frac{7}{15}\right)\sigma \leq \mu+\Phi^{-1}\left(\frac{1-2\epsilon}{2-2\epsilon}\right)\sigma \leq m \leq \mu+\Phi^{-1}\left(\frac{1}{2-2\epsilon}\right)\sigma \leq \mu+\Phi^{-1}\left(\frac{8}{15}\right)\sigma.
$$

In addition, Lemma 22 implies that for $0 \leq \epsilon \leq \frac{1}{16}$, we have

$$d \leq d(\Phi_{\mu,\sigma}) + 4.8\sigma\epsilon \leq \Phi^{-1}(0.75)\sigma + 0.3\sigma \leq \sigma.$$

Therefore, for $c \in [m - d - \frac{t}{2}, m + d + \frac{t}{2}]$ and $0 < t < 1$, we have

$$\begin{aligned} c &\geq m - d - \frac{t}{2} \geq \mu + \Phi^{-1}\left(\frac{7}{15}\right)\sigma - \sigma - 0.5 \geq \mu - 1.1\sigma - 0.5, \quad \text{and} \\ c &\leq m + d + \frac{t}{2} \leq \mu + \Phi^{-1}\left(\frac{8}{15}\right)\sigma + \sigma + 0.5 \leq \mu + 1.1\sigma + 0.5. \end{aligned}$$

Hence,

$$\begin{aligned} \min\left\{\Phi'_{\mu,\sigma}(c) : m - d - \frac{t}{2} \leq c \leq m + d + \frac{t}{2}\right\} &\geq \min\{\Phi'_{\mu,\sigma}(c) : |c - \mu| \leq 1.1\sigma + 0.5\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(1.1\sigma + 0.5)^2}{2\sigma^2}\right). \end{aligned}$$

It follows that

$$\begin{aligned} a(t) = \min\{b_1, b_2, b_3, b_4\} &\geq (1 - \epsilon) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(1.1\sigma + 0.5)^2}{2\sigma^2}\right)\frac{t}{4} \\ &\geq \frac{15}{16\sqrt{2\pi}\sigma} \exp\left(-\frac{(1.1\sigma + 0.5)^2}{2\sigma^2}\right)\frac{t}{4} = c(\sigma)t. \end{aligned}$$

□

## B.4 Auxiliary lemmas

We begin with a lemma describing the behavior of the mean of the Kendall's tau statistic under a contaminated normal distribution. Note that the statement of the lemma does not depend on the variances of the uncontaminated marginals, or the contaminating distribution $H$.

**Lemma 24.** *Let $(X_{k1}, X_{k2})$, for $k = 1, \ldots, n$, be a random sample from*

$$F = (1 - \gamma)\Phi_\rho + \gamma H,$$

*where $\Phi_\rho$ is a bivariate normal distribution with correlation $\rho$ and $H$ is an arbitrary bivariate distribution. Let $\rho^K = E_F(r^K)$, where $r^K$ is Kendall's tau statistic. Then*

$$\rho^K = \frac{2}{\pi}\sin^{-1}(\rho) + R,$$

*where $|R| \leq 12\gamma + 17\gamma^2$.*

*Proof.* Define $a(X) = \mathbb{1}(X > 0)$, and let $\overline{\text{sign}}(X) = 2a(X) - 1$. In particular,

$$\text{sign}(X) = \mathbb{1}(X > 0) - \mathbb{1}(X < 0) = 2a(X) - 1 - \mathbb{1}(X = 0) = \overline{\text{sign}}(X) - \mathbb{1}(X = 0).$$

We may rewrite $\rho^K$ as

$$\begin{aligned}
\rho^K &= E\left[\text{sign}(X_{11} - X_{21})\text{sign}(X_{12} - X_{22})\right] \\
&= E\left[\overline{\text{sign}}(X_{11} - X_{21})\overline{\text{sign}}(X_{12} - X_{22})\right] - E[\mathbb{1}(X_{11} = X_{21})\overline{\text{sign}}(X_{12} - X_{22})] \\
&\quad - E\left[\overline{\text{sign}}(X_{11} - X_{21})\mathbb{1}(X_{12} = X_{22})\right] + E\left[\mathbb{1}(X_{11} = X_{21})\mathbb{1}(X_{12} = X_{22})\right] \\
&:= A + B + C + D.
\end{aligned}$$

In particular,

$$\begin{aligned}
|B| &= \left|E[\mathbb{1}(X_{11} = X_{21})\overline{\text{sign}}(X_{12} - X_{22})]\right| \\
&\leq E[\mathbb{1}(X_{11} = X_{21})] = P(X_{11} = X_{21}),
\end{aligned} \tag{B.30}$$

using the fact that $|\overline{\text{sign}}(X)| = 1$. Furthermore, we have

$$P(X_{11} = X_{21}) \leq \gamma^2,$$

since the normal distribution is absolutely continuous, so we can only have $P(X_{11} = X_{21})$ with positive probability when both $X_1$ and $X_2$ are drawn from the contaminating distribution. Similarly,

$$\begin{aligned} |C| &= \left|E[\overline{\text{sign}}(X_{11} - X_{21})\mathbb{1}(X_{12} = X_{22})]\right| \\ &\leq E[\mathbb{1}(X_{12} = X_{22})] = P(X_{12} = X_{22}) \leq \gamma^2. \end{aligned} \tag{B.31}$$

We also have

$$\begin{aligned} |D| &= |E[\mathbb{1}(X_{11} = X_{21})\mathbb{1}(X_{12} = X_{22})]| \\ &\leq \left(E[\mathbb{1}(X_{11} = X_{21})]\right)^{1/2} \left(E[\mathbb{1}(X_{12} = X_{22})]\right)^{1/2} \leq \gamma^2. \end{aligned} \tag{B.32}$$

Turning to the final term, we have

$$\begin{aligned} A &= E\left[\overline{\text{sign}}(X_{11} - X_{21})\overline{\text{sign}}(X_{12} - X_{22})\right] \\ &= E\big[(2a(X_{11} - X_{21}) - 1)(2a(X_{12} - X_{22}) - 1)\big] \\ &= 4E[a(X_{11} - X_{21})a(X_{12} - X_{22})] - 2E[a(X_{11} - X_{21})] - 2E[a(X_{12} - X_{22})] + 1 \\ &= \big(4E[a(X_{11} - X_{21})a(X_{12} - X_{22})] - 1\big) \\ &\qquad + 2\big(1 - E[a(X_{11} - X_{21})] - E[a(X_{12} - X_{22})]\big) \\ &:= A_1 + A_2. \end{aligned}$$

Here, the expectation is with respect to the joint distribution of $(X_{11}, X_{12}, X_{21}, X_{22})$,

with density

$$\begin{aligned} f &= [(1-\gamma)\phi_1 + \gamma h_1][(1-\gamma)\phi_2 + \gamma h_2] \\ &= (1-\gamma)^2\phi_1\phi_2 + \gamma(1-\gamma)\phi_1 h_2 + \gamma(1-\gamma)\phi_2 h_1 + \gamma^2 h_1 h_2. \end{aligned} \tag{B.33}$$

This follows from the fact that the pairs $(X_{11}, X_{12})$ and $(X_{21}, X_{22})$ are independently drawn from the mixture distribution, where $\phi$ is the joint density of $(X_{k1}, X_{k2})$ under $\Phi_\rho$, and $h$ is the joint density of $(X_{k1}, X_{k2})$ under $H$. Now, let $U = X_{11} - X_{21}$ and $V = X_{12} - X_{22}$. Under the product distribution $\phi_1\phi_2$, the distribution of $(U, V)$ is bivariate normal with mean $\mathbf{0}$ and correlation $\rho$. Hence,

$$E_{\phi_1\phi_2}[a(U)] = E_{\phi_1\phi_2}[a(V)] = \frac{1}{2}, \tag{B.34}$$

and by Lemma 27,

$$E_{\phi_1\phi_2}[a(U)a(V)] = \frac{1}{4}\left[1 + \frac{2}{\pi}\sin^{-1}(\rho)\right]. \tag{B.35}$$

Combining equations (B.33) and (B.34), we then have

$$\begin{aligned} &E_f[a(U)] \\ &= (1-\gamma)^2 E_{\phi_1\phi_2}[a(U)] + \gamma(1-\gamma)E_{\phi_1 h_2}[a(U)] + \gamma(1-\gamma)E_{\phi_2 h_1}[a(U)] + \gamma^2 E_{h_1 h_2}[a(U)] \\ &= \frac{1}{2} - \gamma + \frac{1}{2}\gamma^2 + \gamma(1-\gamma)E_{\phi_1 h_2}[a(U)] + \gamma(1-\gamma)E_{\phi_2 h_1}[a(U)] + \gamma^2 E_{h_1 h_2}[a(U)] \\ &= \frac{1}{2} + \{-1 + E_{\phi_1 h_2}[a(U)] + E_{\phi_2 h_1}[a(U)]\}\,\gamma \\ &\quad + \left\{\frac{1}{2} - E_{\phi_1 h_2}[a(U)] - E_{\phi_2 h_1}[a(U)] + E_{h_1 h_2}[a(U)]\right\}\gamma^2. \end{aligned}$$

Noting that $E_{\phi_1 h_2}[a(U)]$, $E_{\phi_2 h_1}[a(U)]$ and $E_{h_1 h_2}[a(U)]$ are between 0 and 1, we have

$$\left|E_f[a(U)] - \frac{1}{2}\right| \leq \gamma + \frac{3}{2}\gamma^2, \qquad \text{and} \qquad \left|E_f[a(V)] - \frac{1}{2}\right| \leq \gamma + \frac{3}{2}\gamma^2.$$

It follows that

$$|A_2| = 2|1 - E_f[a(U)] - E_f[a(V)]| \leq 4\gamma + 6\gamma^2. \tag{B.36}$$

On the other hand, combining equations (B.33) and (B.35), we have

$$\begin{aligned}
A_1 &= 4E_f[a(U)a(V)] - 1 \\
&= 4\Big\{(1-\gamma)^2 E_{\phi_1\phi_2}[a(U)a(V)] + \gamma(1-\gamma)E_{\phi_1 h_2}[a(U)a(V)] \\
&\qquad + \gamma(1-\gamma)E_{\phi_2 h_1}[a(U)a(V)] + \gamma^2 E_{h_1 h_2}[a(U)a(V)]\Big\} - 1 \\
&= (1-\gamma)^2\left[1 + \frac{2}{\pi}\sin^{-1}(\rho)\right] - 1 \\
&\quad + 4\Big\{\gamma(1-\gamma)E_{\phi_1 h_2}[a(U)a(V)] + \gamma(1-\gamma)E_{\phi_2 h_1}[a(U)a(V)] + \gamma^2 E_{h_1 h_2}[a(U)a(V)]\Big\} \\
&= \frac{2}{\pi}\sin^{-1}(\rho) + (-2\gamma + \gamma^2)\left[1 + \frac{2}{\pi}\sin^{-1}(\rho)\right] \\
&\quad + 4\Big\{\gamma(1-\gamma)E_{\phi_1 h_2}[a(U)a(V)] + \gamma(1-\gamma)E_{\phi_2 h_1}[a(U)a(V)] + \gamma^2 E_{h_1 h_2}[a(U)a(V)]\Big\} \\
&= \frac{2}{\pi}\sin^{-1}(\rho) + \left\{-2 - \frac{4}{\pi}\sin^{-1}(\rho) + 4E_{\phi_1 h_2}[a(U)a(V)] + 4E_{\phi_2 h_1}[a(U)a(V)]\right\}\gamma \\
&\quad + \left\{1 + \frac{2}{\pi}\sin^{-1}(\rho) - 4E_{\phi_1 h_2}[a(U)a(V)] - 4E_{\phi_2 h_1}[a(U)a(V)] + 4E_{h_1 h_2}[a(U)a(V)]\right\}\gamma^2.
\end{aligned}$$

Noting that the quantities

$$-2 - \frac{4}{\pi}\sin^{-1}(\rho) + 4E_{\phi_1 h_2}[a(U)a(V)] + 4E_{\phi_2 h_1}[a(U)a(V)]$$

and

$$1 + \frac{2}{\pi}\sin^{-1}(\rho) - 4E_{\phi_1 h_2}[a(U)a(V)] - 4E_{\phi_2 h_1}[a(U)a(V)] + 4E_{h_1 h_2}[a(U)a(V)]$$

are both bounded in magnitude by 8, we obtain

$$\left| A_1 - \frac{2}{\pi} \sin^{-1}(\rho) \right| \leq 8\gamma + 8\gamma^2. \tag{B.37}$$

Combining inequalities (B.30), (B.31), (B.32), (B.36) and (B.37) then gives

$$\begin{aligned} \left| \rho^K - \frac{2}{\pi} \sin^{-1}(\rho) \right| &= \left| A_1 + A_2 + B + C + D - \frac{2}{\pi} \sin^{-1}(\rho) \right| \\ &\leq \left| A_1 - \frac{2}{\pi} \sin^{-1}(\rho) \right| + |A_2| + |B| + |C| + |D| \\ &\leq 12\gamma + 17\gamma^2. \end{aligned}$$

□

The second lemma provides an analogous result to Lemma 24, this time for the Spearman's rho statistic.

**Lemma 25.** *Let $(X_{k1}, X_{k2})$, for $k = 1, \ldots, n$, be a random sample from*

$$F = (1 - \gamma)\Phi_\rho + \gamma H,$$

*where $\Phi_\rho$ is a bivariate normal distribution with correlation $\rho$, and $H$ is an arbitrary bivariate distribution. Let $r^S$ be the Spearman's rho statistic, and suppose the samples $\{X_{ki} : k = 1, \ldots, n\}$ are unique. Then*

$$E_F(r^S) = \frac{6}{\pi} \sin^{-1}\left(\frac{\rho}{2}\right) + R,$$

*where $|R| \leq 48\gamma + 129\gamma^2 + 88\gamma^3 + \frac{12}{n+1}$.*

*Proof.* Let $\rho^K = E_F(r^K)$ be the population version of Kendall's tau correlation. By

Lemma 26, we have

$$\begin{aligned} E_F(r^S) &= \frac{3(n-2)}{n+1} \cdot E\left[\text{sign}(X_{11}-X_{21})\text{sign}(X_{12}-X_{32})\right] + \frac{3}{n+1}\rho^K \\ &= 3E\left[\text{sign}(X_{11}-X_{21})\text{sign}(X_{12}-X_{32})\right] \\ &\quad + \frac{3}{n+1}(\rho^K - 3E\left[\text{sign}(X_{11}-X_{21})\text{sign}(X_{12}-X_{32})\right]). \end{aligned} \tag{B.38}$$

Note that the second term is clearly bounded in magnitude by $\frac{12}{n+1}$. Now define $a(X) = \mathbb{1}(X > 0)$, and let $\overline{\text{sign}}(X) = 2a(X)-1$. Then $\text{sign}(X) = \overline{\text{sign}}(X) - \mathbb{1}(X = 0)$. It follows that

$$\begin{aligned} &E\left[\text{sign}(X_{11}-X_{21})\text{sign}(X_{12}-X_{32})\right] \\ &= E\left[\overline{\text{sign}}(X_{11}-X_{21})\overline{\text{sign}}(X_{12}-X_{32})\right] - E[\mathbb{1}(X_{11}=X_{21})\overline{\text{sign}}(X_{12}-X_{32})] \\ &\quad - E\left[\overline{\text{sign}}(X_{11}-X_{21})\mathbb{1}(X_{12}=X_{32})\right] + E\left[\mathbb{1}(X_{11}=X_{21})\mathbb{1}(X_{12}=X_{32})\right] \\ &:= A + B + C + D. \end{aligned}$$

A similar argument as in the proof of Lemma 24 yields

$$\max\{|B|, |C|, |D|\} \le \gamma^2, \tag{B.39}$$

and

$$\begin{aligned} A &= \big(4E[a(X_{11}-X_{21})a(X_{12}-X_{32})] - 1\big) \\ &\quad + 2\big(1 - E[a(X_{11}-X_{21})] - E[a(X_{12}-X_{32})]\big) \\ &:= A_1 + A_2. \end{aligned}$$

Here, the expectation is with respect to the joint distribution of $(X_{11}, X_{12}, X_{21},$

$X_{22}, X_{31}, X_{32}$), with density

$$\begin{aligned} f &= [(1-\gamma)\phi_1 + \gamma h_1][(1-\gamma)\phi_2 + \gamma h_2][(1-\gamma)\phi_3 + \gamma h_3] \\ &= (1-\gamma)^3\phi_1\phi_2\phi_3 + \gamma(1-\gamma)^2[\phi_1\phi_2 h_3 + \phi_1\phi_3 h_2 + \phi_2\phi_3 h_1] \\ &\quad + \gamma^2(1-\gamma)[\phi_1 h_2 h_3 + \phi_2 h_1 h_3 + \phi_3 h_1 h_2] + \gamma^3 h_1 h_2 h_3. \end{aligned} \tag{B.40}$$

Now let $U = X_{11} - X_{21}$ and $V = X_{12} - X_{32}$. Under the product distribution $\phi_1\phi_2\phi_3$, the distribution of $(U, V)$ is bivariate normal with mean 0 and correlation $\rho/2$. Hence,

$$E_{\phi_1\phi_2\phi_3}[a(U)] = E_{\phi_1\phi_2\phi_3}[a(V)] = \frac{1}{2}, \tag{B.41}$$

and by Lemma 27,

$$E_{\phi_1\phi_2\phi_3}[a(U)a(V)] = \frac{1}{4}\left[1 + \frac{2}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)\right]. \tag{B.42}$$

Combining equations (B.40) and (B.41), and noting that $E[a(U)]$ is between 0 and 1, we then have

$$\begin{aligned} &E_f[a(U)] \\ &= (1-\gamma)^3 E_{\phi_1\phi_2\phi_3}[a(U)] + \gamma(1-\gamma)^2\big\{E_{\phi_1\phi_2 h_3}[a(U)] + E_{\phi_1\phi_3 h_2}[a(U)] + E_{\phi_2\phi_3 h_1}[a(U)]\big\} \\ &\quad + \gamma^2(1-\gamma)\big\{E_{\phi_1 h_2 h_3}[a(U)] + E_{\phi_2 h_1 h_3}[a(U)] + E_{\phi_3 h_1 h_2}[a(U)]\big\} + \gamma^3 E_{h_1 h_2 h_3}[a(U)] \\ &= \frac{1}{2} - \frac{3}{2}\gamma + \frac{3}{2}\gamma^2 - \frac{1}{2}\gamma^3 + \gamma(1-\gamma)^2\big\{E_{\phi_1\phi_2 h_3}[a(U)] + E_{\phi_1\phi_3 h_2}[a(U)] + E_{\phi_2\phi_3 h_1}[a(U)]\big\} \\ &\quad + \gamma^2(1-\gamma)\big\{E_{\phi_1 h_2 h_3}[a(U)] + E_{\phi_2 h_1 h_3}[a(U)] + E_{\phi_3 h_1 h_2}[a(U)]\big\} + \gamma^3 E_{h_1 h_2 h_3}[a(U)] \\ &= \frac{1}{2} + c\gamma + d\gamma^2 + e\gamma^3, \end{aligned}$$

where $|c| \leq \frac{3}{2}, |d| \leq \frac{9}{2}$, and $|e| \leq \frac{7}{2}$. It follows that

$$\left|E_f[a(U)] - \frac{1}{2}\right| \leq \frac{3}{2}\gamma + \frac{9}{2}\gamma^2 + \frac{7}{2}\gamma^3, \qquad \text{and} \qquad \left|E_f[a(V)] - \frac{1}{2}\right| \leq \frac{3}{2}\gamma + \frac{9}{2}\gamma^2 + \frac{7}{2}\gamma^3,$$

so

$$|A_2| = 2|1 - E_f[a(U)] - E_f[a(V)]| \leq 6\gamma + 18\gamma^2 + 14\gamma^3. \tag{B.43}$$

Furthermore, combining equations (B.40) and (B.42), we have

$$\begin{aligned}
A_1 &= 4E_f[a(U)a(V)] - 1 \\
&= 4\Big\{ (1-\gamma)^3 E_{\phi_1\phi_2\phi_3}[a(U)a(V)] \\
&\quad + \gamma(1-\gamma)^2\big\{E_{\phi_1\phi_2 h_3}[a(U)a(V)] + E_{\phi_1\phi_3 h_2}[a(U)a(V)] + E_{\phi_2\phi_3 h_1}[a(U)a(V)]\big\} \\
&\quad + \gamma^2(1-\gamma)\big\{E_{\phi_1 h_2 h_3}[a(U)a(V)] + E_{\phi_2 h_1 h_3}[a(U)a(V)] + E_{\phi_3 h_1 h_2}[a(U)a(V)]\big\} \\
&\quad + \gamma^3 E_{h_1 h_2 h_3}[a(U)a(V)]\Big\} - 1 \\
&= (1-\gamma)^3\left[1 + \frac{2}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)\right] - 1 \\
&\quad + 4\Big\{\gamma(1-\gamma)^2\big\{E_{\phi_1\phi_2 h_3}[a(U)a(V)] + E_{\phi_1\phi_3 h_2}[a(U)a(V)] + E_{\phi_2\phi_3 h_1}[a(U)a(V)]\big\} \\
&\quad + \gamma^2(1-\gamma)\big\{E_{\phi_1 h_2 h_3}[a(U)a(V)] + E_{\phi_2 h_1 h_3}[a(U)a(V)] + E_{\phi_3 h_1 h_2}[a(U)a(V)]\big\} \\
&\quad + \gamma^3 E_{h_1 h_2 h_3}[a(U)a(V)]\Big\} \\
&= \frac{2}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right) + (-3\gamma + 3\gamma^2 - \gamma^3)\left[1 + \frac{2}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)\right] \\
&\quad + 4\Big\{\gamma(1-\gamma)^2\big\{E_{\phi_1\phi_2 h_3}[a(U)a(V)] + E_{\phi_1\phi_3 h_2}[a(U)a(V)] + E_{\phi_2\phi_3 h_1}[a(U)a(V)]\big\} \\
&\quad + \gamma^2(1-\gamma)\big\{E_{\phi_1 h_2 h_3}[a(U)a(V)] + E_{\phi_2 h_1 h_3}[a(U)a(V)] + E_{\phi_3 h_1 h_2}[a(U)a(V)]\big\} \\
&\quad + \gamma^3 E_{h_1 h_2 h_3}[a(U)a(V)]\Big\} \\
&= \frac{2}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right) + c'\gamma + d'\gamma^2 + e'\gamma^3,
\end{aligned}$$

where $|c'| \leq 10, |d'| \leq 22$, and $|e'| \leq \frac{46}{3}$. Hence, we obtain

$$\left| A_1 - \frac{2}{\pi} \sin^{-1}\left(\frac{\rho}{2}\right) \right| \leq 10\gamma + 22\gamma^2 + \frac{46}{3}\gamma^3. \tag{B.44}$$

Combining inequalities (B.38), (B.39), (B.43) and (B.44), we then obtain

$$\begin{aligned}
\left| E_F(r^S) - \frac{6}{\pi} \sin^{-1}\left(\frac{\rho}{2}\right) \right| &\leq 3 \left| A_1 + A_2 + B + C + D - \frac{2}{\pi} \sin^{-1}\left(\frac{\rho}{2}\right) \right| + \frac{12}{n+1} \\
&\leq 3 \left\{ \left| A_1 - \frac{2}{\pi} \sin^{-1}\left(\frac{\rho}{2}\right) \right| + |A_2| + |B| + |C| + |D| \right\} + \frac{12}{n+1} \\
&\leq 48\gamma + 129\gamma^2 + 88\gamma^3 + \frac{12}{n+1}.
\end{aligned}$$

□

The following lemma comes from Hoeffding (1948):

**Lemma 26.** *Suppose the samples* $\{X_{ki} : k = 1, \ldots, n\}$ *are unique, for* $i = 1, 2$. *The Spearman's rho correlation can be decomposed as*

$$r^S = \frac{n-2}{n+1} U + \frac{3}{n+1} r^K,$$

*where* $\boldsymbol{r}^K$ *is the Kendall's tau correlation, and* $U$ *is a* $U$*-statistic of order 3 with corresponding symmetric kernel*

$$\psi_U(X_1, X_2, X_3) = \frac{1}{3!} \sum_{(i_1,i_2,i_3) \in perm(1,2,3)} 3 \cdot \text{sign}(X_{i_1 1} - X_{i_2 1}) \, \text{sign}(X_{i_1 2} - X_{i_3 2}),$$

*and the summation is taken over all possible permutations of the three arguments.*

The proof of the following lemma is adapted from an argument in Croux & Dehon (2010).

**Lemma 27.** *Suppose $(X, Y)$ follows a bivariate normal distribution with mean 0 and correlation $\rho$. Then*

$$E[a(X)a(Y)] = P(X > 0, Y > 0) = \frac{1}{4}\left[1 + \frac{2}{\pi}\sin^{-1}(\rho)\right].$$

*Proof.* Recall that we may write

$$Y = \rho X + \sqrt{1 - \rho^2}Z,$$

where $(X, Z) \sim N(0, I_2)$. Furthermore, we have the polar coordinate representation

$$(X, Z) = (R\cos\theta, R\sin\theta),$$

where $\theta \sim \text{Uniform}(-\pi, \pi]$, and $R$ follows a Rayleigh distribution. Then

$$Y = R\left(\rho\cos(\theta) + \sqrt{1 - \rho^2}\sin(\theta)\right),$$

which has the convenient representation $Y = R\sin(\alpha + \theta)$, where $\alpha = \sin^{-1}(\rho)$. It follows that

$$\begin{aligned} P(X > 0, Y > 0) &= P(\cos\theta > 0, \sin(\alpha + \theta) > 0) \\ &= P\left(\theta \in \left[-\alpha, \frac{\pi}{2}\right]\right) = \frac{\frac{\pi}{2} + \alpha}{2\pi} = \frac{1}{4}\left[1 + \frac{2}{\pi}\sin^{-1}(\rho)\right]. \end{aligned}$$

□

Finally, we have a simple lemma concerning the Lipschitz behavior of the normal quantile function:

**Lemma 28.** *The standard normal quantile function $\Phi^{-1} : [0, 1] \to \mathbb{R}$, when restricted*

*to the domain* $[0.2, 0.8]$*, is Lipschitz continuous with Lipschitz constant* $3.6$*; i.e.,*

$$|\Phi^{-1}(a) - \Phi^{-1}(b)| \leq 3.6|a - b|, \qquad \forall a, b \in [0.2, 0.8].$$

*Proof.* It suffices to check that $|\frac{d}{dy}\Phi^{-1}(y)| \leq 3.6$, for $y \in [0.2, 0.8]$. Since $[\Phi^{-1}]'(\Phi(x)) \cdot \Phi'(x) = \frac{d}{dx}\Phi^{-1}(\Phi(x)) = \frac{d}{dx}x = 1$, we have

$$[\Phi^{-1}]'(\Phi(x)) = \frac{1}{\Phi'(x)}, \qquad \forall x \in \mathbb{R}.$$

For $y = \Phi(x) \in [0.2, 0.8]$, we have $x \in [-0.8416, 0.8416]$, and for such $x$'s,

$$[\Phi^{-1}]'(\Phi(x)) = \frac{1}{\Phi'(x)} = \sqrt{2\pi}\exp\left(\frac{1}{2}x^2\right) \leq \sqrt{2\pi}\exp\left(\frac{1}{2}\cdot 0.8416^2\right) \leq 3.6.$$

This concludes the proof. □

# C

# Supplement for Chapter 4

This chapter contains supporting materials for Chapter 4. Section C.1 presents the estimation results for the quadratic functional $Q(\mu, \theta)$ when $\mu$ and $\theta$ have different signal strengths, whereas Section C.2 presents the proofs of main theorems given in Section 4.2.

## C.1 Optimal Estimation of $Q(\mu, \theta)$ with Different Signal Strengths

We consider in Chapter 4 the estimation of $Q(\mu, \theta) = \frac{1}{n}\sum_{i=1}^{n} \mu_i^2\theta_i^2$ over the parameter space (4.4) where $j_n = k_n = n^\beta$ and $r_n = s_n = n^b$, with $0 < \epsilon \leq \beta < \frac{1}{2}$ and $b \in \mathbb{R}$. In this section, we present the estimation result for $Q(\mu, \theta)$ with $j_n = k_n = n^\beta$ but allow $r_n$ and $s_n$ to differ. Specifically, we consider the following parameter space

$$\Omega(\beta, \epsilon, a, b) = \{(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}^n : \|\mu\|_0 \leq k_n, \|\mu\|_\infty \leq r_n, \|\theta\|_0 \leq k_n, \|\theta\|_\infty \leq s_n, \\ \|\mu \star \theta\|_0 \leq q_n\}, \tag{C.1}$$

where $k_n = n^\beta, q_n = n^\epsilon$ with $0 < \epsilon \leq \beta < \frac{1}{2}$, and $r_n = n^a, s_n = n^b$ with $a, b \in \mathbb{R}$.

Similar as before, the estimation problem can be divided into three regimes: the

sparse regime ($0 < \epsilon < \frac{\beta}{2}$), the moderately dense regime ($\frac{\beta}{2} \leq \epsilon \leq \frac{3\beta}{4}$), and the strongly dense regime ($\frac{3\beta}{4} < \epsilon \leq \beta$). When $\mu$ and $\theta$ have different signal strengths, the minimax rates of convergence for $Q(\mu, \theta)$ exhibit more elaborate phase transitions, though they still bear the familiar form

$$R^*(n, \Omega(\beta, \epsilon, a, b)) := \inf_{\widehat{Q}} \sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,a,b)} E_{(\mu,\theta)}(\widehat{Q} - Q(\mu, \theta))^2 \asymp \gamma_n(\beta, \epsilon, a, b),$$

where $\gamma_n(\beta, \epsilon, a, b)$ is a function of $n$ indexed by $\beta, \epsilon, a$, and $b$. For readability, we summarize the corresponding $\gamma_n(\beta, \epsilon, a, b)$ in Table C.1 (sparse regime), Table C.2 (moderately dense regime), and Table C.3 (strongly dense regime), respectively. The minimax rates of convergence are attained by the same estimators as before over the respective regimes, as stated in Theorem 16 and Theorem 17 given below.

Although we do not present the result here due to its lengthiness, estimation of $Q(\mu, \theta)$ for the case where no equality constraint is imposed on either sparsity or signal strength of $\mu$ and $\theta$ can be analyzed analogously provided that the magnitude of the simultaneous sparsity $\epsilon$ is compared to $\alpha$ if $a \geq b$, and to $\beta$ if $b \geq a$, for the characterization of the sparse and dense regimes.

**Theorem 16** (Sparse Regime)**.** *Let* $0 < \epsilon < \frac{\beta}{2}$ *and* $0 < \beta < \frac{1}{2}$*. Then* $\widehat{Q}_2$ *defined in* (4.15) *with* $\tau_n = \log n$ *attains the minimax rate of convergence over* $\Omega(\beta, \epsilon, a, b)$ *for* $(a, b) \in \{(a, b) : a \wedge b > 0\}$*. On the other hand,* $\widehat{Q}_0 = 0$ *attains the minimax rate of convergence over* $\Omega(\beta, \epsilon, a, b)$ *for* $(a, b) \in \{(a, b) : a \wedge b \leq 0\}$*.*

**Theorem 17** (Dense Regime)**.** *Let* $\frac{\beta}{2} \leq \epsilon \leq \beta$ *and* $0 < \beta < \frac{1}{2}$*. Then* $\widehat{Q}_4$ *defined in* (4.21) *with* $\tau_n = 4 \log n$ *attains the minimax rate of convergence over* $\Omega(\beta, \epsilon, a, b)$ *for* $(a, b) \in \{(a, b) : a \vee b > 0$ *and* $a \wedge b > \frac{\beta-2\epsilon}{4}\}$*. On the other hand,* $\widehat{Q}_0 = 0$ *attains the minimax rate of convergence over* $\Omega(\beta, \epsilon, a, b)$ *for* $(a, b) \in \{(a, b) : a \vee b \leq 0$ *or* $a \wedge b \leq \frac{\beta-2\epsilon}{4}\}$*.*

The shaded regions in the three tables represent the region where $\widehat{Q}_0$ attains the minimax rate of convergence. Thus, $\{(a,b) : a \wedge b \leq 0\}$ is shaded in Table C.1, while $\{(a,b) : a \vee b \leq 0 \text{ or } a \wedge b \leq \frac{\beta-2\epsilon}{4}\}$ is shaded in Tables C.2 and C.3.

Note that the estimation result for the dense regime turns out to be interesting (and more inspiring) when $r_n$ and $s_n$ can differ. It seems that estimation is desirable whenever the signal strengths of both sequences barely exceed some small threshold ($a \wedge b > \frac{\beta-2\epsilon}{4}$, but $\beta - 2\epsilon \leq 0$ in this case) and at least one sequence has sufficiently strong signal ($a \vee b > 0$). This is in contrast to the sparse regime where estimation is desirable only when the signal strength of both sequences are sufficiently strong ($a \wedge b > 0$). The intuitive explanation is that in the dense regime, knowing that $\mu_i \neq 0$ (because of large $X_i^2$) most often suggests that $\theta_i \neq 0$ too (even if $Y_i^2$ is small), and vice versa, so we cannot afford to estimate $\mu_i^2\theta_i^2$ by 0 with this additional information. On the contrary, in the sparse regime, knowing that $\mu_i \neq 0$ does not entail much about whether $\theta_i \neq 0$ due to the sparseness of simultaneous nonzero coordinates. Therefore it is better to estimate $\mu_i^2\theta_i^2$ by 0 unless both $X_i^2$ and $Y_i^2$ are large.

In fact, the minimax rates of convergence for the sparse regime are relatively simple to describe, when $r_n$ is not necessarily equal to $s_n$:

$$\gamma_n(\beta,\epsilon,a,b) = \begin{cases} n^{2\epsilon+4a+4b-2} & \text{if } a \wedge b \leq 0, \\ n^{2\epsilon+4a\vee b-2}(\log n)^2 & \text{if } 0 < a \wedge b \leq \frac{\epsilon}{2}, \\ n^{\epsilon+4a\vee b+2a\wedge b-2} & \text{if } a \wedge b > \frac{\epsilon}{2}. \end{cases}$$

Unfortunately, we do not have such an easy representation for the minimax rates of convergence in the dense regime. Nonetheless, due to the two-dimensional nature of the estimation problem, we find tables useful not only in presenting the minimax rates of convergence but also in illustrating the regions with weak signals (i.e., the shaded regions).

| | $b \le 0$ | $0 < b \le \frac{\epsilon}{2}$ | $b > \frac{\epsilon}{2}$ |
|---|---|---|---|
| $a \le 0$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ |
| $0 < a \le \frac{\epsilon}{2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a\vee b-2}(\log n)^2$ | $n^{2\epsilon+4b-2}(\log n)^2$ |
| $a > \frac{\epsilon}{2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a-2}(\log n)^2$ | $n^{\epsilon+4a\vee b+2a\wedge b-2}$ |

Table C.1: Minimax rates of convergence in the sparse regime: $0 < \epsilon < \frac{\beta}{2}$.

| | $b \le \frac{\beta-2\epsilon}{4}$ | $\frac{\beta-2\epsilon}{4} < b \le 0$ | $0 < b \le \frac{2\epsilon-\beta}{4}$ | $\frac{2\epsilon-\beta}{4} < b \le \frac{\beta-\epsilon}{2}$ | $b > \frac{\beta-\epsilon}{2}$ |
|---|---|---|---|---|---|
| $a \le \frac{\beta-2\epsilon}{4}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ |
| $\frac{\beta-2\epsilon}{4} < a \le 0$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $\max\{n^{\beta+4b-2}, n^{2\epsilon+4a-2}(\log n)^2\}$ | $n^{\beta+4b-2}$ | $n^{\beta+4b-2}$ |
| $0 < a \le \frac{2\epsilon-\beta}{4}$ | $n^{2\epsilon+4a+4b-2}$ | $\max\{n^{\beta+4a-2}, n^{2\epsilon+4b-2}(\log n)^2\}$ | $n^{2\epsilon-2}(\log n)^4$ | $n^{\beta+4b-2}$ | $n^{\beta+4b-2}$ |
| $\frac{2\epsilon-\beta}{4} < a \le \frac{\beta-\epsilon}{2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{\beta+4a-2}$ | $n^{\beta+4a-2}$ | $n^{\beta+4a\vee b-2}$ | $n^{\beta+4b-2}$ |
| $a > \frac{\beta-\epsilon}{2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{\beta+4a-2}$ | $n^{\beta+4a-2}$ | $n^{\beta+4a-2}$ | $n^{\epsilon+4a\vee b+2a\wedge b-2}$ |

Table C.2: Minimax rates of convergence in the moderately dense regime: $\frac{\beta}{2} \le \epsilon \le \frac{3\beta}{4}$. In this case, we have $\frac{2\epsilon-\beta}{4} \le \frac{\beta-\epsilon}{2}$.

| | $b \le \frac{\beta-2\epsilon}{4}$ | $\frac{\beta-2\epsilon}{4} < b \le 0$ | $0 < b \le \frac{\beta-\epsilon}{2}$ | $\frac{\beta-\epsilon}{2} < b \le \frac{2\epsilon-\beta}{4}$ | $b > \frac{2\epsilon-\beta}{4}$ |
|---|---|---|---|---|---|
| $a \le \frac{\beta-2\epsilon}{4}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ |
| $\frac{\beta-2\epsilon}{4} < a \le 0$ | $n^{2\epsilon+4a+4b-2}$ | $n^{2\epsilon+4a+4b-2}$ | $\max\{n^{\beta+4b-2}, n^{2\epsilon+4a-2}(\log n)^2\}$ | $\max\{n^{\beta+4b-2}, n^{2\epsilon+4a-2}(\log n)^2\}$ | $n^{\beta+4b-2}$ |
| $0 < a \le \frac{\beta-\epsilon}{2}$ | $n^{2\epsilon+4a+4b-2}$ | $\max\{n^{\beta+4a-2}, n^{2\epsilon+4b-2}(\log n)^2\}$ | $n^{2\epsilon-2}(\log n)^4$ | $n^{2\epsilon-2}(\log n)^4$ | $n^{\beta+4b-2}$ |
| $\frac{\beta-\epsilon}{2} < a \le \frac{2\epsilon-\beta}{4}$ | $n^{2\epsilon+4a+4b-2}$ | $\max\{n^{\beta+4a-2}, n^{2\epsilon+4b-2}(\log n)^2\}$ | $n^{2\epsilon-2}(\log n)^4$ | $\max\{n^{2\epsilon-2}(\log n)^4, n^{\epsilon+4a\vee b+2a\wedge b-2}\}$ | $n^{\epsilon+2a+4b-2}$ |
| $a > \frac{2\epsilon-\beta}{4}$ | $n^{2\epsilon+4a+4b-2}$ | $n^{\beta+4a-2}$ | $n^{\beta+4a-2}$ | $n^{\epsilon+4a+2b-2}$ | $n^{\epsilon+4a\vee b+2a\wedge b-2}$ |

Table C.3: Minimax rates of convergence in the strongly dense regime: $\frac{3\beta}{4} < \epsilon \le \beta$. In this case, we have $\frac{\beta-\epsilon}{2} < \frac{2\epsilon-\beta}{4}$.

## C.2 Proofs for Main Results in Section 4.2

This section contains the proofs of main results in Section 4.2. We present the proofs of Theorems 12 and 14 in Section C.2.1, followed by proofs of Theorem 11 and 13 in Sections C.2.2. The proofs of supporting lemmas are given in Section C.2.3.

To simplify notation, in the following we omit the subscripts $n$ in $k_n, q_n, s_n$ and $\tau_n$ that signifies their dependence on the sample size. We denote by $\psi_\mu$ the density of a Gaussian distribution with mean $\mu$ and variance $\sigma^2$, and we denote by $\ell(n,k)$ the class of all subsets of $\{1,\ldots,n\}$ of $k$ distinct elements. We let $\phi(z)$, $\Phi(z) = P(Z \leq z)$, and $\tilde{\Phi}(z) = 1 - \Phi(z)$ be the density, cumulative distribution function, and survival function of a standard normal random variable $Z$, respectively. Finally, $c$ and $C$ denote generic positive constants whose values may vary for each occurrence.

### C.2.1 Proof of Theorems 12 and 14

In this section, we prove Theorems 12 and 14, which constitute the lower bound for the estimation rate of $Q(\mu,\theta)$ in the sparse and the dense regime, respectively. We begin with some technical tools for establishing lower bounds.

#### General Tools

Let $\mathcal{M}$ be a set of probability measures on a measurable space $(\mathcal{X},\mathcal{A})$, and let $\theta : \mathcal{M} \longrightarrow \mathbb{R}$. For $P_f, P_g \in \mathcal{M}$, let $\theta_f = \theta(P_f), \theta_g = \theta(P_g)$, and let $f, g$ denote the density of $P_f, P_g$ with respect to some dominating measure $u$. The chi-square affinity between $P_f$ and $P_g$ is defined as

$$\xi = \xi(P_f, P_g) = \int \frac{g^2}{f}\, du.$$

In particular, for Gaussian distributions, we have

$$\xi(N(\theta_0, \sigma^2), N(\theta_1, \sigma^2)) = e^{(\theta_1-\theta_0)^2/\sigma^2}.$$

Throughout, the proof of lower bounds is established by the construction of two priors which have small chi-square distance but a large difference in the expected values of the resulting quadratic functionals, followed by an application of the Constrained Risk Inequality (CRI) in Brown & Low (1996). Essentially, CRI says that if $P_f$ and $P_g$ are such that $\theta_f, \theta_g \in \Theta$, the parameter space of estimation, with $\xi = \xi(P_f, P_g) < \infty$, then for any estimator $\delta$ of $\theta = \theta(P) \in \Theta$ based on the random variable $X$ with distribution $P$, we have

$$\sup_{\theta\in\Theta} E_\theta(\delta(X) - \theta)^2 \geq \frac{(\theta_g - \theta_f)^2}{(1 + \xi^{1/2})^2}.$$

It follows that to establish lower bound for estimation rate, it suffices to find $P_f$ and $P_g$ such that $(\theta_g - \theta_f)^2$ is as large as possible subject to $\xi(P_f, P_g) < \infty$.

**Proof of Theorem 12**

To prove Theorem 12, it suffices to show that for $0 < \beta < \frac{1}{2}$,

$$\gamma_n(\beta, \epsilon, b) \geq \begin{cases} n^{2\epsilon+4b-2}(\log n)^2 & \text{if } b > 0, \quad \text{for } 0 < \epsilon < \frac{\beta}{2}, \qquad \text{(Case 1)} \\ n^{2\epsilon+8b-2} & \text{if } b \leq 0, \quad \text{for } 0 < \epsilon \leq \beta, \qquad \text{(Case 2)} \\ n^{\epsilon+6b-2} & \text{if } b > 0, \quad \text{for } 0 < \epsilon \leq \beta. \qquad \text{(Case 3)} \end{cases}$$

For individual regions in $\{(\beta, \epsilon, b) : 0 < \epsilon < \frac{\beta}{2}, 0 < \beta < \frac{1}{2}, b \in \mathbb{R}\}$, the minimax rate of convergence is then given by the sharpest rate among all cases in which the region belongs. For instance, the region $\{(\beta, \epsilon, b) : 0 < \epsilon < \frac{\beta}{2}, 0 < \beta < \frac{1}{2}, b > \frac{\epsilon}{2}\}$ is included

in Case 1 and Case 3, hence $\gamma_n(\beta, \epsilon, b) \geq \max\{n^{2\epsilon+4b-2}(\log n)^2, n^{\epsilon+6b-2}\} = n^{\epsilon+6b-2}$.

To establish the desired lower bounds, for each case we construct two priors $f$ and $g$ that have small chi-square distance but a large difference in the expected values of the resulting quadratic functionals, then apply the CRI. The choice of priors $f$ and $g$ is crucial in deriving sharp lower bound for the estimation problem. In fact, the fundamental difference between different phases in the sparse regime for the estimation of $Q(\mu, \theta)$ can be seen from the choices $f$ and $g$.

*Proof of Case 1.* Our proof builds on arguments similar to that used in Cai & Low (2004) and Baraud (2002), who considered the one-sequence estimation problem. We first follow the lines of the proof of Theorem 7 in Cai & Low (2004), and then apply a result from Aldous (1985) as was done in Baraud (2002). Let

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{k} \psi_s(x_i) \prod_{i=k+1}^{n} \psi_0(x_i) \prod_{i=1}^{n} \psi_0(y_i).$$

For $I \in \ell(k, q)$, let

$$g_I(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{k} \psi_s(x_i) \prod_{i=k+1}^{n} \psi_0(x_i) \prod_{i=1}^{k} \psi_{\theta_i}(y_i) \prod_{i=k+1}^{n} \psi_0(y_i),$$

where $\theta_i = \rho \mathbb{1}(i \in I)$ with $\rho > 0$, and let

$$g = \frac{1}{\binom{k}{q}} \sum_{I \in \ell(k,q)} g_I.$$

In both $f$ and $g$, the sequence $\mu = (s, \ldots, s, 0, \ldots, 0)$ is taken to be the same. However, $\theta$ is taken to be all zeros in $f$ but is taken as a mixture in $g$. The nonzero coordinates of $\theta$ are mixed uniformly over the support of $\mu$ at a common magnitude $\rho$, whose value is yet to be determined. Our choice of $f$ and $g$ essentially reduces the two-sequence problem to the case where we only have one Gaussian mean sequence of length $k$ with

$q$ nonzero coordinates, hence explains the correspondence between the sparse regime in the two-sequence case ($q \ll \sqrt{k}$) and the sparse regime in the one-sequence case ($k \ll \sqrt{n}$).

We now compute the chi-square affinity between $f$ and $g$,

$$\int \frac{g^2}{f} = \frac{1}{\binom{k}{q}^2} \sum_{I \in \ell(k,q)} \sum_{J \in \ell(k,q)} \int \frac{g_I g_J}{f}. \tag{C.2}$$

For $I, J \in \ell(k,q)$, let $m = \mathrm{Card}(I \cap J)$. Then

$$\begin{aligned}
\int \frac{g_I g_J}{f} &= \prod_{i=1}^{k} \int \frac{\psi_{\rho \mathbb{1}(i \in I)}(y_i) \cdot \psi_{\rho \mathbb{1}(i \in J)}(y_i)}{\psi_0(y_i)} \, dy_i \\
&= \left[ \int \psi_0(y) \, dy \right]^{k-2q+m} \left[ \int \psi_\rho(y) \, dy \right]^{2q-2m} \left[ \int \frac{\psi_\rho^2(y)}{\psi_0(y)} \, dy \right]^{m} \\
&= \exp\left( \frac{m\rho^2}{\sigma^2} \right).
\end{aligned}$$

It follows that

$$\int \frac{g^2}{f} = E\left[ \exp\left( \frac{M\rho^2}{\sigma^2} \right) \right],$$

where $M$ has the hypergeometric distribution

$$P(M = m) = \frac{\binom{q}{m}\binom{k-q}{q-m}}{\binom{k}{q}}. \tag{C.3}$$

As shown in Aldous (1985), $M$ has the same distribution as the conditional expectation $E(\tilde{M}|\mathcal{B})$, where $\tilde{M}$ is a Binomial($q$, $\frac{q}{k}$) random variable and $\mathcal{B}$ is a suitable $\sigma$-algebra. Coupled with Jensen's inequality, this implies that

$$\int \frac{g^2}{f} \leq E\left[ \exp\left( \frac{\tilde{M}\rho^2}{\sigma^2} \right) \right] = \left( 1 - \frac{q}{k} + \frac{q}{k} e^{\rho^2/\sigma^2} \right)^q.$$

Taking $\rho = \sigma\sqrt{(\beta - 2\epsilon)\log n}$ gives

$$e^{\rho^2/\sigma^2} = n^{\beta-2\epsilon} = \frac{k}{q^2},$$

hence

$$\int \frac{g^2}{f} \leq \left(1 + \frac{1}{q}\right)^q \leq e.$$

Since $Q(\mu, \theta) = 0$ under $f$ and $Q(\mu, \theta) = \frac{1}{n} q s^2 \rho^2$ under $g$, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c\left(\frac{1}{n} q s^2 \rho^2\right)^2 = c n^{2\epsilon+4b-2} (\log n)^2.$$

□

*Proof of Case 2.* Let

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^n \psi_0(x_i) \prod_{i=1}^n \psi_0(y_i)$$

For $I \in \ell(n, q)$, let

$$g_I(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^n \psi_{\mu_i}(x_i) \prod_{i=1}^n \psi_{\theta_i}(y_i),$$

where $\mu_i = \theta_i = \rho \mathbb{1}(i \in I)$ with $\rho > 0$, and let

$$g = \frac{1}{\binom{n}{q}} \sum_{I \in \ell(n,q)} g_I.$$

Contrast the choice of $f$ an $g$ here with that used in the proof of Case 1. Rather than fixing $\mu$ and mixing nonzero coordinates of $\theta$ over the support of $\mu$, in this case mixing is done over all $n$ positions using nonzero coordinates of $\mu$ and $\theta$ simultaneously.

Similar calculation as that used in the proof of Case 1 yields

$$\int \frac{g^2}{f} \leq \left(1 - \frac{q}{n} + \frac{q}{n} e^{2\rho^2/\sigma^2}\right)^q. \tag{C.4}$$

Now take $\rho = s = n^b$. Since $b < 0$, it follows that when $n$ is sufficiently large,

$$e^{2\rho^2/\sigma^2} \leq n^{1-2\epsilon} = \frac{n}{q^2},$$

hence

$$\int \frac{g^2}{f} \leq \left(1 + \frac{1}{q}\right)^q \leq e.$$

Since $Q(\mu, \theta) = 0$ under $f$, and $Q(\mu, \theta) = \frac{1}{n} q\rho^4$ under $g$, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c\left(\frac{1}{n} q\rho^4\right)^2 = cn^{2\epsilon+8b-2}.$$

□

*Proof of Case 3.* The priors used in this case are very different from that considered in the proofs of Case 1 and Case 2. Let

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{q} \psi_s(x_i) \prod_{i=q+1}^{n} \psi_0(x_i) \prod_{i=1}^{q} \psi_s(y_i) \prod_{i=q+1}^{n} \psi_0(y_i),$$
$$g(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{q} \psi_s(x_i) \prod_{i=q+1}^{n} \psi_0(x_i) \prod_{i=1}^{q} \psi_{s-\delta}(y_i) \prod_{i=q+1}^{n} \psi_0(y_i),$$

where $0 < \delta < s$. Note that no mixing is performed in this case. Instead, we fix the sequence $\mu = (s, \ldots, s, 0, \ldots, 0)$ in both $f$ and $g$, and perturb the nonzero entries of $\theta$ by a small amount $\delta$ in $g$. This set of priors provides the sharpest rate for the case when the signal is strong, i.e., $s = n^b$ is large. The intuition is that when $s$ is large, estimation of $Q(\mu, \theta)$ is most difficult due to the indistinguishability between $\theta_i = s$

and $\theta_i = s - \delta$, where $\delta \approx 0$.

The chi-square affinity between $f$ and $g$ is given by

$$\int \frac{g^2}{f} = e^{q\delta^2/\sigma^2}.$$

Let $\delta = \sigma/\sqrt{q} = \sigma n^{-\epsilon/2}$. Then we have

$$\int \frac{g^2}{f} = e < \infty.$$

Since $Q(\mu, \theta) = \frac{1}{n} q s^4$ under $f$ and $Q(\mu, \theta) = \frac{1}{n} q s^2 (s - \delta)^2$ under $g$, it follows from CRI that

$$\begin{aligned} R^*(n, \Omega(\beta, \epsilon, b)) &\geq c\left(\frac{1}{n} q s^2 \big(s^2 - (s-\delta)^2\big)\right)^2 \\ &= c\left(\frac{1}{n}\sqrt{q} s^3\right)^2 (1 + o(1)) = c n^{\epsilon + 6b - 2}(1 + o(1)). \end{aligned}$$

□

**Proof of Theorem 14**

To prove Theorem 14, it is sufficient to show that for $0 < \beta < \frac{1}{2}$,

$$\gamma_n(\beta, \epsilon, b) \geq \begin{cases} n^{2\epsilon + 8b - 2} & \text{if } b \leq 0, \quad \text{for } 0 < \epsilon \leq \beta, \qquad \text{(Case 2)} \\ n^{\epsilon + 6b - 2} & \text{if } b > 0, \quad \text{for } 0 < \epsilon \leq \beta, \qquad \text{(Case 3)} \\ n^{\beta + 4b - 2} & \text{if } b > 0, \quad \text{for } \frac{\beta}{2} \leq \epsilon \leq \beta, \qquad \text{(Case 4)} \\ n^{2\epsilon - 2}(\log n)^4 & \text{if } b > 0, \quad \text{for } 0 < \epsilon \leq \beta. \qquad \text{(Case 5)} \end{cases}$$

The proofs of Case 2 and Case 3 are included in the proof of Theorem 12, hence we will only provide proofs of Case 4 and Case 5 below. For individual regions in

$\{(\beta, \epsilon, b) : \frac{\beta}{2} \leq \epsilon \leq \beta < \frac{1}{2}, b \in \mathbb{R}\}$, the minimax rate of convergence is obtained as the sharpest rate among all cases in which the region belongs to. For instance, the region $\{(\beta, \epsilon, b) : \frac{3\beta}{4} < \epsilon \leq \beta < \frac{1}{2}, b > \frac{\epsilon}{6}\}$ is included in Case 3, Case 4 and Case 5, hence $\gamma_n(\beta, \epsilon, b) \geq \max\{n^{\epsilon+6b-2}, n^{\beta+4b-2}, n^{2\epsilon-2}(\log n)^4\} = n^{\epsilon+6b-2}$.

*Proof of Case 4.* The proof of Case 4 is very similar to the proof of Case 1, besides that a slightly different mixture prior $g$ is employed. Let

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{k} \psi_s(x_i) \prod_{i=k+1}^{n} \psi_0(x_i) \prod_{i=1}^{n} \psi_0(y_i).$$

For $I \in \ell(k, q)$, let

$$\begin{aligned} &g_I(x_1, \ldots, x_n, y_1, \ldots, y_n) \\ &= \prod_{i=1}^{k} \psi_s(x_i) \prod_{i=k+1}^{n} \psi_0(x_i) \prod_{i=1}^{k} \left[\frac{1}{2}\psi_{\theta_i}(y_i) + \frac{1}{2}\psi_{-\theta_i}(y_i)\right] \prod_{i=k+1}^{n} \psi_0(y_i), \end{aligned}$$

where $\theta_i = \rho \mathbb{1}(i \in I)$ with $\rho > 0$, and let

$$g = \frac{1}{\binom{k}{q}} \sum_{I \in \ell(k,q)} g_I.$$

Note that in constructing $g$, mixing is done not only over all possible subsets $\ell(k, q)$ but also over the signs of $\theta_i$'s. This has largely to do with the intuition that when signal is abundant, uncertainty about the signs of $\theta_i$'s further increase the difficulty of the estimation problem. That being said, mixing without sign flips (i.e., simply use the priors $f$ and $g$ as given in the proof of Case 1) does not give us the tightest lower bound. Similar to Case 1, keeping $\mu = (s, \ldots, s, 0, \ldots, 0)$ the same in both $f$ and $g$ essentially reduces the two-sequence problem to a one-sequence problem. Our choice of priors is equivalent to having only one Gaussian mean sequence of length $k$

with $q$ nonzero entries — thus the correspondence between the dense regime in the two-sequence case ($q \gg \sqrt{k}$) and the dense regime in the one-sequence case ($k \gg \sqrt{n}$).

Again, the chi-square affinity between $f$ and $g$ has the form (C.2), where for $I, J \in \ell(k, q)$ with $m = \mathrm{Card}(I \cap J)$,

$$\begin{aligned}
\int \frac{g_I g_J}{f} &= \prod_{i=1}^{k} \int \frac{[\frac{1}{2}\psi_{\rho \mathbb{1}(i\in I)}(y_i) + \frac{1}{2}\psi_{-\rho \mathbb{1}(i\in I)}(y_i)][\frac{1}{2}\psi_{\rho \mathbb{1}(i\in J)}(y_i) + \frac{1}{2}\psi_{-\rho \mathbb{1}(i\in J)}(y_i)]}{\psi_0(y_i)}\, dy_i \\
&= \prod_{i=1}^{k} \int \frac{1}{4}\Bigg\{ \frac{\psi_{\rho \mathbb{1}(i\in I)}(y_i)\psi_{\rho \mathbb{1}(i\in J)}(y_i)}{\psi_0(y_i)} + \frac{\psi_{-\rho \mathbb{1}(i\in I)}(y_i)\psi_{-\rho \mathbb{1}(i\in J)}(y_i)}{\psi_0(y_i)} \\
&\qquad\qquad + \frac{\psi_{\rho \mathbb{1}(i\in I)}(y_i)\psi_{-\rho \mathbb{1}(i\in J)}(y_i)}{\psi_0(y_i)} + \frac{\psi_{-\rho \mathbb{1}(i\in I)}(y_i)\psi_{\rho \mathbb{1}(i\in J)}(y_i)}{\psi_0(y_i)} \Bigg\}\, dy_i \\
&= \prod_{i\in I\cap J} \frac{1}{4}\Big[ \int \frac{\psi_\rho^2(y_i)}{\psi_0(y_i)} + \int \frac{\psi_{-\rho}^2(y_i)}{\psi_0(y_i)} + 2\int \frac{\psi_\rho(y_i)\psi_{-\rho}(y_i)}{\psi_0(y_i)} \Big] \prod_{i\in I^c\cup J^c} 1 \\
&= \prod_{i\in I\cap J} \frac{1}{2}\big[ \exp(\rho^2/\sigma^2) + \exp(-\rho^2/\sigma^2) \big] \\
&= \cosh(\rho^2/\sigma^2)^m.
\end{aligned}$$

It follows that

$$\int \frac{g^2}{f} = E[\cosh(\rho^2/\sigma^2)^M],$$

where $M$ follows hypergeometric distribution as in (C.3). Since $M$ coincides in distribution with the conditional expectation $E(\tilde{M}|\mathcal{B})$ where $\tilde{M}$ is a Binomial($q$, $\frac{q}{k}$) random variable and $\mathcal{B}$ is a suitable $\sigma$-algebra (Aldous, 1985), with Jensen's inequality, we get

$$\int \frac{g^2}{f} \le E[\cosh(\rho^2/\sigma^2)^{\tilde{M}}] = \left(1 + \frac{q}{k}[\cosh(\rho^2/\sigma^2) - 1]\right)^q.$$

Since $\cosh(x) = \frac{1}{2}(e^x + e^{-x}) = 1 + \frac{x^2}{2} + o(x^2)$ when $x \approx 0$, taking $x = \rho^2/\sigma^2$ with $\rho = (\frac{k}{q^2})^{1/4}$ yields

$$\int \frac{g^2}{f} \le \left(1 + \frac{1}{2\sigma^4 q}\right)^q < \infty.$$

Since $Q(\mu, \theta) = 0$ under $f$ and $Q(\mu, \theta) = \frac{1}{n} q s^2 \rho^2$ under $g$, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c \left( \frac{1}{n} q s^2 \rho^2 \right)^2 = c n^{\beta + 4b - 2}.$$

□

*Proof of Case 5.* Let $f$ and $g$ be as given in the proof of Case 2, and take $\rho = \sigma \sqrt{\frac{1}{2}(1 - 2\epsilon) \log n}$ in (C.4). It follows that when $n$ is sufficiently large,

$$e^{2\rho^2/\sigma^2} = n^{1-2\epsilon} = \frac{n}{q^2},$$

hence

$$\int \frac{g^2}{f} \leq \left( 1 + \frac{1}{q} \right)^q \leq e.$$

Since $Q(\mu, \theta) = 0$ under $f$, and $Q(\mu, \theta) = \frac{1}{n} q \rho^4$ under $g$, it follows from CRI that

$$R^*(n, \Omega(\beta, \epsilon, b)) \geq c \left( \frac{1}{n} q \rho^4 \right)^2 = c n^{2\epsilon - 2} (\log n)^4.$$

□

## C.2.2 Proof of Theorems 11 and 13

In this section, we prove Theorems 11 and 13, which constitute the upper bound for the estimation rate of $Q(\mu, \theta)$ in the sparse and the dense regime, respectively.

### Proof of Theorem 11

We need a lemma from Cai & Low (2005) (Lemma 1, page 2939) for proving Theorem 11.

**Lemma 29.** *Let $Y \sim N(\theta, \sigma^2)$ and let $\theta_0 = E(Z^2 - \sigma^2\tau)_+$, where $Z \sim N(0, \sigma^2)$. Then for $\tau \geq 1$ and $\widehat{\theta^2} = (Y^2 - \sigma^2\tau)_+ - \theta_0$,*

$$\begin{aligned} |\theta_0| &\leq \frac{4\sigma^2}{\sqrt{2\pi}\tau^{1/2}e^{\tau/2}}, \\ |E(\widehat{\theta^2}) - \theta^2| &\leq \min\{2\sigma^2\tau, \theta^2\}, \\ \mathrm{Var}\,(\widehat{\theta^2}) &\leq 6\sigma^2\theta^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}. \end{aligned}$$

Lemma 30 is an immediate consequence of Lemma 29.

**Lemma 30.** *Let $Y \sim N(\theta, \sigma^2)$ and let $\theta_0 = E(Z^2 - \sigma^2\tau)_+$, where $Z \sim N(0, \sigma^2)$. Then for $\tau \geq 1$,*

$$(E(Y^2 - \sigma^2\tau)_+ - \theta_0)^2 \leq \max\left\{6\sigma^2\theta^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}, 10\theta^4\right\}. \tag{C.5}$$

*Proof.* Let $B(\theta) = E(Y^2 - \tau\sigma^2)_+ - \theta_0$. We first note that $B(-\theta) = B(\theta) \geq 0$ for $\theta \geq 0$. This follows from

$$\begin{aligned} B'(\theta) &= 2\sigma[\phi(\tau^{1/2} - \theta/\sigma) - \phi(\tau^{1/2} + \theta/\sigma)] \\ &\quad - 2\theta[\Phi(\tau^{1/2} - \theta/\sigma) - \Phi(-\tau^{1/2} - \theta/\sigma) - 1] \\ &\geq 0 \end{aligned}$$

and $B(0) = 0$. So we have $B(\theta) = E(Y^2 - \tau\sigma^2)_+ - \theta_0 \geq 0$ for all $\theta \in \mathbb{R}$. It follows that $(E[(Y^2 - \tau\sigma^2)_+ - \theta_0])^2 \leq (E(Y^2 - \tau\sigma^2)_+)^2 \leq E[(Y^2 - \tau\sigma^2)_+^2]$. To bound the term $E[(Y^2 - \tau\sigma^2)_+^2]$, we consider two cases: $\theta \leq \sigma$ and $\theta > \sigma$. It follows from the proof of Lemma 1 in Cai & Low (2005) that when $\theta \leq \sigma$, then

$$E[(Y^2 - \tau\sigma^2)_+^2] \leq 6\sigma^2\theta^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}.$$

On the other hand, when $\theta > \sigma$, we have

$$E[(Y^2 - \tau\sigma^2)_+^2] \le E[Y^4] = \theta^4 + 6\sigma^2\theta^2 + 3\sigma^4 \le 10\theta^4.$$

If follows that (C.5) holds. □

*Proof of Theorem 11.* We first bound the bias of the estimator $\widehat{Q}_2$ defined in (4.15). Using the equality

$$AB - ab = (A-a)(B-b) + a(B-b) + b(A-a),$$

the independence of $X_i$ and $Y_i$, and the triangle inequality, we get

$$\begin{aligned}
&\left|E_{(\mu_i,\theta_i)}\{[(X_i^2 - \sigma^2\tau)_+ - \mu_0][(Y_i^2 - \sigma^2\tau)_+ - \theta_0]\} - \mu_i^2\theta_i^2\right| \\
&\le \left|E_{\mu_i}[(X_i^2 - \sigma^2\tau)_+ - \mu_0] - \mu_i^2\right| \cdot \left|E_{\theta_i}[(Y_i^2 - \sigma^2\tau)_+ - \theta_0] - \theta_i^2\right| \\
&\qquad + \mu_i^2\left|E_{\theta_i}[(Y_i^2 - \sigma^2\tau)_+ - \theta_0] - \theta_i^2\right| + \theta_i^2\left|E_{\mu_i}[(X_i^2 - \sigma^2\tau)_+ - \mu_0] - \mu_i^2\right| \\
&\le \min\{2\sigma^2\tau, \mu_i^2\}\min\{2\sigma^2\tau, \theta_i^2\} + \mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + \theta_i^2\min\{2\sigma^2\tau, \mu_i^2\} \\
&\le 2\mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + 2\theta_i^2\min\{2\sigma^2\tau, \mu_i^2\},
\end{aligned}$$

the second inequality follows from Lemma 29. It follows that, for $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$ and $\tau \ge 1$,

$$\begin{aligned}
&|E_{(\mu,\theta)}(\widehat{Q}_2) - Q(\mu,\theta)| \\
&= \left|\frac{1}{n}\sum_{i=1}^n E_{(\mu_i,\theta_i)}\{[(X_i^2 - \sigma^2\tau)_+ - \mu_0][(Y_i^2 - \sigma^2\tau)_+ - \theta_0]\} - \frac{1}{n}\sum_{i=1}^n \mu_i^2\theta_i^2\right| \\
&\le \frac{2}{n}\sum_{i=1}^n \left[\mu_i^2\min\{2\sigma^2\tau, \theta_i^2\} + \theta_i^2\min\{2\sigma^2\tau, \mu_i^2\}\right] \\
&\le \frac{4}{n}\min\{2\sigma^2 q s^2\tau, q s^4\},
\end{aligned}$$

the second inequality follows from the fact that, for $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$, there are at most $q$ entries that are simultaneously nonzero for $\mu$ and $\theta$.

We now proceed to bound the variance of $\widehat{Q}_2$. Applying the equality

$$\mathrm{Var}\,(AB) = \mathrm{Var}\,(A)\mathrm{Var}\,(B) + [E(A)]^2\mathrm{Var}\,(B) + [E(B)]^2\mathrm{Var}\,(A),$$

for $\tau \geq 1$, we have

$$\begin{aligned}
&\mathrm{Var}_{(\mu_i,\theta_i)}\{[(X_i^2 - \sigma^2\tau)_+ - \mu_0][(Y_i^2 - \sigma^2\tau)_+ - \theta_0]\} \\
&= \mathrm{Var}_{\mu_i}[(X_i^2 - \sigma^2\tau)_+ - \mu_0]\mathrm{Var}_{\theta_i}[(Y_i^2 - \sigma^2\tau)_+ - \theta_0] \\
&\quad + [E_{\mu_i}(X_i^2 - \sigma^2\tau)_+ - \mu_0]^2\mathrm{Var}_{\theta_i}[(Y_i^2 - \sigma^2\tau)_+ - \theta_0] \\
&\quad + [E_{\theta_i}(Y_i^2 - \sigma^2\tau)_+ - \theta_0]^2\mathrm{Var}_{\mu_i}[(X_i^2 - \sigma^2\tau)_+ - \mu_0] \\
&\leq 3\left[6\sigma^2\mu_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\left[6\sigma^2\theta_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right] \\
&\quad + 10\mu_i^4\left[6\sigma^2\theta_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right] + 10\theta_i^4\left[6\sigma^2\mu_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right],
\end{aligned}$$

the inequality follows from Lemma 29 and Lemma 30. Thus, for $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$ and $\tau \geq 1$,

$$\begin{aligned}
&\mathrm{Var}_{(\mu,\theta)}(\widehat{Q}_2) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_{(\mu_i,\theta_i)}\{[(X_i^2 - \sigma^2\tau)_+ - \mu_0][(Y_i^2 - \sigma^2\tau)_+ - \theta_0]\} \\
&\leq \frac{3}{n^2}\sum_{i=1}^{n}\left[6\sigma^2\mu_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right]\left[6\sigma^2\theta_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right] \\
&\quad + \frac{10}{n^2}\sum_{i=1}^{n}\mu_i^4\left[6\sigma^2\theta_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right] + \frac{10}{n^2}\sum_{i=1}^{n}\theta_i^4\left[6\sigma^2\mu_i^2 + \sigma^4\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right] \\
&\leq \frac{3}{n^2}\left[36\sigma^4 qs^4 + 12\sigma^6 ks^2\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right) + n\sigma^8\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\right]
\end{aligned}$$

$$+\frac{20}{n^2}\left[6\sigma^2 q s^6+\sigma^4 k s^4\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)\right].$$

Combining the bias and variance term, we get, for $\tau \geq 1$,

$$\begin{aligned}
&\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2 - Q(\mu,\theta))^2 \\
&\leq \frac{C}{n^2}\Bigg[\min\{q^2 s^4\tau^2, q^2 s^8\} + \max\Bigg\{qs^4, qs^6, ks^2\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right), \\
&\qquad\qquad ks^4\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right), n\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\Bigg\}\Bigg] \\
&= \frac{C}{n^2}\Bigg[\min\{n^{2\epsilon+4b}\tau^2, n^{2\epsilon+8b}\} + \max\Bigg\{n^{\epsilon+4b}, n^{\epsilon+6b}, n^{\beta+2b}\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right), \\
&\qquad\qquad n^{\beta+4b}\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right), n\left(\frac{4\tau^{1/2}+18}{e^{\tau/2}}\right)^2\Bigg\}\Bigg].
\end{aligned}$$

Suppose that $b > 0$. Then letting $\tau = \log n$ leads to

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_2 - Q(\mu,\theta))^2 \leq C\left[n^{2\epsilon+4b-2}(\log n)^2 + n^{\epsilon+6b-2}\right].$$

□

**Proof of Theorem 13**

The proof of Theorem 13 is based on Lemmas 31 and 32 which bound, respectively, the bias and variance of one term in the estimator $\widehat{Q}_4$ (given in (4.21)). For clarity, we defer the proofs of Lemma 31 and Lemma 32 to Section C.2.3.

**Lemma 31.** *Let $X \sim N(\mu,\sigma^2)$ and $Y \sim N(\theta,\sigma^2)$ be independent. Set $\eta = E[(Z_1^2 - \sigma^2)(Z_2^2-\sigma^2)\mathbb{1}(Z_1^2 \vee Z_2^2 > \sigma^2\tau)]$, where $Z_1, Z_2 \overset{i.i.d.}{\sim} N(0,\sigma^2)$. Then*

$$\eta = -4\sigma^4\tau\phi^2(\tau^{1/2}),$$

*and for* $\tau \geq 1$,

$$\begin{aligned}
&\left|E[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2 \vee Y^2 > \sigma^2\tau)] - \eta - \mu^2\theta^2\right| \\
&\leq \min\{\mu^2, 3\sigma^2\tau\}\min\{\theta^2, 3\sigma^2\tau\} + 2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\mu^2, 3\sigma^2\tau\} \\
&\quad + 2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\theta^2, 3\sigma^2\tau\}.
\end{aligned}$$

**Lemma 32.** *Let* $X \sim N(\mu, \sigma^2)$ *and* $Y \sim N(\theta, \sigma^2)$ *be independent. Then for* $\tau \geq 1$,

$$\begin{aligned}
&\mathrm{Var}\left[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2 \vee Y^2 > \sigma^2\tau)\right] \\
&\leq \begin{cases}
2d^{1/2}\tilde{\Phi}(\tau^{1/2})^{1/2} & \text{if } \mu = \theta = 0, \\
4\sigma^2\mu^4\theta^2 + 4\sigma^2\mu^2\theta^4 + 16\sigma^4\mu^2\theta^2 + 2\sigma^4\mu^4 + 2\sigma^4\theta^4 & \\
\quad +8\sigma^6\mu^2 + 8\sigma^6\theta^2 + 4\sigma^8 + 8\sigma^4\mu^2\theta^2\tau^2 & \text{otherwise},
\end{cases}
\end{aligned}$$

*where* $d = E[(Z_1^2-\sigma^2)^4(Z_2^2-\sigma^2)^4]$ *and* $Z_1, Z_2 \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

*Proof of Theorem 13.* We first compute the bias of $\widehat{Q}_4$. It follows from Lemma 31 that for all $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$ and $\tau \geq 1$, we have

$$\begin{aligned}
&\left|E_{(\mu,\theta)}(\widehat{Q}_4) - Q(\mu, \theta)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|E_{(\mu_i,\theta_i)}[(X_i^2-\sigma^2)(Y_i^2-\sigma^2)\mathbb{1}(X_i^2 \vee Y_i^2 > \sigma^2\tau)] - \eta - \mu_i^2\theta_i^2\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Big[\min\{\mu_i^2, 3\sigma^2\tau\}\min\{\theta_i^2, 3\sigma^2\tau\} + 2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\mu_i^2, 3\sigma^2\tau\} \\
&\qquad + 2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\theta_i^2, 3\sigma^2\tau\}\Big] \\
&\leq \frac{1}{n}\Big[\min\{qs^4, 3\sigma^2 qs^2\tau, 9\sigma^4 q\tau^2\} + 4\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{ks^2, 3\sigma^2 k\tau\}\Big],
\end{aligned}$$

the last inequality follows from the fact that for $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$, there are at most $k$ nonzero entries for either $\mu$ or $\theta$, and there are at most $q$ entries that are

simultaneously nonzero for both $\mu$ and $\theta$.

On the other hand, by Lemma 32, for all $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$ and $\tau \geq 1$, the variance of $\widehat{Q}_4$ satisfies

$$
\begin{aligned}
&\mathrm{Var}_{(\mu,\theta)}(\widehat{Q}_4) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_{(\mu_i,\theta_i)}[(X_i^2 - \sigma^2)(Y_i^2 - \sigma^2)\mathbb{1}(X_i^2 \vee Y_i^2 > \sigma^2\tau)] \\
&\leq \frac{1}{n^2}\Bigg[ \sum_{i:\mu_i=\theta_i=0} 2d^{1/2}\tilde{\Phi}(\tau^{1/2})^{1/2} \\
&\qquad + \sum_{i:\mu_i\neq 0 \text{ or } \theta_i \neq 0} \big(4\sigma^2\mu_i^4\theta_i^2 + 4\sigma^2\mu_i^2\theta_i^4 + 16\sigma^4\mu_i^2\theta_i^2 + 2\sigma^4\mu_i^4 + 2\sigma^4\theta_i^4 \\
&\qquad\qquad + 8\sigma^6\mu_i^2 + 8\sigma^6\theta_i^2 + 4\sigma^8 + 8\sigma^4\mu_i^2\theta_i^2\tau^2\big)\Bigg] \\
&\leq \frac{1}{n^2}\Big[2d^{1/2}n\tilde{\Phi}(\tau^{1/2})^{1/2} + 8\sigma^2 qs^6 + 16\sigma^4 qs^4 + 4\sigma^4 ks^4 + 16\sigma^6 ks^2 + 8\sigma^8 k + 8\sigma^4 qs^4\tau^2\Big] \\
&\leq \frac{C}{n^2}\max\{n\tilde{\Phi}(\tau^{1/2})^{1/2}, qs^4, qs^6, k, ks^2, ks^4, qs^4\tau^2\}.
\end{aligned}
$$

Again, the second to the last inequality follows from the fact that for $(\mu, \theta) \in \Omega(\beta, \epsilon, b)$, there are at most $k$ nonzero entries for either $\mu$ or $\theta$, and there are at most $q$ entries that are simultaneously nonzero for both $\mu$ and $\theta$.

Combining the bias and variance term, we have

$$
\begin{aligned}
&\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_4 - Q(\mu,\theta))^2 \\
&\leq \frac{C}{n^2}\Big[\min\{q^2s^8, q^2s^4\tau^2, q^2\tau^4\} + \tau\phi^2(\tau^{1/2})\min\{k^2s^4, k^2\tau^2\} \\
&\qquad + \max\{n\tilde{\Phi}(\tau^{1/2})^{1/2}, qs^4, qs^6, k, ks^2, ks^4, qs^4\tau^2\}\Big] \\
&= \frac{C}{n^2}\Big[\min\{n^{2\epsilon+8b}, n^{2\epsilon+4b}\tau^2, n^{2\epsilon}\tau^4\} + \tau\phi^2(\tau^{1/2})\min\{n^{2\beta+4b}, n^{2\beta}\tau^2\} \\
&\qquad + \max\{n\tilde{\Phi}(\tau^{1/2})^{1/2}, n^{\epsilon+4b}, n^{\epsilon+6b}, n^{\beta}, n^{\beta+2b}, n^{\beta+4b}, n^{\epsilon+4b}\tau^2\}\Big].
\end{aligned}
$$

Let $\tau = 4\log n$, then we have $\tilde{\Phi}(\tau^{1/2}) \leq C\phi(\tau^{1/2}) = O(n^{-2})$ for some constant $C$. It follows that for $b > 0$,

$$\sup_{(\mu,\theta)\in\Omega(\beta,\epsilon,b)} E_{(\mu,\theta)}(\widehat{Q}_4 - Q(\mu,\theta))^2 \leq C \max\left\{n^{2\epsilon-2}(\log n)^4, n^{\epsilon+6b-2}, n^{\beta+4b-2}\right\}.$$

□

### C.2.3 Proofs of Supporting Lemmas

In this section, we provide the proofs of technical lemmas that are used to establish Theorem 13 in Section 4.2.

#### Proof of Lemma 31

The proof of Lemma 31 is built on Lemma 33 and Lemma 34.

**Lemma 33.** *Let $Y \sim N(\theta, \sigma^2)$. Then for $\tau \geq 1$,*

$$E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \leq \sigma^2\tau)] = \theta^2\left[\tilde{\Phi}(-\tau^{1/2} - \frac{\theta}{\sigma}) - \tilde{\Phi}\Big(\tau^{1/2} - \frac{\theta}{\sigma}\Big)\right] + \phi\Big(\tau^{1/2} + \frac{\theta}{\sigma}\Big)[-\sigma^2\tau^{1/2} + \sigma\theta] + \phi\Big(\tau^{1/2} - \frac{\theta}{\sigma}\Big)[-\sigma^2\tau^{1/2} - \sigma\theta].$$

*In particular, when $\theta = 0$,*

$$E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \leq \sigma^2\tau)] = -2\sigma^2\tau^{1/2}\phi(\tau^{1/2}).$$

*Proof.* Let $\lambda = \tau^{1/2}$. We have

$$E[Y^2\mathbb{1}(Y^2 \leq \sigma^2\tau)] = \int_{-\sigma\lambda}^{\sigma\lambda} y^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\theta)^2/2\sigma^2}\, dy$$

$$= \int_{-\lambda-\theta/\sigma}^{\lambda-\theta/\sigma} (\theta+\sigma z)^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$
$$= \theta^2 \int_{-\lambda-\theta/\sigma}^{\lambda-\theta/\sigma} \phi(z)\, dz + 2\sigma\theta \int_{-\lambda-\theta/\sigma}^{\lambda-\theta/\sigma} z\phi(z)\, dz + \sigma^2 \int_{-\lambda-\theta/\sigma}^{\lambda-\theta/\sigma} z^2\phi(z)\, dz.$$

Using the fact that

$$\int_a^\infty \phi(z)\, dz = \tilde{\Phi}(a), \quad \int_a^\infty z\phi(z)\, dz = \phi(a), \quad \int_a^\infty z^2\phi(z)\, dz = a\phi(a) + \tilde{\Phi}(a),$$

we have

$$\begin{aligned}
&E[Y^2 \mathbb{1}(Y^2 \le \sigma^2\tau)] \\
&= \theta^2[\tilde{\Phi}(-\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)] + 2\sigma\theta[\phi(-\lambda-\theta/\sigma) - \phi(\lambda-\theta/\sigma)] \\
&\quad + \sigma^2[(-\lambda-\theta/\sigma)\phi(-\lambda-\theta/\sigma) + \tilde{\Phi}(-\lambda-\theta/\sigma) \\
&\qquad - (\lambda-\theta/\sigma)\phi(\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)] \\
&= (\theta^2+\sigma^2)[\tilde{\Phi}(-\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)] \\
&\quad + \phi(\lambda+\theta/\sigma)[-\sigma^2\lambda + \sigma\theta] + \phi(\lambda-\theta/\sigma)[-\sigma^2\lambda - \sigma\theta],
\end{aligned}$$

the last equality due to $\phi(-\lambda-\theta/\sigma) = \phi(\lambda+\theta/\sigma)$. The proof is complete since $\sigma^2 E[\mathbb{1}(Y^2 < \sigma^2\tau)] = \sigma^2[\tilde{\Phi}(-\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)]$. □

**Lemma 34.** *Let $Y \sim N(\theta, \sigma^2)$ and set $\theta_0 = E[(Z^2-\sigma^2)\mathbb{1}(Z^2 \le \sigma^2\tau)]$, where $Z \sim N(0,\sigma^2)$. Then for $\tau \ge 1$,*

$$\left|E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] - \theta_0\right| \le \min\{\theta^2, 3\sigma^2\tau\}.$$

*Proof.* Let $B(\theta) = E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] - \theta_0$. We first show that $|B(\theta)| \le 3\sigma^2\tau$.

Define $\lambda = \tau^{1/2}$. Then

$$E[(Y^2 - \sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] \le E[Y^2\mathbb{1}(Y^2 \le \sigma^2\tau)] \le \sigma^2\lambda^2,$$

and

$$\begin{aligned} E[(Y^2 - \sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] &= E(Y^2 - \sigma^2) - E[(Y^2 - \sigma^2)\mathbb{1}(Y^2 > \sigma^2\tau)] \\ &\ge \theta^2 - E(Y^2) = -\sigma^2 \ge -\sigma^2\lambda^2. \end{aligned}$$

By Lemma 33, $\theta_0 = -2\sigma^2\lambda\phi(\lambda)$. It follows that

$$|B(\theta)| \le \left|E[(Y^2 - \sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)]\right| + |\theta_0| \le \sigma^2\lambda^2 + 2\sigma^2\lambda\phi(\lambda) \le 3\sigma^2\lambda^2 = 3\sigma^2\tau.$$

We now show that $|B(\theta)| \le \theta^2$. Straightforward calculation yields for $\theta \ge 0$,

$$\begin{aligned} B'(\theta) &= \sigma(1+\lambda^2)[\phi(\lambda + \theta/\sigma) - \phi(\lambda - \theta/\sigma)] \\ &\quad + 2\theta[\tilde{\Phi}(-\lambda - \theta/\sigma) - \tilde{\Phi}(\lambda - \theta/\sigma)], \qquad \text{(C.6)} \\ B''(\theta) &= \phi(\lambda + \theta/\sigma)[-\lambda^2(\lambda + \theta/\sigma) - \lambda + \theta/\sigma] \\ &\quad + \phi(\lambda - \theta/\sigma)[-\lambda^2(\lambda - \theta/\sigma) - \lambda - \theta/\sigma] \\ &\quad + 2[\tilde{\Phi}(-\lambda - \theta/\sigma) - \tilde{\Phi}(\lambda - \theta/\sigma)]. \qquad \text{(C.7)} \end{aligned}$$

It suffices to only consider $\theta \ge 0$ since $B(\theta) = B(-\theta)$. It follows from (C.6) that for all $\theta \ge 0$, $B'(\theta) \le 2\theta$. Since $B(0) = 0$, this implies that

$$B(\theta) \le \theta^2, \qquad \forall\theta \ge 0. \qquad \text{(C.8)}$$

On the other hand, $\theta_0 \le 0$ immediately gives $B(\theta) \ge -\sigma^2 \ge -\theta^2$ for $\theta \ge \sigma$. For

$0 \leq \theta < \sigma$, we have $\sigma(1+\lambda^2) \geq 2\theta$. For $x > 0$, we have $\tilde{\Phi}(x) < x^{-1}\phi(x)$, so $\tilde{\Phi}(-\lambda - \theta/\sigma) = 1 - \tilde{\Phi}(\lambda + \theta/\sigma) \geq 1 - (\lambda + \theta/\sigma)^{-1}\phi(\lambda + \theta/\sigma)$. It then follows from (C.6) that for $0 \leq \theta < \sigma$,

$$\begin{aligned} B'(\theta) &\geq 2\theta[\phi(\lambda+\theta/\sigma) - \phi(\lambda-\theta/\sigma) + \tilde{\Phi}(-\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)] \\ &\geq 2\theta[1 + (1-(\lambda+\theta/\sigma)^{-1})\phi(\lambda+\theta/\sigma) - \phi(\lambda-\theta/\sigma) - \tilde{\Phi}(\lambda-\theta/\sigma)] \\ &\geq 2\theta\left[1 + (1-(\lambda+\theta/\sigma)^{-1})\phi(\lambda+\theta/\sigma) - \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\right] \geq 0. \end{aligned}$$

Coupled with $B(0) = 0$, this implies that $B(\theta) \geq 0 \geq -\theta^2$ for $0 \leq \theta < \sigma$. Hence,

$$B(\theta) \geq -\theta^2, \qquad \forall \theta \geq 0. \tag{C.9}$$

Since $B(-\theta) = B(\theta)$, combining (C.8) and (C.9), we obtain $|B(\theta)| \leq \theta^2$ for all $\theta \in \mathbb{R}$. □

*Proof of Lemma 31.* Let $Z \sim N(0, \sigma^2)$, and let $\theta_0 = E[(Z^2 - \sigma^2)\mathbb{1}(Z^2 \leq \sigma^2\tau)] = -2\sigma^2\tau^{1/2}\phi(\tau^{1/2})$, the second equality due to Lemma 33. It follows from the expression

$$\begin{aligned} &E[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2 \vee Y^2 > \sigma^2\tau)] \\ &\quad = \mu^2\theta^2 - E[(X^2-\sigma^2)\mathbb{1}(X^2 \leq \sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \leq \sigma^2\tau)] \end{aligned}$$

and

$$\begin{aligned} \eta &= E[(Z_1^2-\sigma^2)(Z_2^2-\sigma^2)\mathbb{1}(Z_1^2 \vee Z_2^2 > \sigma^2\tau)] \\ &= -E[(Z_1^2-\sigma^2)\mathbb{1}(Z_1^2 \leq \sigma^2\tau)]E[(Z_2^2-\sigma^2)\mathbb{1}(Z_2^2 \leq \sigma^2\tau)] = -\theta_0^2 \end{aligned}$$

that we have

$$\begin{aligned}
&\left|E[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2\vee Y^2>\sigma^2\tau)]-\eta-\mu^2\theta^2\right| \\
&=\left|E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]-\theta_0^2\right|. \qquad \text{(C.10)}
\end{aligned}$$

Using the decomposition $AB-ab=(A-a)(B-b)+a(B-b)+b(A-a)$ and the triangle inequality, we get

$$\begin{aligned}
&\left|E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]-\theta_0^2\right| \\
&\le\left|E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)]-\theta_0\right|\left|E[(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]-\theta_0\right| \\
&\quad+\left|\theta_0\right|\left|E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)]-\theta_0\right|+\left|\theta_0\right|\left|E[(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]-\theta_0\right| \\
&\le\min\{\mu^2,3\sigma^2\tau\}\min\{\theta^2,3\sigma^2\tau\}+2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\mu^2,3\sigma^2\tau\} \\
&\quad+2\sigma^2\tau^{1/2}\phi(\tau^{1/2})\min\{\theta^2,3\sigma^2\tau\},
\end{aligned}$$

the last inequality follows from Lemma 34 and substitution of the value of $\theta_0$. □

**Proof of Lemma 32**

We have

$$\begin{aligned}
&\mathrm{Var}\left[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2\vee Y^2>\sigma^2\tau)\right] \\
&=E[(X^2-\sigma^2)^2(Y^2-\sigma^2)^2\mathbb{1}(X^2\vee Y^2>\sigma^2\tau)] \\
&\quad-\left\{E[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2\vee Y^2>\sigma^2\tau)]\right\}^2 \\
&=E[(X^2-\sigma^2)^2(Y^2-\sigma^2)^2]-E[(X^2-\sigma^2)^2\mathbb{1}(X^2\le\sigma^2\tau)(Y^2-\sigma^2)^2\mathbb{1}(Y^2\le\sigma^2\tau)] \\
&\quad-\left\{E[(X^2-\sigma^2)(Y^2-\sigma^2)]-E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]\right\}^2 \\
&=\mathrm{Var}\left[(X^2-\sigma^2)(Y^2-\sigma^2)\right]-E[(X^2-\sigma^2)^2\mathbb{1}(X^2\le\sigma^2\tau)]E[(Y^2-\sigma^2)^2\mathbb{1}(Y^2\le\sigma^2\tau)] \\
&\quad-\left\{E[(X^2-\sigma^2)\mathbb{1}(X^2\le\sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2\le\sigma^2\tau)]\right\}^2
\end{aligned}$$

$$
\begin{aligned}
&\quad + 2\mu^2\theta^2 E[(X^2-\sigma^2)\mathbb{1}(X^2 \le \sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] \\
&\le \mathrm{Var}\,[(X^2-\sigma^2)(Y^2-\sigma^2)] \\
&\quad + 2\mu^2\theta^2 E[(X^2-\sigma^2)\mathbb{1}(X^2 \le \sigma^2\tau)]E[(Y^2-\sigma^2)\mathbb{1}(Y^2 \le \sigma^2\tau)] \\
&\le \mathrm{Var}\,[(X^2-\sigma^2)(Y^2-\sigma^2)] + 8\sigma^4\mu^2\theta^2\tau^2.
\end{aligned}
$$

Straightforward calculation yields

$$
\begin{aligned}
&\mathrm{Var}\,[(X^2-\sigma^2)(Y^2-\sigma^2)] \\
&= \mathrm{Var}\,(X^2-\sigma^2)\mathrm{Var}\,(Y^2-\sigma^2) \\
&\qquad + [E(X^2-\sigma^2)]^2\mathrm{Var}\,(Y^2-\sigma^2) + \mathrm{Var}\,(X^2-\sigma^2)[E(Y^2-\sigma^2)]^2 \\
&= [4\sigma^2\mu^2+2\sigma^4][4\sigma^2\theta^2+2\sigma^4] + \mu^4[4\sigma^2\theta^2+2\sigma^4] + \theta^4[4\sigma^2\mu^2+2\sigma^4] \\
&= 4\sigma^2\mu^4\theta^2 + 4\sigma^2\mu^2\theta^4 + 16\sigma^4\mu^2\theta^2 + 2\sigma^4\mu^4 + 2\sigma^4\theta^4 + 8\sigma^6\mu^2 + 8\sigma^6\theta^2 + 4\sigma^8.
\end{aligned}
$$

Let $d = E[(Z_1^2-\sigma^2)^4(Z_2^2-\sigma^2)^4] < \infty$. Then

$$
\begin{aligned}
&\mathrm{Var}\,[(X^2-\sigma^2)(Y^2-\sigma^2)\mathbb{1}(X^2 \vee Y^2 > \sigma^2\tau)] \\
&\le E[(X^2-\sigma^2)^2(Y^2-\sigma^2)^2\mathbb{1}(X^2 \vee Y^2 > \sigma^2\tau)] \\
&\le \Big(E[(X^2-\sigma^2)^4(Y^2-\sigma^2)^4]P(X^2 \vee Y^2 > \sigma^2\tau)\Big)^{1/2} \\
&= d^{1/2}\Big(1-P(|Z| \le \tau^{1/2})^2\Big)^{1/2}, \qquad \text{where } Z \sim N(0,1) \\
&\le (2d)^{1/2}\Big(1-P(|Z| \le \tau^{1/2})\Big)^{1/2} \\
&= 2d^{1/2}\tilde{\Phi}(\tau^{1/2})^{1/2},
\end{aligned}
$$

the second inequality follows from the Cauchy-Schwarz inequality.

# Bibliography


Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24(3), 441–461.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.* (pp. 1–198). Springer, Berlin.

Alqallaf, F., van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Ann. Statist.*, 37(1), 311–331.

Alqallaf, F. A., Konis, K. P., Martin, R. D., & Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 14–23).: ACM.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68, 337–404.

Bach, F. R. & Jordan, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.*, 3(1), 1–48.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186, 273–289.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9, 485–516.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5), 577–606.

Baraud, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2), 528–551.

Bickel, P. J. (1964). On some alternative estimates for shift in the $p$-variate one sample problem. *Ann. Math. Statist.*, 35, 1079–1090.

Bickel, P. J. & Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, 36(6), 2577–2604.

Bickel, P. J. & Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A*, 50(3), 381–393.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, 80(391), 580–619. With discussion and with a reply by the authors.

Brown, L. D. & Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6), 2524–2535.

Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ace. *Ann. Statist.*, 18(3), 1032–1069.

Buja, A., Hastie, T. J., & Tibshirani, R. J. (1989). Linear smoothers and additive models. *Ann. Statist.*, 17(2), 453–520. With discussions by Breiman, Chen/Gu/Wahba, D.D.Cox, Eubank/Speckman, Gander/Golub, Gasser/Kneip, Kohn/Ansley, Titterington, and a rejoinder by the authors.

Cai, T. & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, 106(496), 1566–1577.

Cai, T., Liu, W., & Luo, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494), 594–607.

Cai, T. T. & Low, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.*, 32(2), 552–576.

Cai, T. T. & Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.*, 33(6), 2930–2956.

Cai, T. T. & Low, M. G. (2006a). Adaptive confidence balls. *Ann. Statist.*, 34(1), 202–228.

Cai, T. T. & Low, M. G. (2006b). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, 34(5), 2298–2325.

Cai, T. T., Ren, Z., & Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields*, 156(1-2), 101–143.

Cai, T. T., Zhang, C.-H., & Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4), 2118–2144.

Cai, T. T. & Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, 40(5), 2389–2420.

Cardoso-Cachopo, A. (2007). Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa. http://ana.cachopo.org/datasets-for-single-label-text-categorization.

Chen, M., Gao, C., & Ren, Z. (2015). Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*.

Collier, O., Comminges, L., & Tsybakov, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv preprint arXiv:1502.00665*.

Consortium, S. P. G.-W. A. S. G. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10), 969–976.

Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., Behrens, T., Cho, J., et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8), e1002254.

Croux, C. & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, 19(4), 497–515.

Donnell, D. J., Buja, A., & Stuetzle, W. (1994). Analysis of additive dependencies and concurvities using smallest additive principal components. *Ann. Statist.*, 22(4), 1635–1673. With a discussion by Bernard D. Flury [Bernhard Flury] and a rejoinder by the authors.

Donoho, D. & Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser. (pp. 157–184). Wadsworth, Belmont, CA.

Donoho, D. & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3), 962–994.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators.* Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455.

Donoho, D. L. & Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity*, 6(3), 290–323.

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.*, 26(1), 288–314.

Efromovich, S. & Low, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, 24(3), 1106–1125.

Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.*, 19(3), 1273–1294.

Fan, J., Liu, H., Ning, Y., & Zou, H. (2014). High dimensional semiparametric latent graphical model for mixed data. *arXiv preprint arXiv:1404.7236*.

Fan, J., Liu, H., & Wang, W. (2015). Large covariance estimation through elliptical factor models. *arXiv preprint arXiv:1507.08377*.

Finegold, M. & Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative $t$-distributions. *Ann. Appl. Stat.*, 5(2A), 1057–1080.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

Fukumizu, K., Bach, F. R., & Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.*, 8, 361–383.

Genovese, C. R. & Wasserman, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.*, 33(2), 698–729.

Giné, E. & Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli*, (pp. 47–61).

Golub, G. H. & Van Loan, C. F. (2013). *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1), 1–15.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (2011). *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69, 383–393.

Han, F. & Liu, H. (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916*.

Han, F. & Liu, H. (2014). Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *J. Amer. Statist. Assoc.*, 109(505), 275–287.

Han, F., Lu, J., & Liu, H. (2015). *Robust scatter matrix estimation for high dimensional distributions with heavy tails*. Technical report, Technical report, Princeton University.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability 43. Boca Raton: Chapman & Hall / CRC, reprint 1999 edition.

Higham, N. J. (2002). Computing the nearest correlation matrixa problem from finance. *IMA journal of Numerical Analysis*, 22(3), 329–343.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19, 293–325.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., & Sustik, M. A. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 2330–2338). Curran Associates, Inc.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35, 73–101.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Ingster, Y. I. & Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer Series in Statistics. New York: Springer-Verlag, second edition.

Kendall, M. G. (1948). *Rank correlation methods*. Griffin.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284), 814–861.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lepski, O. V. & Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2), 333–358.

Leurgans, S. E., Moyeed, R. A., & Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B*, 55(3), 725–740.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3), 1001–1008.

Li, Y. R., Zhao, S. D., Li, J., Bradfield, J. P., Mohebnasab, M., Steel, L., Kobie, J., Abrams, D. J., Mentch, F. D., Glessner, J. T., et al. (2015). Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nature communications*, 6.

Little, R. J. A. & Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc.

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4), 2293–2326.

Maronna, R. A. (1976). Robust $M$-estimators of multivariate location and scatter. *Ann. Statist.*, 4(1), 51–67.

Maronna, R. A. & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307–317.

Oellerer, V. & Croux, C. (2014). Robust high-dimensional precision matrix estimation. *Available at SSRN 2528996*.

Puri, M. L. & Sen, P. K. (1971). *Nonparametric methods in multivariate analysis*. John Wiley & Sons, Inc., New York-London-Sydney.

Rankinen, T., Sarzynski, M. A., Ghosh, S., & Bouchard, C. (2015). Are there genetic paths common to obesity, cardiovascular disease outcomes, and cardiovascular risk factors? *Circulation research*, 116(5), 909–922.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 71(5), 1009–1030.

Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Statist.*, 5, 935–980.

Reed, M. & Simon, B. (1980). *Methods of modern mathematical physics. I.* New York: Academic Press Inc. [Harcourt Brace Jovanovich Publishers], second edition. Functional analysis.

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.*, 43(3), 991–1026.

Rice, J. A. & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, 53(1), 233–243.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)* (pp. 283–297). Reidel, Dordrecht.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388), 871–880.

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.*, 88(424), 1273–1283.

Schölkopf, B. & Smola, A. J. (2002). *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts; London, England: MIT Press, first edition.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comp.*, 10(3), 1299–1319.

Serfling, R. & Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statist. Probab. Lett.*, 79(16), 1767–1773.

Shevlyakov, G. & Vilchevski, N. (2002). *Robustness in Data Analysis: Criteria and Methods*. Modern Probability and Statistics, 6. VSP.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1), 1–24.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., & Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5), 607–618.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., & Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2), 875–891.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Steinwart, I. & Christmann, A. (2008). *Support vector machines*. Information Science and Statistics. Springer, New York.

Swanson, D. (2000). *Signal Processing for Intelligent Sensor Systems*. Signal Processing and Communications. CRC Press.

Tan, K. M., Witten, D., & Shojaie, A. (2015). The cluster graphical lasso for improved estimation of Gaussian graphical models. *Comput. Statist. Data Anal.*, 85, 23–36.

Tarr, G., Müller, S., & Weber, N. C. (2015). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520–525.

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.*, 33, 1–67.

Van Aelst, S. (2016). Stahel-Donoho estimation for high-dimensional data. *International Journal of Computer Mathematics*, 93(4), 628–639.

Vu, V. Q., Cho, J., Lei, J., & Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems* (pp. 2670–2678).

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Wegkamp, M. & Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2), 1184–1226.

Werhli, A. V., Grzegorczyk, M., & Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20), 2523–2531.

Xue, L. & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5), 2541–2571.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1), 19–35.

Zhang, F. (2011). *Matrix Theory: Basic Results and Techniques*. Universitext. Springer.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The `huge` package for high-dimensional undirected graph estimation in `r`. *J. Mach. Learn. Res.*, 13, 1059–1062.