On the Equivalence of Sparse Statistical Problems

by

Sung Min Park

B.S. Cornell University (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

September 2016

Signature redacted

SIGNATURE OF AUTHOR: _____

Department of Electrical Engineering and Computer Science
August 30, 2016

Signature redacted

CERTIFIED BY: _____

Guy Bresler
Assistant Professor of Electrical Engineering and Computer Science

Signature redacted

ACCEPTED BY: _____

Leslie A. Kolodziejski
Chair of the Committee on Graduate Students,
Electrical Engineering and Computer Science

On the Equivalence of Sparse Statistical Problems

by

Sung Min Park

Submitted to the Department of Electrical Engineering and Computer Science on August 30, 2016 in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science

Abstract

Sparsity is a widely used and theoretically well understood notion that has allowed inference to be statistically and computationally possible in the high-dimensional setting.

Sparse Principal Component Analysis (SPCA) and Sparse Linear Regression (SLR) are two problems that have a wide range of applications and have attracted a tremendous amount of attention in the last two decades as canonical examples of statistical problems in high dimension. A variety of algorithms have been proposed for both SPCA and SLR, but their literature has been disjoint for the most part. We have a fairly good understanding of conditions and regimes under which these algorithms succeed. But is there be a deeper connection between computational structure of SPCA and SLR?

In this paper we show how to efficiently transform a blackbox solver for SLR into an algorithm for SPCA. Assuming the SLR solver satisfies prediction error guarantees achieved by existing efficient algorithms such as those based on the Lasso, we show that the SPCA algorithm derived from it achieves state of the art performance, matching guarantees for testing and for support recovery under the single spiked covariance model as obtained by the current best polynomial-time algorithms. Our reduction not only highlights the inherent similarity between the two problems, but also, from a practical standpoint, it allows one to obtain a collection of algorithms for SPCA directly from known algorithms for SLR. Experiments on simulated data show that these algorithms perform well.

Thesis Supervisor: Guy Bresler
Title: Assistant Professor

Acknowledgements

First, I would like to thank my advisor Guy: for welcoming me when I was a lost 2nd year student, for continuous optimism and encouragement–refilling me with hope even in the days when I showed up with nothing, for giving me the right tools, ideas, and ways to think that I came to appreciate only later, and for showing me a little bit of what theoretical research is like. Most of all, I want to thank Guy for reminding me to enjoy the challenge of research and also to have a more positive, playful outlook on life beyond research.

Next, I want to thank my collaborator and friend, Mădălina Persu, who I greatly enjoyed working with over the past year. Thank you for encouraging and motivating me to work harder.

Over the past two years I tried a lot of new things: volleyball, soccer, lifting, skiing, skating, squash, acoustic, cooking, and generally how to be a more interesting person. Thanks to everyone who were part of these memories!

And last but not least, to my family; parents, Ki-On and Eun-Suh, and sister, Jane: Thank you for all your hard work and love throughout the years, and for always believing in and seeing the best in me.

Contents

List of Symbols

Σ	covariance matrix
$\widehat{\Sigma}$	sample covariance matrix
$\mathsf{E}[\cdot]$	expectation over the appropriate sample space
$\mathcal{N}(\mu, \sigma)$	Gaussian distribution with mean vector μ and covariance matrix σ
I_n	$n \times n$ identity matrix
$\mathrm{diag}\{d_1, ..., d_n\}$	diagonal matrix with diagonal entries d_i
$\lesssim, \gtrsim$	inequality up to an absolute constant
$\mathbb{1}\{A\}$	indicator function of the event A
$\mathbb{S}^n$	n-dimensional unit sphere in $\mathbb{R}^{n+1}$
$\mathsf{B}_0(k)$	the set of k-sparse vectors in $\subset \mathbb{R}^d$
$[n]$	$\{1, ..., n\}$
w.p.	"with probability"
w.h.p.	"with high probability"

Chapter 1

Introduction

In modern datasets, we often work in the so called high-dimensional regime, where the dimensionality of the data may be significantly larger than the number of samples. In such settings, classical statistical tools break down, and we need additional assumptions on the data such as sparsity to make the problem statistically or computationally tractable.

Linear regression and principal components analysis have both been widely studied as fundamental problems in supervised learning and unsupervised learning, respectively. In the past decade, their sparse variants have been widely studied in the high-dimensional setting. Numerous techniques and algorithms have been developed for both problems, and conditions necessary or sufficient for the success of these algorithms are well understood.

Beyond the fact that sparsity is a common assumption used in both, could there be a deeper connection between the two problems? Does the difficulty of the two problems arise from the same structural reason? Can one use a blackbox solver for one to solve the other? This thesis makes a step forward in answering the above questions.

1.1 High-dimensional statistics

In classical statistics, we focus on the asymptotic regime where the number of samples n tend to infinity while other parameters are fixed. Classical estimators such as the maximum likelihood estimator are consistent; that is, the sample estimate of a parameter converges to the true population value as we acquire more samples.

In many modern applications, however, the number of samples we have access to is far less than the dimensionality of the data. Hence, it is often unreasonable to assume we have a lot more samples than the number of dimensions. This motivates the study of statistical problems in the high-dimensional setting.

Working in high-dimensions is both a curse and a blessing ([Wai10]). On one hand, exponential blowup in sample complexity or runtime is inevitable in certain cases (the so called "curse of dimensionality"). But on the other, phenomena such as concentration of measure working for

us enable inference under appropriate assumptions.

1.1.1 Sparsity

Due to the curse of dimensionality, problems in high-dimensional settings are intractable without additional assumptions. Often we impose a low-dimensional structure on our models. This not only allows inference to be statistically or computationally possible, but also is guided by our experience that high-dimensional data is often well explained by a much lower dimensional structure.

The need for low dimensional models is further elaborated by rigorous geometric intuition. The survey by [Ver15] gives a beautiful geometric perspective for estimation of high-dimensional signals constrained to a feasible region, unifying a number of results in this area. A general class of efficient convex programs succeed at recovery when the number of samples is on the order of the "effective dimension" of the lower-dimensional feasible region. This low-dimensional structure may be some form of sparsity, or low rank for matrices, for instance.

Sparsity is a simple and natural assumption for many problems, and has been extensively analyzed in theory and applied in practice. Sparse models have low ambient dimension, in the following senses. The set of sparse vectors, i.e., those with few non-zero entries, has low effective dimension as measured by the *Gaussian mean width*. The concept class of sparse linear classifiers also has low VC dimension ([Ney06]). As the ambient dimensionality of these models is low, it plausible that we can recover the parameters of the models using much fewer number of samples. Aside from the mathematical usefulness of such an assumption, real-world data are often sparse in an appropriate basis. For instance, natural images are known be approximately sparse in alternate bases such as wavelet or Fourier, and this fact is used by several compression schemes. In summary, sparse models have proven to be powerful both in theory and practice. See [EK12] or similar for a more extensive history.

We may view the success of sparse models in the light of Occam's razor: that among equivalent explanations, the simplest is best. In the case of linear regression, it is reasonable to expect that only a few of the covariates affect the response variable.

1.1.2 Statistical vs. computational tradeoffs

Despite the flurry of theoretical progress in high-dimensional estimation tasks, our understanding is still lacking in some aspects. For some problems, there still remains a gap between statistically optimal algorithms and known efficient algorithms; the former is often based on a brute-force search over model parameters, while the latter utilizes various convex relaxations and greedy heuristics. In other cases, computationally efficient algorithms require certain restrictive assumptions on the input, and the only known proof techniques rely crucially on those assumptions. Without those assumptions, a much higher signal strength is required in order to do inference–as far as we know. Is there a statistical price to pay for computational efficiency?

In recent years, new evidence for separation between statistical and computational thresholds for specific problems such as Sparse Principal Component Analysis (SPCA) and Submatrix Detection has emerged. But this raises further questions; do these gaps in different sparse problems indicate a common structure? Are different algorithms for these problems actually that different? At a very high level, the hardness on these different sparse problems does seem to arise from the same basic structural reason: the minimax optimal estimators for Sparse Linear Regression (SLR), SPCA, and Submatrix Detection all involve a brute force search over the family of sparse parameters (which is an exponentially large set). One of our goals is to formalize the above intuition.

Average-case hardness This evidence has relied on average-case complexity assumptions. Most known results in complexity work with worst case inputs. This is not the appropriate setting to study statistical problems as they involve data which is generated from processes that are inherently random.

The work on reducing from average-case hard problems to statistical problems was pioneered by [BR13a], and inspired the work in this thesis. They assume the hardness of finding small planted cliques to show that any randomized polynomial time tests must fail if the signal θ is below a certain threshold; this is a constant factor within the threshold at which their SDP relaxation succeeds. Soon after, [MW15] similarly reduced from planted clique showed that submatrix detection is statistically possible but computationally infeasible in a certain regime of parameters. Recently, average-case certification of Restricted Isometery Property, a property of high interest in compressed sensing and statistical learning, was shown to computationally hard based on the hardness of detecting dense subgraphs ([WBP16]), a slightly milder variant of the planted clique hypothesis used in previous works.

The work in thesis was initiated by the goal to show that SLR and SPCA are equivalently hard via a blackbox reduction. While such is not known yet, we make partial progress by giving a blackbox reduction that works in the efficient regime: given a SPCA instance in the computationally tractable regime[1], we solve it using blackbox accesses to an SLR solver.

1.2 Our contributions

We highlight some of our contributions below:

- We give a general and efficient procedure for transforming an SLR blackbox with prediction error guarantees into algorithms for hypothesis testing and support recovery for SPCA under the spiked covariance model. Most known sparse linear regression and sparse recovery algorithms can be used as this blackbox. In experiments, we demonstrate that

[1]In that sense that we already know other direct algorithms that solve SPCA in this regime, and we have evidence that the threshold is tight.

using popular existing SLR algorithms such as Lasso [Tib96] and FoBa [Zha09] for the "blackbox" results in good performance.

- For hypothesis testing, we match state of the art provable guarantee for computationally efficient algorithms; our algorithm successfully distinguishes between isotropic and spiked Gaussian distributions as soon as the signal strength is greater than $\theta \gtrsim \sqrt{\frac{k^2 \log d}{n}}$. This matches the phase transition of diagonal thresholding (DT) [JL09] and Minimal Dual Perturbation (MDP) [BR13b] up to constant factors.

- For support recovery: for general p and n, when each non-zero entry of u is at least $\Omega(1/\sqrt{k})$ (a standard assumption in the literature), our algorithm succeeds with high probability for signal strength $\theta \gtrsim \sqrt{\frac{k^2 \log d}{n}}$. In the scaling limit $d/n \to \alpha$ as $d, n \to \infty$, the recent covariance thresholding algorithm [DM14], theoretically succeeds at a signal strength that is an order of $\sqrt{\log d}$ smaller. However, our experimental results indicate that with an appropriate choice of blackbox, our Q algorithm outperforms covariance thresholding as well as diagonal thresholding.

- We also theoretically and empirically illustrate that our SPCA algorithm is robust to rescaling data, for instance by using a Pearson correlation matrix instead of a covariance matrix. [2]

1.3 Organization

The rest of the thesis is organized as follows. In Chapter 2, we give background on the history and existing literature of SPCA and SLR, with more focus on SPCA. In Chapter 3, we define clear formulations of both problems that will be used in the rest of the thesis. In Chapter 4, we show the reduction and its analysis. In Chapter 5, we present empirical evaluation of our algorithms. In Chapter 6, we conclude with some future directions.

[2] We remark that the idea of seeing if a SPCA algorithm works on the correlation matrix was originally found in [VCLR13].

Chapter 2

Background

We review the historical development and existing literature of Sparse Principal Component Analysis (SPCA) and Sparse Linear Regression (SLR). [1]

2.1 Sparse Principal Component Analysis

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction and compression. PCA is used to project vector data onto a lower dimensional subspace while minimizing reconstruction error in a least squares sense. This subspace is spanned by a few orthogonal directions of maximum variance, called principal components, which are linear combinations of the original variables. The weight of original variables in a principal component are referred to as *loadings*. Principal components also have the advantage that different components are uncorrelated.

The top principal components correspond to top eigenvectors of the covariance matrix, so they can be estimated by computing the eigenvalue (singular value) decomposition of the sample covariance (data) matrix.

More formally, we are given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where d is the input dimension and n is the number of samples. Assume that the data has been appropriately centered to have zero mean. Let Σ denote the population covariance matrix, and $\widehat{\Sigma}$ its sample counterpart. For now, let us focus on just the first principal component. The goal is to solve the following optimization problem:

$$\underset{\|u\|_2=1}{\operatorname{argmax}}\, u^\top \Sigma u$$

If $\widehat{\Sigma}$ is close to Σ in spectral norm[2], then we expect its largest eigenvector to be close to that

[1] While the background on SPCA is more extensive than necessary to understand the context of our reduction, we aim to provide a short comprehensive summary as no satisfactory one was found in the literature.

[2] Spectral norm of a matrix is its largest singular value.

of Σ if the corresponding eigenvalue is unique (and we assume so in rest of this subsection). In the classical setting where the number of samples n tends to infinity for fixed d, the consistency of estimator $\widehat{\Sigma}$ and the continuity of the largest eigenvalue as a function of the matrix entries[3] imply that we can recover the top principal component.

However, in the high-dimensional setting, when d can be (much) larger than n, such methods lead to inconsistent estimates. This is true even when d and n are of the same order. Consider the case when the population covariance matrix is the identity, so its top eigenvalue is just 1. We know that if $p/n \to \alpha > 0$ ([Gem80]),

$$\lambda_{\max}(\widehat{\Sigma}) \to (1 + \sqrt{\alpha})^2 > 1$$

It follows that if the maximum eigenvalue of the population covariance matrix is between 1 and $(1 + \sqrt{\alpha})^2$, the top eigenvector of the sample covariance matrix is not a consistent estimate of the first principal component. Moreover, when $d/n \to \infty$, we have $\lambda_{\max}(\widehat{\Sigma}) \to \infty$, so the above estimate is generally unreliable in the high-dimensional setting.

Sparse PCA As seen above, in the high-dimensional setting one cannot hope to do PCA without additional assumptions. Aside from statistical limitations, principal components that are linear combinations of all of the original variables are also hard to interpret. This motivates the addition of a sparsity constraint, limiting the number of nonzero loadings in a principal component. As the original variables usually have some physical meaning (ex. the expression levels of different genes), a principal component with a few nonzero loadings can be more interpretable. Notice that there is a tradeoff between *statistical fidelity* and *interpretability*. While having fewer nonzero entries aids better physical interpretation, it comes at the price of lower explained variance.

Before the notion of sparse PCA was formalized, ad-hoc methods were used to post-process the ordinary principal components, for instance by simple thresholding of ordinary principal components. But this can be misleading as just looking at the magnitude of the loadings does not take into account of variance or correlation structure between variables, and in general truncating may not yield the best approximation to the original principal component ([CJ95]).

2.1.1 Existing algorithms for sparse PCA

Similar notions of sparse PCA were initially introduced in various works in order to remedy the issues with the ordinary PCA in the high dimensional setting or to achieve sparsity as a goal in itself.

[3]This is because roots of a polynomial are continuous in its coefficients.

More formally, one formulation of sparse PCA is the following:

$$\underset{\|u\|_2=1, \|u\|_0 \leq k}{\operatorname{argmax}} \quad u^\top \widehat{\Sigma} u \tag{2.1}$$

where k is the sparsity or number of non-zeros. The value of the objective above is sometimes referred to as the *k-sparse eigenvalue*. It is not hard to see that the above problem is NP-hard by a trivial reduction to CLIQUE, which suggests that there is no efficient algorithm for worst-case inputs. This means one has to relax the ℓ_0 constraint to an ℓ_1 constraint, or impose certain additional distributional assumptions on the data in order to get guarantees. Below we highlight some of the earlier approaches to this problem.

ℓ_1 penalty Some of the earliest works used different forms of ℓ_1 penalty to heuristically tradeoff between sparsity and explained variance. [JTU03] proposed SCoTLASS (Simplified Component Technique LASSO), which imposes an ℓ_1 penalty on (2.1) to recover sparse loadings; they optimize the objective using a variant of projected gradient descent, which is computationally costly, and is further compounded by the need to optimize over the penalty parameter to control sparsity. [ZHT06] reformulated PCA as a particular regression and use the elastic net (combination of ℓ_2 and ℓ_1 penalty) to induce sparsity, and an algorithm based on alternating minimization was given for the nonconvex optimization problem but no provable guarantees were provided. [MWA05] gave a greedy algorithm with a forward and backward pass to maximize the k-sparse eigenvalue objective, as well as an exact branch-and-bound algorithm using eigenvalue bounds to guide the search.

Diagonal thresholding [JL09] uses a simple procedure called diagonal thresholding (DT) to select a subset of variables with the largest variance, and then runs ordinary PCA on the reduced set of variables. Somewhat surprisingly, this simple algorithm matches the best guarantees for hypothesis testing (and is nearly optimal for support recovery) under the spiked covariance model, as we discuss in more detail in the next section.

SDP relaxations Another way to relax the hard sparsity constraint is to formulate an SDP relaxation to (2.1). [dEGJL07] first introduced a natural SDP relaxation for the problem (DSPCA), and since then this SDP has been used and analyzed in numerous settings. The canonical SDP relaxation replaces $uu^\top$ with any positive semi-definite matrix and drops the rank-1 constraint, and also replaces the ℓ_0 constraint with an ℓ_1 constraint.

Power methods The power method is a popular algorithm for finding the top eigenvector of a given matrix A that simply iteratively applies matrix A to the current estimate. Adaptations of this method to SPCA have been studied.

The Truncated Power method (TPower) of [YZ13] performs very well in practice; this extremely simple iterative method can recover the top eigenvector (which is assumed to be sparse) in ℓ_2 norm, but this requires starting with a sufficiently close initial direction.

The Generalized Power method (GPower) of [JNRS10] applies a gradient based iterative procedure for maximizing convex functions on compact set to reformulations of the SPCA objective 2.1 with ℓ_1 penalty, but lack guarantees on quality of the solution as the gradient procedure can be only shown to reach a stationary point (not even a local maximum) in general.

2.1.2 Spiked covariance model

These earlier works did not give completely satisfactory provable guarantees on the quality of the solution found. As the most general SPCA problem is NP-hard, additional assumptions are needed in order to derive better guarantees. One popular and successfully analyzed distributional setting has been the *spiked covariance model.* We focus on the single spike model due to [Joh01].

In the spiked covariance model for r spikes, $X \in \mathbb{R}^d$ is generated by the formula:

$$X = VDU^\top + Z$$

where V is $n \times r$ random effects matrix with i.i.d. $\mathcal{N}(0,1)$ entries, $D = \mathrm{diag}(\lambda_1^{1/2}, ..., \lambda_r^{1/2})$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$, U is $d \times r$ orthonormal and Z has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries independent of V. Equivalently, X has rows independently drawn from $\mathcal{N}(0, \Sigma)$, where $\Sigma = U\Lambda U^\top + \sigma^2 I_d$ and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_r)$.

In this thesis, we focus on the single spike case, where we just have $U = u$. We introduce a signal strength parameter θ, defined as $\theta = \lambda_1/\sigma$. For the discussion below, we comment that some works treat signal strength θ as a variable parameter and compare against a fixed threshold, while others fix θ to be a constant and analyze the required number of samples n as a function of dimension d and sparsity k.

Below we review state of the art guarantees for two different goals.

Support recovery The goal of support recovery is to find the support S of the spike u. Notice that if we exactly recover the support of S, then we can recover u by finding the top eigenvector of the restricted covariance matrix $\Sigma_{S,S}$ because this puts us back in the low-dimensional or classical regime. In general, a lower bound on the entries of u is needed in order to guarantee successful recovery (see [FRG09, Wai07] for related lower bounds for sparse recovery). Under the spiked covariance model, for a subcase when the spike is uniform in all k coordinates, [AW09] analyzed both diagonal thresholding and SDP for support recovery. They showed that the SDP requires an order of k fewer samples when the SDP optimal solution is rank one. However, [KNV13] showed that the rank one condition does not happen in general, particularly in the regime approaching the information theoretic limit ($\sqrt{n} \lesssim k \lesssim \frac{n}{\log d}$). This is consistent with computational lower

bounds from [BR13a] ($k \gtrsim \sqrt{n}$), but a small gap remains (diagonal thresholding, SDP's succeed only up to $k \lesssim \sqrt{n/\log d}$). The state of the art for support recovery that closes the above gap is the covariance thresholding algorithm, first suggested by [KNV13] and analyzed by [DM14], that succeeds in the regime $\sqrt{n/\log d} \lesssim k \lesssim \sqrt{n}$, although the theoretical guarantee is limited to the regime when $d/n \to \alpha$ due to relying on techniques from random matrix theory.

Hypothesis testing Some works [BR13b, AW09, dBEG14] have focused on the problem of detection. Here one only wants to distinguish between $u = 0$ and $\|u\|_2 = 1$ (with u still k-sparse). In this case, [BR13b] observed that it suffices to work with the much simpler dual of the standard SDP, called Minimal Dual Perturbation (MDP). In the dual problem, the goal is to perturb the sample covariance matrix to minimize the max eigenvalue subject to a penalty proportional to the entries of the perturbation and the sparsity level k. Diagonal thresholding (DT) and MDP work up to the same signal threshold θ as for support recovery, but MDP seems to outperform DT on simulated data [BR13b]. Note that MDP works at the same signal threshold as the standard SDP relaxation for SPCA.

[dBEG14] analyze a statistic based on an SDP relaxation and its approximation ratio to the optimal statistic. In the regime where k, n are proportional to d, their statistic suceeds at a signal threshold for θ that is independent of d, unlike the MDP. However, their statistic is quite slow to compute; runtime is at least a high order polynomial in d.

2.1.3 Beyond the single spike covariance model

Multiple spikes While this thesis focuses on the single spike case, in practice it is obviously desirable to obtain multiple principal components. For finding more than one principal component, a popular strategy is to simply iterate the algorithm for finding one component, while using deflating techniques in between iterations to remove the contribution of existing components. The performance of such a procedure depends on the deflation technique used; see [Mac09] for more details.

Subspace recovery A contrasting strategy is to estimate the entire principal subspace spanned by the first r components at once. Settings vary in whether they allow supports to be disjoint or identical across different components. More recent work such as [Ma13] analyze minimax bounds for this problem. [CMW13] analyzed optimal rates for estimating the principal subspace along with an efficient adaptive procedure that works for a more restrictive set of parameters.

Alternative guarantees Aside from the spiked covariance model, different works have managed to give efficient algorithms by parameterizing the ambient dimension of the problem in different ways.

In [APKD15], they instead focus on finding sparse components with disjoint supports that jointly capture the maximum variance. Through a clever reformulation of the objective, they reduce the problem to multiple instances of bipartite maximum weight matching, and the algorithm's complexity is polynomial in the ambient dimension of the input data, though exponential in rank.

In [PDK13], they give an algorithm that guarantees good approximation as long the spectral profile of the covariance matrix has sufficient decay. Their algorithm first obtains Σ_r, a rank-r approximation of Σ, then uses Σ_r to obtain $O(n^r)$ candidates supports (which is potentially much fewer than the naive enumeration of $O(n^k)$).

[VCLR13] generalize the DSPCA approach of [dEGJL07] to recovering sparse principal subspaces by formulating an SDP with a Fantope constraint, and also give an efficient alternating direction method of multipliers (ADMM) algorithm to solve the SDP. Their subspace recovery guarantee is in terms of the spectral profile of the population covariance matrix. To the best of our knowledge, their work was also the first to point out that diagonal thresholding trivially fails after rescaling all variables to have equal variance, indirectly hinting at the limitation of the spiked covariance model.

Probabilistic approaches Some line of work has focused on probabilistic formulations of (sparse) PCA. PCA can be formulated as a maximum likelihood solution to a latent variable model ([TB99]). [SB08] gave an Expectation-Maximization procedure for finding principal components, where they add ℓ_1 or non-negativity constraints in the M-step to enforce sparsity or non-negativity. Because EM only guarantees local optima, the performance of their algorithm depends on good initialization. [GD09] provide a more complete Bayesian solution to sparse probabilistic PCA, using different priors to induce sparsity on the coefficients of the latent variables.

2.2 Sparse Linear Regression

Linear regression is one of the most fundamental statistical tools and a canonical problem in supervised learning. In linear regression, we observe a response vector $y \in \mathbb{R}^n$ and a design matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ that are linked by the linear model $y = \mathbb{X}\beta^* + w$. The vector $w \in \mathbb{R}^n$ is some form of observation noise, and our goal is to recover β^* given noisy observations y. We focus on the standard Gaussian model, where the entries of w are i.i.d. $\mathcal{N}(0, \sigma^2)$. We also work with deterministic design; while the matrices $\mathbb{X}$ we consider arise from a (random) correlated Gaussian design (as analyzed in [Wai07], [Wai09]), it will make no difference to assume the matrices are deterministic (by conditioning). Most of the relevant results on sparse linear regression pertain to deterministic design.

Analogous to the setting for PCA, linear regression in the high-dimensional setting is meaningless without further constraints. In general, when $n < d$, the system is under-determined and

there is a whole subspace of solutions minimizing reconstruction error.

In sparse linear regression, we additionally assume that β^* is sparse, or has only a small number, $k < d$, of non-zero entries. This makes the problem well posed in the high dimensional setting, though computationally more challenging. Beyond mathematical necessity, sparsity has been found to be a very suitable assumption, as often real world signals are suitable in an appropriate basis; that is, the intrinsic dimensionality of the data is often much lower than the dimensionality of the original dataset.

Commonly used performance measures for SLR are tailored to prediction error, support recovery (recovering support of β^*), or parameter estimation (estimating β^* under some norm). We focus on prediction error, defined as $\frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2$, and analyzed over random realizations of the noise.

The ℓ_0 estimator, which minimizes the reconstruction error $\|y - \mathbb{X}\widehat{\beta}\|_2^2$ over all k-sparse regression vectors, achieves prediction error bound of form ([BTW07a], [RWY11]):

$$\frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2 \lesssim \frac{\sigma^2 k \log d}{n}$$

The runtime of this estimator is $O(n^k)$, which is both theoretically and practically as soon as k is larger than a constant.

Efficient methods Various efficient methods have been proposed to circumvent the computational intractability of the above estimator: basis pursuit, Lasso[Tib96], and the Dantzig selector [CT07] are some of initial approaches. Greedy pursuit methods such as OMP [MZ93], IHT[BD09], CoSaMP[NT09], and FoBa[Zha09] among others offer more efficient alternatives. [4] Many of the optimization-based approaches relax the ℓ_0 penalty to some form of ℓ_1 penalty or an equivalent constraint. These algorithms achieve the same prediction error guarantee as ℓ_0 up to a constant, but under the assumption that $\mathbb{X}$ satisfies certain properties, such as restricted eigenvalue ([BRT09]), compatibility ([vdG07]), restricted isometry property ([CT05]), and (in)coherence ([BTW07b]). In this work, we focus on the restricted eigenvalue (see Definition 3.3 for a formal definition). We remark that restricted eigenvalue is among the weakest, and is only slightly stronger than the compatibility condition. Moreover, [ZWJ14] give complexity-theoretic evidence for the necessity of dependence on the RE constant for certain worst case instances of the design matrix. See [VDGB+09] for implications between various conditions. In the next subsection, we give more intuition for some of these conditions.

Slow rate Without such conditions on $\mathbb{X}$, the best known guarantees provably obtain only a $1/\sqrt{n}$ decay rather than a $1/n$ decay in prediction error as number of samples increase. [ZWJ15] give some evidence that this gap may be unavoidable by showing that the family of M-estimators

[4] Note that some of these algorithms were presented for compressed sensing; nonetheless, their guarantees can be converted appropriately.

based on minimizing the sum of squared loss and a coordinate-wise decomposable regularizer cannot achieve a rate better than $1/\sqrt{n}$.

Optimal estimators The SLR estimators we consider are efficiently computable. Another line of work considers arbitrary estimators that are not necessarily efficiently computable. These include BIC [BTW07a], Exponential Screening [RT11], and Q-aggregation [DRXZ14]. Such estimators achieve strong guarantees regarding minimax optimality in the form of oracle inequalities on MSE.

2.2.1 Properties of design matrix

We give some intuition for why certain properties of the design matrix are desirable and natural for signal recovery (though not necessarily in the sense of minimizing prediction error).

One commonly used property is *incoherence* and (its generalization) *restricted isometry property (RIP)*. The intuition behind incoherence is that we want the error due to under-sampling[5] to look roughly like noise. To put it another way, incoherence measures the tendency of linear reconstruction to leak energy from the true underlying source to other sources; we want to spread this out as uniformly as possible over all sources.

While incoherence just looks at pairs of vectors, RIP generalizes by looking at subsets of k vectors. Though it is hopeless for a $n \times d$ matrix to be well-conditioned[6] for $n < d$, RIP of k means that we only need the matrix to be well-conditioned if we look at any submatrix spanned by k columns.

It turns out that a much weaker condition such as *restricted eigenvalue* suffices for the guarantee of certain SLR algorithms. RE says that the design matrix has bounded eigenvalue in a restricted set of directions. In fact, even more generally, this property corresponds to *restricted strong convexity* for real-valued functions. Strong convexity here means that the Hessian of the function we are optimizing is strictly positive; this implies the function being sufficiently well-conditioned. Restricted means we only need strong convexity to hold in certain restricted set of directions; usually what suffices is the cone of directions spanned by "roughly" sparse[7] vectors. In the case of SLR, strong convexity of the ℓ_2 reconstruction error is exactly equivalent to the design matrix having a restricted eigenvalue.

It turns out RSC together with a property called *decomposability* of the regularizer is sufficient to imply very general results on the performance of certain class of M-estimators for high-dimensional statistical tasks with a low-dimensional structure. See [NYWR09] for a lengthier discussion and general results along this line.

[5]This expression is from compressed sensing, but basically means the same high-dimensional setting we have been discussing when the linear system is under-determined.

[6]Generally, a function is said to be well-conditioned if output value varies less relative to the change in input value; condition number is also the ratio or largest to smallest singular value.

[7]In the sense that ℓ_1-weight outside a sparse support is not much larger than the ℓ_1-weight on the support

2.2.2 Connections to Compressed Sensing

SLR is very closely related to compressed sensing (CS), which studies the recovery of signals based on a small number of measurements when the signal is known to be (approximately) sparse in some suitable basis. CS grew out of a sequence of seminal papers (the core theory was developed in [CRT06b, CRT06a, Don06], but some its ideas were hinted in [FB96, BF96, VB98, VMB02, CDS01]), and since then its ideas and related techniques have been taken up by researchers in various communities spanning statistics, machine learning, theoretical computer science, and information theory. The main difference in the settings of SLR and CS is that in SLR the design matrix is a given, with less control on the distributional or independence properties, while in CS the goal is to *design* a measurement matrix with similar yet usually stronger properties, often with its design influenced by physical constraints such as those arising from magnetic resonance imaging (MRI). Recovery objective also varies depending on the community; in signal processing, the goal is often to recover the signal up to some error: in SLR, prediction error and variable selection, which are weaker and (usually) stronger, respectively, are also popular.

2.3 Prior work

Despite the similarity of SPCA and SLR, the literature for each problem has been mostly disjoint. Some prior work has also drawn connections between SPCA and SLR, though in ways different from the work in this thesis.

Regression based approaches Though some previous works ([ZHT06]) have used specific algorithms for SLR such as Lasso as a subroutine, to the best of our knowledge our work is the first to give a general framework that uses SLR in a blackbox fashion, while matching state of the art theoretical guarantees.

A similar regression-based approach has been used in [MB06] as applied to a restricted class of graphical models. The goal there is to recover the neighborhood of each node in a graphical model; they do is by regressing the random variable corresponding each node on the observations from the rest of the graph. While our regression setup is similar, their statistic is different and their analysis depends directly on the particulars of Lasso. Further, their algorithm requires extraneous conditions on the data. In particular, their Assumption 5 on minimum partial correlation requires $\theta^2 \gtrsim k$, compared to $\theta^2 \gtrsim \frac{k^2 \log d}{n}$ in our work.

[CMW13] also uses reduction to linear regression for their sparse subspace estimation, but is different from our approach in several ways: First, their algorithm depends crucially on a good initialization done by a diagonal thresholding-like pre-processing step, whereas our algorithm does not. This further implies that under rescaling of data[8], their initialization fails. Second,

[8] See Section 4.4.3 for more discussion on rescaling.

their framework uses regression for the specific case of orthogonal design, whereas our design matrix can be more general as long as it satisfies a condition similar to the restricted eigenvalue condition. On the other hand, their setup allows for more general ℓ_q-based sparsity as well as the estimation of an entire subspace as opposed to a single component.

Sparsity inducing priors Connections between SPCA and SLR has been noted in the probabilistic setting, albeit in an indirect manner. At a high level, the same sparsity-inducing priors can be used for either problem.

[KKGP14] consider the problem of given a base prior, finding another distribution closest to it in KL divergence ("information projection") while satisfying some constraints. They look at domain constraints (limiting domain to a particular subset) in particular, and show that the desired optimal distribution is just the base distribution restricted to the subset and rescaled appropriately. Now, if we want to do information projection on all the distributions that k-sparse support S, then it turns out that the cost function (KL divergence) is submodular in S, so one can achieve $(1 - 1/e)$-approximation to the optimal objective.

[KGPK15] look at the probabilistic formulation of PCA along with the EM algorithm for it. In the E-step, they optimize over distribution Q that has sparse support. Since the E-step minimizes KL divergence between the distribution of latent variables (principal components) and Q, the technique from [KKGP14] can be readily applied. However, as based on EM they can only guarantee local optimality.

Beyond SPCA and SLR Beyond SPCA and SLR, a wide variety of models have been studied in the high-dimensional setting. Notably, many of the approaches share a common pool of techniques and analyses. This was formalized in the work of [NYWR09], who gave a general framework and explanation for why a family of M-estimators with an appropriate loss function and regularizer has been so successful in a variety of high-dimensional statistical tasks with low dimensional structure, including sparse regression, structured covariance estimation, low rank matrix approximation, sparse principal component analysis, and discrete Markov random fields.

Chapter 3

Preliminaries

We give the precise setup for SPCA and SLR that we will study, and review some notation that is used throughout the thesis.

3.1 Problem formulation for SPCA

Hypothesis testing Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be n i.i.d. copies of a Gaussian random variable X in $\mathbb{R}^d$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix whose rows are $X^{(i)}$. The objective of the SPCA detection problem is to distinguish whether there is some distinguished (sparse) direction u along which X has higher variance. In SPCA, we also assume that this direction is sparse. This motivates the following null and alternate hypotheses:

$$H_0 : X \sim \mathcal{N}(0, I_d) \text{ and } H_1 : X \sim \mathcal{N}(0, I_d + \theta uu^\top),$$

where u has unit norm and at most k nonzero entries. The distribution under H_1 is known as the *spiked covariance* model. As smaller θ makes the problem only harder, we assume $\theta \leq 1$ for ease of computation and as standard in literature.

We say that a test discriminates between H_0 and H_1 with probability $1 - \delta$ if both type I and II errors have a probability smaller than δ. The goal is therefore to find a statistic $\phi(\mathbf{X})$ and a threshold τ depending on d, n, k, δ such that for the test $\psi(\mathbf{X}) = \mathbb{1}\{\phi(\mathbf{X}) > \tau\}$

$$\mathbf{P}_{H_0}(\psi(X) = 1) \leq \delta \text{ and } \mathbf{P}_{H_1}(\psi(X) = 0) \leq \delta.$$

We assume the following additional condition on the spike u.

Assumption 3.1. $\frac{c_{min}^2}{k} \leq u_i^2 \leq 1 - \frac{c_{min}^2}{k}$ *for at least one* $i \in [d]$ *where* $c_{min} > 0$ *is some constant.*

While in general we always have at least one $i \in [d]$ s.t. $u_i^2 \geq \frac{1}{k}$, this is not enough for our regression setup, since we want at least one other coordinate j to have sufficient correlation with coordinate i. We remark that the above condition is a very mild technical condition. If it were

violated, almost all of the mass of u is on a single coordinate, so a simple procedure for testing the variance (which is akin to diagonal thresholding) would suffice. Furthermore, for u drawn uniformly random from $\mathbb{S}^{k-1}$, we in fact expect a constant fraction of the coordinates to have mass at least $\frac{c_{min}}{k}$, in which case the above assumption is immediately satisfied.

Support recovery The goal of support recovery is to identify the support of u when X_i's are drawn from the spiked distribution under H_1. More precisely, we say that a support recovery algorithm succeeds if the recovered support $\widehat{S}$ equals S, the support of u. As standard in the literature [AW09, MB06], we need to assume a minimal bound on the size of entries of u in the support.

Though the settings are a bit different, this minimal bound also is consistent with lower bounds known for sparse recovery. These lower bounds ([FRG09, Wai07]; bound of [FRG09] is a factor of k weaker) imply that the number of samples (or measurements in their language) must grow roughly as $n \gtrsim \frac{1}{u_{min}^2} k \log d$ where u_{min} is the smallest entry of our signal u normalized by $1/\sqrt{k}$.

For our support recovery algorithm, we will make the following assumption (note that it implies Assumption 3.1 and is much stronger):

Assumption 3.2. *$|u_i| \geq c_{min}/\sqrt{k}$ for some constant $0 < c_{min} < 1$ $\forall i \in [d]$*

This is nearly optimal in comparison to the lower bounds mentioned above. If the smallest entries is smaller by a factor of some constant C, then signal strength θ needs to be stronger by a factor of C for our recovery algorithm to succeed, which is consistent with the lower bounds.

Unknown sparsity Note that throughout the paper we assume that the sparsity level k is known. However, if k is unknown, standard techniques could be used to adaptively find approximate values of k. For hypothesis testing for instance, we can start with an initial overestimate k', and keep halving until we get enough coordinates i with Q_i that passes the threshold for the given k'.

3.2 Problem formulation for SLR

We are given $(y, \mathbb{X})$ where $y \in \mathbb{R}^n$ and $\mathbb{X} \in \mathbb{R}^{n \times d}$ that are linked by the linear model $y = \mathbb{X}\beta^* + w$. The vector $w \in \mathbb{R}^n$ is i.i.d. $\mathcal{N}(0, \sigma^2)$, and our goal is to compute $\widehat{\beta}$ that minimizes the prediction error (or MSE) $\frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2$.

We define the *restricted eigenvalue* constant that will be important in our analysis. Many variants exist in the literature. Below, we give a definition from [ZWJ14].

Definition 3.3. First define the cone

$$\mathbb{C}(S) = \{\beta \in \mathbb{R}^d \mid \|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1\}$$

where S^c denotes the complement, β_T is β restricted to the subset T. The *restricted eigenvalue* (RE) constant of $\mathbb{X}$, denoted $\gamma(\mathbb{X})$, is defined as the largest constant γ s.t.

$$\frac{1}{n}\|\mathbb{X}\beta\|_2^2 \geq \gamma\|\beta\|_2^2 \quad \text{for all } \beta \in \bigcup_{|S|=k, S\subseteq[d]} \mathbb{C}(S)$$

Blackbox condition We now define the condition to require on our SLR blackbox, which is invoked as $\mathsf{SLR}(y, \mathbb{X}, k)$. This is similar to the guarantees achieved by known results for SLR.

Condition 3.4. Condition A. *Let $\gamma(\mathbb{X})$ denote the restricted eigenvalue of $\mathbb{X}$. There are universal constants c, c', c'' such that $\mathsf{SLR}(y, \mathbb{X}, k)$ outputs $\widehat{\beta}$ that is k-sparse and satisfies:*

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq \frac{c}{\gamma(\mathbb{X})^2} \cdot \frac{\sigma^2 k \log d}{n} \quad \forall \beta^* \in \mathsf{B}_0(k) \text{ w.p. } \geq 1 - c' \exp(-c'' k \log d)$$

3.3 Notation

We describe some of the notation used in this thesis. See page 7 for a more extensive list of symbols used.

Capital letters such as $\mathbf{X}, \mathbb{X}$ are used to denote matrices, and lowercase letters such as y for vectors. For clarity, we reserve $\mathbf{X}$ for the data matrix in SPCA and $\mathbb{X}$ for the design matrix SLR. $\mathbf{X}_i$ is the ith column of $\mathbf{X}$, and $\mathbf{X}_{-i}$ is the submatrix obtained by deleting the ith column from $\mathbf{X}$. Similarly, u_i denotes denotes the ith coordinate of u and $u_{-i} \in \mathbb{R}^{d-1}$ is u with ith coordinate removed.

$\Sigma_{S,T}$ is $\Sigma = \mathsf{E}[\mathbf{X}\mathbf{X}^\top]$ restricted to rows in S and columns in T; if $S = T$, we abbreviate it as Σ_S. For example, $\Sigma_{2:d}$ is Σ restricted to coordinates $2, ..., d$.

All vector norms are 2-norms unless specified otherwise.

C, C' to denote constants that may change from line to line.

Chapter 4

Reduction

We first discuss the underlying linear structure in the data generated from the spiked covariance model. The specifics of this linear structure will be used in defining our statistic and algorithms. We then state and prove the guarantees for our algorithms.

4.1 The linear model

We now set up linear regression for the data $\mathbf{X}$ from the SPCA problem. One natural way to apply regression to our samples $\mathbf{X}$ from SPCA is to regress one column or coordinate on the remaining columns. More formally, let $\mathbf{X}_{-i}$ denote the matrix of samples in the SPCA model with the ith column removed. For each column i, let us take as input to the blackbox SLR the design matrix $\mathbb{X} = \mathbf{X}_{-i}$ and the response variable $y = \mathbf{X}_i$.

Under the alternate hypothesis H_1, if $i \in S$, then X_i is correlated with X_j where $j \in S, j \neq i$. Using properties of multivariate Gaussians, we can write $y = \mathbb{X}\beta^* + w$ where $\beta^* = \frac{\theta u_i}{1+(1-u_i^2)\theta} u_{-i}$ and $w \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1 + \frac{\theta u_i^2}{1+(1-u_i^2)\theta}$. By theory of LMMSE, this β^* minimizes the error σ^2. (See Appendix A.1, A.2 for details of this calculation.) If $i \notin S$, and for any $i \in [d]$ under the null hypothesis, $y = w$ where $w = \mathbf{X}_i \sim \mathcal{N}(0, 1)$ (implicitly $\beta^* = \mathbf{0}$).

Because population covariance $\Sigma = \mathsf{E}[\mathbb{X}\mathbb{X}^\mathsf{T}]$ has minimum eigenvalue 1, with high probability the sample design matrix $\mathbb{X}$ has constant restricted eigenvalue value given enough samples n (see Appendix A.2.1 for more details), and the prediction error guarantee of Condition 3.4 will be good enough for our analysis.

Though the dimension and the sparsity of our SLR instances are $d - 1$ and $k - 1$ (since we remove one column from the SPCA data matrix $\mathbf{X}$ to obtain the design matrix $\mathbb{X}$), for ease of exposition we just use d, k in their place since it only affects our analysis up to small constant factors.

4.2 Algorithms and main results

4.2.1 Intuition of test statistic

Consider a matrix $\mathbf{X}$ of samples generated from the single spiked covariance model. The intuition behind the algorithm is that if i is in the support of the spike, then the rest of the support should allow to provide a nontrivial prediction for $\mathbf{X}_i$ since variables in the support are correlated. Conversely, for i not in the support (or under the isotropic null hypothesis), all of the variables are independent and other variables are useless for predicting $\mathbf{X}_i$. So we regress $\mathbf{X}_i$ onto the rest of the variables and our goal is to measure the reduction in noise.

How much predictive power do we gain by using X_{-i}? The linear minimum mean-square-error (LMMSE) estimate[1] of X_i conditioned on X_{-i} (when i is on support) turns out to put approximately θ/k weight on all the other coordinates on support.[2] A calculation shows that the variance in X_i is reduced by approximately θ^2/k. We want to measure this reduction in noise to detect when i is on support or not.

Suppose for instance that we have access to β^* rather than $\widehat{\beta}$ (note that this is not possible in practice since we do not know the support!). Since we want to measure the reduction in noise when the variable is on support, as a first step we might employ the following statistic:

$$Q_i = \frac{1}{n}\|y - \mathbb{X}\beta^*\|_2^2$$

Unfortunately this statistic will not be able to distinguish the two hypotheses, as the reduction in LMMSE is miniscule (on the order of θ^2/k compared to order of $1+\theta$), so deviation due to random sampling will mask the reduction in noise.

We can fix this by adding the variance term $\|y\|^2$:

$$Q_i = \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\beta^*\|_2^2$$

Notice that since $y = X\beta^* + w$, the noise term $\|w\|_2^2$ cancels out nicely. This effectively shifts the mean of the statistic, and now we are left with a statistic that is close to 0 under H_0 and is larger by about θ^2/k under H_1, so distinguishing using this statistic is more effective. On a more intuitive level, including $\|y\|_2^2$ allows us to measure the relative *gain* in predictive power without being penalized by a possibly large variance in y. Fluctuations in y due to noise will typically be canceled out in the difference of terms in Q_i, minimizing the variance of our statistic.

We have to add one final fix to the above estimator. We obviously do not have access to β^*, so we must use the estimate $\widehat{\beta} = \mathsf{SLR}(y, \mathbb{X}, k)$ ($y, \mathbb{X}$ are as defined in Section 4.1) which we get from our blackbox. The bulk of the analysis is showing that this substitution does not affect much of the discriminative power of Q_i.

[1] See Appendix A.1 for more details.

[2] For illustrative purposes, we consider the case where u is uniform on all k coordinates on support

This gives our final statistic:

$$Q_i = \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\widehat{\beta}\|_2^2.$$

4.2.2 Algorithms

Below we give two algorithms based on the Q statistic, one for hypothesis testing and one for support recovery:

Algorithm 1 Q-hypothesis testing

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}, k$
Output: ψ
for $i = 1, \ldots, d$ **do**
  $\widehat{\beta}_i = \mathsf{SLR}(\mathbf{X}_i, \mathbf{X}_{-i}, k)$
  $Q_i = \frac{1}{n}\|\mathbf{X}_i\|_2^2 - \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\beta}_i\|_2^2$
  if $Q_i > \frac{13k\log\frac{d}{k}}{n}$ **then**
    return $\psi = 1$
  end if
end for
Return $\psi = 0$

Algorithm 2 Q-support recovery

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}, k$
$\widehat{S} = \varnothing$
for $i = 1, \ldots, d$ **do**
  $\widehat{\beta}_i = \mathsf{SLR}(\mathbf{X}_i, \mathbf{X}_{-i}, k)$
  $Q_i = \frac{1}{n}\|\mathbf{X}_i\|_2^2 - \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\beta}_i\|_2^2$
  if $Q_i > \frac{13k\log\frac{d}{k}}{n}$ **then**
    $\widehat{S} := \widehat{S} \cup \{i\}$
  end if
end for
Return $\widehat{S}$

Below we summarize our guarantees for the above algorithms.

Theorem 4.1 (Hypothesis test). *Given* SLR *that satisfies Condition 3.4 and with runtime* $T(d, n, k)$ *per instance, and given Assumption 3.1, there exist universal constants* c_1, c_2, c_3, c_4 *s.t. if* $\theta^2 > \frac{c_1}{c_{min}^2}\frac{k^2\log d}{n}$ *and* $n > c_2 k\log d$ *then Algorithm 1 outputs* ψ *s.t.*

$$\mathbf{P}_{H_0}(\psi(X) = 1) \vee \mathbf{P}_{H_1}(\psi(X) = 0) \leq c_3\exp(-c_4 k\log d)$$

in time $O(dT + d^2 n)$.

Theorem 4.2 (Support recovery). *Under the same condition on SLR and given Assumption 3.2, if* $\theta^2 > \frac{c_1}{c_{min}^2}\frac{k^2\log d}{n}$, *Algorithm 2 above finds* $\widehat{S} = S$ *with probability at least* $1 - c_3\exp(-c_4 k\log d)$ *in time* $O(dT + d^2 n)$.

Remark 4.3. Though both guarantees involve bounding the signal strength θ in terms of c_{min}, Assumption 3.2 on u in Theorem 4.2 is much stronger as all entries in the support of u need to be minimally bounded for Assumption 3.2 to hold.

4.3 Analysis

In this section we analyze the distribution of Q_i under both H_0 and H_1 on our way to proving Theorems 4.1 and 4.2.

4.3.1 Analysis of Q_i under H_1

Without loss of generality assume the support of u, denoted S, is $\{1, ..., k\}$ and consider the first coordinate. We expand Q_1 by using $y = \mathbb{X}\beta^* + w$ as follows:

$$\begin{aligned} Q_1 &= \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\widehat{\beta}\|_2^2 = \frac{1}{n}\|\mathbb{X}\beta^* + w\|_2^2 - \frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2 - \frac{2}{n}w^\top(\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}) - \frac{1}{n}\|w\|_2^2 \\ &= \frac{1}{n}\|\mathbb{X}\beta^*\|_2^2 - \frac{2}{n}w^\top\mathbb{X}\beta^* - \frac{1}{n}(\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2) - \frac{2}{n}w^\top(\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}) \end{aligned}$$

Observe that the noise term $\|w\|_2^2$ cancels conveniently.

Before bounding each of these four terms, we introduce a useful lemma to bound cross terms involving noise w:

Lemma 4.4 (Lemmas 8 and 9, [RWY11]). *For any fixed $\mathbb{X} \in \mathbb{R}^{n\times d}$ and independent noise vector $w \in \mathbb{R}^n$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries:*

$$\frac{|w^\top\mathbb{X}\theta|}{n} \le 9\sigma\frac{\|\mathbb{X}\theta\|_2}{n}\sqrt{k\log\frac{d}{k}}$$

for all $\theta \in \mathsf{B}_0(2k)$ w.p. at least $\ge 1 - 2\exp(-40k\log(d/k))$

We bound each term as follows:

Term 1. The first term $\frac{\|\mathbb{X}\beta^*\|_2^2}{n}$ contains the signal from the spike; notice its resemblance to the k-sparse eigenvalue statistic. Rewritten in another way,

$$(\beta^*)^\top\frac{\mathbb{X}^\top\mathbb{X}}{n}\beta^* = (\beta^*)^\top\widehat{\Sigma}_{2:d}\beta^*$$

Hence, we expect this to concentrate around $(\beta^*)^\top\Sigma_{2:d}\beta^*$, which simplifies to (see Appendix A.2 for the full calculation):

$$(\beta^*)^\top\Sigma_{2:d}\beta^* = (\Sigma_{1,2:d}\Sigma_{2:d}^{-1})\Sigma_{2:d}(\Sigma_{2:d}^{-1}\Sigma_{2:d,1}) = \frac{\theta^2u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta}$$

For concentration, observe that we may rewrite

$$(\beta^*)^\top\widehat{\Sigma}_{2:d}\beta^* = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}^{(i)}\beta^*)^2$$

where $\mathbb{X}^{(i)}$ is the ith row, representing the ith sample. This is just an appropriately scaled

chi-squared random variable with n degrees of freedom (since each $\mathbb{X}^{(i)}\beta^*$ is i.i.d. normal), and the expected value of each term in the sum is the same as computed above. Applying a lower tail bound on χ^2 distribution (see Appendix), with probability at least $1-\delta$ we have

$$(\beta^*)^\top \widehat{\Sigma}_{2:d}\beta^* \geq \frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta} \cdot \left(1 - 2\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Choosing $\delta = \exp(-k\log d)$,

$$\begin{aligned}\frac{\|\mathbb{X}\beta^*\|_2^2}{n} &\geq \frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta} \cdot \left(1 - 2\sqrt{\frac{k\log d}{n}}\right) \\ &\overset{(a)}{\geq} \frac{1}{2} \cdot \frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta} \\ &\overset{(b)}{\gtrsim} c_{min}^2 \frac{\theta^2}{k} \end{aligned} \tag{4.1}$$

where (a) as long as $n > 16k\log d$ and (b) since $\theta \leq 1$ and $u_1^2(1-u_1^2) \gtrsim c_{min}^2/k$ under Assumption 3.1.

Term 2. The absolute value of the second term $\frac{2}{n}w^\top \mathbb{X}\beta^*$ can be bounded by $18\frac{\|\mathbb{X}\beta^*\|_2}{n}\sqrt{k\log\frac{d}{k}}$ using Lemma 4.4. From (4.1) as long as $\theta^2 > \frac{C}{c_{min}^2}\frac{k^2\log d}{n}$,

$$\frac{\|\mathbb{X}\beta^*\|_2^2}{n} \gtrsim c_{min}^2\frac{\theta^2}{k} \gtrsim C\frac{k\log d}{n}$$

so the first two terms together are lower bounded by:

$$\frac{\|\mathbb{X}\beta^*\|_2}{n}(\|\mathbb{X}\beta^*\|_2 - 18\sqrt{k\log d/k}) \geq C\frac{\|\mathbb{X}\beta^*\|^2}{n}, \tag{4.2}$$

constant fraction of the first term.

Term 3. The third term, which is the prediction error $\frac{\|\mathbb{X}\beta^* - \mathbb{X}\hat{\beta}\|_2^2}{n}$, is upper bounded by $\frac{C}{\gamma(X)^2}\frac{\sigma^2 k\log d}{n}$ with probability at least $1 - C\exp(-C'k\log d)$ by Condition 3.4 on our SLR black-box. Note $\sigma^2 < 2$ as we assume $\theta \leq 1$ (see Section 4.1). Now, $\gamma(X) \geq \frac{1}{8}$ with probability at least $1 - C\exp(-C'n)$ if $n > C''k\log d$ since $\theta \leq 1$ (see Appendix A.2.1 for more details). Then,

$$\frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\hat{\beta}\|_2^2 \leq C\frac{k\log d}{n}$$

Term 4. The contribution of the last cross term $\frac{2}{n}w^\top\mathbb{X}(\beta^* - \widehat{\beta})$ can also bounded by Lemma 4.4 w.h.p. (note $\beta^* - \widehat{\beta} \in \mathsf{B}_0(2k)$)

$$\frac{|w^\top\mathbb{X}(\beta^* - \widehat{\beta})|}{n} \leq 9\sigma\frac{\|\mathbb{X}(\beta^* - \widehat{\beta})\|_2}{n}\sqrt{k\log\frac{d}{k}}.$$

Combined with the above bound for prediction error, this bounds the cross term's contribution by at most $C\frac{k\log d}{n}$.

Putting the bounds on four terms together, we get the following lower bound on Q.

Lemma 4.5. *There exists constants c_1, c_2, c_3, c_4 s.t. if $\theta^2 > \frac{c_1}{c_{min}^2}\frac{k^2\log d}{n}$ and $n > c_2 k\log d$, with probability at least $1 - c_3\exp(-c_4 k\log d)$, for any $i \in S$ that satisfies the size bound in Assumption 3.1,*

$$Q_i > \frac{13k\log d}{n}$$

Proof. From 1-4 above, by union bound, all four bounds fail to hold with probability at most $c_3\exp(-c_4 k\log d)$ for appropriate constants if $\theta^2 > \frac{c_1}{c_{min}^2}\frac{k^2\log d}{n}$ (required by Term 2) and $n > c_2 k\log d$ for some $c_2 > 0$ (note that both Terms 1 and 3 require sufficient number of samples n). That is, we have:

$$Q_i > c_{min}^2 C\frac{\theta^2}{k} - C'\frac{k\log d}{n}$$

So if c_1 is sufficiently large, the above bound is greater than $\frac{13k\log d}{n}$. □

4.3.2 Analysis of Q_i under H_0

We could proceed by decomposing Q_i the same way as in H_1; all the error terms including prediction error are still bounded by $O(k\log d/n)$ in magnitude, and the signal term is gone now since $\beta^* = 0$. This will give the same upper bound (up to a constant) as the following proof is about to show. However, we find the following direct analysis more informative and intuitive.

Since our goal is to upper bound Q_i under H_0, we may let $\widehat{\beta}$ be the optimal possible choice given y and $\mathbb{X}$ (one that minimizes $\|y - \mathbb{X}\widehat{\beta}\|_2^2$, and hence maximizes Q_i). We further break this into two steps. We enumerate over all possible subsets S of size k, and conditioned on each S, choose the optimal $\widehat{\beta}$.

Fix some support S of size k. The span of $\mathbb{X}_S$ is at most a k-dimensional subspace of $\mathbb{R}^n$. Hence, we can consider some unitary transformation U of $\mathbb{R}^n$ that maps the span of $\mathbb{X}_S$ into the subspace spanned by the first k standard basis vectors. Since U is an isometry by definition,

$$nQ_i = \|y\|_2^2 - \|y - \mathbb{X}\widehat{\beta}_S\|_2^2 = \|Uy\|_2^2 - \|Uy - U\mathbb{X}\widehat{\beta}_S\|_2^2$$

Let $\tilde{y} = Uy$. Since $U\mathbb{X}\widehat{\beta}_S$ has nonzero entries only in the first k coordinates, the optimal choice (in the sense of maximizing the above quantity) of $\widehat{\beta}_S$ is to choose linear combinations of the first k columns of $\mathbb{X}$ so that $U\mathbb{X}\widehat{\beta}_S$ equals the first k coordinates of $\tilde{y}$. Then, nQ_i is just the squared norm of the first k coordinates of $\tilde{y}$. Since U is some unitary matrix that is independent of y (being a function of $\mathbb{X}_S$ which is independent of y), $\tilde{y}$ still has i.i.d. $\mathcal{N}(0,1)$ entries, and hence nQ_i is a χ^2-var with k degrees of freedom.

Now we apply an upper tail bound on the χ^2 distribution (See Appendix A.3.1). Choosing $t = 3\log\frac{d}{k}$, and after union bounding over all $\binom{d}{k} \le \left(\frac{de}{k}\right)^k$ supports S, $nQ_i > k + 12k\log\frac{d}{k}$, or

$Q > \frac{13k \log \frac{d}{k}}{n}$ with probability at most $\exp(-3k \log \frac{d}{k} + k \log \frac{de}{k}) \leq \exp(-k \log \frac{d}{k})$ if $\frac{d}{k} \geq e$.

Lemma 4.6. *Under H_0, $\forall i$ $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ w.p. at least $1 - \exp(-k \log \frac{d}{k})$.*

Remark 4.7. Union bounding over all S is necessary for the analysis. For instance, we cannot just fix S to be $S(\widehat{\beta})$ (this denotes the support of $\widehat{\beta}$) since $\widehat{\beta}$ is a function of y, so fixing S changes the distribution of y.

Remark 4.8. Observe that this analysis of Q_i for H_0 also extends immediately to H_1 when coordinate i is outside the support. The reason the analysis cannot extend to when $i \in S$ is because U is not independent of y in this case.

Corollary 4.9. *Under H_1, if $i \notin S$, $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ w.p. at least $1 - \exp(-k \log \frac{d}{k})$.*

4.3.3 Proof of Theorem 4.1

Proof. Proof follows immediately from Lemma 4.6 and Lemma 4.5. Now, we can use our estimators Q_i to separate H_0 and H_1. Under H_0, applying Lemma 4.6 to each coordinate i and union bounding, $\forall i$, $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ with probability at least $1 - \exp(-Ck \log d)$. Meanwhile, under H_1, if we consider any coordinate i that satisfies Assumption 3.1, Lemma 4.5 gives

$$Q_i > \frac{13k \log d}{n}$$

with probability at least $1 - c_3 \exp(-c_4 k \log d)$. Since ψ tests whether $Q_i > \frac{13k \log \frac{d}{k}}{n}$ for at least one i, ψ distinguishes H_0 and H_1 successfully, with bound on type I and type II error probability $c_3 \exp(-c_4 k \log d)$ for appropriate constants c_3, c_4 (note, these may be different from those of Lemma 4.5). For runtime, note that we make d oracle calls to SLR and work with matrices of size $n \times d$. □

4.3.4 Proof of Theorem 4.2

Proof. As long as every u_i for $i \in S$ has magnitude $c_{min}/\sqrt{k}$ as in Assumption 3.2, we can repeat the same analysis from above to all coordinates in the support. If θ meets the same threshold, $Q_i > 13k \log \frac{d}{k}/n$ for all $i \in S$ with probability at least $1 - C \exp(-C' k \log d)$ by union bound. Also, recall $Q_i > 13k \log \frac{d}{k}/n$ for any $i \notin S$ with probability at most $C \exp(-C' k \log d)$ by Corollary 4.9. By union bound over all $d - k$ coordinates outside the support, the error probability is at most $d \cdot C \exp(-C' k \log d) \leq C \exp(-C'' k \log d)$. We showed that with high probability we exactly recover the support S of u.

Runtime analysis is identical to that for the hypothesis test. □

4.4 Discussion

4.4.1 Running time

The runtime of both Algorithms 1 and 2 is $\tilde{O}(nd^2)$,[3] if we assume the SLR blackbox takes nearly linear time in input size, $\tilde{O}(nd)$, which is achieved by known existing algorithms. This seems a bit expensive at first, but computing the sample covariance matrix alone takes $O(nd^2)$ time.[4]

For a broad comparison, we consider spectral methods and SDP-based methods, though there are methods that do not fall in either category.

Spectral methods such as covariance thresholding or truncated power method have an iteration cost $O(d^2)$ due to operating on $d \times d$ matrices, and hence have total running time $\tilde{O}(d^2)$ ($\tilde{O}(\cdot)$ hiding precise convergence rate) in addition to the same $O(nd^2)$ initialization time.

SDP-based methods in general take $\tilde{O}(d^3)$ time, the time taken by interior point methods to optimize. So overall, Algorithms 1 and 2 are competitive choices for (single spiked) SPCA, at least theoretically.

4.4.2 Alternate blackbox

The above algorithms seem rather wasteful because there is a lot of overlapping information between the different $\widehat{\beta}$'s we get for the Q_i's on support. For instance, it is plausible that $\widehat{\beta}$ contains a good fraction of the entries in the support if the coordinate we are regressing on happens to be on support. In such case, it is unnecessary to compute Q_j's for the j's we already are confident to be on support.

We may be able to utilize such information more easily if instead of prediction error we consider support recovery or parameter estimation (say in ℓ_2-norm) guarantees for our SLR blackbox.

4.4.3 Robustness of Q statistic to rescaling

A natural and simple way to make diagonal thresholding fail is to rescale all the variables so that their variance is equal. Intuitively, we expect our algorithms based on Q to be robust to rescaling, since it should be possible to predict one variable in the support from the others in the support even after some rescaling.

We can more precisely justify this intuition as follows.. Let $\widetilde{X} \leftarrow DX$ be the rescaling of X, where D is some diagonal matrix. Let D_S be D restricted to rows and columns in S. Note that $\widetilde{\Sigma}$, the covariance matrix of the rescaled data, is just $D\Sigma D$ by expanding the definition. Similarly, note $\widetilde{\Sigma}_{2:d,1} = D_1 D_{2:d} \Sigma_{2:d,1}$ where $D_{2:d}$ denotes D without row and column 1. Now,

[3] In what follows $\tilde{O}(\cdot)$ hides possible log and accuracy parameter ϵ factors.

[4] Assuming one is using naive implementation of matrix multiplication.

recall the term which dominated our analysis of Q_i under H_1, $(\beta^*)^\top \Sigma_{2:d}\beta^*$, which was equal to

$$\Sigma_{1,2:d}\Sigma_{2:d}^{-1}\Sigma_{2:d,1}$$

We replace the covariances by their rescaled versions to obtain:

$$\tilde{\beta^*}^\top \tilde{\Sigma}\tilde{\beta^*} = (D_1\Sigma_{1,2:d}D_{2:d})D_{2:d}^{-1}\Sigma_{2:d}^{-1}D_{2:d}^{-1}(D_{2:d}\Sigma_{2:d,1}D_1) = D_1^2 \cdot (\beta^*)^\top \Sigma_{2:d}\beta^*$$

For the spiked covariance model, rescaling variances to one amount to rescaling with $D_1 = \frac{1}{1+\theta}$. Thus, we see that our signal strength is affected only by constant factor (assuming $\theta \leq 1$).

We should note though that after normalizing variances, the variance term $\|y\|_2^2$ loses its effect in the Q statistic, and Q is essentially affected by just the reconstruction error $\|y - \mathbb{X}\widehat{\beta}\|_2^2$.

A new model for SPCA? This robustness to rescaling is an attractive property because intuitively, our algorithms for detecting correlated structure in data should be invariant to rescaling of data; the precise scale or units for which one variable is measured should not have an impact on our ability to find meaningful structure underlying the data. Perhaps this suggests an avenue for exploring new alternative models for SPCA that are more flexible and robust.

Chapter 5

Experiments

On randomly simulated synthetic data we demonstrate the performance of our algorithm compared to other existing algorithms for SPCA. The code was implemented in Python using standard libraries. We refer to both hypothesis and support recovery variants of our algorithm from Section 4.2 as Q.

5.1 Support recovery

We randomly generate a spike u by choosing uniformly among all k-sparse vectors that are uniform on all coordinates (with random signs). In order for comparison with the work of [DM14], we use the same parameter setting of $n = d$. We study how the performance of four algorithms (diagonal thresholding, covariance thresholding, Q with thresholded Lasso with $\lambda = 0.1$, and Q with FoBa with $\epsilon = 0.1$) vary over various values of k for fixed $n = d$. For covariance thresholding, we tried various levels of their parameter τ and indeed it performed best at [DM14]'s recommended value of $\tau \approx 4$, which is what is shown. We modified each algorithm to return the top k most likely coordinates in the support (rather than thresholding based on a cutoff), and we count the fraction of planted support recovered. This is averaged over $T = 50$ trials. On the horizontal axis we measure $k/\sqrt{n}$; our metric on the vertical axis is the fraction of support correctly recovered. We observe that across almost all regimes of k both versions of Q algorithms outperform covariance thresholding. It is an interesting question to investigate whether the $\log d$ factor in our analysis can be removed. Diagonal thresholding is outperformed by all other methods across most values of k.

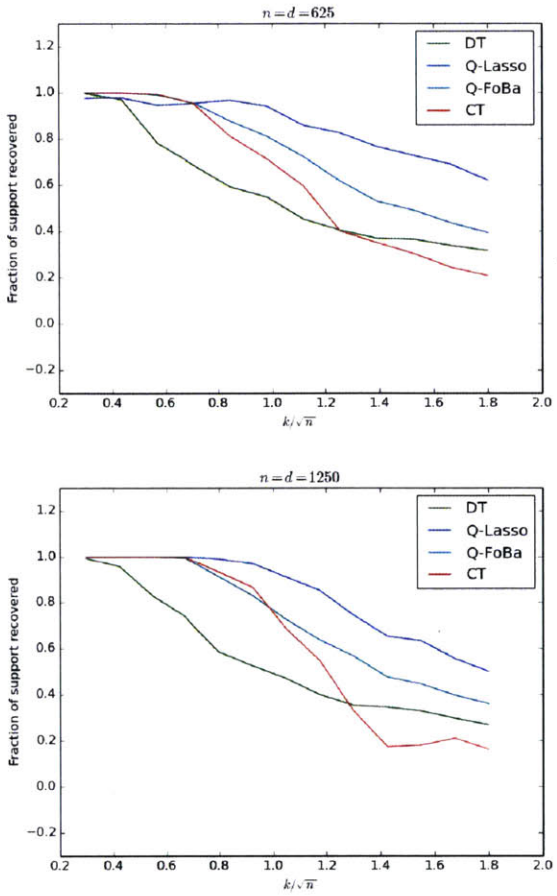

Figure 5.1: Performance of diagonal thresholding (DT), covariance thresholding (CT), and Q for support recovery at $n = d = 625, 1250$, varying values of k, and $\theta = 4$

5.2 Hypothesis testing

Here we instead generate a spike u by sampling a uniformly random direction from the k-dimensional unit sphere, and embedding the vector at a random subset of k coordinates among d coordinates. For hypothesis testing, in a single trial, we compute various statistics (diagonal thresholding (DT), Minimal Dual Perturbation (MDP), and Q) after drawing n samples from $\mathcal{N}(0, I_d + \theta uu^\top)$. We repeat for $T = 50$ trials, and plot the resulting empirical distribution for each statistic. We observe similar performance of DT and Q, while MDP seems slightly more effective at distinguishing H_0 and H_1 at the same signal strength (that is, the distributions of the statistics under H_0 vs. H_1 are more well-separated).

Rescaling variables As discussed in Section 4.4.3, our algorithms should be robust to rescaling the covariance matrix to the correlation matrix. As illustrated in Figure 5.2 (right), DT fails while Q appears to be still effective for distinguishing hypotheses the same regime of parameters. Other methods such as MDP and CT also appear to be robust to such rescaling

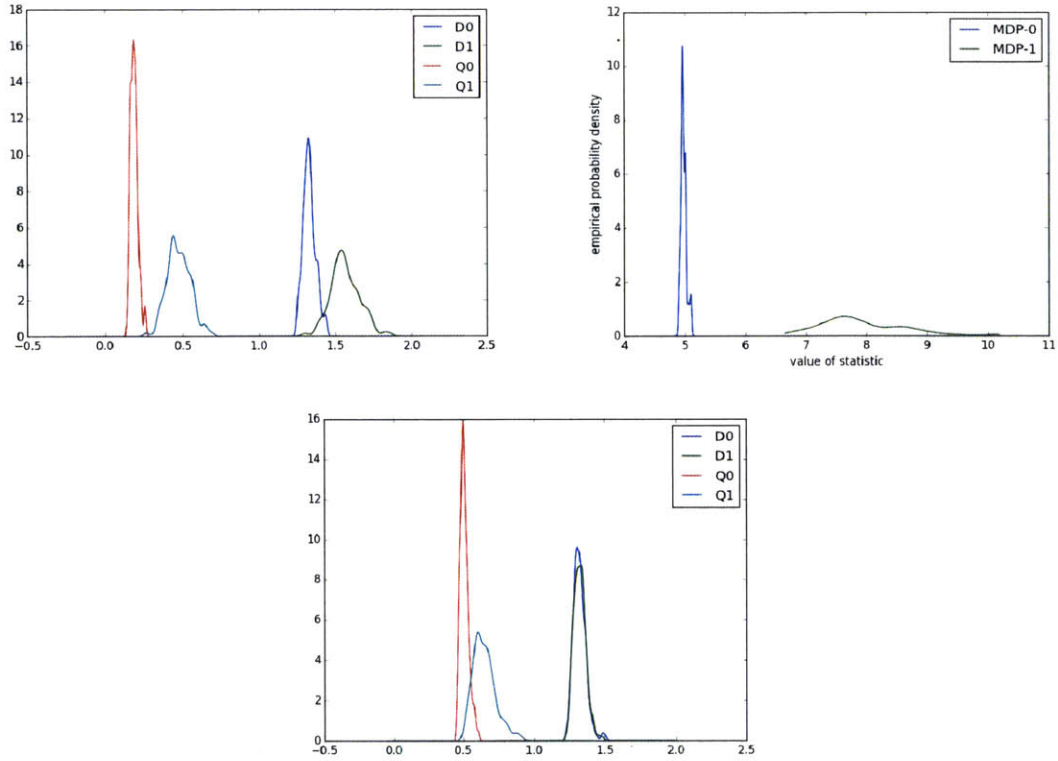

Figure 5.2: Performance of diagonal thresholding (D), MDP, and Q for hypothesis testing at $n = 200, d = 500, k = 30, \theta = 4$ (left and center). T0 denotes the statistic T under H_0, and similarly for T1. Effect of rescaling covariance matrix to make variances indistinguishable is demonstrated (right)

(not shown). This suggests that more modern algorithms for SPCA may be more appropriate than diagonal thresholding in practice, particularly on instances where the relative scales of the variables may not be accurate or knowable in advance, but we still want to be able to find a correlational structure between the variables.

Chapter 6

Conclusion

We gave a reduction from SPCA to SLR that works up to the computational threshold for SPCA that we believe based on average-case hardness assumptions. One obvious question is if there is a different reduction that extends all the way down to the statistical threshold; this would imply the average-case hardness of SLR under some conditions.

Another limitation of the current reduction is that it does not readily extend to the sub-Gaussian setting. Such an extension would be a more robust result as Gaussian assumption is highly restrictive while sub-Gaussian random variables capture other useful classes of random variables such as those that are bounded. A related question is formulating a model more robust than the gaussian spiked covariance model for SPCA, yet still amenable to analysis.

It would also be interesting to see if the reduction can be done in the other direction, from SLR to SPCA. One would probably have to restrict the design matrix to a certain class in order to have sufficient control on its distribution.

Bibliography

[APKD15] Megasthenis Asteris, Dimitris Papailiopoulos, Anastasios Kyrillidis, and Alexandros G Dimakis. Sparse pca via bipartite matchings. In *Advances in Neural Information Processing Systems*, pages 766–774, 2015.

[AW09] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2009.

[BD09] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[BF96] Yoram Bresler and Ping Feng. Spectrum-blind minimum-rate sampling and reconstruction of 2-d multiband signals. In *Proc. Int. Conf. Image Processing (ICIP)*. IEEE, 1996.

[BR13a] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[BR13b] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[BTW07a] Florentina Bunea, Alexandre B Tsybakov, and Marten H Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[BTW07b] Florentina Bunea, Alexandre B Tsybakov, and Marten H Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning theory*, pages 530–543. Springer, 2007.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[CJ95] Jorge Cadima and Ian T Jolliffe. Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214, 1995.

[CMW13] T Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[CRT06a] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[CRT06b] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[CT05] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[CT07] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[dBEG14] Alexandre dâĂŹAspremont, Francis Bach, and Laurent El Ghaoui. Approximation bounds for sparse principal component analysis. *Mathematical Programming*, 148(1-2):89–110, 2014.

[dEGJL07] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[DM14] Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

[Don06] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[DRXZ14] Dong Dai, Philippe Rigollet, Lucy Xia, and Tong Zhang. Aggregation of affine estimators. *Electronic Journal of Statistics*, 8(1):302–327, 2014.

[EK12] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[FB96] Ping Feng and Yoram Bresler. Spectrum-blind minimum-rate sampling and reconstruction of multiband signals. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1688–1691. IEEE, 1996.

[FRG09] Alyson K Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.

[GD09] Yue Guan and Jennifer G Dy. Sparse probabilistic principal component analysis. In *AISTATS*, pages 185–192, 2009.

[Gem80] Stuart Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, pages 252–261, 1980.

[JL09] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 2009.

[JNRS10] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553, 2010.

[Joh01] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[JTU03] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

[KGPK15] Rajiv Khanna, Joydeep Ghosh, Russell A Poldrack, and Oluwasanmi Koyejo. Sparse submodular probabilistic pca. In *AISTATS*, 2015.

[KKGP14] Oluwasanmi O Koyejo, Rajiv Khanna, Joydeep Ghosh, and Russell Poldrack. On prior distributions and approximate inference for structured variables. In *Advances in Neural Information Processing Systems*, pages 676–684, 2014.

[KNV13] Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations really solve sparse pca. Technical report, Technical report, Weizmann Institute of Science, 2013.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[Ma13] Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[Mac09] Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[MW15] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

[MWA05] Baback Moghaddam, Yair Weiss, and Shai Avidan. Spectral bounds for sparse pca: Exact and greedy algorithms. In *Advances in neural information processing systems*, pages 915–922, 2005.

[MZ93] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[Ney06] Tyler Neylon. *Sparse solutions for linear prediction problems.* PhD thesis, New York University, 2006.

[NT09] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[NYWR09] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[PDK13] Dimitris S Papailiopoulos, Alexandros G Dimakis, and Stavros Korokythakis. Sparse pca through low-rank approximations. In *ICML (3)*, pages 747–755, 2013.

[RT11] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[RWY10] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[RWY11] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[SB08] Christian D Sigg and Joachim M Buhmann. Expectation-maximization for sparse and non-negative pca. In *Proceedings of the 25th international conference on Machine learning*, pages 960–967. ACM, 2008.

[TB99] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[VB98] Raman Venkataramani and Yoram Bresler. Further results on spectrum blind sampling of 2d signals. In *International Conference on Image Processing (ICIP)*, volume 2, pages 752–756. IEEE, 1998.

[VCLR13] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678, 2013.

[vdG07] Sara van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

[VDGB+09] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[Ver15] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*, pages 3–66. Springer, 2015.

[VMB02] Martin Vetterli, Pina Marziliano, and Thierry Blu. Sampling signals with finite rate of innovation. *IEEE transactions on Signal Processing*, 50(6):1417–1428, 2002.

[Wai07] Martin Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *2007 IEEE International Symposium on Information Theory*, pages 961–965. IEEE, 2007.

[Wai09] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[Wai10] Martin Wainwright. High-dimensional statistics: some progress and challenges ahead. Winedale Workshop, 2010.

[WBP16] Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. *arXiv preprint arXiv:1605.09646*, 2016.

[YZ13] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *The Journal of Machine Learning Research*, 14(1):899–925, 2013.

[Zha09] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[ZWJ14] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*, 2014.

[ZWJ15] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *arXiv preprint arXiv:1503.03188*, 2015.

Appendix A

Supplement

A.1 Linear minimum mean-square-error estimation

Given random variables Y and X (this can be a vector more generally), a natural question is what is the best prediction for Y conditioned on knowing $X = x$? What is considered "best" can vary, but usually we consider the mean-square-error. That is, we want to come up with $\hat{y}(x)$ s.t.

$$\mathsf{E}[(Y - \hat{y})^2]$$

is minimized.

It is not hard to show that $\hat{y}$ is just the conditional expectation of Y conditioned on X. The minimum mean-square-error estimate can be a highly nontrivial function of X.

The linear minimum mean-square-error (LMMSE) estimate instead restricts the attention to estimators of the form $\hat{Y} = AX + b$. Notice here that A and b are fixed and are not functions of X.

One can show that the LMMSE estimator is given by: $A = (\Sigma_{XX})^{-1})\Sigma_{XY}$, where $\Sigma_\cdot$ is the appropriately indexed covariance matrix, and b is chosen in the obvious way to make our estimator unbiased.

A.2 Calculations for linear model from Section 4.1

To recap our setup, we input the design matrix $\mathbb{X} = \mathbf{X}_{-i}$ and the response variable $y = \mathbf{X}_i$ as inputs to an SLR blackbox. Our goal is to express y as a linear function of $\mathbb{X}$ plus some independent noise w. Without loss of generality let $i = 1$, and for our discussion below assume $S = \{1, ..., k\}$. For illustration, at times we will simplify our calculation further for the uniform case where $u_i = \frac{1}{\sqrt{k}}$ for $1 \leq i \leq k$ and $u_i = 0$ for $i > k$.

For the moment, just consider one row of $\mathbf{X}$, corresponding to one particular sample X of the original SPCA distribution. Since X is jointly Gaussian, we can express (the expectation of) $y = X_1$ as a linear function of the other coordinates:

$$\mathsf{E}[X_1 | X_{2:d} = x_{2:d}] = \Sigma_{1,2:d}(\Sigma_{2:d})^{-1} x_{2:d}$$

Hence we can write

$$X_1 = \Sigma_{1,2:d}(\Sigma_{2:d})^{-1} X_{2:d} + w$$

where $w \sim \mathcal{N}(0, \sigma^2)$ for some σ to be determined and $w \perp\!\!\!\perp X_i$ for $i = 2, .., d$.

By directly computing the variance of the above expression for X_1, we deduce an expression for the noise level:

$$\sigma^2 = \Sigma_{11} - \Sigma_{1,2:d}(\Sigma_{2:d})^{-1}\Sigma_{2:d,1}$$

Note that σ^2 is just Σ_{11} under H_0. We proceed to compute σ^2 under H_1, when $\Sigma = I_d + \theta uu^\top$. To compute $(\Sigma_{2:d})^{-1}$, we use (a special case of) the Sherman-Morrison formula: $(I + wv^\top)^{-1} = I - \frac{wv^\top}{1+v^\top w}$.

$$\Sigma_{2:d}^{-1} = \left(I_{d-1} + \theta u_{-1}u_{-1}^\top\right)^{-1} = I_{d-1} - \frac{\theta}{1 + (1-u_1^2)\theta}u_{-1}u_{-1}^\top$$

where $u_{-1} \in \mathbb{R}^{d-1}$ is u restricted to coordinates $2, ..., d$.

$$\begin{aligned}
\Sigma_{1,2:d}(\Sigma_{2:d})^{-1}\Sigma_{2:d,1} &= \left(\frac{\theta u_1}{1 + (1-u_1^2)}\right)^2 u_{-1}^\top(I + \theta u_{-1}u_{-1}^\top)u_{-1} \\
&= \frac{\theta^2 u_1^2(1-u_1^2)}{1 + (1-u_1^2)\theta} \\
&\text{(specializing to uniform case again)} \\
&= \frac{\theta^2}{k}\left(1 - \frac{1}{k}\right)\frac{1}{1 + \frac{k-1}{k}\theta} \approx \frac{\theta^2}{k(1+\theta)}
\end{aligned}$$

Finally, substituting into the expression for σ^2

$$\begin{aligned}
\sigma^2 &= 1 + \theta u_1^2 - \frac{\theta^2 u_1^2(1-u_1^2)}{1 + (1-u_1^2)\theta} \\
&= 1 + \frac{\theta u_1^2}{1 + (1-u_1^2)\theta} \\
&\leq 2 \quad \text{if } \theta \leq 1
\end{aligned}$$

We remark that the noise level of column 1 has been reduced by roughly $\tau := \frac{\theta^2}{k(1+\theta)}$ by regressing on correlated columns.

In summary, under H_1 (and if $1 \in S$) we can write

$$y = \mathbb{X}\beta^* + w$$

where

$$\begin{aligned}
\beta^* &= (\Sigma_{2:d})^{-1}\Sigma_{2:d,1} \\
&= (I - \frac{\theta}{1 + (1-u_1)^2\theta}u_{-1}u_{-1}^\top)\theta u_1 u_{-1} \\
&= \theta u_1\left(1 - \frac{\theta}{1 + (1-u_1^2)\theta}(1-u_1^2)\right)u_{-1} \\
&= \frac{\theta u_1}{1 + (1-u_1^2)\theta}u_{-1}
\end{aligned}$$

(technically, the definition of β^* on the RHS is a $k-1$ dimensional vector, but we augment it with zeros to make it $d-1$ dimensional) and $w \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 1 + \frac{\theta u_1^2}{1+(1-u_1^2)\theta}$. Note that in the uniform case, $\beta^* \to \frac{1}{k-1}\mathbb{1}_{k-1}$ as $\theta \to \infty$ where $\mathbb{1}_{k-1}$ is uniform 1 on first $k-1$ coordinates, as expected.

A.2.1 Properties of design matrix $\mathbb{X}$

Restricted eigenvalue (RE) Here we check that $\mathbb{X}$ defined as in Section 4.1 has constant restricted eigenvalue constant. This allows us to apply Condition 3.4 for the SLR blackbox with good guarantee on prediction error.

The rows of $\mathbb{X}$ are drawn from $\mathcal{N}(0, I_{d-1\times d-1} + \theta u_{-1}u_{-1}^\top)$ where u_{-1} is u restricted to coordinates $2, ..., d$ wlog. [1]

Let $\Sigma = I_{d-1\times d-1} + \theta u_{-1}u_{-1}^\top$. We can show that $\Sigma^{1/2}$ satisfies RE with $\gamma = 1$ by bounding Σ's minimum eigenvalue. First, we compute the eigenvalues of $\theta u_{-1}u_{-1}^\top$. $\theta u_{-1}u_{-1}^\top$ has a nullspace of dimension $d-2$, so eigenvalue 0 has multiplicity $d-2$. u_{-1} is a trivial eigenvector with eigenvalue $\theta u_{-1}^\top u_{-1} = \theta\frac{k-1}{k}$. Therefore, Σ has eigenvalues 1 and $1 + \theta\frac{k-1}{k}$.

Now we can extend this to the sample matrix $\mathbb{X}$ by applying Corollary 1 of [RWY10] (also see Example 3 therein), and conclude that as soon as $n > C''\frac{\max_j \Sigma_{jj}}{\gamma^2} k \log d = C(1 + \frac{\theta}{k})k \log d$ or $n = \Omega(k \log p)$ the matrix $\mathbb{X}$ satisfies RE with $\gamma(\mathbb{X}) = 1/8$.

We remark that the following small technical condition also appears in known bounds on prediction error:

Column normalization This is a condition on the scale of $\mathbb{X}$ relative to the noise in SLR, which is always σ^2.

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} \leq \|\theta\|_2^2$$

for all $\theta \in \mathsf{B}_0(2k)$

We can always rescale $\mathbf{X}$ (and hence $\mathbb{X}$) to satisfy this, which would also rescale the noise level σ in our linear model since the noise is derived from coming $\mathbf{X}$ from the SPCA generative model, rather than added independently as in the usual SLR setup.

Hence, since all scale dependent quantities are scaled by the same amount when we scale the original data $\mathbf{X}$, wlog we may continue to use the same $\mathbb{X}$ and σ in our analysis. As the column normalization condition does not affect us, we drop it from Condition 3.4 of our blackbox assumption.

[1] We assume here that $1 \in S$ as in the previous section

A.3 Tail inequalities

A.3.1 Chi-squared

Lemma A.1 (Concentration on upper and lower tails of the χ^2 distribution ([LM00], Lemma 1)). *Let Z be the χ^2 random variable with k degrees of freedom. Then,*

$$\Pr(Z - k \geq 2\sqrt{kt} + 2t) \leq \exp(-t)$$
$$\Pr(Z - X \geq 2\sqrt{kt}) \leq \exp(-t)$$

We can simplify the upper tail bound as follows for convenience:

Corollary A.2. *For χ^2 r.v. Z with k degrees of freedom and deviation $t \geq 1$,* $\Pr\left(\frac{Z-k}{k} \geq 4t\right) \leq \exp(-kt)$.