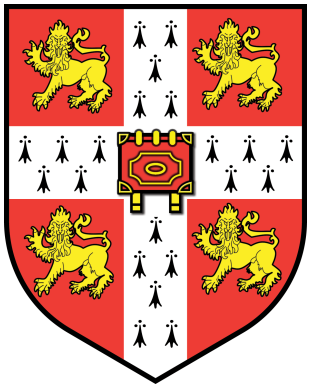
Statistical inference in high-dimensional matrix models

Matthias Löffler

Darwin College

University of Cambridge



A dissertation submitted for the degree of

Doctor of Philosophy

June 2019

Abstract

Matrix models are ubiquitous in modern statistics. For instance, they are used in finance to assess interdependence of assets, in genomics to impute missing data and in movie recommender systems to model the relationship between users and movie ratings. Typically such models are either high-dimensional, meaning that the number of parameters may exceed the number of data points by many orders of magnitudes, or nonparametric in the sense that the quantity of interest is an infinite dimensional operator. This leads to new algorithms and also to new theoretical phenomena that may occur when estimating a parameter of interest or functionals of it or when constructing confidence sets. In this thesis, we will exemplarily consider three such matrix models and develop statistical theory for them: Matrix completion, Principal Component Analysis (PCA) with Gaussian data and transition operators of Markov chains.

We start with matrix completion and investigate the existence of adaptive confidence sets in the 'Bernoulli' and 'trace-regression' models. In the 'Bernoulli' model we show that adaptive confidence sets do not exist when the variance of the errors is unknown, whereas we give an explicit construction in the 'trace-regression' model. Finally, in the known variance case, we show that adaptive confidence sets do also exist in the 'Bernoulli' model based on a testing argument.

Next, we consider PCA in a Gaussian observation model with complexity measured by the effective rank, the reciprocal of the percentage of variance explained by the first principal component. We investigate estimation of linear functionals of eigenvectors and prove Berry-Essen type bounds. Due to the high-dimensionality of the problem we discover a new phenomenon: The plug-in estimator based on the sample eigenvector can have non-negligible bias and hence may be not $\sqrt{n}$-consistent anymore. We show how to de-bias this estimator, achieving $\sqrt{n}$-convergence rates, and prove exact matching minimax lower bounds.

Finally, we consider nonparametric estimation of the transition operator of a Markov chain and its transition density. We assume that the singular values of the transition operator decay exponentially. For example, this assumption is fulfilled by discrete, low frequency observations of periodised, reversible stochastic differential equations. Using penalization techniques from low rank matrix estimation we develop a new algorithm and show improved convergence rates.

Acknowledgements

First and foremost, I am deeply grateful to my supervisor, Richard Nickl, for his continuous support, enlightening discussions, guidance, induction into the statistics community and kind words of encouragement throughout the last four years. In particular, I would like to thank him for his support and kindness throughout the last few months which have been much more challenging than I ever imagined.

In addition, I am very grateful to Rajen Shah and Markus Reiß for acting as examiners for my defense, which was a truly enriching experience, giving rise to new potential avenues of research.

Moreover, I would like to thank Harry Zhou for inviting me to visit him at Yale University, his hospitality and the fruitful discussions we had there. I am also grateful to my collaborators Alexandra Carpentier, Olga Klopp, Vladimir Koltchinskii, Antoine Picard, Anderson Zhang and Harry Zhou and what they have taught me.

I would also like to thank all people in CCA, the Statslab and DPMMS which have made the last four years much more fun and interesting. In particular, I would like to thank Mo Dick for many discussions and dinners, his help and a great number of cups of tea.

Finally, I would like to give special thanks to Christiane and my parents, Sabine and Kurt, for their encouragement and support throughout the years. Without them I would not be the person I am today.

Lastly, I gratefully acknowledge the financial support of ERC grant UQMSI/647812 and EPSRC grant EP/L016516/1 which have made this thesis possible.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.
It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution.
Chapter 1 is the product of my own work and gives an overview of the use of mathematical statistics in high and infinite dimensional inference and outlines the results of this thesis. Chapter 2 is published as [CKLN18] and is joint work with A. Carpentier, O. Klopp and R. Nickl. Chapter 3 is published as [KLN19] and is joint work with V. Koltchinskii and R. Nickl and finally chapter 4 is joint work with A. Picard. A previous version of chapter 4 was submitted to arXiv as [LP19] and the present version will be submitted for publication to a journal soon.

Contents

Chapter 1

Introduction

Mathematical models build the foundation of statistics. For a given data set, a model describes approximately how this data was generated and often contains the necessary information to draw meaningful conclusions from it. The mathematical analysis of a model leads to the extraction of this information in the form of statistical inference by considering estimators, confidence sets and tests. A mathematical statistician considers such models and investigates how to pursue inference in an 'optimal' way, attempting to solve many intriguing mathematical puzzles on his or her search for an answer to this question. This thesis is concerned with the analysis of three such models in high,- and infinite dimensions.

The basic assumption in statistics is the following: we observe data Y which is distributed according to some probability measure $\mathbb{P}$ acting on a polish space Ω. Usually the exact distribution $\mathbb{P}$ is not known and hence a statistician considers a multitude of possible probability measures in some broader family $\mathcal{P}$ such that $\mathbb{P} \in \mathcal{P}$. Since dealing directly with a probability measure can be difficult, we usually take a slightly different point of view: we parameterize each probability measure $\mathbb{P} \in \mathcal{P}$ with some parameter $\theta \in \Theta$, where Θ is contained in some other polish space, meaning that we find a bijection between $\mathcal{P}$ and Θ. To account for this relationship, we often write $\mathbb{P}_\theta$ for the measure generated by θ through the aforementioned mapping. For instance, if $\mathcal{P}$ denotes the class of univariate Gaussian distributions with unknown mean and variance one, then $\mathcal{P} = \{\mathcal{N}(\theta, 1), \ \ \theta \in \mathbb{R}\}$ and $\Theta = \mathbb{R}$. We call a family $\mathcal{P}$ *parametric* if Θ is finite dimensional, $\dim(\Theta) < \infty$, and *nonparametric* otherwise.

To enable statistical analysis in parametric statistics, it is typically assumed that the number of available data points, n, is of larger magnitude than the (constant) dimensionality of Θ. However, for many modern phenomena this assumption is too rigid and hence parametric statistical theory does not provide the necessary tools for these data sets. On the other hand, a completely nonparametric model might throw away too much

information and the number of data points might not be sufficient to allow for meaningful inference. For example, in the famous Netflix prize data set [BL07] 100 480 507 movie ratings were provided. These were given by 480 189 users to 17 770 movies. The goal was to predict the remaining 8 432 478 023 unknown ratings as well as possible. Hence, without further assumptions the number of free parameters is approximately 85 times the number of data points and therefore the required assumption of parametric statistics is not fulfilled. In fact, typically used estimators such as the Maximum-Likelihood-Estimator (MLE) may not even be defined in similar situations. This kind of problem has spawned a new area of statistics in recent years, known as *high-dimensional* statistics, where the dimensionality may depend on the number of available data points and exceed it by many magnitudes, see for example the monograph by Bühlmann and van de Geer [BvdG11] for a comprehensive exposition.

In this thesis, we will consider problems from this area and at its intersection with non-parametric statistics. We will focus on three models where the parameters are operators acting on a separable Hilbert space $\mathbb{H}$. If $\mathbb{H}$ is finite dimensional such an operator can be directly identified with a matrix. Many phenomena can be modelled by such operator or high-dimensional matrix models. First, we will consider the problem described above, *matrix completion*, and the existence of adaptive confidence sets in it. Another instance of a high-dimensional matrix model is inference about the main sources of variability in a data set. This is frequently used in Finance or Psychometrics and known in the statistics literature as factor analysis or *principal component analysis (PCA)*. The main parameter here is the covariance operator Σ of a collection of centred Gaussian random variables, but the object of interest are *functionals* of Σ such as its first eigenvector. In chapter 3 we will develop results for this model. Finally, in chapter 4 we will consider a Markov chain model where the observed data is not independent and how this data evolves over time is described by the *transition operator* P.

1.0.1 Structural assumptions

We are first going to introduce some typically used statistical models in high-dimensional and nonparametric statistics and discuss possible structural assumptions with an emphasis on matrix models.

The first prototypical model we consider is regression:

$$Y_i = \theta(X_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

with X_i being some known *design vectors* or *sensing matrices*, θ in some function class Θ and ε_i independent, mean zero errors. For instance, *linear (parametric) regression*

amounts to considering $\Theta = \{\theta(x) = \theta^T x, \quad \theta, x \in \mathbb{R}^p\}$, *sparse high-dimensional linear regression* to $\Theta = \{\theta(x) = \theta^T x, \quad \theta, x \in \mathbb{R}^p, \ \sum \mathbf{1}_{\theta_i \neq 0} \leq s\}$ for some monotone growing functions $s = s(n) : \mathbb{N} \to \mathbb{N}$ and $p = p(n) : \mathbb{N} \to \mathbb{N}$ and nonparametric regression to setting $\Theta = C^\beta$, or $\Theta = H^\beta$, $\beta > 0$, for some bounded Hölder or Sobolev balls of functions $\mathbb{R}^p \to \mathbb{R}$. Nonparametric regression and sparse (vector) linear regression have an extremely rich history and for conciseness we refer to the books [Tsy08, GN16] and [BvdG11], respectively, and references therein for an overview.

If θ is a matrix, two low rank regression models have been mostly considered: The reduced rank model and trace regression. The reduced rank model assumes a multivariate linear regression model with a $n \times p_2$ response variable Y and a $p_1 \times p_2$ regression matrix θ with possibly small rank. Historically, this model has been widely been considered in the econometrics literature from the 1950's onwards, assuming a parametric model, [RV98, Ize75, DT82, And51]. Relaxing the parametric assumption has led to the discovery of new phenomena and estimators, see for instance [YELM07, BSW11].
In chapter 2 we will be particularly interested in the case where in the regression model (1.1) X_i is a matrix in $\mathbb{R}^{p_1 \times p_1}$ and $\Theta = \{\theta(x) = \langle \theta, x \rangle, \quad \text{rank}(\theta) \leq r, \quad \theta \in \mathbb{R}^{p_1 \times p_2}\}$, where p_1, p_2 and r may again depend on n. This model is known as *trace regression* [RT11] and includes *matrix completion* [KLT11] and *quantum state tomography* [Kol11]. Other structural assumptions in matrix regression models have focussed mostly on sparsity. For instance, row or column sparsity [KT15, CW19], group sparsity [LPVDGT11] or submatrix localization [BI13, CLR17]. Moreover, combining these lines of work, simultaneous sparsity and low rank assumptions have been considered in [ANW12, MMS14, BSW12, LSHM10, YMB16, KLT17].

Covariance estimation is another typical matrix inference problem and closely tied to regression: considering the model

$$X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma), \tag{1.2}$$

for some covariance operator Σ, we define the sample covariance $\hat{\Sigma} = \frac{1}{n} \sum X_i X_i^T$. We can then rewrite the model as

$$Y = \Sigma + (\hat{\Sigma} - \Sigma) \tag{1.3}$$

and may view $\hat{\Sigma} - \Sigma$ as centred, non-i.i.d. noise. For an extensive treatment of parametric estimation of Σ we refer to the monograph by Eaton [Eat83]. Sparsity assumptions have been considered in the estimation of Σ, its inverse and its eigenvectors, too. For instance, row sparsity assumptions for Σ have been proposed and analyzed by Bickel and Levina [BL08a] and El Karoui [EK08] and sparsity of the leading eigenvectors of a 'spiked' covariance matrix by Johnstone and Lu [JL09] and Ma [Ma13]. Assuming spar-

sity of Σ^{-1} goes back to the analysis of graphical models by Dempster [Dem72], as zero entries of the precision matrix Σ^{-1} allow to conclude conditional independence of underlying variables and are thus highly desirable. His model has been extended to the modern high-dimensional setting with $p \to \infty$ by [MB06, YL07, CLL12, CLZ16]. Another assumption, motivated by longitudinal data, particularly autoregressive processes, is bandedness where the entries of Σ or Σ^{-1} decay polynomially when moving away from the diagonal [BL08b, CZZ10, WP03]. Other structural assumptions are Toeplitz structure, $\Sigma_{ij} = \sigma(|i-j|)$, as considered by Cai et al. [CRZ13] and approximate low rank and missing observations [Lou14].

The second prototypical model is density estimation where we assume that data is generated as

$$X_i \overset{\text{identically distributed}}{\sim} \theta, \qquad i = 1, \dots, n \tag{1.4}$$

for some probability density function θ. The classic nonparametric approach is to assume that in addition the X_i are independent and that $\theta \in H^\beta$ or C^β. We refer again to [GN16] for an extensive, contemporary treatment. Density estimation models and their combination with matrices are natural in the context of *Markov chains*. Indeed, consider a Markov chain with discrete state space $\{1, 2, \dots, p\}$ and presume that $(X_i)_{i \in \mathbb{N}}$ are discrete observations of this chain, i.e. $\mathbb{E}[X_i | X_{i-1}, \dots, X_1] = \mathbb{E}[X_i | X_{i-1}]$. The transition probabilities of going from one state to another form a $p \times p$ matrix, the *transition matrix*. The statistical literature about inference for the transition matrix goes back to Bartlett and Anderson and Goodman in the 1950's [Bar51, AG57]. Only recently, Zhang and Wang [ZW18] and Zhang et al. [LWZ18] have proposed a high-dimensional framework for Markov chains and have additionally imposed and motivated low rank assumptions.
Assuming that the state space of the Markov chain is not countable anymore, leads to the nonparametric equivalent of the transition matrix, the transition operator P, and its kernel, the transition density p:

$$Pf(x) = \mathbb{E}[f(X_2) | X_1] = \int p(x, y) f(y) dy, \qquad f \in L^2. \tag{1.5}$$

Assuming smoothness, estimating p nonparametrically is similar to density estimation and has been first considered in 1969 by Roussas [Rou69]. More recently, Zhang and Wang [ZW18] proposed to assume additionally that P has a known low rank.

Furthermore, it has been proven that in case $\Theta = H^\beta$ or $\Theta = C^\beta$, $\beta > p/2$, a multitude of statistical models are *asymptotically equivalent*: the density model with independent observations from above if constrained to $[0, 1]$ and bounded from below, nonparametric regression, the *Gaussian White noise model*, spectral density estimation, continuous time diffusion processes and Poisson processes with varying intensity [Nus96, BL96, BCLZ02,

Rei08, BCLZ04, DR06, GNZ10, RSH18]. Asymptotic equivalence means that, when passing from one model to the other by some transformation, no information is lost asymptotically. Asymptotic equivalence allows to convey lower bounds for minimax rates and gives indications how to construct estimators in these models. Historically, many phenomena have been first explored assuming the Gaussian white noise model,

$$dY(t) = \theta(t)dt + \frac{1}{\sqrt{n}}dW_t, \quad t \in [0,1]^p,$$

where W_t is a Brownian motion, as it is the easiest to handle and the equivalence results from the aforementioned references give some indication that the Gaussian white noise model may be viewed as the archetype of a nonparametric model. Asymptotic equivalence results for other parameter spaces are more sparse in the literature. One notable example in the context of high-dimensional matrix models is the article by Wang [Wan13] where asymptotic equivalence between quantum state tomography and noisy matrix completion is investigated.

1.0.2 Basic definitions: estimators, loss functions and minimax rates

In this section we review the basic notion of an estimator and ways to measure its quality.

Given a data set and having specified an appropriate family of models, the goal is to perform inference to answer open questions about phenomena which have given rise to this data. For instance, the statistician might be interested in investigating the relationship between mutation of certain genes in a person's DNA and his or her likelihood of developing a certain type of disease. This amounts to analyzing which specific models and parameters are appropriate and how uncertain these conclusions are.

Inferring which models are appropriate leads to the construction of an estimator for the unknown parameter θ. Sometimes also only a specific aspect of θ may be of interest. This can be described by a measurable map $\Psi : \Theta \to \mathcal{T}, \theta \mapsto \Psi(\theta)$. We call such a map *functional*. An estimator for θ is a measurable map $\hat{\theta} : \Omega \to \Theta,\ \omega \mapsto \hat{\theta}(\omega)$. In principle any such map is an estimator, but the statistician would also like to make sure that his or her choice of $\hat{\theta}$ is reasonable. This leads to the mathematical analysis of $\hat{\theta}$.

We will measure the performance of an estimator by how close it is to the truth on average or with high probability. Naturally, this performance depends on how is it measured. Such a measuring function is called a loss function and is often a metric induced by a norm on Θ. In nonparametric statistics, the Euclidean and ℓ_∞ norms are used, and in matrix inference the ℓ_∞ norm, the spectral norm and the Hilbert-Schmidt (Frobenius) norm. In addition, in the case of regression models (1.1) often the (squared) *prediction*

error

$$\|\theta - \hat{\theta}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\theta(X_i) - \hat{\theta}(X_i))^2 \tag{1.6}$$

is considered. Note that this is not necessarily a norm on Θ. However, under appropriate conditions on the design X_i, it can often be shown to be equivalent to the Euclidean loss for well behaved regression functions up to some, usually small, error. We will always consider the Euclidean distance as loss function.

Our key definition here is the minimax rate, the best possible performance in the worst case. Naturally, it depends on the model under investigation and the loss function used. We also emphasize the dependence on n, the number of observations or the signal to noise level.

Definition 1.0.1. *Given a family of models $\mathcal{P} = \mathcal{P}_n$ with corresponding parameter space Θ and a loss function $\ell : \Theta \times \Theta \to [0, \infty)$ the minimax rate over Θ is defined by*

$$r_{n,\Theta} := \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\tilde{\theta}, \theta), \tag{1.7}$$

where the infimum is taken over all measurable estimators $\tilde{\theta}$.

The definition above is nonasymptotic and holds for a fixed number of observations, n. However, constructing estimators that achieve this rate exactly for a finite n is often difficult. Moreover, it can be argued that in many cases an asymptotic analysis reflects sufficiently well the nonasymptotic behaviour, too, and hence most theoretical developments have been taken place under the assumption that n goes to ∞. This means that it is only desired that for an estimator $\hat{\theta}$

$$\limsup_n \sup_{\theta \in \Theta} \frac{\mathbb{E}_\theta \ell(\hat{\theta}, \theta)}{r_{n,\Theta}} = 1. \tag{1.8}$$

We call an estimator that fulfills (1.8) asymptotically exact minimax. In parametric statistics sufficient answers to asymptotic exact minimaxity are provided by the Hájek-Le Cam asymptotic minimax Theorem, yielding an asymptotic lower bound for $r_{n,\Theta}$, and, moreover, under mild regularity conditions it can be shown that the maximum-likelihood-estimator (MLE) attains the above bound asymptotically (see chapter 8 in [vdV98]).

In the context of nonparametric and high-dimensional models, obtaining estimators that are asymptotically exact minimax is much more difficult and is performed on a case by case basis. The first proof into this direction is due to Pinsker in 1980, where he determined the exact constant in the nonparametric Gaussian white noise model under squared error loss [Pin80]. His result was extended by Tsybakov [Tsy98] to pointwise and supremum

loss.
In the context of high-dimensional models asymptotic exact minimaxity was mostly studied in the normal means model where $Y = \theta + \varepsilon,\ \theta \in \mathbb{R}^n$. For instance, Donoho and Johnstone [DJ94b] constructed hard and soft thresholding estimators that asymptotically achieve the exact rate over nearly sparse parameter spaces. Other approaches that achieve the asymptotic exact rate in this model are empirical Bayes [JS04] and false discovery rate thresholding [ABDJ06]. Considering exactly sparse balls instead, only recently Wu and Zhou [WZ13] showed that a penalized least squares estimator achieves asymptotically the exact rate. Finally, in context of the sparse high-dimensional linear regression model, Su and Candès [SC16] extended the false discovery rate approach and proved asymptotically exact rates for sparse parameter spaces.

In many situations it is viewed as sufficient that only the order of the rate in (1.7) is matched. This means that asymptotically, as the number of observations n goes to infinity, an estimator $\hat{\theta}$ fulfills

$$\limsup_n \sup_{\theta\in\Theta} \frac{\mathbb{E}_\theta \ell(\hat{\theta}, \theta)}{r_{n,\Theta}} < \infty. \tag{1.9}$$

This is the notion of minimaxity we will employ in this thesis and which has been prevalent in the statistical literature. Moreover, as customary in high-dimensional statistics (compare e.g. [BvdG11]), we relax this even further and instead of attaining the minimax rate in expectation we will usually only prove that our procedures attain the minimax rate with high probability.

1.0.3 Minimax estimators

In this section we discuss some principles how to construct estimators for the whole parameter θ. In principle, for every problem there are different ways to construct estimators that attain the minimax rate (1.8). However, in case the loss function is given by the prediction error (1.6) a generalized least-squares estimator, named *minimum contrast estimator* by Pfanzagl [Pfa69] and first used by Huber [Hub67], can be shown to be minimax optimal in many cases as proven by Birgé and Massart in their seminal paper [BM93].

For illustration of this principle, consider the least squares estimator in the general regression model (1.1) from above:

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \left(Y_i - \theta(X_i)\right)^2. \tag{1.10}$$

Its prediction error can be closely related to the richness of the geometry of Θ. In particular, the key quantity is the local covering number $N(\epsilon, \delta, \Theta, \|\cdot\|_n)$, the smallest number of

ϵ-balls in $\|\cdot\|_n$ distance required to cover a ball of size δ in Θ. The logarithm of the covering number is known as *metric entropy*. Provided that Θ is not too complex such that the entropy integral below converges, van de Geer [vdG90] proved the following theorem which in this form can be found as Theorem 9.1 in her monograph [vdG00].

Theorem 1.0.1. *Assume that ε_i are centred, independent sub-Gaussian random variables. Suppose $\Psi(\delta)$ is such that*

$$\Psi(\delta) \geq \int_0^\delta \sqrt{\log N(u, \delta, \Theta_n, \|\cdot\|_n)} du \vee \delta$$

and that $\Psi(\delta)/\delta^2$ is non-decreasing in δ. Moreover, suppose δ_n is such that for a constant $c > 0$

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n).$$

Then, for another constant $C > 0$

$$\mathbb{P}\left(\|\hat{\theta} - \theta\|_n > \delta_n\right) \leq e^{-Cn\delta_n^2}. \tag{1.11}$$

This approach has been generalized to other statistical models such as density estimation by Birgé and Massart in their seminal article [BM93]. They also considered the case where the entropy integral above does not converge and showed that the obtained convergence rates for the least squares estimator (1.10) may be strictly sub-optimal. To circumvent this issue they proposed subsequently to restrict the minimization problem in (1.10) to a finite dimensional approximation space $\Theta^{(m_n)} \subset \Theta$, a *sieve* [BM98]. Sieves for the purpose of nonparametric density estimation had already been considered in 1962 by Cencov [Cen62], and were brought back to the attention of the nonparametric statistics community by Grenander in his monograph in 1981 [Gre81], who also coined the term sieve. Some other key references are [Sto80, Cox88] for nonparametric regression and [GH82, SW94, WS95] for nonparametric maximum-likelihood estimation. Traditionally, spans of wavelets, the Fourier basis or local polynomials have been used as sieves.
Some other popular approaches in nonparametric estimation include kernel density estimators (e.g. [GN16] for an overview), regression trees [Don97], random forests [Bre01, ATW19, MGS18] and sieves of neural networks [GBC16, BK19, SH19].
Moreover, in recent years, the Bayesian approach has been investigated in more depth in the frequentist nonparametric and high-dimensional statistics literature. In particular, if the prior is chosen well it can often be shown that the posterior measure concentrates around the true value with speed of convergence given by the minimax rate. For an extensive treatment of the subject we refer to the recent monograph by Ghosal and van der Vaart [GvdV17].

Having constructed an estimator which achieves the minimax rate of convergence for

a specific parameter space is usually just the first step. For instance, in the normal means model

$$Y_{ij} = \theta_{ij} + \varepsilon_i, \qquad i,j = 1,\dots,n, \qquad \operatorname{rank}(\theta) = r \tag{1.12}$$

the true rank r is usually not known. We call r a *hyperparameter*. Taking $\Theta = \mathbb{R}^{n\times n}$ in the least squares estimator (1.10) results in a sub-optimal Euclidean norm convergence rate of n^2, whereas if r were known it would be possible to achieve the rate rn. On the other hand, assuming r to be smaller than the true r yields an estimator whose convergence cannot be guaranteed at all. Hence, the statistician's goal is to construct estimators which do not require the knowledge of the true hyperparameter but still achieve the best possible convergence rate.

Definition 1.0.2. *Given a family of statistical models $\mathcal{P}_n$ with parameter space Θ and minimax rate $r_{n,\Theta}$ and a subfamily $\mathcal{P}_{n,0} \subset \mathcal{P}_n$ with parameter space $\Theta_0 \subset \Theta$ and minimax rate r_{n,Θ_0} and a loss function ℓ, we call an estimator $\hat\theta$ adaptive to Θ_0 if for some constant C not depending on n*

$$\sup_{\theta\in\Theta} \frac{\mathbb{E}_\theta \ell(\hat\theta,\theta)}{r_{n,\Theta}} \le C$$
$$\sup_{\theta\in\Theta_0} \frac{\mathbb{E}_\theta \ell(\hat\theta,\theta)}{r_{n,\Theta_0}} \le C.$$

In general it is not possible to construct adaptive estimators, see e.g. the recent article by Liu and Gao [LG18] for a counterexample. However, for many statistical models and loss functions adaptive estimators do exist. If for some model the construction of adaptive estimators is theoretically possible, there are several approaches to the construction. In the theoretical community the two most prevalent methods are penalization of minimum contrast estimators and Lepski's method. Other approaches to constructing adaptive estimators include (generalized) cross-validation [Sto74, Gei75, WW75, CW79, HM85], early stopping [ZY05, Büh06, BHR18a, BHR18b], aggregation of estimators (e.g. [BTW07, DGP18, DT08] and the recent review by Tsybakov [Tsy14]) or, when considering a Bayesian approach, empirical Bayes methods [Zha05, JS04, RS17, KSvdVvZ16] and hyperpriors (e.g. [Hua04, vdVvZ09, BG03, CvdV12, GvdVZ15]).

Lepski's method, proposed by Lepski in a series of papers in the beginning of the 1990's [Lep90, Lep91, Lep92] is based on multiple comparisons of a family of estimators. Its advantage is that it is fairly versatile, applicable in many statistical settings and for many estimators and loss functions and that it directly acts on the object of interest. Some examples include kernel estimators in the nonparametric white noise model [LMS97], nonparametric density estimation [Efr08], Tikhonov regularizers in inverse problems [BH05, Mat06], high-dimensional sparse linear regression [Zha13] and tail index esti-

mation in extreme value problems [BT15]. Moreover, Lepski's method has been extended to multivariate hyperparameters by Lepski and Goldenshluger [GL08, GL11, GL14], giving rise to the name 'Lepski-Goldenshluger-method'. We refer to chapter 8 in [GN16] for a thorough explanation of Lepski's method. However, in simulations it has been observed that the performance of Lepski's method is highly dependant on the choice of additional tuning parameters and hence one has to use it carefully in practice. As a partial remedy, Lacour and Massart [LM15] extend the concept of a minimal penalty (due to [BM07]) to the Lepski-Goldenshluger method and argue that choosing the tuning parameter close to the theoretically smallest possible is optimal. They also give some advice on how to implement their approach for real data sets [LM15]. Another possibility is to use a multiplier bootstrap as calibration tool as proposed by Chernozhukov et al. [CCK14a].

The second, widely considered approach to adaptation is penalization, particularly in the analysis of high-dimensional models. For illustration, consider again the regression model (1.1). Barron, Birgé and Massart [BBM99] showed very generally how to combine the method of sieves with penalization. Particularly, suppose that we are given a sequence of finite dimensional, linear sieves $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(m_n)} \subset \Theta$ with respective least squares estimators $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(m_n)}$. Now, for some constant $\lambda > 0$ to be specified, pick the estimator $\theta^{(\hat{k})}$ where

$$\hat{k} \in \underset{k \in \{1,\dots,m_n\}}{\arg\min} \left(\sum_{i=1}^{n} (Y - \hat{\theta}^{(k)}(X_i))^2 + 24\lambda \dim(\Theta^{(k)})/n \right). \tag{1.13}$$

If all sieves are contained in $\Theta^{(m_n)}$, this construction is equivalent to solving the following penalized problem:

$$\hat{\theta}^{(\hat{k})} \in \underset{\theta \in \Theta^{m_n}}{\arg\min} \left(\sum_{i=1}^{n} (Y - \hat{\theta}(X_i))^2 + 24\lambda \mathrm{Pen}(\theta)/n \right), \tag{1.14}$$

where $\mathrm{Pen}(\theta) := \min_{\Theta^{(k)}:\ \theta \in \Theta^{(k)}} \dim(\Theta^{(k)})$. This slightly different point of view has been taken by van de Geer [vdG01]. The statistical idea is that the penalty term acts as a countermeasure to overfitting by preventing overly complex models. Barron et al. [BBM99] proved the following theorem and generalized versions which apply to nonlinear sieves, too.

Theorem 1.0.2. *Assume that ε_i are centred, independent Gaussian random variables with unit variance. Suppose that the sieves $\Theta^{(i)}$ are linear and $\lambda \geq 1$ is such that*

$$\sum_{k=1}^{m_n} \exp(-\lambda \dim(\Theta^{(k)})) \leq C$$

for some constant $C > 0$. Then $\hat{\theta}^{(\hat{k})}$ satisfies for another constant $C' > 0$

$$\mathbb{E}\|\hat{\theta}^{(\hat{k})} - \theta\|_n^2 \leq C' \inf_k \left(\inf_{\theta_k \in \Theta^{(k)}} \|\theta - \theta_k\|_n^2 + \frac{\lambda \dim(\Theta^{(k)} + 1}{n} \right). \tag{1.15}$$

In nonparametric models the sieves can usually be picked in a subsequent manner, $\Theta^{(i)} \subset \Theta^{(i+1)}$, and λ of constant order. In high-dimensional sparse regression the sieves are not necessarily nested and it is necessary to pick $\lambda \asymp \log(p)$ to account for the fact that for each sparsity level s there are $\binom{p}{s} \leq p^s$ s-dimensional linear sieves. Inequalities such as (1.15) above are called *oracle inequalities* as they balance the first term, the bias, and the second term, the variance, as well as an oracle that would tell the statistician the optimal sieve $\Theta^{(k_{\text{opt}})}$. The representation (1.13) shows that the penalized least squares approach is closely related to traditional model selection procedures such as Mallows' CP [Mal73], AIC [Aka73] or BIC [Sch78] and, moreover, Birgé and Massart [BM01] proved that in the case of the Gaussian white noise model it is also equivalent to cross-validation. While Mallow's CP, AIC and BIC are based on direct considerations of the likelihood, [BBM99] argue that the right criterion for the choice of a penalty is the complexity of the sieves.

However, while being extremely general and also optimal from a theoretical perspective, this method may suffer in terms of practical applicability. Particularly, computing solutions of the least-squares estimators in (1.10) may be not feasible at all, and a total of m_n estimators have to be computed, adding an additional computational burden.
If the $\Theta^{(k)}$ are convex, computation of the least squares estimators (1.10) is usually feasible, either by methods from convex optimization, such as gradient descent, or there are explicit solutions of the optimization problem. For instance, in case of the nonparametric white noise model, if $\Theta^{(k)}$ is chosen as the span of the first k Fourier basis elements the solution of (1.10) are the first k empirical Fourier coefficients.

In high-dimensional problems, such as sparse linear regression or low rank matrix regression, the sieves $\Theta^{(k)}$ are usually not convex and the least square estimators in (1.10) have an explicit solution only in special cases . In particular, in sparse linear regression such an explicit solution for (1.10) usually does not exist unless $s = p$ or the design is orthogonal. Moreover, in sparse linear regression the penalty function in (1.14) equals $\text{Pen}(\theta) = \|\theta\|_0 := \sum \mathbf{1}_{\theta_i \neq 0}$ and hence the optimization problem in (1.14) is also not convex.
A remedy for this issue is to substitute the non-convex penalty function by a convex relaxation of it. For instance, in the sparse linear regression model Tibshirani [Tib96] proposed the *lasso* where the ℓ_0 penalty is substituted by the ℓ_1 norm. Thus, computation becomes more feasible, for example by (stochastic) sub-gradient descent or specialized algorithms

such as LARS [EHJT04]. In the case of generalized sparse linear models, including sparse linear regression, this convex relaxation can be proven to still attain the minimax optimal rates while being efficiently computable [BRT09, vdG08, Zha09]. Consequently, this methodology has also been brought forward in low rank regression problems such as trace regression [CP10, KMO10, CP11] and linear matrix regression [YELM07]. It can be shown to attain the optimal rates of convergence using the trace (nuclear) norm as relaxation of the rank [KLT11, RT11]. However, in the particular case of rank constrained models this relaxation seems not to be necessary as the non-convex optimization problem (1.14) has often an explicit solution. For instance, the solution of the rank penalized optimization problem proposed by Bunea et al. [BSW11] is given by a hard thresholded singular value decomposition (SVD) of a linear estimator. In fact, in models where explicit solutions to (1.14) exist it can often be seen that they and their convex relaxation counterpart are similar: the solution to (1.14) is usually a hard thresholded estimator whereas its convex relaxation via the ℓ_1 norm or the nuclear norm is soft thresholded. Besides matrix linear regression this is for instance the case in nonparametric regression with wavelet approximation spaces [DJ94a], the normal means model and linear regression with orthogonal design [DJ94b, Tib96] and trace regression [KLT11]. Hard thresholding has the practical benefit that the kept coefficients or singular values are not shrunk, hence reducing bias. Alternative, computationally feasible, approaches to reduce bias that have been proposed are to use the convex relaxed estimator as an initial estimator and then to fit an unbiased estimator on the selected model [BC13], to hard threshold the debiased lasso [vdG16] or to use the adaptive lasso [HMZ08, Zou07] where each coefficient in the lasso ℓ_1 penalty is penalized differently by an adaptively chosen weight.

Cai, Liang and Rakhlin [CLR16] showed how to extend the convex relaxation approach to linear models under general structural assumptions such as sparsity or low rank. Hence, one might be tempted to conjecture that convex relaxation always works in the sense that the corresponding estimators achieve the minimax rate. However, this is not the case. Pioneering work by Berthet and Rigollet [BR13] showed that minimaxity and computational efficiency do not have to go hand in hand. In particular, they considered sparse PCA and showed that it is necessary to pay an additional factor of $s^{1/2}$ in the minimax rate of testing when considering tests that are not NP-hard to compute. Similar phenomena have now been discovered in many other high-dimensional models. For instance, computational gaps appear in (sparse) sub-matrix localization [MW15, CW19, CLR17], sparse CCA [GMZ17], sparse and low rank logistic regression [BB18] and community detection [HWX15].

The choice of numerical constants in the penalization (tuning) parameter is also a debated subject as they influence (impact) largely the practical performance of estimators.

In their work on Gaussian model selection Birgé and Massart [BM01, BM07] observed that the tuning parameter must have at least some given numerical value to prevent inferior performance and that this minimal value is observable in the data. Moreover, they showed that the theoretically optimal tuning parameter is roughly twice the minimal parameter, leading to a feasible algorithm for tuning parameter selection. For further discussion and some numerical examples we refer to the recent survey by Arlot [Arl19]. Other approaches to prevent inferior numerical performance in high-dimensional sparse linear regression include the square root lasso [Owe07, BCW11, BCW14, Klo14] for which the tuning parameter does not depend on the error variance, sub-sampling (coined stability selection) [MB10], estimation of the variance and otherwise a design dependant fixed value [BC13] (see also [CCK13] for further justification for heavy-tailed errors and a combination with the bootstrap and heteroscedastic noise) and cross-validation [Tib96, ZHT07].

1.0.4 Functionals

A function of a parameter defining the model is a *functional*, i.e. $\Psi(\theta)$ for some map Ψ. For instance, in nonparametric density estimation, we might be interested in the value of the density function θ at a certain value X_0, $\Psi(\theta) = \theta(X_0)$ or the second moment of θ, $\Psi(\theta) = \int_0^1 \theta(x)^2 dx$. Another example would be the value of a high-dimensional vector θ at one specific coordinate θ_i, the average of θ or the Euclidean norm of θ, $\|\theta\|$.

In parametric statistics, assuming differentiability of Ψ, the Δ-method in combination with asymptotic efficiency of the MLE $\hat{\theta}$ provides a sufficient answer to the problem of optimally estimating functionals: the estimator $\Psi(\hat{\theta})$ is asymptotically efficient with limiting Gaussian distribution, too. In nonparametric and high-dimensional statistics, many more phenomena appear in the inference for functionals.

In particular, the minimax rate of a functional may differ from the minimax rate of the estimator for θ. In fact, in many cases even asymptotic efficient estimators with Gaussian limits exist. Particularly, this means that in such cases no bias-variance trade off has to be performed as there exist estimators with asymptotic bias $o(n^{-1/2})$. In nonparametric statistics such situations are well explored and for the sake of conciseness we refer to [vdV98]. The key realization is that a similar phenomena to the parametric case may hold, namely that nonparametric estimators exist that achieve the minimax rate of convergence measured by the Euclidean norm loss function and simultaneously satisfy central limit theorems. This was coined 'plug-in' property by Bickel and Ritov [BR03]. Nickl [Nic07] and Giné and Nickl [GN08b, GN09a, GN09b] showed that some commonly used estimators such as the nonparametric MLE, kernel density estimators and wavelet estimators fulfill this property in density estimation. For nonparametric integral functionals, which

are sufficiently smooth, we obtain by a second order functional Taylor expansion, that

$$\Psi(\hat{\theta}) - \Psi(\theta) = D\Psi(\theta)[\hat{\theta} - \theta] + O_{\mathbb{P}}(\|\hat{\theta} - \theta\|^2).$$

If the convergence rate in Euclidean norm is at least $n^{-1/4}$ (corresponding to a smoothness index larger than $p/2$) the second term vanishes asymptotically and hence, when $\hat{\theta}$ satisfies a (uniform) CLT, asymptotic efficiency and a Gaussian limiting distribution can be established. If the achievable convergence rate in Euclidean norm is less than $n^{-1/4}$ asymptotically efficient estimators may still exist, but a more careful case by case analysis has to be performed. For instance, estimating the second moment $\int \theta^2(x)dx$ at $\sqrt{n}$-rate with Gaussian limiting distribution is possible even when the achievable Euclidean norm convergence rate is slower than $n^{-1/4}$ [BR88, GN08a].

In high-dimensional statistics such 'plug-in' estimators unfortunately seem to exist only in situations without intrinsic low-dimensional structure such as sparsity or low rank. In these cases the typical assumption is that $p = o(n)$ and M-estimators are considered. Similarly to the nonparametric setting, it turns out that usually functionals of $\hat{\theta}$ are asymptotically efficient if $p^2 = o(n)$ and consequently Euclidean norm convergence rates are at least $n^{-1/4}$. This has already been noted by Huber in his seminal paper on robust regression [Hub73]. When $p^2/n \to \infty$ it is necessary to perform a case by case analysis, too. For instance, linear functionals of M-estimators in linear regression were considered in [YM79, Por84, Por85, Mam89] where it was shown that $p^{3/2}\log(n)^{2/3} = o(n)$ is a sufficient condition to establish $\sqrt{n}$-rates, asymptotic Gaussianity and efficiency. More recently, [KZ18, Kol17, JHW18] considered estimation of general non-linear functionals in regression, covariance estimation and a binomial model, respectively. They propose a novel, iterative 'bootstrap-chain' debiasing method which allows the construction of $\sqrt{n}$-consistent, asymptotically Gaussian and efficient estimators, even when $p^2/n \to \infty$ and only $p = o(n)$.

When the model has some intrinsic structure such as sparsity or a low rank, estimators such as the lasso are biased in possibly every coordinate and hence linear functionals of θ are not asymptotically Gaussian. As a remedy for this issue one step debiasing procedures were independently proposed by various authors [ZZ14, JM14, vdGBRD14]. For instance, consider the high-dimensional sparse linear regression model and assume that $\hat{\theta}$ is the lasso estimator and that the design vectors are Gaussian distributed with covariance matrix Ξ^{-1}. Then the debiased lasso is defined as

$$\tilde{\theta} := \hat{\theta} + \frac{\Xi X^T(Y - X\hat{\theta})}{n} = \theta + \Xi^T X^T \varepsilon/n + (I - \Xi^T X^T X/n)(\hat{\theta} - \theta).$$

Assuming a convergence rate in Euclidean norm of the initial estimator $\hat{\theta}$ of at least $n^{-1/4}$ the third term on the right-hand side above can be shown to be asymptotically negligible (in ℓ_∞ norm) and asymptotic Gaussianity and efficiency for linear functionals can be established. Moreover, it is possible to replace Ξ by a consistent estimator. If the convergence rate of $\hat{\theta}$ with respect to Euclidean loss is less than $n^{-1/4}$ it turns out that efficient estimation of linear functionals is still possible if and only if the covariance matrix of the design vectors X_i can be estimated at a sufficiently high convergence rate [JM18, CG17]. Carpentier and Kim [CK18] extended this methodology to low rank trace regression. Moreover, debiased estimators were also considered in generalized linear models [NL17], sparse precision matrices [JvdG15], sparse PCA [Jv18] and z-estimation [BCCW18]. Other, possibly asymptotically efficient, functionals were also investigated in the sparse setting, e.g. the correlation between two regressors [GWCL19] and explained variance [CG18b].

On the other hand, if no $\sqrt{n}$-convergence rates are achievable, similar phenomena occur in nonparametric and high-dimensional theory. Instead of using debiased estimators it is necessary to perform a careful bias-variance trade-off, leading to penalized (thresholded) estimators and the use of Lepski's method, see for instance [CCTV18, FRW15, VG18] for sparse high-dimensional models. Additionally, it has been observed that adapting to an unknown hyperparameter such as sparsity or smoothness may come at a price. In particular, this issue has been noted first by Lepski [Lep90, Lep92] when estimating $\theta(X_0)$ in the Gaussian white noise model where he showed that one has to pay a logarithmic price in the minimax rate. Cai and Low [CL05] extended this result to more general linear functionals and also gave examples where the rate penalty may even be of polynomial order. More recently, in the context of estimating $\theta(X_0)$ in a contaminated nonparametric density model Liu and Gao [LG18] have shown that adaptation to the contamination level *and* the smoothness simultaneously is impossible at all. In the context of high-dimensional models this phenomena has also been observed, for example in the sparse normal means model when estimating $\sum \theta_i$ [CCTV18].

1.0.5 Confidence sets

Confidence sets and statistical tests both help to quantify the uncertainty of the process of drawing conclusions from the data. A confidence set $C = C(Y)$ is a measurable, data dependant subset of Θ, such that for the true parameter θ the probability that θ lies in the confidence set has at least a prescribed value,

$$\mathbb{P}_\theta(\theta \in C) \geq 1 - \alpha, \qquad 0 < \alpha < 1. \tag{1.16}$$

If the inequality above holds for any arbitrary $\theta \in \Theta$ we call C *honest* at level $1-\alpha$ and if the inequality holds asymptotically we call C *asymptotically honest*.
Similar to the case of estimators we measure the performance of a confidence set by its diameter in some given metric.

In parametric statistics, asymptotically honest confidence sets can be constructed by using asymptotic normality of the MLE. Indeed, in parametric models under some regularity conditions (e.g. [vdV98])

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, I(\theta)^{-1}),$$

where $I(\theta)$ denotes the Fisher-information at θ. Assuming continuity of $I(\theta)$ at the true θ, it is possible to estimate it consistently with the plug-in estimator $I(\hat{\theta})$ and hence to construct an asymptotic $1-\alpha$ confidence set with diameter of order $n^{-1/2}$. Similar ideas often work for functionals for which estimation at $\sqrt{n}$-rate is possible and in chapter 3 we will consider such a case in detail.

In situations where no $\sqrt{n}$-rate estimation is possible constructing confidence sets may be more difficult. A lower bound for the diameter of a confidence set is the minimax rate of estimation (see section 6.4.1. in [GN16]) and so it is desirable that a confidence set attains this bound.

For the purpose of visualization confidence bands, confidence sets defined through the ℓ_∞ norm, are widely considered in nonparametric statistics. One possibility for their construction is to use results from extreme value theory. Indeed, it is possible to prove extreme value type limit Theorems for the supremum of minimax optimal estimators centred around their expectation with normalizing factor $r_{n,\Theta}^{-1}$. Consequently, by slightly undersmoothing, it is possible to build asymptotic confidence bands with almost minimax optimal diameter [BR73, GN10]. However, Hall [Hal91] proved that convergence to the limit is slow and thus the confidence bands are honest only for extremely large n . Hence, to construct honest confidence bands the bootstrap may be used for smaller values of n [CCK14b, CCK14a, CvK03, HH13]. Other possibilities to construct confidence bands in nonparametric problems are the use of Rademacher complexities and multiscale approaches (see chapter 6.4.2 in [GN16]). Similarly, in high-dimensional statistics, uniform confidence sets have been considered and the bootstrap has been advocated for practical implementation. Building on the seminal work by Chernozhukov et al. [CCK13] on high-dimensional Gaussian approximation, Belloni et al. [BCK15] first proposed to bootstrap debiased estimators in high-dimensional sparse linear regression models. Furthermore, modified versions of the bootstrap in this setting have been proposed in [DBZ17] and [ZC17].

Confidence sets with respect to the Euclidean metric ('confidence balls') are at a first glance less appealing in high and infinite dimensions as they are harder to visualize. However, the Euclidean diameter of a confidence set centred at some estimator may be viewed as a proxy for the Euclidean loss of this estimator and is thus a useful performance measure. Estimating the Euclidean loss for the purpose of constructing confidence balls has been proposed by Juditsky and Lambert-Lacroix [JLL04] and by Robins and van der Vaart [RVDV07] in nonparametric models and has been further investigated in high-dimensional sparse linear regression by Nickl and van de Geer [Nv13] and Cai and Guo [CG18a].

If the hyperparameters defining the model are unknown, constructing *adaptive* confidence sets with diameter adapting to the minimax risk of some smaller sub-model may be impossible at all. This is due to the fact that the minimax rate of testing between two models is a lower bound for any honest confidence set [JLL04, HN11, GN16] and will be explained in more detail in chapter 2.2.

1.0.6 Minimax lower bounds

In this section we discuss approaches to derive lower bounds for the minimax risk (1.7) and the minimax diameter of confidence sets in high and infinite dimensions. For more details and proofs we refer to chapter 2 in the monograph by Tsybakov [Tsy08].

For deriving minimax lower bounds of estimation, the trinity of Le Cam's method [LeC73], Assouad's Lemma [Ass83] and Fano's inequality provides sufficient tools for most problems. Fano's inequality is originally due to Fano [Fan61] but was developed into its modern version by [IK82, KT93, Tsy08]. The three methods are closely related as discussed for example by Yu [Yu97] and more extensively by Guntuboyina in his PhD thesis [Gun11].

The basic principle of all three methods is to lower bound the minimax risk via a testing argument. Consequently, it is necessary to construct a null hypothesis and a subspace of alternative hypotheses. The alternative hypotheses are supposed to appear close to the null hypothesis in the data with reference to some information theoretic distance such as the Kullback-Leibler distance, but are actually as far apart as possible from each other when their distance is measured with respect to the loss function.

For estimating the whole parameter θ Fano's inequality and Assouad's Lemma have been prevalent in nonparametric and high-dimensional statistics. In nonparametric statistics wavelets are particularly easy to handle in the construction of these alternative hypotheses. For this reason they have found multiple applications in the construction of lower bounds, see for example Kerkyacharian and Picard [KP92] and Donoho et al. [DJKP96] for lower bounds over Besov balls in density estimation by using Fano's inequality or Gobet

et al. [GHR04] for lower bounds for scalar diffusion processes observed in low frequency by using Assouad's Lemma.
In high-dimensional statistics Fano's inequality and a modified version of Assouad's Lemma have also been frequently used. For instance, Fano's inequality has been used to derive lower bounds in trace regression [KLT11], high-dimensional sparse linear regression [RWY11, RT11], sparse PCA [VL13] and for the estimation of sparse covariance matrices [RT12]. Moreover, Cai et al. [CZZ10] combined Le Cam's method with Assouad's Lemma to obtain bounds for banded covariance matrices. Using this new technique, Cai and Zhou proved lower bounds for sparse covariance matrices [CZ12] and Cai et al. for sparse precision matrices [CLZ16].

Constructing the space of alternative hypotheses is performed on a case by case basis. However, the resulting minimax bounds are usually closely tied to the geometric structure of the parameter space Θ. Hence, similar to the general theory for deriving upper bounds [vdG90, BM93], there have also been attempts to formulate a general theory for lower bounds based on the geometry of Θ. The close connection between entropy and rates had already been noticed by Birgé [Bir83] in 1983 in his consideration of the nonparametric density model. Building on his work Yang and Barron [YB99] formulated a general, geometric lower bound: it can be derived from the relationship between the entropy of a Kullback-Leibler ball and the entropy of a ball in the loss-function. Unfortunately, as Yang and Barron already noted, their bound appears to be tight only in situations where the entropy grows polynomially, i.e. in nonparametric settings. Thus, Ma and Wu [MW15] extended this work to high-dimensional matrices by using tools from convex geometry. They considered models where the loss-function is given by an arbitrary unitarily invariant norm and showed that this new method provides tight lower bounds for matrix completion with low rank constraint, sparse submatrix estimation and covariance matrix estimation.

To prove lower bounds for nonlinear functionals, such as $\|\theta\|_2^2$, Fano's inequality is also used, too, see for instance Collier et al. [CCT17] for lower bounds for quadratic functionals in the sparse normal means model. An alternative to the use of Fano's inequality is the same approach as for testing lower bounds. It is based on constructing a set of alternative distributions that are close to a null hypothesis in total variation distance but for which the functionals are far apart. The main difference lies in the proofs: it is necessary to bound the χ^2-distance between a null distribution and a mixture measure of alternatives instead of the KL-distance between null and each alternative separately. This approach has been successfully employed for instance in the proof of lower bounds for a linear functional by Collier et al. [CCT17] and for a quadratic functional of a sparse covariance matrix by Fan et al. [FRW15].

For linear functionals in convex parameter spaces there is a more involved theory available: for nonparametric densities Donoho and Liu [DL91a, DL91b] proved general lower bounds in terms of the interplay between functional and geometry of Θ by using Le Cam's method. Since for functionals adaptive rates may differ from the nonadaptive rates [Lep90, Lep92], in subsequent work Brown and Low [BL96] and Cai and Low [CL05, CL04a] developed new lower bound techniques, showing different general lower bounds. In high-dimensional statistics the only work regarding adaptivity of linear functionals that we are aware of is Collier et al. [CCTV18] who prove that when estimating a linear functional in the sparse normal means model one *may* have to pay a price for adaptation depending on the asymptotic regime: either no price, a log log-factor or a log-factor.

If estimation of functionals at a $\sqrt{n}$-rate is possible, sharp semiparametric information bounds, indicating asymptotic efficiency, are of interest, too. For functionals in nonparametric statistics these bounds are similar to the parametric case. The main difference is that the nonparametric nature of the model may deteriorate the achievable variance, see for example chapter 25 in [vdV98]. The bounds obtained this way are asymptotic in nature. An alternative is to use van Trees' inequality [vT68, GL95] which provides a non-asymptotic minimax lower bound containing the Fisher Information. For instance, Dalalyan et al. [DGT06] used van Trees' inequality to prove sharp lower bounds in the estimation of a shift parameter in the Gaussian white noise model.
In high-dimensional statistics parametric asymptotic lower bounds do not apply anymore since the models are not fixed as n grows. In high-dimensional sparse linear regression van de Geer et al. [vdGBRD14] embedded their high-dimensional model into an infinite dimensional model where the classic semi-parametric bounds apply. To circumvent this embedding, allowing for less stringent assumptions, Janková and van de Geer [JvdG18] extended the parametric Cramér-Rao inequality and Le Cam's lower bounds (see e.g. [vdV98]) for asymptotically unbiased estimators to the high-dimensional case.

Lower bounds for adaptive confidence sets are determined by the addition of estimation lower bounds and testing lower bounds for composite hypotheses (see chapter 6.4 in [GN16]). The concept of the minimax rate of testing and the development of testing lower bounds were pioneered by Ingster, see for instance the monograph Ingster and Suslina [IS03] for an overview. Typically, testing lower bounds are more difficult to derive than estimation lower bounds and no general theory is available. The commonly used approach due to Ingster is to construct a product measure of alternative hypotheses, then to lower bound the minimax risk by the Bayes risk over this prior and then to bound the χ^2-distance bet-ween the null-hypothesis and the product measure of alternatives. In high-dimensional statistics this approach was successfully used to construct testing lower bounds for the simple null hypothesis $H_0 : \theta = 0$ against a composite alternative

for instance by Ingster et al. [ITV10] and Arias-Castro et al. [ACCP11] in the sparse high-dimensional regression model, by Mukherjee et al. [MPL15] in a sparse binary regression model and by Carpentier and Nickl [CN15] in low rank trace regression. Usually, the testing lower bounds for the simple null hypothesis $H_0 : \theta = 0$ can be extended to composite null hypotheses, thus providing lower bounds for the diameter of adaptive confidence sets, see for instance Nickl and van de Geer [Nv13] for such an extension in the high-dimensional sparse linear regression model. Using Ingster's strategy, testing lower bounds for functionals are proven in the same way. For instance, Berthet and Rigollet [BR13] proved lower bounds for testing the value of the largest eigenvalue in sparse PCA and Cai and Guo [CG17] developed lower bounds for the diameter of confidence sets of linear functionals in high-dimensional sparse linear regression.

1.0.7 Contributions

In this thesis, we consider three problems from the areas of constructing estimators, estimating functionals and constructing confidence sets at the intersection of high-dimensional and nonparametric matrix and operator inference. In the following, we briefly summarize our main contributions.

Adaptive confidence sets for matrix completion

In chapter 2, we consider the construction of adaptive confidence balls in two closely related matrix completion models, the trace regression model and Bernoulli model, assuming in both cases an unknown low rank of the target matrix $M_0 \in \mathbb{R}^{m_1 \times m_2}$. Additionally, we assume that M_0 is bounded by some $a > 0$. The typical application of matrix completion models are recommender systems [BL07, GNOT92], but they have also been proposed in genomics [CZC+13] and sensor localization [Sin08].
In the trace regression model noisy entries from the target matrix are sampled with replacement. This means, that the data is given as

$$Y_i = \langle X_i, M_0 \rangle + \varepsilon_i, \qquad i = 1, \dots, n,$$

where the ε_i are independent, bounded, mean zero random variables and the X_i are chosen uniformly at random from the set of matrices with exactly one nonzero entry which equals one. Particularly, this means that in the trace regression model, it is possible to observe an entry of the unknown matrix M_0 multiple times.
On the other hand, the Bernoulli model is a normal means model with entries missing at random. Here, data is given as

$$Y_{ij} = B_{ij}((M_0)_{ij} + \varepsilon_{ij}), \quad 1 \le i \le m_1,\ 1 \le j \le m_2,$$

where the B_{ij} are independent Bernoulli random variables with success probability $p = n/m_1m_2$. Here sampling an entry twice is impossible. The main assumption in both models is that M_0 is a matrix of rank r with r being much smaller than $\min(m_1, m_2)$. In both models it can be shown that the matrix lasso estimator proposed in the trace regression model by Koltchinskii et al. [KLT11] achieves the optimal rate of convergence in Frobenius norm up to a logarithmic factor with high probability (for a proof in the Bernoulli model see Proposition 2.8.3).

For the trace regression model, we give an explicit construction of a confidence ball, even when the error variance is unknown. If the variance is known, we use a χ^2-statistic which has been also used in trace regression models with Gaussian design by Carpentier et al. [CEGN15] and prove in Theorem 2.4.1 that it is honest and has Frobenius diameter adapting to the minimax risk up to a log-factor.
The unknown variance case is more challenging. Inspired by Robins and van der Vaart [RVDV07], we propose to estimate the unknown risk via the U-statistic (2.19). In our construction we split the sample into two equally sized batches. We use the first sample to construct a minimax optimal estimator such as the matrix lasso $\hat{M}$ and the second sample to estimate the risk unbiasedly. In the construction of the U-statistic, we consider only entries that have been sampled at least twice. The number of such samples is random, but assuming the uniform sampling regime, we show in Lemma 2.4.1 that with high probability there are sufficiently many for our construction. In Theorem 2.4.2 we show that the constructed confidence set is asymptotically honest, and, that its Frobenius diameter adapts up to a log-factor to the minimax rate of estimation.

The information geometry of the Bernoulli model turns out to be completely different: in Corollary 2.5.1 and Corollary 2.5.2 we show that adaptive confidence sets do only exist if the error variance is known. The proofs of these results are both based on testing arguments:
If the variance is known, we give a theoretical existence proof based on an upper bound for the minimax rate of testing for a composite null hypotheses containing all matrices with bounded rank and the existence of an oracle estimator. To bound the minimax rate of testing, we employ in Theorem 2.5.1 an 'infimum'-type test (see e.g. [BN13, Nv13]). In its proof, we bound the type II error probability by using Talagrand's inequality [Tal96] in combination with sharp bounds for the spectral norm of random matrices due to Bandeira and van Handel [BvH16]. Moreover, we show in Proposition 2.8.3 that the matrix lasso estimator by Koltchinskii et al. [KLT11] fulfills the required oracle inequality in the Bernoulli model, employing again the bounds due to [BvH16].
In the unknown variance case, we prove in Theorem 2.5.2 a lower bound for the minimax rate of testing, which implies the non-existence of adaptive confidence sets. In our

proof, we follow the approach of Nickl and van de Geer [Nv13]. We first reduce the testing problem to the simple null hypothesis $H_0 : M_0 = 0$ and then prove testing lower bounds following Ingster's approach. The main conceptual difference to previous works (e.g. [ITV10]) is that we allow for any arbitrary, bounded, mean zero noise instead of using strictly Gaussian distributed error random variables. This might be more realistic in the matrix completion setting, as in applications such as movie recommender systems [BL07] all observed entries are bounded.

Efficient Estimation of Linear Functionals of Principal Components

In chapter 3, we focus on principal component analysis (PCA), which is a commonly used dimension reduction technique for high-dimensional data sets. Assuming a general framework where the data lies in a Hilbert space $\mathbb{H}$, PCA and our results are applicable to a wide range of problems, including functional data analysis [RS05] and kernel PCA in machine learning [BBZ07].
We assume a model where for some covariance operator Σ Gaussian data is given as

$$X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma),$$

and we measure the model complexity by the *effective rank*,

$$\mathbf{r}(\Sigma) := \frac{\operatorname{trace}(\Sigma)}{\|\Sigma\|}.$$

The effective rank is the reciprocal of percentage of variance explained by the first principal component and captures the difficulty of estimating Σ in spectral norm as proven by Koltchinskii and Lounici [KL17a]. Particularly, the effective rank may be of a much smaller order than the dimension of $\mathbb{H}$, for instance if the eigenvalues of Σ decay polynomially or exponentially. A small effective rank compared to the dimensionality of the data is also observed empirically in some applications, see for instance Novembre et al. [NJB+08] for an example in genomics where the data lies in $\mathbb{R}^{500.568}$, but the effective rank is roughly 3.
Since Σ is symmetric and positive definite, it has an eigendecomposition,

$$\Sigma = \sum_{i \geq 1} \lambda_i \left(\theta_i \otimes \theta_i\right).$$

We study estimation of linear functionals of eigenvectors of Σ, $\langle \theta_r, u \rangle$, by proving Berry-Esseen type bounds and bounds for the convergence of moments.

We start by proving in Theorem 3.3.2 that the plug-in estimator $\langle \hat{\theta}_r, u \rangle$ is asymptotically Gaussian with $\sqrt{n}$-rate of convergence when centred around its expectation. The

main tools for the proof are a representation of the bias term and a concentration inequality for higher order perturbation terms, which were both proven by Koltchinskii and Lounici in [KL16]. As typical in high-dimensional statistics [Hub73, vdGBRD14, GZ16], we are able to establish asymptotic Gaussianity if the complexity parameter is of smaller order than $\sqrt{n}$: if $\mathbf{r}(\Sigma) \ll \sqrt{n}$, Corollary 3.3.3 shows that the bias is of smaller order than $n^{-1/2}$ and hence that the plug-in estimator $\langle \hat{\theta}_r, u \rangle$ attains a $\sqrt{n}$-convergence rate with Gaussian limit.
Moreover, in Theorem 3.3.4 we give a sharp minimax lower bound in quadratic loss. This shows that the plug-in estimator is in fact asymptotically exact minimax and efficient if $\mathbf{r}(\Sigma) \ll \sqrt{n}$ and thus implies semiparametric optimality of $\hat{\theta}_r$. The proof is based on an application of van Tree's inequality [vT68]. As van Tree's inequality is non-asymptotic, we prevent possible issues with high-dimensional asymptotic formulations.

If $\mathbf{r}(\Sigma) \gg \sqrt{n}$, the bias is non-negligible. We calculate the value of the second order perturbation term of $\langle \hat{\theta}_r, u \rangle$ and prove that it is of order $\mathbf{r}(\Sigma)/n$. Hence, in this case the plug-in estimator does not even achieve a $\sqrt{n}$-rate of convergence (as proven in Proposition 3.3.5).
Nevertheless, contrary to the situation in sparse linear regression [CG17, JM18], efficient estimation at $\sqrt{n}$-rate is possible when $\mathbf{r}(\Sigma) \gg \sqrt{n}$. We develop a novel method of bias reduction by estimating the bias and correcting the plug-in estimator. In our construction we use two samples of size $o(n)$ to estimate the bias parameter and a third sample containing the majority of the data for estimating the plug-in estimator. In Theorem 3.3.6 we show that this newly constructed estimator attains a $\sqrt{n}$-rate of convergence rate with optimal variance in the Gaussian limit as long as $\mathbf{r}(\Sigma) = o(n)$.
Finally, by using perturbation theory for linear operators, we show in Corollary 3.6 that the asymptotic variance can be substituted by its empirical version to obtain a standard Gaussian limit.

Spectral thresholding for the estimation of Markov chain transition operators

In chapter 4, we consider the estimation of the transition operator and the corresponding transition density of discrete observations $X_1, \dots, X_n$ of a Markov chain. The distribution of the Markov chain is characterized by the transition operator

$$Pf(x) = \mathbb{E}[f(X_1)|X_0 = x] = \int f(y)p(x,y)\mathrm{d}y, \qquad f \in L^2.$$

In contrast to previous works in nonparametric statistics (e.g. [Lac07, Clé00]) the central object of our assumptions is the transition operator P instead of the transition density p.

Motivated by low frequency observations of periodised, reversible diffusion processes, we

assume an approximatively low rank structure of the transition operator with exponentially decaying singular values. In addition, we assume that the corresponding left and right singular functions have finite Sobolev norm.
In Lemma 4.2.1 we show that these assumptions are indeed met for periodised, reversible diffusion processes. The proof follows after an application of Weyl's law [Wey11] for operators with non-smooth coefficients due to Ivrii [Ivr00] and also uses elliptic regularity results for p.d.e's developed in a recent article by Nickl and Ray [NR19].

We propose a new algorithm that combines a Galerkin-estimator by Gobet et al. [GHR04] with low rank matrix estimation techniques by Klopp [Klo11] by hard thresholding the singular values of the initial estimator. This is closely related to approaches used in molecular dynamics (e.g. [SMP14, CSP+07, CKL+08]), where only the first few singular pairs are kept.

In the first part of Theorem 4.2.1 we show improved convergence rates of this estimator compared to situations without singular value decay. Particularly, the dependance of the dimension d on the nonparametric rate increases, up to log-factors, from $2d$ to d. The proof is constructed in several steps. We first derive novel spectral norm bounds for various quantities appearing in our estimator by using a Bernstein inequality for non-reversible Markov chains due to Jiang et al. [JSF18]. Afterwards, we apply the general proof for rank penalized estimators developed by Klopp [Klo11], bound the low rank approximation error and the smoothness approximation error. Additionally, we obtain in Lemma 4.2.3 that the rank of our estimator is, with high probability, of polylogarithmic order.

In the second part of Theorem 4.2.1 we prove a tight matching lower bound. For the proof we use Fano's inequality and construct the alternative hypotheses by combining wavelets with projection matrices, adapting an idea by Koltchinskii and Xia [KX15] in quantum state tomography to our setting.

Finally, in section 4.2.5 we illustrate our theoretical results with simulated data. We simulate discrete observations of two reversible diffusion processes, an Ohrnstein-Uhlenbeck process and a Cox-Ingersoll-Ross (CIR) process. Comparing our estimator to the Galerkin-estimator [GHR04] without hard thresholding of singular values, we observe a considerably improved performance.

Chapter 2

Adaptive confidence sets for matrix completion

2.1 Introduction

In matrix completion we observe n noisy entries of a data matrix $M = (M_{ij}) \in \mathbb{R}^{m_1 \times m_2}$, and we aim at doing inference on M. In a typical situation of interest, n is much smaller than $m_1 m_2$, the total number of entries. This problem arises in many applications such as recommender systems and collaborative filtering [BL07, GNOT92], genomics [CZC+13] or sensor localization [Sin08]. Two statistical models have been proposed in the matrix completion literature: the *trace-regression model* (e.g. [CZ16, Klo14, KLT11, NW12, RT11]) and the *'Bernoulli model'* (e.g. [CR09, Cha15, Klo15]).
In the *trace-regression model* we observe n pairs (X_i, Y_i^{tr}) satisfying

$$Y_i^{tr} = \langle X_i, M \rangle + \epsilon_i = \mathrm{tr}(X_i^T M) + \epsilon_i, \quad i = 1, \dots, n, \tag{2.1}$$

where (ϵ_i) is a noise vector. The random matrices $X_i \in \mathbb{R}^{m_1 \times m_2}$ are independent of the ϵ_i's, chosen uniformly at random from the set

$$\mathcal{B} = \left\{ e_j(m_1) e_k^T(m_2),\ 1 \le j \le m_1,\ 1 \le k \le m_2 \right\}, \tag{2.2}$$

were the $e_j(s)$ are the canonical basis vectors of $\mathbb{R}^s$. In this model Y_i^{tr} returns the noisy value of the entry corresponding to the random position X_i.

In the *Bernoulli model* each entry of $M + E$, where $E = (\epsilon_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of random errors, is observed independently of the other entries with probability $p = n/(m_1 m_2)$. More precisely, if $n \le m_1 m_2$ is given and B_{ij} are i.i.d. Bernoulli random variables of parameter p independent of the ϵ_{ij}'s, we observe

$$Y_{ij}^{Ber} = B_{ij} \left(M_{ij} + \epsilon_{ij} \right), \quad 1 \le i \le m_1, 1 \le j \le m_2. \tag{2.3}$$

The major difference between these models is that in the trace-regression model multiple sampling of a particular entry is possible whereas in the Bernoulli model each entry can be sampled at most once. A further difference is that in the trace regression model the number of observations, n, is fixed whereas in the Bernoulli model the number of observations $\hat{n} := \sum_{ij} B_{ij}$ is random with expectation $E\hat{n} = n$. Despite these differences, the results on minimax optimal recovery using computationally efficient algorithms for these two models in the literature are very similar and from a 'parameter estimation' point of view the models appear to be effectively equivalent (see, e.g., [CZ16, CT10, CP11, Cha15, Gro11, KMO10, Klo14, KLT11, NW12, Rec11]).

In this chapter we investigate questions that go beyond mere 'estimation' of the matrix parameter, namely about the existence of confidence sets for estimators $\hat{M}$ that adapt to the unknown rank of M. This question is in non-parametric and high-dimensional models a delicate matter (see e.g. [Bar04, BN13, CL04b, GN10, HN11, JLL04, Low97, NS16, Nv13, RVDV07, SvdVvZ15] and Chapter 8.3 in [GN16]) that depends on a rather subtle interaction of certain 'information geometric' properties of the model – the material relevant for this chapter is reviewed in Section 2.2. Many of these results reveal limitations by showing that confidence regions that adapt to the whole parameter space do not exist unless one makes specific 'signal strength' assumptions. For example, Low [Low97] and Giné and Nickl [GN10] investigated this question in nonparametric density estimation and Nickl and van de Geer [Nv13] in the sparse high-dimensional regression model.
Construction of confidence sets in the matrix completion setting is difficult mainly due to two reasons. Firstly, the Restricted Isometry Property (RIP) does not hold, requiring a more involved analysis than in a standard trace regression setting such as in [CEGN15]. Moreover, in most practical applications of matrix completion such as movie recommender systems [BL07, GNOT92] the variance of the errors is not known. Typical constructions of non-asymptotic confidence sets such as χ^2-confidence sets (e.g. [RVDV07, CEGN15]) require explicit knowledge of the variance and are thus not feasible. Particularly in the 'Bernoulli model', the problem of unknown variance can be expected to be potentially severe: for the related standard normal means model (without low rank structure and without missing observations) Baraud [Bar04] has shown that in the unknown variance case honest confidence sets of shrinking diameter do not exist, even if the true model is low dimensional. Similarly, in high-dimensional regression Cai and Guo [CG18a] prove the impossibility of constructing adaptive confidence sets for the l_q-loss, $1 \leq q \leq 2$, of adaptive estimators if the variance is unknown.
Our main contribution is that we show that in the case of unknown noise variance the information-theoretic structure of the two models considered is fundamentally different: in the trace regression model, even if only an upper bound for the variance of the noise is known, honest confidence sets exist that have Frobenius norm diameter that adapts to the unknown rank of M. Contrary to this, we prove that such confidence regions cannot

exist in the Bernoulli model when the noise variance is unknown. The construction of the confidence set in the trace-case is explicit and allows for an computationally efficient implementation. This is a stark contrast to the situation of *estimating* the unknown M_0 where results and algorithms (e.g. those in [KMO10, KLT11]) can often be conveyed from one model to the other.

To complement our findings we also show how to construct adaptive honest confidence sets for these two models in the case of known noise variance. However, even in this simpler case the algorithms used in each case do not convey to the other model. The construction for the trace-model is again easily implementable whereas we have to use a different, only theoretical approach in the Bernoulli model which is not computationally efficient.

This chapter is organized as follows: in Subsection 2.1.1 we formulate the assumptions and collect notation which we use throughout the paper. Then, in Section 2.2, we review and present general results about the existence of honest and adaptive confidence sets in terms of some information-theoretic quantities that determine the complexity of the adaptation problem at hand. Afterwards we review the literature on minimax estimation in matrix completion problems. In Section 2.4 we give an explicit construction of honest and adaptive confidence sets in the trace-regression case, adapting a U-statistic approach inspired by Robins and van der Vaart [RVDV07] (see also [GN16], Section 6.4, and [CEGN15]). Finally, we present our results for the Bernoulli model in Section 2.5. First, we derive an upper bound for the minimax rate of testing a low rank hypothesis and deduce from it the existence of honest and adaptive confidence regions in the known variance case. Then, we show that in the Bernoulli model, contrary to the trace-regression case, honest and adaptive confidence sets over the whole parameter space do not exist if the variance of the errors is not known a priori. Sections 2.7-2.8 contain the proofs of our results.

2.1.1 Notation & assumptions

By construction, in the Bernoulli model (2.3) the expected number of observations, n, is smaller than the total number of matrix entries, i.e. $n \leq m_1 m_2$. To provide a meaningful comparison we will assume throughout that $n \leq m_1 m_2$ also holds in the trace regression model (2.1). In many applications of matrix completion, such as recommender systems (e.g. [BL07, GNOT92]) or sensor localization (e.g. [BLWY06, Sin08]) the noise is bounded but not necessarily identically distributed. This is the assumption which we adopt in this chapter. More precisely, we assume that the ϵ_ι are independent random variables that are homoscedastic, have zero mean and are bounded:

Assumption 2.1.1. *In the models (2.1) and (2.3) with index $\iota = i$ and $\iota = (i,j)$, respectively, we assume $\mathbb{E}(\epsilon_\iota) = 0$, $\mathbb{E}(\epsilon_\iota^2) = \sigma^2$, $\epsilon_\iota \perp \epsilon_\eta$ for $\iota \neq \eta$ and that there exists a positive*

constant $U > 0$ such that almost surely

$$\max_{\iota} |\epsilon_\iota| \leq U.$$

We denote by $M = (M_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ the unknown matrix of interest and define

$$m = \min(m_1, m_2),$$
$$d = m_1 + m_2.$$

For any $l \in \mathbb{N}$ we set $[l] = \{1, \dots, l\}$. Let A, B be matrices in $\mathbb{R}^{m_1 \times m_2}$. We define the matrix scalar product as $\langle A, B \rangle := \mathrm{tr}(A^T B)$. The trace norm of the matrix A is defined as $\|A\|_* := \sum \sigma_j(A)$, the operator norm as $\|A\| := \sigma_1(A)$ and the Frobenius norm as $\|A\|_F^2 := \sum_i \sigma_i^2 = \sum_{i,j} A_{ij}^2$ where $(\sigma_j(A))$ are the singular values of A arranged in decreasing order. Finally $\|A\|_\infty = \max_{i,j} |A_{ij}|$ denotes the largest absolute value of any entry of A. Given a semi-metric $\mathcal{D}$ we define the diameter of a set S by

$$|S|_{\mathcal{D}} := \sup\{\mathcal{D}(x, y) : \quad x, y \in S\}.$$

Furthermore, for $k \in \mathbb{N}_0$ we define the parameter space of rank k matrices with entries bounded by a in absolute value as

$$\mathcal{A}(a, k) := \{A \in \mathbb{R}^{m_1 \times m_2} : \|A\|_\infty \leq a \quad \text{and} \quad \mathrm{rank}(A) \leq k\}. \tag{2.4}$$

Finally, for a subset $\Sigma \subset (0, U]$ we define

$$\mathcal{A}(a, k) \otimes \Sigma := \{(A, \sigma) : \ A \in \mathcal{A}(a, k), \ \sigma \in \Sigma\}.$$

As usual, for sequences a_n and b_n we say $a_n \lesssim b_n$ if there exists a constant C independent of n such that $a_n \leq C \cdot b_n$ for all n. We write $\mathbb{P}_{M,\sigma}$ (and $\mathbb{E}_{M,\sigma}$ for the corresponding expectation) for the distribution of the observations in the models (2.1) or (2.3), respectively.

2.2 Minimax theory for adaptive confidence sets

In this section we present results about existence of honest and adaptive confidence sets in a general minimax framework. To this end, let $Y = Y^n \sim \mathbb{P}_f^n$ on some measure space $(\Omega_n, \mathcal{B}), n \in \mathbb{N}$, where f is contained in some parameter space $\mathcal{A}$, endowed with a semi-metric $\mathcal{D}$. Let r_n denote the minimax rate of estimation over $\mathcal{A}$, i.e.

$$\inf_{\tilde{f}_n : \Omega_n \to \mathcal{A}} \sup_{f \in \mathcal{A}} \mathbb{E}_f \mathcal{D}(\tilde{f}, f) \asymp r_n(\mathcal{A}).$$

We consider an 'adaptation hypothesis' $\mathcal{A}_0 \subset \mathcal{A}$ characterised by the fact that the minimax rate of estimation in $\mathcal{A}_0$ is of asymptotically smaller order than in $\mathcal{A}$: $r_n(\mathcal{A}_0) = o(r_n(\mathcal{A}))$ as $n \to \infty$. In our matrix inference setting we will choose for $\mathcal{D}$ the distance induced by $\|\cdot\|_F$, for $\mathcal{A}_0, \mathcal{A}$ the parameter spaces $\mathcal{A}(a, k_0) \otimes \Sigma$, $\mathcal{A}(a, k) \otimes \Sigma$ from above, $k_0 = o(k)$ as $\min(n, m) \to \infty$, and data (Y_i, X_i) or (Y_{ij}, B_{ij}) arising from equation (2.1) or (2.3), respectively.

Definition 2.2.1 (Honest and adaptive confidence sets)**.** *Let $\alpha, \alpha' > 0$ be given. A set $C_n = C_n(Y, \alpha) \subset \mathcal{A}$ is a honest confidence set at level α for the model $\mathcal{A}$ if*

$$\liminf_n \inf_{f \in \mathcal{A}} \mathbb{P}_f^n(f \in C_n) \geq 1 - \alpha. \tag{2.5}$$

Furthermore, we say that C_n is adaptive for the sub-model $\mathcal{A}_0$ at level α' if there exists a constant $K = K(\alpha, \alpha') > 0$ such that

$$\sup_{f \in \mathcal{A}_0} \mathbb{P}_f^n \left(|C_n|_{\mathcal{D}} > K r_n(\mathcal{A}_0) \right) \leq \alpha' \tag{2.6}$$

while still retaining

$$\sup_{f \in \mathcal{A}} \mathbb{P}_f^n \left(|C_n|_{\mathcal{D}} > K r_n(\mathcal{A}) \right) \leq \alpha'. \tag{2.7}$$

We next introduce certain composite testing problems.

Definition 2.2.2 (Minimax rate of testing & uniformly consistent tests)**.** *Consider the testing problem*

$$H_0 : f \in \mathcal{A}_0 \quad \textit{against} \quad H_1 : f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq \rho_n \tag{2.8}$$

where $(\rho_n : n \in \mathbb{N})$ is a sequence of non-negative numbers. We say that ρ_n is the minimax rate of testing for (2.8) if

(i) $\forall \beta > 0$ $\exists$ a constant $L = L(\beta) > 0$ and a test $\Psi_n = \Psi_n(\beta)$, $\Psi_n : \Omega_n \to \{0, 1\}$ such that

$$\sup_{f \in \mathcal{A}_0} \mathbb{E}_f[\Psi_n] + \sup_{f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq L\rho_n} \mathbb{E}_f\left[1 - \Psi_n\right] \leq \beta. \tag{2.9}$$

We say that such a test Ψ_n is β-uniformly consistent.

(ii) For some $\beta_0 > 0$ and any sequence $\rho_n^* = o(\rho_n)$ we have

$$\liminf_{n \to \infty} \inf_{\Psi_n : \Omega_n \to \{0,1\}} \left[\sup_{f \in \mathcal{A}_0} \mathbb{E}_f[\Psi_n] + \sup_{f \in \mathcal{A},\ \mathcal{D}(f, \mathcal{A}_0) \geq \rho_n^*} \mathbb{E}_f\left[1 - \Psi_n\right] \right] \geq \beta_0 > 0. \tag{2.10}$$

Theorem 2.2.1. *Let ρ_n be the minimax rate of testing for the testing problem (2.8) and suppose that $\beta_0 > 0$ is as in (2.10). Suppose that*

$$r_n(\mathcal{A}_0) = o(\rho_n).$$

Then a honest and adaptive confidence set C_n that satisfies (2.5)-(2.7) for any $\alpha, \alpha' > 0$ such that $0 < 2\alpha + \alpha' < \beta_0$ does not exist. In fact if $3\alpha < \beta_0$, then for any honest confidence set C_n that satisfies (2.5) we have that

$$\sup_{f \in \mathcal{A}_0} \mathbb{E}_f |C_n|_{\mathcal{D}} \geq c\rho_n. \tag{2.11}$$

for a constant $c = c(\alpha) > 0$.

The first claim of this theorem is Proposition 8.3.6 in [GN16]. The lower bound (2.11) also follows from that proof, arguing as in the proof of Theorem 4 in [CN15].

A converse of Theorem 2.2.1 also exists, as can be extracted from Proposition 8.3.7 in [GN16] and an observation in Carpentier (see [Car13], proof of Theorem 3.5 in Section 6). For this we need the notion of an *oracle*-estimator.

Definition 2.2.3 (Oracle estimator)**.** *Let $\beta > 0$ be given. We say that an estimator $\hat{f}$ satisfies an oracle inequality at level β if there exists a constant C such that for all $f \in \mathcal{A}$ we have with $\mathbb{P}_f^n$-probability at least $1 - \beta$,*

$$\mathcal{D}(\hat{f}, f) \leq C \inf_{\tilde{\mathcal{A}} \in \{\mathcal{A}, \mathcal{A}_0\}} \left(\mathcal{D}(f, \tilde{\mathcal{A}}) + r_n(\tilde{\mathcal{A}}) \right). \tag{2.12}$$

This is a typical property of adaptive estimators, and is for example in the trace-regression setting fulfilled by the soft-thresholding estimator proposed by Koltchinskii et.al. [KLT11]. The following theorem proves that if the minimax rate of testing is no larger than the minimax rate of estimation in the adaptation hypothesis, then honest adaptive confidence sets do exist. The proof is constructive and yields a confidence set of non-asymptotic coverage at least $1 - \alpha$.

Theorem 2.2.2. *Let $\alpha, \alpha' > 0$ be given. Let ρ_n be the minimax rate of testing for the problem (2.8) such that a $\min(\alpha/2, \alpha')$-uniformly consistent test exists. Assume that $\rho_n \leq C' r_n(\mathcal{A}_0)$ for some constant $C' = C(\alpha, \alpha') > 0$. Moreover, assume that an oracle estimator $\hat{f}$ at level $\alpha/2$ fulfilling (2.12) exists. Then there exists a confidence set C_n that adapts to the sub-model $\mathcal{A}_0$ at level α' satisfying (2.6), (2.7) and that is honest at level α, i.e.,*

$$\sup_{f \in \mathcal{A}} \mathbb{P}_f^n \left(f \notin C_n \right) \leq \alpha.$$

2.3 Minimax matrix completion

First results on noisy matrix completion where obtained by Candès and Plan [CP10]. Using a semidefinite programming approach they construct an estimator $\widehat{M}_{SDP}$ and prove a by a $\sqrt{n}$-factor sub-optimal upper bound on the estimation error in Frobenius norm. Noisy matrix completion was further studied in several papers (see e.g. [CP10, CP11,

KMO10, KLT11, NW12, Klo14, Cha15, CZ16, Klo15, Rec11]). Optimal rates have first been achieved by Negahban and Wainwright [NW12] and Keshavan et al. [KMO10]. However, the construction of the estimator and the achievable upper bound in [NW12] requires knowledge of the 'spikiness' ratio of the unknown matrix and leads to sub-optimal rates in cases of matrices M with $\|M\|_F^2 = o(m_1 m_2)$ and $\max(|M_{ij}|) \geq c$. The bounds due to Keshavan et al. [KMO10] are only optimal for almost square matrices and are off by a factor of $\sqrt{\max(m_1, m_2)/\min(m_1, m_2)}$ in the case of arbitrary rectangular matrices. If the true rank is not known the best bounds in terms of the size of the set of matrices under consideration and the achieved convergence rates have up until now been obtained by Koltchinskii. et al. [KLT11] and Klopp [Klo14] for the trace-regression model and by Klopp [Klo15] for the Bernoulli model. For example, in the trace-regression setting, Klopp [Klo14] shows that a constrained Matrix Lasso estimator $\hat{M} := \hat{M}(a, \sigma)$ satisfies with $\mathbb{P}_{M_0,\sigma}$-probability at least $1 - 2/d$

$$\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} \leq C \frac{kd \log(d)}{n} \quad \text{and} \quad \|M_0 - \hat{M}\|_\infty \leq 2a \tag{2.13}$$

as long as $m \log(d) \leq n \leq d^2 \log(d)$ and where $C = C(\sigma, a) > 0$. Similarly, in the Bernoulli model with noise bounded by U it has been shown in Klopp [Klo15] that an iterative soft thresholding estimator $\hat{M} := \hat{M}(a, \sigma)$ satisfies with $\mathbb{P}_{M_0,\sigma}$-probability at least $1 - 8/d$

$$\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} \leq C \frac{kd}{n} \quad \text{and} \quad \|M_0 - \hat{M}\|_\infty \leq 2a \tag{2.14}$$

for $n \geq m \log(d)$ and for a constant $C = C(\sigma, a, U) > 0$. Matching lower bounds have also been shown by Koltchinskii. et a. [KLT11] and Klopp [Klo15]. In the trace-regression model with Gaussian noise we have for constants $\beta \in (0, 1)$ and $c = c(\sigma, a) > 0$ that

$$\inf_{\hat{M}} \sup_{M_0 \in \mathcal{A}(a,k)} \mathbb{P}_{M_0,\sigma} \left(\frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} > c \frac{kd}{n} \right) \geq \beta.$$

A similar lower bound can be obtained in the Bernoulli setting (see Klopp [Klo15]). These lower and upper bounds imply that for the Frobenius loss and the parameter space $\mathcal{A}(a, k)$ the minimax rate $r_{n,m}(\mathcal{A}(a, k))$ is (at most up to a log-factor) of order

$$\sqrt{m_1 m_2 kd/n}. \tag{2.15}$$

2.4 Trace Regression Model

We first consider the trace regression model. For the sake of precision we sometimes write M_0 for the 'true parameter' M that has generated the equation (2.1).
For notational simplicity we assume that n is even. Then we can split our observations in

two independent sub-samples of equal size $n/2$. In what follows all probabilistic statements are under the distribution $\mathbb{P}$ (with corresponding expectation written $\mathbb{E}$) of the first sub-sample $(Y_i^{tr}, X_i)_{i\leq n/2}$ of size $n/2 \in \mathbb{N}$, conditional on the second sub-sample $(Y_i^{tr}, X_i)_{i>n/2}$, i.e. we have $\mathbb{P}(.) = \mathbb{P}_{M_0,\sigma}(\,\cdot\,|(Y_i^{tr}, X_i)_{i>n/2})$.

2.4.1 A non-asymptotic confidence set in the trace regression model with known variance of the errors.

In this case we can adapt the construction of Carpentier et al. [CEGN15]. That is, we estimate the risk $\|\hat{M} - M_0\|_F^2/(m_1 m_2)$ by the χ^2-statistic (2.16) and center the confidence set around a minimax optimal estimator $\hat{M}$. In (2.16) subtraction of σ^2 is crucial since without this $\hat{R}_n$ would have a bias of $\sigma^2 = \Omega(1)$ and the diameter of C_n would be dominated by σ^2. The other quantities in the definition of (2.17) are required to make sure that the right hand side is with large enough probability an upper bound on $\|\hat{M} - M_0\|_F^2/(m_1 m_2)$ such that the confidence set has the desired coverage probability.
More precisely, using only the second sub-sample $(Y_i^{tr}, X_i)_{i>n/2}$ we compute the matrix lasso estimator from Klopp [Klo14] which achieves the bound (2.13) with probability at least $1 - 2/d$. Then, we freeze $\hat{M}$ and the second sub-sample. We define the following residual sum of squares statistic:

$$\hat{R}_n = \frac{2}{n} \sum_{i\leq n/2} (Y_i^{tr} - \langle X_i, \hat{M}\rangle)^2 - \sigma^2. \tag{2.16}$$

Given $\alpha > 0$, let $\xi_{\alpha,\sigma,U} = \sqrt{2}\sigma U \log(\alpha)$, $z_\alpha = \log(3/\alpha)$ and, for a $z > 0$, a fixed constant to be chosen, define the confidence set

$$C_n = \left\{ A \in \mathbb{R}^{m_1\times m_2} : \frac{\|A - \hat{M}\|_F^2}{m_1 m_2} \leq 2\left(\hat{R}_n + z\frac{d}{n} + \frac{\bar{z} + \xi_{\alpha,\sigma,U}}{\sqrt{n}}\right)\right\}, \tag{2.17}$$

where

$$\bar{z}^2 = \bar{z}^2(\alpha, d, n, \sigma, z) = z_\alpha \sigma^2 \max\left(\frac{3\|A - \hat{M}\|_2^2}{m_1 m_2}, 4zd/n\right).$$

It is not difficult to see (using that $x^2 \lesssim y + x/\sqrt{n}$ implies $x^2 \lesssim y + 1/n$) that

$$\mathbb{E}_{M_0,\sigma}\left[\frac{|C_n|_F^2}{m_1 m_2}\middle|\hat{M}\right] \lesssim \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} + \frac{zd + \sigma^2 z_{\alpha/3}}{n} + \frac{\xi_{\alpha,\sigma,U}}{\sqrt{n}}. \tag{2.18}$$

Markov's inequality, (2.18) and that $\hat{M}$ is minimax optimal (up to a log-factor) with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 2/d$ as long as $m\log(d) \leq n \leq d^2\log(d)$ imply that C_n has an adaptive and up to a log-factor minimax optimal squared diameter with probability $1 - \alpha'$ for any $\alpha' > 2/d$. The following theorem shows that C_n is also a honest confidence set:

Theorem 2.4.1. *Let $\alpha > 0$, $\alpha' > 2/d$ and suppose that $m\log(d) \leq n \leq d^2\log(d)$, that Assumption 2.1.1 is satisfied and that $\sigma > 0$ is known. Let $C_n = C_n(Y, \alpha, \sigma)$ be given by (2.17) with $z > 0$. Then, for every $n \in \mathbb{N}$ and every $M_0 \in \mathcal{A}(a, m)$,*

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \in C_n\right) \geq 1 - \frac{2\alpha}{3} - 2e^{-zd/(11a^2)}.$$

Hence, for any $1 \leq k_0 < k \leq m$, C_n is a honest and (up to a log-factor) adaptive confidence set at the level α for the model $\mathcal{A}(a,k) \otimes \{\sigma\}$ and adapts to the sub-model $\mathcal{A}(a,k_0) \otimes \{\sigma\}$ at level α'.

The proof of Theorem 2.4.1 follows the lines of the proof of Theorem 2 in [CEGN15] and we omit it here as the unknown variance results considered in the next section straightforwardly imply the known variance results.

2.4.2 A non-asymptotic confidence set in the trace regression model with unknown error variance.

In this subsection we assume, that the precise knowledge of the noise variance σ is *not* available, although the quantities a, U are available to the statistician (i.e. upper bounds on the matrix entries and on the noise). Instead, we assume that σ belongs to a known set $\Sigma \subset (0, U]$. In applications of matrix completion this is usually a realistic assumption, since the entries of M_0 are bounded, too. For example in a movie recommender system (e.g. [BL07, GNOT92]) the entries of the observations Y and consequently M_0 and ϵ_i are bounded from above by the best possible rating and below from the worst possible rating. As the variance is now assumed to be unknown the construction from (2.17) is not feasible anymore since we can not compute the test statistic (2.16). Instead we use a U-statistic approach which only requires an upper bound on the variance for using Markov's inequality.

As in the previous section, we use the second half of the sample, $(Y_i^{tr}, X_i)_{n/2<i\leq n}$, for constructing a minimax optimal estimator $\hat{M}$ of M that fulfills $\|\hat{M}\|_\infty \leq a$. Since σ is bounded by U we use again the matrix lasso estimator from Klopp [Klo14] with σ replaced by U which achieves the bound (2.13) with probability at least $1 - 2/d$. In order to construct the confidence set, we will be interested in all pairs of observations (Y_l^{tr}, X_l) and (Y_s^{tr}, X_s) in the first sub-sample with $1 \leq l < s \leq n/2$ such that $X_l = X_s$ (that is, independent measurements of the same matrix entry). For each $(i,j) \in [m_1] \times [m_2]$, let $\mathcal{S}_{(i,j)} = \{k \in \{1, \dots, n/2\} : X_k = e_i(m_1)e_j^T(m_2)\} =: \{a_1 < \dots < a_{p_{(i,j)}}\}$ where $p_{(i,j)}$ is the number of times that we observe the entry (i,j). For all indices (i,j) such that $S_{(i,j)} \neq \emptyset$, we form the $\lfloor p_{(i,j)}/2 \rfloor$ couples $(X_{a_1}, X_{a_2}), (X_{a_3}, X_{a_4}), \dots$ etc. We denote by $\mathcal{N}$ the set of all these pairs and let $|\mathcal{N}| = N$ be their number. Re-ordering, we can write $(\tilde{X}_k, Z_k, Z'_k)_{k\leq N}$ where $\tilde{X}_k = X_l = X_s$ for some couple $(X_l, X_s) \in \mathcal{N}$ and $Z_k = Y_l^{tr}$ and $Z'_k = Y_s^{tr}$. That is, using two different samples of the same entry $\tilde{X}_k = X_l = X_s$ we form

the observation triples $(\tilde{X}_k, Z_k, Z'_k)$. We use $(\tilde{X}_k, Z_k, Z'_k)_{k\leq N}$ to construct a U-Statistic to estimate the squared Frobenius loss. Contrary to the construction in (2.16) this does not require knowledge of the variance of the errors. We define:

$$\hat{R}_N := \frac{1}{N}\sum_{k=1}^{N}(Z_k - \langle \hat{M}, \tilde{X}_k\rangle)(Z'_k - \langle \hat{M}, \tilde{X}_k\rangle), \tag{2.19}$$

and we set $\hat{R}_N = 0$ if $N = 0$. Note that

$$\mathbb{E}_{M_0,\sigma}\left[\hat{R}_N \middle| \hat{M}, N \geq 1\right] = \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2}. \tag{2.20}$$

We define the confidence set

$$C_n := \left\{ A \in \mathcal{A}(a,m) : \quad \frac{\|A - \hat{M}\|_F^2}{m_1 m_2} \leq \hat{R}_N + z_{\alpha,N} \right\} \tag{2.21}$$

where the random quantile constant $z_{\alpha,N}$ is defined as

$$z_{\alpha,N} := \frac{U^2 + 4a^2}{\sqrt{N\alpha}} \quad \text{if} \quad N \neq 0 \quad \text{and} \quad z_{\alpha,N} = 4a^2 \quad \text{if} \quad N = 0.$$

The quantity N is random but we can bound it from below with high probability by $n^2/(64m_1m_2)$ as proven in the following lemma.

Lemma 2.4.1. *For $n \leq m_1 m_2$ we have with probability at least $1 - \exp\left(-n^2/(372 m_1 m_2)\right)$ that:*

$$N \geq \frac{n^2}{64 m_1 m_2}.$$

Markov's inequality, (2.20), Lemma 2.4.1 and that $\hat{M}$ achieves the nearly optimal rate (2.13) with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 2/d$ imply for any $k \leq m$, any $M_0 \in \mathcal{A}(a,k)$, any $\sigma \leq U$, any $\alpha' > 2/d + \exp(-n^2/(372 m_1 m_2))$ and a large enough constant $C = C(\alpha, \alpha', \sigma, a, U) > 0$ that

$$\mathbb{P}_{M_0,\sigma}\left(\frac{|C_n|_F^2}{m_1 m_2} > C\frac{kd\log(d)}{n}\right) \leq \alpha'. \tag{2.22}$$

Since k is arbitrary this implies that C_n is a confidence set whose $\|\cdot\|_F^2$-diameter adapts to the unknown rank of M_0 without requiring the knowledge of $\sigma \in \Sigma$. The following theorem implies that C_n is also a honest confidence set. Note that our result is non-asymptotic and holds for any triple $(n, m_1, m_2) \in \mathbb{N}^3$ as long as $m \log d \leq n \leq m_1 m_2$.

Theorem 2.4.2. *Let $\alpha > 0$ be given, assume $m\log(d) \leq n \leq m_1 m_2$ and that Assumption 2.1.1 is fulfilled. Let $C_n = C_n(Y, \alpha)$ as in (2.21). Then C_n satisfies for any*

$M_0 \in \mathcal{A}(a,m)$ and any $\sigma \in \Sigma$

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \in C_n\right) \geq 1-\alpha.$$

Hence, for any $\alpha' > 2/d + \exp(-n^2/(372 m_1 m_2))$ and any $1 \leq k_0 < k \leq m$, C_n is a honest confidence set at level α for the model $\mathcal{A}(a,k) \otimes \Sigma$ that adapts (up to a log-factor) to the rank k_0 of any sub-model $\mathcal{A}(a,k_0) \otimes \Sigma$ at level α'.

2.5 Bernoulli Model

In this section we consider the Bernoulli model (2.3). As before we let $\mathbb{P}_{M,\sigma}$ (and $\mathbb{E}_{M,\sigma}$ for the corresponding expectation) denote the distribution of the data when the parameters are M and σ, and we sometimes write M_0 for the 'true' parameter M for the sake of precision.

2.5.1 A non-asymptotic confidence set in the Bernoulli model with known variance of the errors.

Here we assume again that $\sigma > 0$ is known. In case of the Bernoulli model we are not able to obtain two independent samples and cannot use the risk estimation approaches from the trace-regression setting. Instead we use the duality between testing and honest and adaptive confidence sets laid out in Section 2.2. We first determine an upper bound for the minimax rate $\rho = \rho_{n,m}$ of testing the low rank hypothesis

$$H_0 : M \in \mathcal{A}(a,k_0) \quad \text{against} \quad H_1 : M \in \mathcal{A}(a,k), \quad \|M - \mathcal{A}(a,k_0)\|_F^2 \geq \rho^2, \tag{2.23}$$

and then apply Theorem 2.2.2. As test statistic, we propose an infimum-test which has previously been used by Bull and Nickl [BN13] and Nickl and van de Geer [Nv13] in density estimation and high-dimensional regression, respectively (see also Section 6.2.4. in [GN16]). Since $\sigma^2 = \mathbb{E}\epsilon_{ij}^2$ is known we can define the statistic

$$T_n := \inf_{A \in \mathcal{A}(a,k_0)} \left| \frac{1}{\sqrt{2n}} \sum_{i,j} B_{ij} \left((Y_{ij} - A_{ij})^2 - \sigma^2 \right) \right| = \inf_{A \in \mathcal{A}(a,k_0)} \left| \frac{1}{\sqrt{2n}} \sum_{i,j} \left((Y_{ij} - B_{ij}A_{ij})^2 - B_{ij}\sigma^2 \right) \right| \tag{2.24}$$

and choose the quantile constant u_α such that

$$\mathbb{P}_\sigma \left(\frac{1}{\sqrt{2n}} \left| \sum_{i,j} B_{ij} (\epsilon_{ij}^2 - \mathbb{E}\epsilon_{ij}^2) \right| > u_\alpha \right) \leq \alpha/3. \tag{2.25}$$

For example, using Markov's inequality, we get

$$\mathbb{P}_\sigma\left(\frac{1}{\sqrt{2n}}\left|\sum_{i,j}B_{ij}(\epsilon_{ij}^2-\sigma^2)\right|>u_\alpha\right)\leq\frac{1}{2nu_\alpha^2}\sum_{i,j}\mathrm{Var}_\sigma\left(B_{ij}(\epsilon_{ij}^2-\sigma^2)\right)\leq\frac{\sigma^2(U^2-\sigma^2)}{2u_\alpha^2}$$

so $u_\alpha=\sigma\sqrt{\big(3(U^2-\sigma^2)\big)/(2\alpha)}$ is an admissible choice.

Theorem 2.5.1. *Let $\alpha\geq 12\exp(-100d)$ be given. Consider the Bernoulli model (2.3) and the two parameter spaces $\mathcal{A}(a,k)$ and $\mathcal{A}(a,k_0)$, $1\leq k_0<k\leq m$. Furthermore assume that Assumption 2.1.1 is fulfilled, that $\sigma>0$ is known, that $n\geq m\log(d)$ and consider the testing problem (2.23). Suppose*

$$\rho^2\geq C\frac{m_1m_2k_0d}{n}\asymp r_{n,m}^2(\mathcal{A}(a,k_0))$$

where $C=C(\alpha,a,U,\sigma)>0$ is a constant. Then the test $\Psi_n:=\mathbf{1}_{\{T_n>u_\alpha\}}$ where u_α is the quantile constant in (2.25) and T_n is as in (2.24) fulfills

$$\sup_{M\in\mathcal{A}(a,k_0)}\mathbb{E}_{M,\sigma}[\Psi_n]+\sup_{M\in\mathcal{A}(a,k),\ \|M-\mathcal{A}(a,k_0)\|_F^2\geq\rho^2}\mathbb{E}_{M,\sigma}[1-\Psi_n]\leq\alpha.$$

Now in order to apply Theorem 2.2.2 we use the soft-thresholding estimator proposed by Koltchinskii et al. [KLT11] which satisfies the oracle inequality (2.12) up to a log-factor in the trace regression model. That this holds in the Bernoulli-model as well with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1-1/d$ can be proven in a similar way and we sketch this in Proposition 2.8.3, removing the log-factor by using stronger bounds on the spectral norm of the noise matrix $(B_{ij}\epsilon_{ij})_{i,j}$.

This and Theorem 2.5.1 imply, using Theorem 2.2.2, that there exist honest and adaptive confidence sets in the Bernoulli model if the variance of the errors is known.

Corollary 2.5.1. *Let $\alpha\geq 2/d$ and $\alpha'\geq 12\exp(-100d)$ be given. Suppose that $\sigma>0$ is known, that Assumption 2.1.1 is fulfilled and that $n\geq m\log(d)$. Then, for any $1\leq k_0<k\leq m$, there exists a honest confidence set C_n at the level α for the model $\mathcal{A}(a,k)\otimes\{\sigma\}$, i.e., for any $M_0\in\mathcal{A}(a,k)$,*

$$\mathbb{P}_{M_0,\sigma}\left(M_0\in C_n\right)\geq 1-\alpha,$$

and C_n adapts to the sub-model $\mathcal{A}(a,k_0)\otimes\{\sigma\}$ at level α'.

2.5.2 The case of the Bernoulli model with unknown error variance.

In this subsection we assume again, as in Subsection 5.2, that the precise knowledge of the error variance σ is *not* available. Whereas in this case for the trace-regression model the construction of honest and adaptive confidence set was seen to be possible, we will now show that this is not the case for the Bernoulli model. We use again the duality

between testing and confidence sets, this time applying Theorem 2.2.1. The next theorem gives a lower bound for the minimax rate of testing for the composite null hypothesis $H_0 : M \in \mathcal{A}(a, k_0)$ of M having rank at most k_0 against a rank-k alternative. To simplify the exposition we will consider only square matrices and also an asymptotic 'high-dimensional' framework where $\min(n, m) \to \infty$ and $k_0 = o(k)$. We formally allow for $k_0 = 0$, thus including the 'signal detection problem' when $H_0 : M = 0, \sigma^2 = 1$.

Theorem 2.5.2. *Suppose that Assumption 2.1.1 is satisfied for some $U \geq 2$ and assume $m = m_1 = m_2$. Furthermore, let $k = k_{n,m} \to \infty$ be such that $0 < k \leq m^{1/3}$ and $k^{1/4}\sqrt{m/n} < \min(1, a)/2$. For $0 \leq k_0 < k$ satisfying $k_0 = o(k)$ and a sequence $\rho = \rho_{n,m} \in (0, 1/2)$ consider the testing problem*

$$H_0 : M \in \mathcal{A}(a, k_0),\ \sigma^2 = 1 \quad vs \quad H_1 : M \in \mathcal{A}(a, k),\ \|M - \mathcal{A}(a, k_0)\|_F^2 \geq m^2\rho^2,\ \sigma^2 = 1 - 4\rho^2. \tag{2.26}$$

If, as $\min(n, m) \to \infty$,

$$\rho^2 = o\Big(\frac{\sqrt{k}m}{n}\Big), \tag{2.27}$$

then for any test Ψ we have that

$$\liminf_{\min(n,m)\to\infty} \left[\sup_{M \in \mathcal{A}(a,k_0)} \mathbb{E}_{M,1}[\Psi] + \sup_{M \in \mathcal{A}(a,k),\ \|M - \mathcal{A}(a,k_0)\|_F^2 \geq m^2\rho^2} \mathbb{E}_{M,\sqrt{1-4\rho^2}}[1 - \Psi] \right] \geq 1. \tag{2.28}$$

In particular, if $\Sigma \subset (0, U]$ contains the interval $[\sqrt{1 - 4\tau}, 1]$ where $\tau = \limsup_{n,m} k^{1/4}\sqrt{m/n}$, then (2.10) holds for the choices $\mathcal{A}_0 = \mathcal{A}(a, k_0) \otimes \Sigma, \mathcal{A} = \mathcal{A}(a, k) \otimes \Sigma$ and $\beta_0 = 1, \rho^* = \rho$.

Using Theorem 2.2.1 this implies the non-existence of honest and adaptive confidence sets in the model (2.3) if the variance of the errors is unknown and $k_0 = o(\sqrt{k})$. In particular adaptation to a constant rank k_0, $k_0 = O(1)$, is never possible if $k \to \infty$ as $\min(m, n) \to \infty$.

Corollary 2.5.2. *Assume that the conditions of Theorem 2.5.2 are fulfilled and that $k_0 = o(\sqrt{k})$. Then for any $\alpha, \alpha' > 0$ satisfying $0 < 2\alpha + \alpha' < 1$ a honest confidence set for the model $\mathcal{A}(a, k) \otimes \Sigma$ at level α that adapts to the sub-model $\mathcal{A}(a, k_0) \otimes \Sigma$ at level α' does not exist. In fact if $\alpha < 1/3$, we have for every honest confidence set C_n for the model $\mathcal{A}(a, k) \otimes \Sigma$ at level α and constant $c = c(a, U, \alpha)$ that*

$$\sup_{(M_0,\sigma) \in \mathcal{A}(a,k_0) \otimes \Sigma} \mathbb{E}_{M_0,\sigma} |C_n|_F^2 \geq c\frac{m^3\sqrt{k}}{n}.$$

2.6 Conclusions

We have investigated confidence sets in two matrix completion models: the Bernoulli model and the trace regression model. In the trace regression model the construction of

adaptive confidence sets is possible, even if the variance is unknown. Contrary to this we have shown that the information theoretic structure in the Bernoulli model is different; in this case the construction of adaptive confidence sets is not possible if the variance is unknown.

One interpretation is that in practical applications (e.g. recommender systems such as Netflix [BL07]) one should incentivise users to perform multiple ratings of every product they rate, to justify the use of the trace regression model and the proposed U-statistic confidence set.

In the case of the Bernoulli model a few questions remain open: Our proof only shows that one can not adapt to a low rank hypothesis $k_0 = o(\sqrt{k})$ if the variance is unknown. It remains an open question whether the lower bound ρ in Theorem 2.5.2 is tight, as well as whether adaptation over 'non-low-rank parameter spaces' when $k_0 \gg \sqrt{k}$ or $k > m^{1/3}$ is possible.

2.7 Proofs

2.7.1 Proof of Theorem 2.2.2

Proof. Let Ψ_n be a test that attains the rate ρ with error probabilities bounded by $\min(\alpha/2, \alpha')$ and let $L = L(\min(\alpha/2, \alpha'))$ be the corresponding constant in (2.9). Let $\hat{f}$ denote an estimator that satisfies the oracle inequality (2.12) with probability of at least $1 - \alpha/2$. Define a confidence set

$$C_n := \{f \in \mathcal{A} : \ \mathcal{D}(\hat{f}, f) \leq K\left(r_n(\mathcal{A})\Psi_n + r_n(\mathcal{A}_0)(1 - \Psi_n)\right)\}$$

where $K > 0$ is a constant to be chosen.
We first prove that C_n is adaptive: If $f \in \mathcal{A}\backslash\mathcal{A}_0$ there is nothing to prove, and if $f \in \mathcal{A}_0$ we have

$$\mathbb{P}_f^n\left(|C_n|_{\mathcal{D}} > K r_n(\mathcal{A}_0)\right) = \mathbb{P}_f^n(\Psi_n = 1) \leq \alpha'.$$

For coverage we investigate three distinct cases and note that

$$\sup_{f \in \tilde{\mathcal{A}}} \mathbb{P}_f^n\left(\mathcal{D}(\hat{f}, f) > C r_n(\tilde{\mathcal{A}})\right) \leq \alpha/2 \tag{2.29}$$

where $C > 0$ is as in (2.12) and where $\tilde{\mathcal{A}} \in \{\mathcal{A}_0, \mathcal{A}\}$. Hence $\hat{f}$ is, by the oracle inequality, an adaptive estimator.
Then for $f \in \mathcal{A}_0$, by (2.29)

$$\mathbb{P}_f^n(f \notin C_n) \leq \mathbb{P}_f^n\left(\mathcal{D}(\hat{f}, f) > K r_n(\mathcal{A}_0)\right) \leq \alpha/2 \leq \alpha$$

for $K \geq C$.
If $f \in \mathcal{A}\backslash\mathcal{A}_0$ and $\mathcal{D}(f, \mathcal{A}_0) \geq L\rho_n$, then for $K \geq C$

$$\begin{aligned}\mathbb{P}_f^n(f \notin C_n) &= \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}), \Psi_n = 1) + \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}), \Psi_n = 0) \\ &\leq \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A})) + \mathbb{P}_f^n(\Psi_n = 0) \leq \alpha.\end{aligned}$$

If $f \notin \mathcal{A}\backslash\mathcal{A}_0$ but $\mathcal{D}(f, \mathcal{A}_0) < L\rho_n$, then by the oracle inequality and since $\rho_n \leq C' r_n(\mathcal{A}_0)$ we have with probability at least $1 - \alpha/2$ for such f that

$$\mathcal{D}(\hat{f}, f) \leq C(\mathcal{D}(f, \mathcal{A}_0) + r_n(\mathcal{A}_0)) \leq CL\rho_n + Cr_n(\mathcal{A}_0) \leq C(LC' + 1)r_n(\mathcal{A}_0).$$

Thus we still have

$$\mathbb{P}_f^n\left(f \notin C_n\right) = \mathbb{P}_f^n(\mathcal{D}(\hat{f}, f) > Kr_n(\mathcal{A}_0)) \leq \alpha/2 \leq \alpha$$

for $K \geq C(LC' + 1)$. □

2.7.2 Proof of Theorem 2.4.2

Proof. Recall that

$$\mathbb{E}_{M_0,\sigma}\Big(\hat{R}_N | N, N > 0\Big) = \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} =: r. \tag{2.30}$$

Thus using Markov's inequality we have for $N > 0$ that

$$\begin{aligned}\mathbb{P}_{M_0,\sigma}\left(M_0 \notin C_n | N, N > 0\right) &\leq \mathbb{P}_{M_0,\sigma}\left(|\hat{R}_N - r| > z_{\alpha,N} | N, N > 0\right) \\ &\leq \frac{\mathrm{Var}_{M_0,\sigma}\Big(\hat{R}_N \big| N, N > 0\Big)}{z_{\alpha,N}^2}.\end{aligned} \tag{2.31}$$

Using equation (2.30) we compute

$$\begin{aligned}\mathrm{Var}_{M_0,\sigma}\Big(\hat{R}_N \big| N, N > 0\Big) &= \frac{1}{N}\mathbb{E}_{M_0,\sigma}\Big(\big((Z_k - \langle\hat{M}, \tilde{X}_k\rangle)(Z_k' - \langle\hat{M}, \tilde{X}_i\rangle) - r\big)^2 \Big| N, N > 0\Big) \\ &\leq \frac{1}{N}\left[\Big(\mathbb{E}\langle M_0 - \hat{M}, X_1\rangle^4\Big) + 2\sigma^2 r + \sigma^4\right] \\ &= \frac{1}{N}\left[\frac{\|\hat{M} - M_0\|_{L^4}^4}{m_1 m_2} + 2\sigma^2 r + \sigma^4\right] \\ &\leq \frac{U^4 + 8U^2a^2 + 16a^4}{N} = \alpha z_{\alpha,N}^2\end{aligned}$$

since $\|\hat{M} - M_0\|_\infty \leq 2a$ and where we define $\|\hat{M} - M_0\|_{L^4}^4 := \sum_{i,j}(\hat{M}_{ij} - M_{ij})^4$. Hence

(2.31) implies

$$\mathbb{P}_{M_0,\sigma}\left(M_0 \notin C_n | N > 0\right) \leq \alpha.$$

Moreover, as $\|\hat{M} - M_0\|_\infty \leq 2a$ and $z_{\alpha,0} = 4a^2$, we have that $\mathbb{P}\left(M_0 \notin C_n | N = 0\right) = 0$.

□

2.7.3 Proof of Theorem 2.5.1

Proof. If $M \in \mathcal{A}(a, k_0)$, then by definition of the infimum and u_α we have

$$\mathbb{E}_{M,\sigma}[\Psi] = \mathbb{P}_{M,\sigma}\left(T_n > u_\alpha\right) \leq \mathbb{P}_\sigma\left(\frac{1}{\sqrt{2n}}\left|\sum_{ij} B_{ij}(\epsilon_{ij}^2 - \sigma^2)\right| > u_\alpha\right) \leq \alpha/3.$$

The case $M \in \mathcal{A}(a,k), \|M - \mathcal{A}(a,k_0)\|_F^2 \geq \rho^2$ requires more elaborate arguments. Let A^* be a minimizer in (2.24). Then

$$\begin{aligned}\mathbb{E}_{M,\sigma}[1-\Psi] &= \mathbb{P}_{M,\sigma}\left(T_n < u_\alpha\right) \\ &= \mathbb{P}_\sigma\left(\left|\sum_{ij} B_{ij}[(A^*_{ij} - M_{ij})^2 - 2\epsilon_{ij}(A^*_{ij} - M_{ij}) + (\epsilon_{ij}^2 - \sigma^2)]\right| < \sqrt{2n}u_\alpha\right).\end{aligned} \tag{2.32}$$

For $\rho \geq 8072a\sqrt{k_0 d/p} = 8072a\sqrt{m_1 m_2 k_0 d/n}$ we can apply Lemma 2.8.1 which yields a weaker version of the Restricted Isometry Property (RIP). Namely, Lemma 2.8.1 implies that the event

$$\Xi := \left\{\sum_{i,j} B_{ij}(A_{ij} - M_{ij})^2 \geq \frac{p}{2}\|A - M\|_F^2 \quad \forall A \in \mathcal{A}(a,k_0)\right\}, \quad M \in H_1,$$

occurs with probability of at least $1 - 2\exp(-100d)$. We can thus bound (2.32) by

$$\mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0)}\left[2\left|\sum_{i,j} B_{ij}\epsilon_{ij}(A_{ij} - M_{ij})\right| - \frac{\sum_{i,j} B_{ij}(A_{ij} - M_{ij})^2}{2}\right] > -\sqrt{n}u_\alpha, \Xi\right) \tag{2.33}$$

$$+\mathbb{P}_\sigma\left(\left|\sum_{i,j} B_{ij}(\epsilon_{ij}^2 - \sigma^2)\right| > \frac{\sum_{i,j} B_{ij}(A^*_{ij} - M_{ij})^2}{2} - \sqrt{n}u_\alpha, \Xi\right) + 2\exp(-100d). \tag{2.34}$$

The stochastic term (2.34) can be bounded using $d^2 \geq 3n$ and that ρ is large enough. Indeed, on the event Ξ we have that

$$\frac{\sum_{i,j} B_{ij}(A^*_{ij} - M_{ij})^2}{2} \geq p\rho^2/4 \geq (1+\sqrt{2})/\sqrt{3}du_\alpha \geq (1+\sqrt{2})\sqrt{n}u_\alpha$$

for $\rho \geq 2\sqrt{u_\alpha d/p}$ which implies together with the definition of u_α in (2.25) that (2.34) can be bounded by $\alpha/3 + 2\exp(-100d)$. For the cross term (2.33) we use the two following inequalities which, just as before, hold on the event $\Xi \ \forall \ A \in \mathcal{A}(a, k_0)$

$$\frac{\sum_{i,j} B_{ij}(A_{ij} - M_{ij})^2}{4} \geq \sqrt{n}u_\alpha \quad \text{and} \quad \frac{\sum_{i,j} B_{ij}(A_{ij} - M_{ij})^2}{8} \geq \frac{p\|A-M\|_F^2}{16}.$$

Hence, using also a peeling argument, (2.33) can be bounded by

$$\begin{aligned}
&\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty} \mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0),\ 2^s\leq p\|A-M\|_F^2\leq 2^{s+1}} \frac{\left|\sum_{i,j} B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|}{p\|A-M\|_F^2} > \frac{1}{16}\right) \\
\leq &\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty} \mathbb{P}_\sigma\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2^{s+1}} \left|\sum_{i,j} B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right| > \frac{2^s}{16}\right) \\
= &\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty} \mathbb{P}_\sigma\left(Z(s) > \frac{2^s}{16}\right) \qquad (2.35)
\end{aligned}$$

where we set the corresponding probability to 0 if the supremum is taken over an empty set and where we define

$$Z(s) := \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2^{s+1}} \left|\sum_{i,j} B_{ij}\epsilon_{ij}(A_{ij}-M_{ij})\right|.$$

Lemma 2.8.2 (with choices $z = 16^2$, $\xi_{ij} = \epsilon_{ij}$, $t = 2^s$ and $q = 1$ there) implies for $\rho \geq 16144U\sqrt{k_0 d/p}$ and for $2^s \geq p\rho^2/2$ that

$$\mathbb{P}_\sigma\left(Z(s) > \frac{2^s}{16}\right) \leq \exp\left(\frac{-2^s}{2097152U^2 + 517120aU}\right)$$

Hence, (2.35) can be upper bounded by

$$\begin{aligned}
\sum_{s\in\mathbb{N}:\ p\rho^2/2\leq 2^s<\infty} \exp\left(\frac{-2^s}{2097152U^2+517120aU}\right) \leq &2\exp\left(-\frac{p\rho^2}{2097152U^2+517120aU}\right) \qquad (2.36) \\
\leq &2\exp(-100d)
\end{aligned}$$

for $\rho \geq 16169U(a \vee U)\sqrt{d/p}$. Consequently (2.32) can be bounded by $\alpha/3+4\exp(-100d) \leq 2\alpha/3$ since $\alpha \geq 12\exp(-100d)$. □

2.7.4 Proof of Theorem 2.5.2

Proof. **Step I : Reduction to an easier testing problem between two distributions**

Assume without loss of generality that m is divisible by k. Suppose

$$\rho = \rho_{n,m} = \frac{vk^{1/4}\sqrt{m}}{\sqrt{n}} \tag{2.37}$$

where $v = v_{n,m}$ is a sequence such that $v = o(1)$, and assume w.l.o.g. that $0 < v \leq 1$. Moreover we denote $u = 2\rho$. For $1 \leq i \leq m,\ 1 \leq \kappa \leq k,\ 1 \leq j \leq m$ let

$$B_{ij} \overset{i.i.d.}{\sim} \mathcal{B}(p) \quad \text{and} \quad U_i^{\kappa} \overset{i.i.d.}{\sim} \mathcal{R} \quad \text{and} \quad V_j \overset{i.i.d.}{\sim} \mathcal{R},$$

where $\mathcal{B}(p)$ is a Bernoulli distribution of parameter $p = n/m^2$ and $\mathcal{R}$ is the standard Rademacher distribution $\Pr(V_1 = \pm 1) = 1/2$. Let $\mathcal{P}$ be a uniform random partition of $\{1, \ldots, m\}$ in k groups of size m/k, and denote by K_j, $K_j \in \{1, ..., k\}$, the label of element j of $\mathcal{P}$. Consider the following testing problem:

$$H_0' : M = 0 \quad \text{and} \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{R}$$

against

$$\begin{aligned} H_1' : M_{ij} &= uU_i^{K_j}V_j \\ \text{and} \quad \epsilon_{ij} &\sim \delta_{\{1-M_{ij}\}}(1+M_{ij})/2 + \delta_{\{-1-M_{ij}\}}(1-M_{i,j})/2 \end{aligned} \tag{2.38}$$

Note that the variance of ϵ_{ij} under H_0 is 1 and the variance of the noise under H_1 is

$$(1-M_{ij})^2(1+M_{ij})/2 + (-1-M_{ij})^2(1-M_{ij})/2 = (1-M_{ij})(1+M_{ij}) = 1-4\rho^2,$$

so the noise variables are homoscedastic across the (i,j)'s and $|\epsilon_{ij}| \leq 2 \leq U$. Let π be the distribution of M under H_1' and write ν_0 and ν_1 for the distribution of Y under H_0' and H_1', respectively.

Since the prior M in (2.38) consists of k i.i.d. scaled Rademacher vectors that each form m/k columns of M we have $\text{rank}(M) \leq k$ and $\|M\|_\infty = u = 2\rho \leq a$ for v small enough and since $k^{1/4}\sqrt{m/n} \leq a/2$. Thus $M \in \mathcal{A}(a,k)$. Then, reordering the columns of M we have

$$\|M - \mathcal{A}(a,k_0)\|_F^2 = \|M_{ord} - \mathcal{A}(a,k_0)\|_F^2$$

where M_{ord} is a $m \times m$ matrix with the $(((i-1)m/k)+1)$-th to the (im/k)-th columns each given by ur_i where r_i are $i.i.d$ Rademacher vectors of length m, $i = 1, ..., k$. Then (as in the proof of Theorem 1 in [CN15]) we transform M_{ord} into the $m \times k$ matrix $M_{ord}P = u\sqrt{m/k}R$ consisting of k column vectors $u\sqrt{m/k}r_i$, $i = 1, ..., k$. The $m \times k$ projection matrix P consists of k column vectors, the i-th having zero entries except for the indices $s \in [((i-1)m/k)+1, \ldots, im/k]$ where it equals $\sqrt{k/m}$. Hence P is an orthonormal projection matrix and we obtain

$$\|M - \mathcal{A}(a, k_0)\|_F^2 \geq \|(M_{ord} - \mathcal{A}(a, k_0))P\|_F^2 = \|u\sqrt{m/k}R - \mathcal{A}(a\sqrt{m/k}, k, k_0)\|_F^2$$

where we define

$$\mathcal{A}(a, k, k_0) := \{A \in \mathbb{R}^{m\times k} : \ \|A\|_\infty \leq a \ \text{ and } \ \text{rank}(A) \leq k_0\}.$$

Therefore, if $\sigma_{min}(A)$ denotes the minimal singular value of a matrix A, we have that

$$\begin{aligned}
\|M - \mathcal{A}(a, k_0)\|_F^2 &\geq \frac{m^2}{k}\|uR/\sqrt{m} - \mathcal{A}(a/\sqrt{m}, k, k_0)\|_F^2 \\
&\geq \frac{m^2u^2}{k}(k - k_0)(\sigma_{min}(R/\sqrt{m}))^2 \\
&\geq \frac{m^2u^2}{2}(\sigma_{min}(R/\sqrt{m}))^2 \geq \frac{m^2u^2}{4} = m^2\rho^2
\end{aligned} \tag{2.39}$$

with probability going to 1, where we have used that $k - k_0 \geq k/2$ for m large enough (recall $k_0 = o(k)$) as well as the variational characterisation of minimal eigenvalues combined with Corollary 1 in [Nv13] (with choices $n = m$, $p = k_1 = k$, $\theta = 0$ and $\Lambda_{min} = 1$ there) to lower bound $\sigma_{min}^2(R/\sqrt{m})$ by $1/2$.
To conclude, π is concentrated on H_1 and the primed testing problem above is, asymptotically, strictly easier than the testing problem (2.26) since H_0' is contained in H_0 and H_1' is asymptotically contained in H_1. Thus, we have for any test Ψ by a standard lower bound (as, e.g., in (6.23) in [GN16]) that for all $\eta > 0$

$$\mathbb{E}_{H_0}\Psi + \sup_{H_1}\mathbb{E}_{H_1}(1-\Psi) \geq \mathbb{E}_{H_0'}\Psi + \mathbb{E}_{H_1'}(1-\Psi) - o(1) \geq (1-\eta)\left(1 - \frac{d_{\chi^2}(\nu_0, \nu_1)}{\eta}\right) - o(1),$$

where $d_{\chi^2}(\nu_0, \nu_1)$ denotes the χ^2-distance between ν_0 and ν_1, which remains to be bounded.

Step II : Expectation over censored data
We define $I = [m] \times [m]$ and observe that the likelihood of the data under ν_0 is

$$L(Y_1, ...Y_{m,m}) = \prod_{(i,j)\in I}\left((1-p)\mathbf{1}_{\{Y_{ij}=0\}} + \frac{p}{2}\mathbf{1}_{\{Y_{ij}=1\}} + \frac{p}{2}\mathbf{1}_{\{Y_{ij}=-1\}}\right)$$

and that the likelihood of the data under ν_1 is

$$L(Y_1,...Y_{m,m}) = \mathbb{E}_{M\sim\pi} \prod_{(i,j)\in I} \Big((1-p)\mathbf{1}_{\{Y_{ij}=0\}}+p(1/2+M_{ij}/2)\mathbf{1}_{\{Y_{ij}=1\}}+p(1/2-M_{ij}/2)\mathbf{1}_{\{Y_{ij}=-1\}}\Big).$$

Thus, the likelihood ratio $\mathcal{L}$ between these two distributions is given by

$$\mathcal{L} = \mathbb{E}_{M\sim\pi} \prod_{(i,j)\in I} \Big(\mathbf{1}_{\{Y_{ij}=0\}} + (1+M_{ij})\mathbf{1}_{\{Y_{ij}=1\}} + (1-M_{ij})\mathbf{1}_{\{Y_{ij}=-1\}}\Big).$$

So we have that

$$\begin{aligned} d_{\chi^2}(\nu_0,\nu_1)^2 + 1 &= \mathbb{E}_{Y\sim\nu_0}\mathcal{L}^2 \\ &= \mathbb{E}_{Y\sim\nu_0}\Big[\mathbb{E}_{M\sim\pi} \prod_{(i,j)\in I} \Big(\mathbf{1}_{\{Y_{ij}=0\}} + (1+M_{ij})\mathbf{1}_{\{Y_{ij}=1\}} + (1-M_{ij})\mathbf{1}_{\{Y_{ij}=-1\}}\Big)\Big]^2 \\ &= \mathbb{E}_{M,M'\sim\pi} \prod_{i,j} \Big[\Big(1-p+\frac{p}{2}(1+M_{ij})(1+M'_{ij}) + \frac{p}{2}(1-M_{ij})(1-M'_{ij})\Big)\Big] \\ &= \mathbb{E}_{M,M'\sim\pi} \prod_{i,j} \Big[1+pM_{ij}M'_{ij}\Big]. \end{aligned} \tag{2.40}$$

where M' is an independent copy of M.

Step III : Conditioning over the cross information

Let $N_{r,r'}$ be the number of times where the couple $K_j = r, K'_j = r'$ occurs. That is,

$$N_{r,r'} := \sum_{j=1}^{m} \mathbf{1}_{\{K_j=r,K'_j=r'\}}.$$

We enumerate the elements inside these groups from 1 to $N_{r,r'}$. We write $\tilde{V}_j^{r,r'}$ for the corresponding enumeration of the V_j. Setting $\mathbf{N} = (N_{r,r'})_{r,r'}$ and using the definition of the prior, we compute

$$\begin{aligned} \mathbb{E}_{M,M'\sim\pi} \prod_{i,j} \Big[1+pM_{ij}M'_{ij}\Big] &= \mathbb{E}_{\mathbf{N},U,\tilde{V},U',\tilde{V}'} \prod_{i=1}^{m} \prod_{r,r'\in\{1,...,k\}^2} \prod_{j=1}^{N_{r,r'}} \Big[1+pu^2U_i^r\tilde{V}_j^{r,r'}(U_i^{r'})'(\tilde{V}_j^{r,r'})'\Big] \\ &=: \mathbb{E}_{\mathbf{N}} \prod_{r,r'\in\{1,...,k\}^2} \mathcal{I}(N_{r,r'}) \end{aligned} \tag{2.41}$$

where we define for any $N = N_{r,r'} > 0$

$$\mathcal{I}(N) = \mathbb{E}_{X,W,X',W'} \prod_{i=1}^{m}\prod_{j=1}^{N} \Big[1+pu^2X_iW_jX'_iW'_j\Big]$$

and where $(X_i)_{i\le m}, (X'_i)_{i\le m}, (W_i)_{j\le N}, (W'_i)_{j\le N}$ are *i.i.d.* Rademacher random variables.

Moreover, we set $\mathcal{I}_{r,r'}(0) = 0$.

Step IV : Bound on $\mathbb{E}_{\mathbf{N}} \prod_{r,r' \in \{1,\dots,k\}^2} \mathcal{I}(N_{r,r'})$.
In order to bound $\mathcal{I}(N)$ we use the following lemma proved below

Lemma 2.7.1. *Let $N = N_{r,r'}$. There exist constants $C_1, C_2, C_3 > 0$ such that for v small enough*

$$\mathcal{I}(N) \leq \exp\Big(C_1 v^4 N/m\Big) \exp\Big(\frac{C_2 v^4 k^2 N}{m^2}\Big) \exp\Big(C_3 v^4 N^2 k^2/m^2\Big). \tag{2.42}$$

Using (2.40), (2.41) and (2.42) we have that

$$\begin{aligned}
&d_{\chi^2}(\nu_0, \nu_1)^2 + 1 \\
&= \mathbb{E}_{\mathbf{N}} \prod_{r,r' \in \{1,\dots,k\}^2} \mathcal{I}(N_{r,r'}) \qquad (2.43) \\
&\leq \mathbb{E}_{\mathbf{N}} \left[\left(\exp\left(\frac{C_2 v^4 k^2}{m^2} \sum_{r,r'} N_{r,r'} \right) \right) \left(\exp\left(\frac{C_1 v^4}{m} \sum_{r,r'} N_{r,r'} \right) \right) \left(\prod_{r,r' \in \{1,\dots,k\}^2} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \right) \right] \\
&= \exp\left(C_2 v^4 \frac{k^2}{m} + C_1 v^4 \right) \mathbb{E}_{\mathbf{N}} \left[\prod_{r,r' \in \{1,\dots,k\}^2} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \right], \qquad (2.44)
\end{aligned}$$

since $\sum_{r,r'} N_{r,r'} = m$. We bound the expectation of the stochastic term in (2.44) using the following lemma proved below:

Lemma 2.7.2. *There exists a constant $C' > 0$ such that for v small enough we have*

$$\mathbb{E}_N\Big[\prod_{r,r'} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big) \Big] \leq 1 + 2C' v^4 + \exp\Big(- m/k^2 \Big). \tag{2.45}$$

Inserting (2.45) into (2.44) and summarizing all the steps we obtain

$$0 \leq d_{\chi^2}(\nu_0, \nu_1)^2 \leq C\left(v^2 + \exp\big(- m/k^2\big)\right) = o(1)$$

for a constant $C > 0$ and therefore, letting $\eta \to 0$,

$$\mathbb{E}_0[\Psi] + \sup_{H_1} \mathbb{E}_{H_1}[1 - \Psi] \geq (1 - \eta)\left(1 - \frac{d_{\chi^2}(\nu_0, \nu_1)}{\eta}\right) - o(1) = 1 - o(1).$$

□

Proof of Lemma 2.7.1. Note that, by construction of $\mathcal{P}$, we have that

$$N = N_{r,r'} \leq m/k$$

since the number of j where $M_{\cdot,j}$ corresponds to $K_j = r$ is bounded by m/k. As the

product of two independent Rademacher random variables is again a Rademacher random variable, we have

$$\mathcal{I}(N) = \mathbb{E}_{R,R'} \prod_{i=1}^{m} \prod_{j=1}^{N} \left[1 + pu^2 R_i R'_j\right],$$

where $R = (R_i)_{i=1}^m, R' = (R'_i)_{i=1}^N$ are independent Rademacher vectors of length m and N, respectively. The usual strategy to use $1+x \leq e^x$ and then to bound iterated exponential moments of Rademacher variables (as in the proof of Theorem 1 of [CN15]) only works when $k = const$, and a more refined estimate is required for growing k, as relevant here. We now bound $\mathcal{I}(N)$ for a fixed $N, m/k \geq N > 0$. Using the binomial theorem twice we have

$$\begin{aligned}
\mathcal{I}(N) &= \mathbb{E}_{R'}\left[\left[\frac{1}{2}\prod_{j=1}^{N}\left[1+pu^2R'_j\right] + \frac{1}{2}\prod_{j=1}^{N}\left[1-pu^2R'_j\right]\right]^m\right] \\
&= \frac{1}{2^m}\sum_{s=1}^{m}\binom{m}{s}\left[\frac{1}{2}\left[1+pu^2\right]^s\left[1-pu^2\right]^{m-s} + \frac{1}{2}\left[1-pu^2\right]^s\left[1+pu^2\right]^{m-s}\right]^N \\
&= \frac{1}{2^m 2^N}\sum_{s=1}^{m}\binom{m}{s}\sum_{q=1}^{N}\binom{N}{q}\left[1+pu^2\right]^{sq+(m-s)(N-q)}\left[1-pu^2\right]^{(m-s)q+s(N-q)} \\
&= \mathbb{E}_{Q,S}\left[\left[1+pu^2\right]^{SQ+(m-S)(N-Q)}\left[1-pu^2\right]^{(m-S)Q+S(N-Q)}\right]
\end{aligned}$$

with independent Binomial random variables $S \sim \mathcal{B}(1/2, m), Q \sim \mathcal{B}(1/2, N)$. If $A := \frac{1-pu^2}{1+pu^2}$, we obtain

$$\begin{aligned}
\mathcal{I}(N) &= \mathbb{E}_{Q,S}\left[\left[1+pu^2\right]^{mN}\left[\frac{1-pu^2}{1+pu^2}\right]^{SN+mQ-2SQ}\right] \\
&= \left[1+pu^2\right]^{mN}\mathbb{E}_Q\left[A^{mQ}\mathbb{E}_S A^{S(N-2Q)}\right] \\
&= \left[1+pu^2\right]^{mN}\mathbb{E}_Q\left[A^{mQ}2^{-m}\left(A^{(N-2Q)}+1\right)^m\right] \\
&= \left[1+pu^2\right]^{mN}\mathbb{E}_Q\left[A^{Nm/2}\left(\frac{1}{2}A^{(N/2-Q)} + \frac{1}{2}A^{(-N/2+Q)}\right)^m\right] \\
&= \left[1-p^2u^4\right]^{mN/2}\mathbb{E}_Q\left(\frac{1}{2}A^{Q-N/2} + \frac{1}{2}A^{N/2-Q}\right)^m.
\end{aligned}$$

Now, we denote $x := pu^2 = 4vk^{1/2}/m \leq 1/2$ for v small enough. Furthermore, we Taylor expand $\log(A)$ about 1 up to second order, i.e.

$$\log(A) = \log(1-x) - \log(1+x) = -2x - \frac{1}{2}\left(\frac{1}{\xi_1^2} - \frac{1}{\xi_2^2}\right)x^2 =: -2x - c(x)x^2$$

for $\xi_1 \in [1/2, 1]$, $\xi_2 \in [1, 3/2]$ and where $c(x) \in [0, 16/9]$ since $x \leq 1/2$. Hence, using also

the inequality $e^x \le 1 + x + x^2/2 + x^3/6 + 2x^4$ we deduce

$$\begin{aligned}
\mathcal{I}(N) &\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q\Big[\frac{1}{2}\exp\big(-2x(Q-N/2) - c(x)(Q-N/2)x^2)\big) \\
&\quad + \frac{1}{2}\exp\big(-2x(N/2-Q) - c(x)(N/2-Q)x^2)\big)\Big]^m \\
&\le \exp\big[-mNx^2/2\big] \\
&\quad \cdot \mathbb{E}_Q\Bigg[\frac{1}{2}\Big(1 - 2x(Q-N/2) - c(x)(Q-N/2)x^2 + (-2x(Q-N/2) - c(x)(Q-N/2)x^2)^2/2 \\
&\quad + (-2x(Q-N/2) - c(x)(Q-N/2)x^2)^3/6 + 2(-2x(Q-N/2) - c(x)(Q-N/2)x^2)^4\Big) \\
&\quad + \frac{1}{2}\Big(1 - 2x(N/2-Q) - c(x)(N/2-Q)x^2 + (-2x(N/2-Q) - c(x)(N/2-Q)x^2)^2/2 \\
&\quad + (-2x(N/2-Q) - c(x)(N/2-Q)x^2)^3/6 + 2(-2x(N/2-Q) - c(x)(N/2-Q)x^2)^4\Big)\Bigg]^m.
\end{aligned}$$

Since $x \le 1/2$ and $|N/2 - Q|x \le 1/4$ there exist two constants $c_2 = c_2(x) = c(x)/2 + c(x)^2/32 \le 1$ and $c_1 = c_1(x) = 32 + 32c(x) + 12c(x)^2 + 2c(x)^3 + c(x)^4/8 \le 140$ such that the last equation above can be bounded by

$$\begin{aligned}
&\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q\Big[1 + 2x^2(Q-N/2)^2 + c_1|Q-N/2|^4x^4 + c_2|Q-N/2|x^2\Big]^m \\
&\le \exp\big[-mNx^2/2\big]\mathbb{E}_Q\exp\Big[mx^2(N-2Q)^2/2 + c_1m(Q-N/2)^4x^4 + c_2m|Q-N/2|x^2\Big] \\
&= \mathbb{E}_Q\Big[\exp\Big(\frac{m}{2}\big(x^2(2Q-N)^2 - Nx^2\big)\Big)\exp\Big(c_1m(Q-N/2)^4x^4 + c_2m|Q-N/2|x^2\Big)\Big].
\end{aligned}$$

Using the Cauchy-Schwarz inequality twice, this implies that

$$\begin{aligned}
\mathcal{I}(N) \le \sqrt{\mathbb{E}_Q\Big[\exp\Big(mx^2N\big((2Q-N)^2/N - 1\big)\Big)\Big]}\Bigg[\mathbb{E}_Q\Big[\exp\Big(c_1mx^4(N-2Q)^4/4\Big)\Big] \\
\cdot \mathbb{E}_Q\Big[\exp\Big(2c_2m|2Q-N|x^2\Big)\Big]\Bigg]^{1/4} =: \sqrt{(I)}(II)^{1/4}(III)^{1/4}.
\end{aligned}$$

Step 1 : Bound on term (III)
Since $Q \sim \mathcal{B}(1/2, N)$, since $(2Q - N)$ is symmetric and since $2c_2mx^2 \le 1/2$ we have that

$$\begin{aligned}
(III) &= \mathbb{E}_Q\Big[\exp\Big(2c_2m|2Q-N|x^2\Big)\Big] \le 2\mathbb{E}_Q\Big[\exp\Big(2c_2m(2Q-N)x^2\Big)\Big] \\
&= 2\Big[\exp\Big(2c_2mx^2\Big) + \exp\Big(-2c_2mx^2\Big)\Big]^N \le 2\Big[1 + 8c_2^2m^2x^4\Big]^N \\
&\le \exp\big(8c_2^2m^2x^4N\big) \le \exp\big(\frac{C_2v^4k^2N}{m^2}\big). \qquad (2.46)
\end{aligned}$$

Step 2 : Term (II)

We use $mN^2x^4 \leq 64v^4/m$, $(N-2Q)^2 \leq N^2$ and $N \leq m/k$ to obtain

$$(II) \leq \mathbb{E}_Q\Big[\exp\Big(64c_1v^4N/m \cdot (N-2Q)^2/N\Big)\Big].$$

Since $Q \sim \mathcal{B}(1/2, N)$ the Rademacher average $Z = (N-2Q)/\sqrt{N}$ is sub-Gaussian with sub-Gaussian constant at most 1. It hence satisfies (e.g., equation (2.24) in [GN16]) for $c > 2$

$$\mathbb{E}\exp\{Z^2/c^2\} \leq 1 + \frac{2}{c^2/4-1} \leq e^{c_3c^{-2}},$$

which for v small enough and the choice $c^{-2} = 64c_1v^4N/m$ implies for some constant C_1 that

$$(II) \leq \exp\left(\frac{4C_1v^4N}{m}\right).$$

Step 3 : Term (I)

We have that

$$\begin{aligned}(I) &= \mathbb{E}_Q\Big[\exp\Big(mNx^2\left[\frac{(2Q-N)^2}{N}-1\right]\Big)\Big] \\ &= \mathbb{E}\left[\exp\left(\frac{16v^2Nk}{m}\left[\frac{1}{N}\left(\sum_{i=1}^N\varepsilon_i\right)^2-1\right]\right)\right] = \mathbb{E}\left[\exp\left(\frac{16v^2k}{m}\sum_{i\neq j, i,j\leq N}\varepsilon_i\varepsilon_j\right)\right],\end{aligned}$$

where ε_i are *i.i.d.* Rademacher random variables. If $A = (a_{ij})$ is a symmetric matrix with all elements on the diagonal equal to zero, then for the Laplace transform of an order-two Rademacher chaos $Z = \sum_{i,j} a_{ij}\varepsilon_i\varepsilon_j$ we have the inequality

$$\mathbb{E}e^{\lambda Z} \leq \exp\left\{\frac{16\lambda^2\|A\|_F^2}{2\left(1-64\|A\|\lambda\right)}\right\}, \quad \lambda > 0,$$

see, e.g., Exercise 6.9 on p.212 in [BLM13] with $\mathcal{T} = \{A\}$. Now take $A = (\delta_{i\neq j})_{i,j\leq N}$ so that we have $\|A\| \leq N$ and for v small enough $16v^2kN/m \leq 16v^2 \leq 1/128$.

$$\mathbb{E}\Big[\exp\Big(\frac{16v^2k}{m}\sum_{i\neq j, i,j\leq N}\varepsilon_i\varepsilon_j\Big)\Big] \leq \exp\left(\frac{16^3v^4k^2\|A\|_F^2}{2m^2(1-1024v^2k\|A\|/m)}\right) \leq \exp\left(\frac{16^3v^4k^2N^2}{m^2}\right)$$

and therefore we conclude for a constant $C_3 > 0$ that

$$(I) \leq \exp\Big(2C_3v^4k^2N^2/m^2\Big). \tag{2.47}$$

Step 4 : Conclusion on $\mathcal{I}(N)$

Combining the bounds for (I), (II) and (III) with the bound on $\mathcal{I}(N)$ we have that

$$\mathcal{I}(N) \leq \exp\Big(C_2v^4k^2N/m^2\Big)\exp\Big(C_1v^4N/m\Big)\exp\Big(C_3v^4k^2N^2/m^2\Big).$$

□

Proof of Lemma 2.7.2. We bound the expectation by bounding it separately on two complementary events. For this we consider the event ξ where all $N_{r,r'}$ are upper bounded by $\tau := 15m/k^2$, assumed to be an integer (if not replace it by its integer part plus one in the argument below). More precisely we define

$$\xi = \Big\{\forall r \leq k, \forall r' \leq k \ : \ N_{r,r'} \leq \tau\Big\}.$$

Note that $\{N_{r,r'} > \tau\}$ occurs only if the size of the intersection of the class r of partition $\mathcal{P}$ with the class r' of partition $\mathcal{P}'$ is larger than τ. This means that at least τ elements among m/k elements of the class r', must belong to the class r. The positions of these τ elements can be taken arbitrarily within the m/k elements. For the first element, among those τ, the probability to belong to the class r is $\frac{m/k}{m}$. For the second element this probability is $\frac{m/k}{m-1}$ or $\frac{(m/k)-1}{m-1}$ and so on. All these probabilities are smaller than $(m/k)/(m-m/k+1)$. Therefore we have

$$\mathbb{P}_{\mathbf{N}}(N_{r,r'} > \tau) \leq \binom{m/k}{\tau}\left(\frac{m/k}{m-m/k+1}\right)^{\tau} \leq \frac{(m/k)^{\tau}}{\tau!}(2/k)^{\tau} \leq 2^{\tau}(m/k^2)^{\tau}\tau^{-\tau}e^{\tau} \leq e^{-\tau},$$

where we use $\binom{m/k}{\tau} \leq \frac{(m/k)^{\tau}}{\tau!}$ and Stirling's formula. Using a union bound this implies that the probability of ξ is lower bounded by $1 - k^2\exp(-15m/k^2)$.
We have on the event ξ

$$\begin{aligned}
&\mathbb{E}_{\mathbf{N}}\Big[\mathbf{1}\{\xi\}\prod_{r,r'\in\{1,\dots,k\}^2}\exp\Big(C_3v^4N_{r,r'}^2k^2/m^2\Big)\Big]\\
&\leq \exp\Big(C_3v^4k^2\cdot 15^2(m/k^2)^2k^2/m^2\Big)\Big]\\
&\leq \exp\Big(C'v^4\Big) \leq 1 + 2C'v^4.
\end{aligned}$$

for $C' = 225C_3$ and for v small enough. Moreover, by definition of $N_{r,r'}$, we have that $N_{r,r'} \leq m/k$ and $\sum_{r,r'} N_{r,r'} = m$. Hence

$$\sum_{r,r'} N_{r,r'}^2 \leq km^2/k^2 = m^2/k$$

which implies that on ξ^C

$$\begin{aligned}
&\mathbb{E}_{\mathbf{N}}\Big[\mathbf{1}\{\xi^C\}\prod_{r,r'\in\{1,\dots,k\}^2}\exp\Big(C_3v^4N_{r,r'}^2k^2/m^2\Big)\Big]\\
&\leq \mathbb{P}_{\mathbf{N}}(\xi^C)\exp\Big(C_3v^4k\Big)\\
&\leq k^2\exp\Big(-15m/k^2 + C_3v^4k\Big)
\end{aligned}$$

$$\leq k^2 \exp\Big(- 3m/k^2\Big) \leq \exp\Big(- m/k^2\Big),$$

for v small enough and since $k^3 \leq m$. Thus, combining the bounds on ξ and ξ^C, we have that

$$\mathbb{E}_N\Big[\prod_{r,r'} \exp\Big(C_3 v^4 N_{r,r'}^2 k^2/m^2\Big)\Big] \leq 1 + 2C'v^4 + \exp\Big(- m/k^2\Big).$$

□

2.8 Auxiliary results

2.8.1 Proof of Lemma 2.4.1

Proof. Assume that among the first $n/4$ samples we have less than $n/8$ entries that are sampled twice - otherwise the result holds since $n/8 \geq n^2/64m_1m_2$ for $n \leq m_1m_2$. Then, among the first $n/4$ samples, there are at least $n/8$ distinct elements of $\mathcal{B}$, the set of all standard basis matrices in $\mathbb{R}^{m_1 \times m_2}$, that have been sampled at least once. We write $\mathcal{S}$ for the set of *distinct* elements of $\{X_i\}_{i \leq n/4}$ and obviously have $|\mathcal{S}| \geq n/8$. Hence, by definition of the sampling scheme, we have that

$$\mathbb{P}(X_i \in \mathcal{S}) \geq \frac{n}{8m_1m_2}, \qquad n/4 < i \leq n/2.$$

Furthermore, when sampling an element from $\mathcal{S}$ we have to remove this element from $\mathcal{S}$ as we have to use the entry that is stored in $\mathcal{S}$ to form a pair of entries. Hence the probability to sample another element from $\mathcal{S}$ decreases and is bounded by

$$\mathbb{P}(X_j \in \mathcal{S}\backslash\{X_i\} \big| X_i \in \mathcal{S}) \geq \frac{n-1}{8m_1m_2}$$

for $n/4 < i < j < n/2$. We deduce by induction for $j > i + k$ and $k \leq n/2 - i - 1$ that

$$\mathbb{P}(X_j \in \mathcal{S}\backslash\{X_i, ..., X_{i+k}\} \big| X_i, ..., X_{i+k} \in \mathcal{S}) \geq \frac{n-k}{8m_1m_2}$$

which yields

$$\begin{aligned}\mathbb{P}\left(N \geq \frac{n^2}{64m_1m_2}\right) &\geq \mathbb{P}\left(\sum_{n/4<i\leq n/2} \mathbf{1}_{\{X_i \in \mathcal{S}\}} \geq \frac{n^2}{64m_1m_2}\right) \\ &\geq \mathbb{P}\left(\sum_{n/4<i\leq n/2} \mathbf{Z}_i \geq \frac{n^2}{64m_1m_2}\right) \end{aligned} \tag{2.48}$$

where $\mathbf{Z}_i$ can be taken to be Bernoulli random variables with success probability

$$p' = \frac{n - \frac{n^2}{64m_1m_2}}{8m_1m_2}.$$

Then, Bernstein's inequality for bounded random variables (see e.g. Theorem 3.1.7 in [GN16]), (2.48) and the estimates

$$\mathbb{E}\left[\sum_{n/4<i\le n/2} \mathbf{Z}_i\right] \ge \frac{n^2}{33m_1m_2}$$

which holds for $n \le m_1m_2$ and

$$\mathrm{Var}\left(\sum_{n/4<i\le n/2} \mathbf{Z}_i\right) \le \frac{n^2}{32m_1m_2}$$

imply that

$$\mathbb{P}\left(N \ge \frac{n^2}{64m_1m_2}\right) \ge 1 - \mathbb{P}\left(\sum_{n/4<i\le n/2} \mathbf{Z}_i - \mathbb{E}\left[\sum_{n/4<i\le n/2} \mathbf{Z}_i\right] \le \frac{-n^2}{72m_1m_2}\right) \ge 1 - \exp\left(\frac{n^2}{372m_1m_2}\right).$$

□

2.8.2 Lemma 2.8.1

Lemma 2.8.1. *Consider the Bernoulli model (2.3) and assume $n \ge m\log(d)$. Then, with probability at least $1 - 2\exp(-100d)$ we have for any given $M \in \mathcal{A}(a,m)$ that*

$$\sup_{A\in\mathcal{A}(a,m),\ \|M-A\|_F\ge Ca\sqrt{(\mathrm{rank}(A)\vee 1)d/p}} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|M_0-A\|_F^2\right] \le 0$$

where $C = 8072$.

Proof. We have, using a union bound, that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{A}(a,m),\ \|M-A\|_F\ge Ca\sqrt{(\mathrm{rank}(A)\vee 1)d/p}} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|M_0-A\|_F^2\right] > 0\right) \\
&\le \sum_{k=1}^{m} \mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ p\|M-A\|_F^2\ge C^2a^2kd} \left[\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right| - \frac{p}{2}\|A-M\|_F^2\right] > 0\right).
\end{aligned} \tag{2.49}$$

Then, using a peeling argument each of the terms in (2.49) can be bounded by

$$\sum_{s\in\mathbb{N}:\ C^2a^2kd/2\le 2^s<\infty} \mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ 2^s\le p\|A-M\|_F^2\le 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right|>2^s/2\right)$$
$$\le \sum_{s\in\mathbb{N}:\ C^2a^2kd/2\le 2^s<\infty} \mathbb{P}\left(\sup_{A\in\mathcal{A}(a,k),\ p\|A-M\|_F^2\le 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right|>2^s/2\right) \tag{2.50}$$

with the convention that if the supremum is taken over an empty set the corresponding probability is set equal to 0. For the cases where the supremum is not taken over an empty set, we apply Lemma 2.8.2 (with choices $\xi_{ij}=1$, $q=2$, $z=4$, $U=1$ and $t=2^s$ there) and obtain for

$$Z(s) := \sup_{A\in\mathcal{A}(a,k),\ p\|A-M\|_F^2\le 2^{s+1}}\left|\sum_{i,j}(B_{ij}-p)(A_{ij}-M_{ij})^2\right|$$

that we can bound

$$\mathbb{P}\left(Z(s)>2^s/2\right)\le\exp\left(\frac{-2^s}{260352a^2}\right)$$

Hence, (2.50) can be upper bounded by

$$\sum_{s\in\mathbb{N}:\ Ca^2kd/2\le 2^s<\infty}\exp\left(\frac{-2^s}{260352a^2}\right)\le 2\exp\left(-\frac{C^2kd}{260352}\right)\le 2\exp(-101d).$$

The result then follows by noting that $\log(m)\le d$. □

2.8.3 Lemma 2.8.2

Lemma 2.8.2. *Consider the Bernoulli model (2.3). Suppose that ξ_{ij} are independent random variables with $\max_{ij}|\xi_{ij}|\le U$ and that $m\log(d)\le n$. Let $z>0$, $q\in\{1,2\}$, $M\in\mathcal{A}(a,m)$ and $1\le k_0<m$ be given. Finally, for $C=1009$ suppose that $t\in\mathbb{R}_+$ is such that $t\ge C^2z(4a)^{2q-2}U^2k_0d/2$ and that the supremum in*

$$Z(t) := \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\le 2t}\left|\sum_{i,j}\left[(B_{ij}\xi_{ij}-\mathbb{E}B_{ij}\xi_{ij})(A_{ij}-M_{ij})^q\right]\right|$$

is not empty. Then,

$$\mathbb{P}\left(Z(t)>\frac{t}{\sqrt{z}}\right)\le\exp\left(\frac{-t}{32^2(8(2a)^{2q-2}U^2z+505(2a)^qU\sqrt{z}/32)}\right) \tag{2.51}$$

Proof. We first bound $\mathbb{E}Z(t)$ and then apply Talagrand's [Tal96] inequality. Using sym-

metrization (e.g. Theorem 3.1.21 in [GN16]) and two contraction inequalities (e.g. Theorems 3.1.17 and 3.2.1 in [GN16]), we obtain that

$$\begin{aligned}
\mathbb{E}Z(t) &\leq 2U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t}\left|\sum_{i,j}B_{ij}\varepsilon_{ij}(A_{ij}-M_{ij})^q\right|\right)\\
&\leq 2(4a)^{q-1}U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t}\left|\sum_{i,j}B_{ij}\varepsilon_{ij}(A_{ij}-M_{ij})\right|\right)\\
&\leq 2(4a)^{q-1}U\mathbb{E}\left(\sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t}|\langle\Sigma_R,A-A_0\rangle|\right)+2(4a)^{q-1}U\mathbb{E}\,|\langle\Sigma_R,A_0-M\rangle|\\
&\leq 8(4a)^{q-1}U\sqrt{k_0t/p}\mathbb{E}\,\|\Sigma_R\|+2(4a)^{q-1}U\mathbb{E}\,|\langle\Sigma_R,A_0-M\rangle|\,. \qquad (2.52)
\end{aligned}$$

where ε_{ij} are independent Rademacher random variables, $\Sigma_R := \left(B_{ij}\varepsilon_{ij}\right)_{ij}$ and where A_0 is an arbitrary element in $\mathcal{A}(a,k_0)$ such that $p\|A_0-M\|_F^2\leq 2t$. Such an A_0 exists as soon as the supremum is not taken over an empty set. An extension of Corollary 3.6 in [BvH16] to rectangular matrices by self-adjoint dilation (e.g. section 3.1. in [BvH16]) implies (with choices $\xi_{ij}=B_{ij}\varepsilon_{ij}/\sqrt{p}$, $b_{ij}=\sqrt{p}$, $\alpha=3$ and $\sigma=\max\left(\max_j\sqrt{\sum_{i=1}^{m_1}b_{ij}^2},\max_i\sqrt{\sum_{j=1}^{m_2}b_{ij}^2}\right)\leq\sqrt{pd}$ there) that

$$\mathbb{E}\,\|\Sigma_R\|\leq e^{2/3}(2\sqrt{pd}+42\sqrt{\log(d)})\leq 86\sqrt{pd}$$

since $m\log(d)\leq n$. For the second term in (2.52) we have

$$\begin{aligned}
\mathbb{E}|\langle\Sigma_R,A_0-M\rangle| &\leq (\mathrm{Var}(\langle\Sigma_R,A_0-M\rangle))^{1/2}\\
&= \left(p\|A_0-M\|_F^2\right)^{1/2}\leq\sqrt{2t}.
\end{aligned}$$

Hence, for $C^2z(4a)^{2q-2}U^2k_0d/2\leq t$ and since $C=1009$ we have that

$$\mathbb{E}Z(t)\leq 688(4a)^{q-1}U\sqrt{k_0td}+2(4a)^{q-1}U\sqrt{2t}\leq 31t/(32\sqrt{z}). \qquad (2.53)$$

We now make use of the following inequality due to Talagrand [Tal96], which in the current form with explicit constants can be obtained by inverting the tail bound in Theorem 3.3.16 in [GN16].

Theorem 2.8.1. *Let $(S,\mathcal{S})$ be a measurable space and let $n\in\mathbb{N}$. Let X_k, $k=1,\dots,n$ be independent S-valued random variables and let $\mathcal{F}$ be a countable set of functions $f=(f_1,\dots,f_n):S^n\to[-K,K]^n$ such that $\mathbb{E}f_k(X_k)=0$ for all $f\in\mathcal{F}$ and $k=1,\dots,n$. Set*

$$Z:=\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f_k(X_k).$$

Define the variance proxy

$$V_n := 2K\mathbb{E}Z + \sup_{f\in\mathcal{F}} \sum_{k=1}^{n} \mathbb{E}\left[(f_k(X_k))^2\right].$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(Z - \mathbb{E}Z \geq t\right) \leq \exp\left(\frac{-t^2}{4V_n + (9/2)Kt}\right).$$

The functional $A \to \|A - M\|_F^2$ is continuous on the compact set of matrices $\{A \in \mathcal{A}(a, k_0) : \|A-M\|_F^2 \leq 2t\}$, hence by continuity and compactness the supremum is attained over a countable subset. Thus we may apply Talagrand's inequality to $Z(t)$. We have for our particular case, since $\sup_{f\in\mathcal{F}} |f(X)| = \sup_{f\in\{\mathcal{F}\bigcup\{-\mathcal{F}\}\}} f(x)$, that

$$X_{ij} = B_{ij}\xi_{ij} - \mathbb{E}B_{ij}\xi_{ij}, \quad S = [-2U, 2U]$$

$$\mathcal{F} = \Big\{ f : S^{m_1\times m_2} \to [-2(2a)^q U, 2(2a)^q U]^{m_1\times m_2},\ f_{ij}(X_{ij}) = (-1)^l X_{ij}(A_{ij} - M_{ij})^q,$$
$$A \in \mathcal{A}(a, k_0),\ p\|A - M\|_F^2 \leq 2t,\ l \in \{0, 1\} \Big\}$$

and moreover

$$\begin{aligned}
&\sup_{(A,l),\ A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t,\ l\in\{0,1\}} \sum_{i,j} \mathbb{E}\left[\Big((-1)^l(B_{ij}\xi_{ij} - \mathbb{E}B_{ij}\xi_{ij})(A_{ij} - M_{ij})^q\Big)^2\right] \\
\leq &(2a)^{2q-2} \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t} \sum_{i,j} \mathrm{Var}(B_{ij}\xi_{ij})(A_{ij} - M_{ij})^2 \\
\leq &(2a)^{2q-2}U^2 \sup_{A\in\mathcal{A}(a,k_0),\ p\|A-M\|_F^2\leq 2t} \sum_{i,j} p(A_{ij} - M_{ij})^2 \leq 2(2a)^{2q-2}U^2 t.
\end{aligned}$$

Therefore, using our previous estimate in (2.53) for $\mathbb{E}Z(t)$ as well, we have for the variance proxy $V_{m_1m_2}$ that

$$V_{m_1m_2} \leq 2(2a)^{2q-2}U^2 t + 31(2a)^q Ut/(8\sqrt{z}).$$

Hence, using (2.53) and Talagrand's inequality, we obtain

$$\mathbb{P}\left(Z(t) > \frac{t}{\sqrt{z}}\right) \leq \mathbb{P}\left(Z(t) - \mathbb{E}Z(t) > \frac{t}{32\sqrt{z}}\right) \leq \exp\left(\frac{-t}{32^2(8(2a)^{2q-2}U^2 z + 505(2a)^q U\sqrt{z}/32)}\right).$$

□

2.8.4 An oracle estimator in the Bernoulli model

Here we prove that the soft-thresholding estimator proposed by Koltchinskii et al. [KLT11] for the trace-regression setting fulfills the oracle inequality (2.12) in the Bernoulli model. Their estimator is defined as

$$\hat{M} \in \operatorname*{arg\,min}_{A \in \mathbb{R}^{m_1 \times m_2}} \left(\frac{\|A\|_F^2}{m_1 m_2} - \frac{2}{n} \langle Y, A \rangle + \lambda \|A\|_* \right) \tag{2.54}$$

where λ is a tuning parameter which we choose as

$$\lambda = 3 \left(\frac{3\sqrt{2}\sigma + \sqrt{2C}U}{\sqrt{mn}} \right) \tag{2.55}$$

where $C > 0$ is the constant in Corollary 3.12 in [BvH16].

Proposition 2.8.3. *Consider the Bernoulli model (2.3). Assume $n \geq m \log(d)$ and that Assumption 2.1.1 is fulfilled. Let $\hat{M}$ be given as in (2.54) with a choice of λ as in (2.55). Then, with $\mathbb{P}_{M_0,\sigma}$-probability of at least $1 - 1/d$ we have for any $M_0 \in \mathcal{A}(a, m)$ that*

$$\begin{aligned} \frac{\|\hat{M} - M_0\|_F^2}{m_1 m_2} &\leq \inf_{A \in \mathbb{R}^{m_1 \times m_2}} \left(\frac{\|M_0 - A\|_F^2}{m_1 m_2} + C \frac{\operatorname{drank}(A)}{n} \right) \\ &\leq \inf_{k \in \{0, \dots, m\}} \left(\frac{\|M_0 - \mathcal{A}(a,k)\|_F^2}{m_1 m_2} + C \frac{dk}{n} \right) \end{aligned}$$

for a constant $C = C(a, \sigma, U) > 0$.

Proof. Going through the proof of Theorem 2 and Corollary 2 in [KLT11] line by line we see that we only need to bound the spectral norm of the matrix

$$\Sigma := \frac{1}{n} (B_{ij} \epsilon_{ij})_{i,j}$$

by $\lambda/3$ with high probability. Using self-adjoint dilation to generalize Corollary 3.12 and Remark 3.13 in [BvH16] for rectangular matrices (with choices $\varepsilon = 1/2$, $\tilde{\sigma}_* = U$ and

$$\tilde{\sigma} = \max \left(\max_j \sqrt{\sum_{i=1}^{m_1} \mathbb{E}_\sigma B_{ij}^2 \epsilon_{ij}^2}, \max_i \sqrt{\sum_{j=1}^{m_2} \mathbb{E}_\sigma B_{ij}^2 \epsilon_{ij}^2} \right) = \sigma \sqrt{n/m}$$

there) we obtain

$$\mathbb{P}_\sigma \left(\left\| \sum_{i=1}^n \varepsilon_i X_i \right\| > 3\sqrt{2}\sigma \sqrt{\frac{n}{m}} + t \right) \leq d \exp\left(-\frac{-t^2}{C_1 U^2} \right)$$

for a constant $C_1 > 0$. Choosing $t = \sqrt{2C_1} U \sqrt{\frac{n}{m}}$ and using that $n \geq m \log(d)$ yields that Ξ occurs with $\mathbb{P}_\sigma$-probability at least $1 - 1/d$. □

Chapter 3

Efficient Estimation of Linear Functionals of Principal Components

3.1 Introduction

Principal Component Analysis (PCA) is commonly used as a dimension reduction technique for high-dimensional data sets. Assuming a general framework where the data lies in a Hilbert space $\mathbb{H}$, PCA can be applied to a wide range of problems such as functional data analysis [RS05, LAS16] or machine learning [BBZ07].
The parametric setting has been well understood since the 1960's (e.g. [And63] and [DPR82]) and the asymptotic distribution of sample eigenvalues and sample eigenvectors is well known. For high-dimensional data, where the dimension $p = p(n) \to \infty$ with the sample size n, the spiked covariance model introduced by Johnstone in [Joh01] has been the most common framework to study the asymptotic properties of principal components. In this model, it is assumed that the covariance matrix is given by a 'spike' and a noise part, that is

$$\Sigma = \sum_{j=1}^{l} s_i(\theta_i \otimes \theta_i) + \sigma^2 I_p,$$

where $\sum_{j=1}^{l} s_i(\theta_i \otimes \theta_i)$ is a low rank covariance matrix involving several orthonormal components ('spikes') θ_i and $\sigma^2 I_p$ is the covariance of the noise. Error bounds in this model, based on perturbation analysis, were studied in [Nad08]. Moreover, if $\frac{p}{n} \to c \in (0, 1]$ the asymptotic distribution of sample eigenvectors was derived in [Pau07] and in more general asymptotic regimes in [WF17]. Assuming sparsity of the eigenvectors (sparse PCA), inference is possible even when $\frac{p}{n} \to \infty$. This model has recently received substantial attention, e.g. [CMW13, BR13, VL13, WBS16, GZ15].

More recently, a so-called 'effective rank' setting for PCA has been considered, for example, in [KL16, KL17a, KL17c, Ver12, RW16, NSU18]. In this dimension-free setting, it is assumed that the covariance Σ is an operator acting in a Hilbert space $\mathbb{H}$, no structural assumptions are made about Σ and its 'complexity' is characterized by the *effective rank* $\mathbf{r}(\Sigma) := \mathrm{trace}(\Sigma)/\|\Sigma\|$, $\mathrm{trace}(\Sigma)$ denoting the trace and $\|\Sigma\|$ denoting the operator (spectral) norm of Σ. In a series of papers [KL17a, KL16, KL17c, KL17b], Koltchinskii and Lounici derived sharp bounds on the spectral norm loss of estimation of Σ by the sample covariance $\hat{\Sigma}$ that provide complete characterization of the size of $\|\hat{\Sigma}-\Sigma\|$ in terms of $\|\Sigma\|$ and $\mathbf{r}(\Sigma)$, and obtained error bounds and limiting results for empirical spectral projection operators and eigenvectors of $\hat{\Sigma}$ under the assumption that $\mathbf{r}(\Sigma) = o(n)$ as $n \to \infty$. In a recent paper [NSU18], Naumov et al. constructed bootstrap confidence sets for spectral projections in a lower dimensional regime where $\mathbf{r}(\Sigma) = o(n^{1/3})$. In [RW16], Reiss and Wahl considered the reconstruction error for spectral projections.
In this paper, we further develop the results of [KL16] and [KL17c] in the direction of semi-parametric statistics. In particular, we develop a bias reduction method in the problem of estimation of linear functionals of principal components (eigenvectors of Σ) and show asymptotic normality of the resulting debiased estimators under the assumption that $\mathbf{r}(\Sigma) = o(n)$. We prove a non-asymptotic risk lower bound that asymptotically exactly matches our upper bounds, thus establishing rigorously the semi-parametric optimality of our estimator in a general dimension-free setting (as long as $\mathbf{r}(\Sigma) = o(n)$).

The problem of $\sqrt{n}$-consistent estimation of low-dimensional functionals of high-dimensional parameters has received increased attention in recent years, and in various models semi-parametric efficiency of regularisation-based estimators has been studied, see for instance [vdGBRD14, JvdG18, RSZZ15, NL17]. Moreover, the paper [GZ16] develops Bernstein-von-Mises (BvM) results for functionals of covariance matrices in situations where bias is asymptotically negligible. While formal calculations of the Fisher information in such models indicate optimality of these procedures, a rigorous interpretation of such efficiency claims requires some care: the standard asymptotic setting for semi-parametric efficiency [vdV98] can not be straightforwardly applied because parameters in high-dimensional models are not fixed but vary with sample size n, so that establishing LAN expansions to apply Le Cam theory is not always possible or even desirable. In [JvdG18] some non-asymptotic techniques have been suggested under conditions that ensure asymptotic negligibility of the bias of candidate estimators. We take here a different approach, based on using the van Trees' inequality [GL95] to construct non-asymptotic lower bounds for the minimax risk in our estimation problem that match the upper bound *exactly* in the large sample limit.

3.2 Preliminaries

3.2.1 Notation and conventions

Let $\mathbb{H}$ be a separable Hilbert space. In what follows, $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathbb{H}$ and also, with a little abuse of notation, the Hilbert–Schmidt inner product between Hilbert–Schmidt operators acting on $\mathbb{H}$. Similarly, the notation $\| \cdot \|$ is used both for the norm of vectors in $\mathbb{H}$ and for the operator (spectral) norm of bounded linear operators in $\mathbb{H}$. For a nuclear operator A, $\mathrm{trace}(A)$ denotes its trace. We use the notation $\| \cdot \|_p$, $1 \le p \le \infty$ for the Schatten-p norms of operators in $\mathbb{H}$: $\|A\|_p := (\mathrm{trace}(|A|^p))^{1/p}$, where $|A| = \sqrt{A^* A}$, A^* being the adjoint operator of A. For $p = 1$, $\|A\|_1$ is the nuclear norm; for $p = 2$, $\|A\|_2$ is the Hilbert–Schmidt norm; for $p = \infty$, $\|A\|_\infty = \|A\|$ is the operator norm.

Given vectors $u, v \in \mathbb{H}$, $u \otimes v$ denotes the tensor product of u and v :

$$(u \otimes v) : \mathbb{H} \mapsto \mathbb{H}, (u \otimes v)w := \langle v, w \rangle u.$$

Given bounded linear operators $A, B : \mathbb{H} \mapsto \mathbb{H}$, $A \otimes B$ denotes their tensor product:

$$(A \otimes B)(u \otimes v) = Au \otimes Bv, \ u, v \in \mathbb{H}.$$

Note that $A \otimes B$ can be extended (by linearity and continuity) to a bounded operator in the Hilbert space $\mathbb{H} \otimes \mathbb{H}$, which could be identified with the space of Hilbert–Schmidt operators in $\mathbb{H}$. It is easy to see that, for a Hilbert–Schmidt operator C, we have $(A \otimes B)C = ACB^*$ (in the finite-dimensional case, this defines the so called Kronecker product of matrices). On a couple of occasions, we might need to use the tensor product of Hilbert–Schmidt operators A, B, viewed as vectors in the space of Hilbert–Schmidt operators. For this tensor product, we use the notation $A \otimes_v B$.

Throughout the paper, the following notations will be used: for nonnegative a, b, $a \lesssim b$ means that there exists a numerical constant $c > 0$ such that $a \le cb$; $a \gtrsim b$ is equivalent to $b \lesssim a$; finally, $a \asymp b$ is equivalent to $a \lesssim b$ and $b \lesssim a$. Sometimes, constant c in the above relationships could depend on some parameter γ. In this case, we provide signs $\lesssim$, $\gtrsim$ and $\asymp$ with subscript γ. For instance, $a \lesssim_\gamma b$ means that there exists a constant $c_\gamma > 0$ such that $a \le c_\gamma b$.

In many places in the proofs, we use exponential bounds for some random variables, say, ξ of the following form: for all $t \ge 1$ with probability at least $1 - e^{-t}$, $\xi \le Ct$. In some cases, it would follow from our arguments that the inequality holds with a slightly different probability, say, at least $1 - 3e^{-t}$. In such cases, it is easy to rewrite the bound again as $1 - e^{-t}$ by adjusting the value of constant C. Indeed, for $t \ge 1$ with probability at least $1 - e^{-t} = 1 - 3e^{-t - \log(3)}$, we have $\xi \le C(t + \log(3)) \le 2\log(3)Ct$. We will use such an adjustment of the constants in many proofs, often, without further notice.

3.2.2 Bounds on sample covariance

Let X be a Gaussian vector in $\mathbb{H}$ with mean $\mathbb{E}X = 0$ and covariance operator $\Sigma := \mathbb{E}(X \otimes X)$. Given i.i.d. observations $X_1, \dots, X_n$ of X, let $\hat{\Sigma} = \hat{\Sigma}_n$ be the sample (empirical) covariance operator defined as follows:

$$\hat{\Sigma} := n^{-1} \sum_{j=1}^{n} X_j \otimes X_j.$$

Definition 3.2.1. *The effective rank of the covariance operator Σ is defined as*

$$\mathbf{r}(\Sigma) := \frac{\operatorname{trace}(\Sigma)}{\|\Sigma\|}.$$

The role of the effective rank as a complexity parameter in covariance estimation is clear from the following result proved in [KL17a].

Theorem 3.2.2. *Let X be a mean zero Gaussian random vector in $\mathbb{H}$ with covariance operator Σ and let $\hat{\Sigma}$ be the sample covariance based on i.i.d. observations $X_1, \dots, X_n$ of X. Then*

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n} \right). \tag{3.1}$$

This result shows that the size of the properly rescaled operator norm deviation of $\hat{\Sigma}$ from Σ, $\frac{\mathbb{E}\|\hat{\Sigma}-\Sigma\|}{\|\Sigma\|}$, is characterized up to numerical constants by the ratio $\frac{\mathbf{r}(\Sigma)}{n}$. In particular, the condition $\mathbf{r}(\Sigma) = o(n)$ is necessary and sufficient for operator norm consistency of $\hat{\Sigma}$ as an estimator of Σ. In addition to this, the following concentration inequality for $\|\hat{\Sigma} - \Sigma\|$ around its expectation was also proved in [KL17a].

Theorem 3.2.3. *Under the conditions of the previous theorem, for all $t \geq 1$ with probability at least $1 - e^{-t}$*

$$\left| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \right| \lesssim \|\Sigma\| \left(\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee 1 \right) \sqrt{\frac{t}{n}} \bigvee \frac{t}{n} \right). \tag{3.2}$$

It immediately follows from the bounds (3.1) and (3.2) that, for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$\|\hat{\Sigma} - \Sigma\| \lesssim \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n} \right) \tag{3.3}$$

and, for all $p \in [1, \infty)$,

$$\mathbb{E}^{1/p}\|\hat{\Sigma} - \Sigma\|^p \lesssim_p \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n} \right). \tag{3.4}$$

3.2.3 Perturbation theory and empirical spectral projections

The covariance operator Σ is self-adjoint, positively semidefinite and nuclear. It has spectral decomposition

$$\Sigma = \sum_{r\geq 1} \mu_r P_r,$$

where μ_r are distinct strictly positive eigenvalues of Σ arranged in decreasing order and P_r are the corresponding spectral projection operators. For $r \geq 1$, P_r is an orthogonal projection on the eigenspace of the eigenvalue μ_r. The dimension of this eigenspace is finite and will be denoted by m_r. The eigenspaces corresponding to different eigenvalues μ_r are mutually orthogonal. Denote by $\sigma(\Sigma)$ the spectrum of operator Σ and let $\lambda_j = \lambda_j(\Sigma), j \geq 1$ be the eigenvalues of Σ arranged in a non-increasing order and repeated with their multiplicities. Denote $\Delta_r := \{j : \lambda_j = \mu_r\}, r \geq 1$. Then $\mathrm{card}(\Delta_r) = m_r$. The r-th spectral gap is defined as

$$g_r = g_r(\Sigma) := \mathrm{dist}(\mu_r; \sigma(\Sigma) \setminus \{\mu_r\}).$$

Let $\bar{g}_r = \bar{g}_r(\Sigma) := \min_{1\leq s\leq r} g_s$.

We turn now to the definition of empirical spectral projections of sample covariance $\hat{\Sigma}$ that could be viewed as estimators of the true spectral projections $P_r, r \geq 1$. In [KL16], the following definition was used: let $\hat{P}_r$ be the orthogonal projection on the direct sum of eigenspaces of $\hat{\Sigma}$ corresponding to its eigenvalues $\{\lambda_j(\hat{\Sigma}) : j \in \Delta_r\}$. This is not a perfect definition of a statistical estimator since the set Δ_r is unknown and it has to be recovered from the spectrum $\sigma(\hat{\Sigma})$ of $\hat{\Sigma}$.

When $\hat{\Sigma}$ is close to Σ in the operator norm, the spectrum $\sigma(\hat{\Sigma})$ of $\hat{\Sigma}$ is a small perturbation of the spectrum $\sigma(\Sigma)$ of Σ. This could be quantified by the following inequality that goes back to H. Weyl:

$$\sup_{j\geq 1} |\lambda_j(\hat{\Sigma}) - \lambda_j(\Sigma)| \leq \|\hat{\Sigma} - \Sigma\|. \tag{3.5}$$

It easily follows from this inequality that, if $\|\hat{\Sigma} - \Sigma\|$ is sufficiently small, then the eigenvalues $\lambda_j(\hat{\Sigma})$ of $\hat{\Sigma}$ form well separated clusters around the eigenvalues $\mu_1, \mu_2, \dots$ of Σ. To make the last claim more precise, consider a finite or countable bounded set $A \subset \mathbb{R}_+$ such that $0 \in A$ and 0 is the only limit point (if any) of A. Given $\delta > 0$, define $\lambda_\delta(A) := \max\big\{\lambda \in A : (\lambda-\delta, \lambda)\cap A = \emptyset\big\}$ and let $T_\delta(A) := A\setminus[0, \lambda_\delta(A))$. The set $T_\delta(A)$ will be called the top δ-cluster of A. Let $A_1 := T_\delta(A), A_2 := T_\delta(A \setminus A_1), A_3 := T_\delta(A \setminus (A_1 \cup A_2)), \dots$ and $\nu = \nu_\delta := \min\{j : A_{j+1} = \emptyset\}$. Obviously, $\nu < \infty$. We will call the sets $A_1, \dots, A_\nu$ the δ-clusters of A. They provide a partition of A into sets separated by the gaps of length at least δ and such that the gaps between the points inside each of the clusters are smaller than δ.

The next lemma easily follows from inequality (3.5).

Lemma 3.2.4. *Let $\delta > 0$ be such that, for some $r \geq 1$,*

$$\|\hat{\Sigma} - \Sigma\| < \delta/2 \quad \text{and} \quad \delta < \frac{\bar{g}_r}{2}.$$

Let $\hat{A}_1^\delta, \dots, \hat{A}_\nu^\delta$ be the δ-clusters of the set $\sigma(\hat{\Sigma})$. Then $\nu \geq r$ and, for all $1 \leq s \leq r$

$$\hat{A}_s^\delta \subset (\mu_s - \delta/2, \mu_s + \delta/2) \quad \text{and} \quad \{j : \lambda_j(\hat{\Sigma}) \in \hat{A}_s^\delta\} = \Delta_s.$$

Given $\delta > 0$ and δ-clusters $\hat{A}_1^\delta, \dots, \hat{A}_\nu^\delta$ of $\sigma(\hat{\Sigma})$, define, for $1 \leq s \leq \nu$, the empirical spectral projection $\hat{P}_s^\delta$ as the orthogonal projection on the direct sum of eigenspaces of $\hat{\Sigma}$ corresponding to its eigenvalues from the cluster $\hat{A}_s^\delta$. It immediately follows from Lemma 3.2.4 that, under its assumptions on δ, $\hat{P}_s^\delta = \hat{P}_s, s = 1, \dots, r$.

In the following sections, we will be interested in the problem of estimation of spectral projections in the case when the true covariance Σ belongs to certain subsets of the following class of covariance operators:

$$\mathcal{S}^{(r)}(\mathfrak{r}; a) := \Big\{\Sigma : \mathbf{r}(\Sigma) \leq \mathfrak{r}, \frac{\|\Sigma\|}{\bar{g}_r(\Sigma)} \leq a\Big\},$$

where $a > 1, \mathfrak{r} > 1$. We will allow the effective rank to be large, $\mathfrak{r} = \mathfrak{r}_n \to \infty$, but not too large such that $\mathfrak{r}_n = o(n)$ as $n \to \infty$. For $\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)$, we take $\delta := \tau\|\hat{\Sigma}\|$ for a sufficiently small value of the constant $\tau > 0$ in the definition of spectral projections $\hat{P}_s^\delta$.

The following lemma is an easy consequence of the exponential bound (3.3).

Lemma 3.2.5. *Suppose $a > 1$ and $\mathfrak{r}_n = o(n)$ as $n \to \infty$. Take $\tau \in \big(0, \frac{1}{4a} \wedge 2\big)$ and $\delta := \tau\|\hat{\Sigma}\|$. Then, there exists a numerical constant $\beta > 0$ such that, for all large enough n,*

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)} \mathbb{P}_\Sigma\{\exists s = 1, \dots, r : \hat{P}_s^\delta \neq \hat{P}_s\} \leq e^{-\beta\tau^2 n}.$$

Proof. By (3.3) with $t := \beta\tau^2 n$, we obtain that

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)} \mathbb{P}_\Sigma\Big\{\|\hat{\Sigma} - \Sigma\| \geq C\|\Sigma\|\Big(\sqrt{\frac{\mathfrak{r}_n}{n}} \bigvee \sqrt{\frac{\beta\tau^2 n}{n}}\Big)\Big\} \leq e^{-\beta\tau^2 n},$$

where $C > 0$ is a numerical constant. Take $\beta = \frac{1}{16C^2}$ and note that, for all large enough n, $C\sqrt{\frac{\mathfrak{r}_n}{n}} \leq \tau/4$ to obtain that

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)} \mathbb{P}_\Sigma\{\|\hat{\Sigma} - \Sigma\| \geq (\tau/4)\|\Sigma\|\} \leq e^{-\beta\tau^2 n},$$

Since $\tau/4 \leq 1/2$, we easily obtain that, for all $\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)$ and for all n large enough with probability at least $1 - e^{-\beta\tau^2 n}$, $(1/2)\|\Sigma\| \leq \|\hat{\Sigma}\| \leq 2\|\Sigma\|$. This implies that with the

same probability (and on the same event)

$$\|\hat{\Sigma} - \Sigma\| < (\tau/4)\|\Sigma\| \le (\tau/2)\|\hat{\Sigma}\| = \delta/2.$$

On the other hand, for all $\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}; a)$,

$$\delta = \tau\|\hat{\Sigma}\| \le 2\tau\|\Sigma\| < \frac{1}{2a}\|\Sigma\| \le \frac{\bar{g}_r(\Sigma)}{2}.$$

It remains to use Lemma 3.2.4 to complete the proof. □

In the proofs of the main results of the paper, we deal for the most part with spectral projections $\hat{P}_r$ that were studied in detail in [KL16]. We use Lemma 3.2.5 to reduce the results for $\hat{P}_r^\delta$ to the results for $\hat{P}_r$.

3.3 Main Results

Our main goal is to develop an efficient estimator of the linear functional $\langle \theta_r, u\rangle$, where $u \in \mathbb{H}$ is a given vector and $\theta_r = \theta_r(\Sigma)$ is a unit eigenvector of the unknown covariance operator Σ corresponding to its r–th eigenvalue μ_r, which is assumed to be simple (that is, of multiplicity $m_r = 1$). The corresponding spectral projection P_r is one-dimensional: $P_r = \theta_r \otimes \theta_r$. A "naive" plug-in estimator of P_r is the empirical spectral projection $\hat{P}_r^\delta$ with $\delta = \tau\|\hat{\Sigma}\|$ for a suitable choice of a small constant τ, as described in Lemma 3.2.5. According to this lemma and under its assumptions, $\hat{P}_r^\delta$ coincides with a high probability with the one-dimensional empirical spectral projection $\hat{P}_r := \hat{\theta}_r \otimes \hat{\theta}_r$, where $\hat{\theta}_r$ is the corresponding unit eigenvector of $\hat{\Sigma}$. As an estimator of θ_r, we can use an arbitrary unit vector $\hat{\theta}_r^\delta$ from the eigenspace $\mathrm{Im}(\hat{P}_r^\delta)$, which with a high-probability coincides with $\pm\hat{\theta}_r$ (under conditions of Lemma 3.2.5). In case $r = 1$, when the top eigenvalue $\mu_1 = \|\Sigma\|$ of Σ is simple and the goal is to estimate a linear functional of the top principal component θ_1, there is no need to use δ-clusters to define an estimator of θ_1 since $\hat{\theta}_1$ (a unit eigenvector in the eigenspace of the top eigenvalue $\|\hat{\Sigma}\|$ of $\hat{\Sigma}$) is already a legitimate estimator.

Note that both θ_r and $-\theta_r$ are unit eigenvectors of Σ, so, strictly speaking, $\langle \theta_r, u\rangle$ can be estimated only up to its sign. In what follows, we assume that $\hat{\theta}_r^\delta$ and θ_r (or, whenever is needed, $\hat{\theta}_r$ and θ_r) are *properly aligned* in the sense that $\langle \hat{\theta}_r^\delta, \theta_r\rangle \ge 0$ (which is always the case either for θ_r, or for $-\theta_r$). This allows us to view $\langle \hat{\theta}_r^\delta, u\rangle$ as an estimator of $\langle \theta_r, u\rangle$.

It was shown in [KL16] that "naive" plug-in estimators of the functional $\langle \theta_r, u\rangle$, such as $\langle \hat{\theta}_r^\delta, u\rangle$ or $\langle \hat{\theta}_r, u\rangle$, are biased with the bias becoming substantial enough to affect the efficiency of the estimator or even its convergence rates as soon as the effective rank is large enough, namely, $\mathbf{r}(\Sigma) \gtrsim n^{1/2}$. Moreover, it was shown that the quantity

$$b_r = b_r(\Sigma) := \mathbb{E}_\Sigma \langle \hat{\theta}_r, \theta_r\rangle^2 - 1 \in [-1, 0]$$

plays the role of a bias parameter. In particular, the results of [KL16] imply that the random variable $\langle\hat{\theta}_r, u\rangle$ concentrates around $\sqrt{1+b_r}\langle\theta_r, u\rangle$ (rather than around $\langle\theta_r, u\rangle$) with the size of the deviations of order $O(n^{-1/2})$ provided that $\mathbf{r}(\Sigma) = o(n)$ as $n \to \infty$. Thus, the bias of $\langle\hat{\theta}_r, u\rangle$ as an estimator of $\langle\theta_r, u\rangle$ is of the order $(\sqrt{1+b_r} - 1)\langle\theta_r, u\rangle \asymp b_r\langle\theta_r, u\rangle$. It was shown in [KL16] that $|b_r| \lesssim \frac{\mathbf{r}(\Sigma)}{n}$ and it will be proved below in this paper that, in fact, $|b_r| \asymp \frac{\mathbf{r}(\Sigma)}{n}$ (see Lemma 3.4.9 and bounds (3.39), (3.40)). This fact implies that, indeed, the bias of $\langle\hat{\theta}_r, u\rangle$ (and of $\langle\hat{\theta}_r^{\delta}, u\rangle$) is not negligible and affects the convergence rate as soon as $\frac{\mathbf{r}(\Sigma)}{n^{1/2}} \to \infty$. This resembles the situation in sparse regression (see e.g. [JM14, vdGBRD14, ZZ14]): If p denotes the dimension of the model and s its sparsity and if $s\log(p) = o(n^{1/2})$, the bias of a desparsified LASSO estimator for the regressor β is negligible, which makes it possible to prove asymptotic normality of linear forms of β. On the other hand, if $s\log(p) \gg n^{1/2}$, Cai and Guo [CG17] proved that adaptive confidence sets for linear forms do not exist in general. This implies that any attempt to further de-bias the desparsified LASSO or any other estimator to prove asymptotic normality is deemed to fail. Contrary to this, in our case estimation of the bias parameter b_r is possible (as will be shown below).

We will state a uniform (and somewhat stronger) version of some of the results of [KL16] on asymptotic normality of linear forms

$$\sqrt{n}(\langle\hat{\theta}_r^{\delta}, u\rangle - \sqrt{1+b_r(\Sigma)}\langle\theta_r(\Sigma), u\rangle), u \in \mathbb{H}$$

under the assumption that $\mathbf{r}(\Sigma) = o(n)$. To this end, define the following operator

$$C_r := \sum_{s\neq r} \frac{1}{\mu_r - \mu_s} P_s,$$

which is bounded with $\|C_r\| = \frac{1}{g_r}$. Denote

$$\sigma_r^2(\Sigma; u) := \langle\Sigma\theta_r, \theta_r\rangle\langle\Sigma C_r u, C_r u\rangle = \mu_r\langle\Sigma C_r u, C_r u\rangle.$$

Clearly,

$$\sigma_r^2(\Sigma; u) \leq \frac{\|\Sigma\|^2}{g_r^2}\|u\|^2. \tag{3.6}$$

Note that, if $\mathbb{H}$ is finite-dimensional (with a fixed dimension) and Σ is non-singular, then the Fisher information for the model $X \sim N(0;\Sigma)$ is $\mathbb{I}(\Sigma) = \frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})$ (see, e.g., [Eat83]). The maximum likelihood estimator $\hat{\Sigma}$ based on n i.i.d. observations of X (the sample covariance) is then asymptotically normal with $\sqrt{n}$-rate and limit covariance $\mathbb{I}(\Sigma)^{-1} = 2(\Sigma\otimes\Sigma)$. An application of the Delta Method to the smooth function $g(\Sigma) := \langle\theta_r(\Sigma), u\rangle$ shows that $g(\hat{\Sigma})$ is also asymptotically normal with limiting variance $\big\langle(\mathbb{I}(\Sigma)^{-1}g'(\Sigma), g'(\Sigma)\big\rangle$, which turns out to be equal to $\sigma_r^2(\Sigma; u)$.

For $u \in \mathbb{H}$, $\mathfrak{r} > 1$, $a > 1$ and $\sigma_0 > 0$, consider the following class of covariance operators

in $\mathbb{H}$:

$$\mathcal{S}^{(r)}(\mathfrak{r}, a, \sigma_0, u) := \big\{\Sigma : \mathbf{r}(\Sigma) \le \mathfrak{r}, \frac{\|\Sigma\|}{\bar{g}_r(\Sigma)} \le a, \sigma_r^2(\Sigma; u) \ge \sigma_0^2\big\}.$$

We emphasize here that we regard a and σ_0 as fixed constants, but $\mathfrak{r}, \|\Sigma\|$ and $\bar{g}_r$ may all possibly depend on n. For example, this allows that $\|\Sigma\| \to \infty$ as long as $\bar{g}_r \to \infty$ at the same rate as it is the case in factor models as considered in [WF17]. Note that some additional conditions on $\mathfrak{r}, a, \sigma_0, u$ are needed for the class $\mathcal{S}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ to be nonempty. Say, bound (3.6) implies that it is necessary for this that $\sigma_0^2 \le a^2\|u\|^2$. It is also obvious that there should be $a > r$ (since $\|\Sigma\| \ge r g_r(\Sigma)$).

We will also need the following assumption on the loss function ℓ.

Assumption 3.3.1. *Let $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ be a loss function satisfying the following conditions: $\ell(0) = 0$, $\ell(u) = \ell(-u), u \in \mathbb{R}$, ℓ is nondecreasing and convex on $\mathbb{R}_+$ and, for some constants $c_1, c_2 > 0$*

$$\ell(u) \le c_1 e^{c_2 u}, u \ge 0.$$

The proofs to all our theorems are in fact non-asymptotic and often can be expressed by Berry-Esseen type bounds. However, for a more concise presentation we present asymptotic statements.

In what follows, Z denotes a standard Gaussian random variable and Φ denotes its distribution function.

Theorem 3.3.2. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $\mathfrak{r}_n > 1$ and $\mathfrak{r}_n = o(n)$ as $n \to \infty$. Let $\delta = \tau\|\hat{\Sigma}\|$ for some $\tau \in \big(0, \frac{1}{4a} \wedge 2\big)$. Then*

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \sup_{x \in \mathbb{R}} \Big|\mathbb{P}_\Sigma\Big\{\frac{\sqrt{n}(\langle\hat{\theta}_r^\delta, u\rangle - \sqrt{1 + b_r(\Sigma)}\langle\theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)} \le x\Big\} - \Phi(x)\Big| \to 0 \text{ as } n \to \infty.$$

Moreover, under Assumption 3.3.1,

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \Big|\mathbb{E}_\Sigma \ell\Big(\frac{\sqrt{n}(\langle\hat{\theta}_r^\delta, u\rangle - \sqrt{1 + b_r(\Sigma)}\langle\theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)}\Big) - \mathbb{E}\ell(Z)\Big| \to 0 \text{ as } n \to \infty.$$

The proof of this theorem will be given in Section 3.4 that also includes a number of auxiliary statements used in the proofs of our main results on efficient estimation of linear functionals.

Corollary 3.3.3. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $\mathfrak{r}_n > 1$ and $\mathfrak{r}_n = o(\sqrt{n})$ as $n \to \infty$. Let $\delta = \tau\|\hat{\Sigma}\|$ for some $\tau \in \big(0, \frac{1}{4a} \wedge 2\big)$. Then*

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \sup_{x \in \mathbb{R}} \Big|\mathbb{P}_\Sigma\Big\{\frac{\sqrt{n}(\langle\hat{\theta}_r^\delta, u\rangle - \langle\theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)} \le x\Big\} - \Phi(x)\Big| \to 0 \text{ as } n \to \infty.$$

Moreover, under Assumption 3.3.1,

$$\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)}\Big|\mathbb{E}_\Sigma\ell\Big(\frac{\sqrt{n}(\langle\hat\theta_r^\delta,u\rangle-\langle\theta_r(\Sigma),u\rangle)}{\sigma_r(\Sigma;u)}\Big)-\mathbb{E}\ell(Z)\Big|\to 0 \text{ as } n\to\infty.$$

Our next goal is to provide a minimax lower bound on the risk of an arbitrary estimator of the linear functional $\langle\theta_r(\Sigma),u\rangle$ in the case of quadratic loss $\ell(t)=t^2, t\in\mathbb{R}$. The proof is based on van Trees' inequality and will be given in Section 3.7. Define

$$\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u):=\big\{\Sigma:\mathbf{r}(\Sigma)<\mathfrak{r},\frac{\|\Sigma\|}{\bar g_r(\Sigma)}<a,\sigma_r^2(\Sigma;u)>\sigma_0^2\big\},\mathfrak{r}>1,a>1,\sigma_0^2>0,$$

the interior of the set $\mathcal{S}^{(r)}(\mathfrak{r},a,\sigma_0,u)$.

Theorem 3.3.4. *Let $\mathfrak{r}>1$, $a>1$ and $\sigma_0>0$. Suppose $\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u)\neq\emptyset$. Then, for all statistics $T_n(X_1,\dots,X_n)$,*

$$\liminf_{n\to\infty}\inf_{T_n}\sup_{\Sigma\in\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u)}\frac{n\mathbb{E}_\Sigma(T_n(X_1,...,X_n)-\langle\theta_r(\Sigma),u\rangle)^2}{\sigma_r^2(\Sigma;u)}\geq 1.$$

Moreover, for any $\Sigma_0\in\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u)$

$$\lim_{\varepsilon\to 0}\liminf_{n\to\infty}\inf_{T_n}\sup_{\Sigma\in\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u),\|\Sigma-\Sigma_0\|_1\leq\varepsilon}\frac{n\mathbb{E}_\Sigma(T_n(X_1,...,X_n)-\langle\theta_r(\Sigma),u\rangle)^2}{\sigma_r^2(\Sigma;u)}\geq 1.$$

It follows from Corollary 3.3.3 and Theorem 3.3.4 that the estimator $\langle\hat\theta_r^\delta,u\rangle$ is efficient in a semi-parametric sense for quadratic loss under the assumption that $\mathfrak{r}_n=o(n^{1/2})$. It turns out, however, that if $\frac{\mathfrak{r}_n}{n^{1/2}}\to\infty$, then not only the efficiency, but even the $\sqrt{n}$–convergence rate of this estimator fails in the class of covariance operators $\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)$.

Proposition 3.3.5. *Let $a>r$ and let σ_0^2 be sufficiently small, say,*

$$\sigma_0^2\leq\frac{1}{2}\Big[\frac{a^2}{(r-1)^2}-\frac{a}{r-1}\Big].$$

Let $\mathfrak{r}_n=o(n)$ and $\frac{\mathfrak{r}_n}{n^{1/2}}\to\infty$ as $n\to\infty$. Then, for some constant $c=c(r;a;\sigma_0)>0$

$$\lim_{n\to\infty}\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)}\mathbb{P}_\Sigma\big\{|\langle\hat\theta_r^\delta,u\rangle-\langle\theta_r(\Sigma),u\rangle|\geq c\|u\|\frac{\mathfrak{r}_n}{n}\big\}=1.$$

The reason for the loss of the $\sqrt{n}$–convergence rate of plug-in estimators of linear functionals of principal components is their large bias in the case when the complexity of the problem is even moderately high (that is, $\frac{\mathfrak{r}_n}{n^{1/2}}\to\infty$). In [KL16], a method of bias reduction in this problem was suggested that led to $\sqrt{n}$-consistent estimation of linear functionals. The estimator is, however, not efficient, since the basic sample split

employed in its construction gives a limiting variance that is twice as large as the optimal one. Since the bias parameter depends itself on sample size in a subtle way, modifying the algorithm in [KL16] to obtain an efficient estimator is not straightforward, and we describe below a construction that yields an asymptotically normal estimator of $\langle \theta_r(\Sigma), u\rangle$ with optimal variance in the class of covariance operators $\mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)$ with $\mathfrak{r}_n = o(n)$. The idea is to use only a small portion of the data (of size $o(n)$) to estimate the bias parameters and to use most of the data for the estimator of the target eigenvector.

For some $m < n/3$, we split the sample $X_1, \dots, X_n$ into three disjoint subsamples, one of size $n' := n - 2m > n/3$ and two others of size m each. In Theorem 3.3.6 below, we choose $m = m_n = o(n)$ as $n \to \infty$, which implies $n' = n'_n = (1 + o(1))n$ as $n \to \infty$. Denote by $\hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)}, \hat{\Sigma}^{(3)}$ the sample covariances based on these three subsamples and let $\hat{\theta}_r^{\delta_j, j}, j = 1, 2, 3$ be the corresponding empirical eigenvectors with parameters $\delta_j = \tau \|\hat{\Sigma}^{(j)}\|$ for a proper choice of τ (see Lemma 3.2.5). Let

$$\check{d}_r := \frac{\langle \hat{\theta}_r^{\delta_1,1}, \hat{\theta}_r^{\delta_2,2}\rangle}{\langle \hat{\theta}_r^{\delta_2,2}, \hat{\theta}_r^{\delta_3,3}\rangle^{1/2}} \quad \text{and} \quad \check{\theta}_r := \frac{\hat{\theta}_r^{\delta_1,1}}{\check{d}_r \vee (1/2)}.$$

Our main goal is to prove the following result showing the efficiency of the estimator $\langle \check{\theta}_r, u\rangle$ of the linear functional $\langle \theta_r(\Sigma), u\rangle$. Its proof will be given in Section 3.5.

Theorem 3.3.6. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $\mathfrak{r}_n > 1$ and $\mathfrak{r}_n = o(n)$ as $n \to \infty$. Take $m = m_n$ such that $m_n = o(n)$ and $n\mathfrak{r}_n = o(m_n^2)$ as $n \to \infty$. Then*

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \sup_{x \in \mathbb{R}} \Big| \mathbb{P}_\Sigma \Big\{ \frac{\sqrt{n}(\langle \check{\theta}_r, u\rangle - \langle \theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)} \le x \Big\} - \Phi(x) \Big| \to 0 \text{ as } n \to \infty. \tag{3.7}$$

Moreover, under Assumption 3.3.1 on the loss ℓ,

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \Big| \mathbb{E}_\Sigma \ell\Big(\frac{\sqrt{n}(\langle \check{\theta}_r, u\rangle - \langle \theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)} \Big) - \mathbb{E}\ell(Z) \Big| \to 0 \text{ as } n \to \infty.$$

Remark 3.3.1. The assumption $\mathfrak{r}_n = o(n)$ is not necessary for the existence of a $\sqrt{n}$-consistent estimator of $\langle \theta_r(\Sigma), u\rangle$. In fact, the estimator $\langle \check{\theta}_r, u\rangle$ (say, with $m = n/4$) is $\sqrt{n}$-consistent provided that $\mathfrak{r}_n \le cn$ for a sufficiently small constant $c > 0$. This fact easily follows from (3.66) of Corollary 3.5.3 in Section 3.5. This is also the case for a somewhat simpler estimator (based on splitting the sample into two parts) considered earlier by Koltchinskii and Lounici [KL16] (see Proposition 3). However, it is not clear whether asymptotically efficient estimators (in the sense of Theorem 3.3.6) of linear functionals $\langle \theta_r(\Sigma), u\rangle$ of the eigenvector $\theta_r(\Sigma)$ with $\sqrt{n}$-rate and optimal limit variance $\sigma_r(\Sigma; u)$ exist when the condition $\mathfrak{r}_n = o(n)$ does not hold. In this case, the linear term of the perturbation series, that determines the limit variance $\sigma_r(\Sigma; u)$, is no longer dominant, which makes the existence of such estimators unlikely. However, asymptotically normal

estimators of functionals $\langle \theta_r(\Sigma), u\rangle$ might still exist (but with a larger limit variance). It could be easier to develop such estimators in the case of spiked covariance models rather than in the more general framework of the current paper. The solution of this problem would rely on the tools of random matrix theory (see, [Pau07] as well as the more recent paper [BKYY16]) rather than perturbation theory, and, possibly, it would require the development of minimax lower bound techniques different from those employed in the present paper.

Remark 3.3.2. It is not hard to develop similar asymptotically efficient estimators for l-dimensional "functionals" of the form $A\theta_r(\Sigma)$, where A is a linear operator from $\mathbb{H}$ into $\mathbb{R}^l$ for a fixed (small) dimension l. This is equivalent to the problem of estimation of $(\langle \theta_r(\Sigma), u_1\rangle, \dots, \langle \theta_r(\Sigma), u_l\rangle)$ for several linear functionals $u_1, \dots, u_l \in \mathbb{H}$. The bias reduction method developed in this paper can be extended to this case and the proof of asymptotic normality of the resulting estimators follows along the same lines as in the case when $l = 1$ with asymptotic covariance matrix equal to

$$(\mu_r \langle \Sigma C_r u_i, C_r u_j\rangle)_{i,j=1,\dots,p}\,.$$

Similarly, our approach can be extended to linear functionals of multiple eigenvectors of multiplicity 1 each, e.g. $(\langle \theta_r(\Sigma), u\rangle, \langle \theta_s(\Sigma), v\rangle)$, $u, v \in \mathbb{H}$. In this case the asymptotic covariance equals

$$-\frac{\mu_r \mu_s}{(\mu_r - \mu_s)^2}\langle \theta_r(\Sigma), v\rangle \langle \theta_s(\Sigma), u\rangle.$$

In this case the debiasing strategy in Theorem 3.3 can be adjusted by using the second and third part of the sample to estimate the bias for both $\theta_r(\Sigma)$ and $\theta_s(\Sigma)$.
However, note that when $\mathbf{r}(\Sigma)$ is large, the asymptotic normality of random vectors $n^{1/2}(\check{\theta}_r - \theta_r(\Sigma))$ holds only in the sense of finite-dimensional distributions, not in the sense of weak convergence in the Hilbert space $\mathbb{H}$ (indeed, the norm $\|\check{\theta}_r - \theta_r(\Sigma)\|$ is of order $\sqrt{\mathbf{r}(\Sigma)/n} \gg 1/\sqrt{n}$).

Remark 3.3.3. Our method of bias reduction does not seem to have an easy extension to the problem of estimation of linear functionals of spectral projections P_r for an eigenvalue of multiplicity > 1. In part, this was a motivation for the first author to develop a more general approach to bias reduction (a so called "bootstrap chain" method) and to study the problem of efficient estimation for more general smooth functionals of covariance of the form $\langle f(\Sigma), B\rangle$, where f is a smooth function on the real line (see [Kol17]). So far, the asymptotic efficiency for the resulting "bootstrap chain" estimators has been proved under more restrictive assumptions on the underlying covariance Σ. In particular, it was assumed that $\mathbb{H}$ is a space of finite (high) dimension p and that the spectrum of Σ is both upper and lower bounded away from 0 by constants which implies that $\mathbf{r}(\Sigma) \asymp p$.

Remark 3.3.4. Lemma 3.5.4 of Section 3.5 provides explicit bounds on the accuracy of

the normal approximation in Theorem 3.3.6. Using these bounds, it is possible to state somewhat more complicated conditions under which the normal approximation holds if $a = a_n \to \infty$ or $\sigma_0 = \sigma_0^{(n)} \to 0$ as $n \to \infty$. In particular, the normal approximation (3.7) still holds uniformly in $\mathcal{S}^{(r)}(\mathfrak{r}_n, a_n, \sigma_0^{(n)}, u)$ provided that $m_n = o(n)$ and

$$\frac{a_n^2}{\sigma_0^{(n)}}\left(\sqrt{\frac{n\mathfrak{r}_n}{m_n^2}\log\frac{m_n^2}{n\mathfrak{r}_n}} \bigvee \sqrt{\frac{n\log^2\frac{m_n^2}{n\mathfrak{r}_n}}{m_n^2}}\right) \to 0 \text{ as } n \to \infty.$$

Finally, we show that $\sigma_r(\Sigma; u)$ can be consistently estimated by $\sigma_r(\hat{\Sigma}; u)$, which allows us to replace the standard deviation $\sigma_r(\Sigma; u)$ in the normal approximation (3.7) by its empirical version. This yields the following result that can be used for hypotheses testing of linear functionals of θ_r. See Section 3.6 for its proof.

Corollary 3.3.7. *Under the conditions of Theorem 3.3.6,*

$$\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)} \sup_{x\in\mathbb{R}} \left|\mathbb{P}_\Sigma\left\{\frac{\sqrt{n}(\langle\check{\theta}_r,u\rangle - \langle\theta_r(\Sigma),u\rangle)}{\sigma_r(\hat{\Sigma};u)} \le x\right\} - \Phi(x)\right| \to 0 \text{ as } n \to \infty.$$

3.4 Proof of Theorem 3.3.2

We will prove the result for empirical eigenvectors $\hat{\theta}_r$ rather than for $\hat{\theta}_r^\delta$. The reduction to this case is based on Lemma 3.2.5 which immediately implies that

$$\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)} \mathbb{P}_\Sigma\{\hat{\theta}_r^\delta \ne \hat{\theta}_r\} \le e^{-\beta\tau^2 n}.$$

Therefore, denoting

$$\xi_n(\Sigma) := \frac{\sqrt{n}(\langle\hat{\theta}_r^\delta, u\rangle - \sqrt{1+b_r(\Sigma)}\langle\theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)}$$

and

$$\eta_n(\Sigma) := \frac{\sqrt{n}(\langle\hat{\theta}_r, u\rangle - \sqrt{1+b_r(\Sigma)}\langle\theta_r(\Sigma), u\rangle)}{\sigma_r(\Sigma; u)},$$

we obtain

$$\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)} \sup_{x\in\mathbb{R}} |\mathbb{P}_\Sigma\{\xi_n(\Sigma) \le x\} - \mathbb{P}_\Sigma\{\eta_n(\Sigma) \le x\}| \le e^{-\beta\tau^2 n} \to 0 \text{ as } n \to \infty.$$

Also, since $\xi_n(\Sigma) \le \frac{2\sqrt{n}\|u\|}{\sigma_r(\Sigma;u)}$ and $\eta_n(\Sigma) \le \frac{2\sqrt{n}\|u\|}{\sigma_r(\Sigma;u)}$, we obtain that

$$\sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)} |\mathbb{E}_\Sigma\ell(\xi_n(\Sigma)) - \mathbb{E}_\Sigma\ell(\eta_n(\Sigma))|$$

$$\leq \sup_{\Sigma\in\mathcal{S}^{(r)}(\mathfrak{r}_n,a,\sigma_0,u)} \mathbb{E}_\Sigma|\ell(\xi_n(\Sigma)) - \ell(\eta_n(\Sigma))|I(\hat\theta_r^\delta \neq \hat\theta_r) \leq 2\ell\left(\frac{2\sqrt{n}\|u\|}{\sigma_0}\right)e^{-\beta\tau^2 n} \to 0,$$

under Assumption 3.3.1.

We will prove more explicit bounds for the estimator $\hat\theta_r$ stated below in Lemma 3.4.8 that immediately implies the result.

Our starting point is the first order perturbation expansion of the empirical spectral projection operator $\hat P_r$:

$$\hat P_r = P_r + L_r(E) + S_r(E) \tag{3.8}$$

with a linear term $L_r(E) = P_rEC_r + C_rEP_r$ and a remainder $S_r(E)$, where $E := \hat\Sigma - \Sigma$. It was proved in [KL16] that, under the assumption

$$\mathbb{E}\|\hat\Sigma - \Sigma\| \leq \frac{(1-\gamma)g_r}{2} \tag{3.9}$$

for some $\gamma \in (0,1)$, the bilinear form of the remainder $S_r(E)$ satisfies the following concentration inequality: for all $u, v \in \mathbb{H}$ and for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$\left|\langle (S_r(E) - \mathbb{E}S_r(E))u, v\rangle\right| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\sqrt{\frac{t}{n}}\|u\|\|v\|. \tag{3.10}$$

Under the same assumption, it was also proved in [KL16] that the following representation holds for the bias $\mathbb{E}\hat P_r - P_r$ of empirical spectral projections $\hat P_r$:

$$\mathbb{E}\hat P_r - P_r = P_r(\mathbb{E}\hat P_r - P_r)P_r + T_r, \tag{3.11}$$

where the main term $P_r(\mathbb{E}\hat P_r - P_r)P_r$ is aligned with the spectral projection P_r and is of order

$$\|P_r(\mathbb{E}\hat P_r - P_r)P_r\| \lesssim \frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n} \tag{3.12}$$

and the remainder T_r satisfies the bound

$$\|T_r\| \lesssim_\gamma \frac{m_r\|\Sigma\|^2}{g_r^2}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\frac{1}{\sqrt{n}}. \tag{3.13}$$

Representation (3.11) is especially simple in the case when P_r is of rank 1 ($m_r = 1$), which also implies that $\hat P_r$ is of rank 1. In this case, $P_r = \theta_r \otimes \theta_r$, $\hat P_r = \hat\theta_r \otimes \hat\theta_r$ for unit eigenvectors $\theta_r, \hat\theta_r$ of covariance operators $\Sigma, \hat\Sigma$, respectively, and

$$P_r(\mathbb{E}\hat P_r - P_r)P_r = b_rP_r$$

for a "bias parameter" $b_r = b_r(\Sigma)$:

$$b_r = \mathbb{E}\langle\hat\theta_r, \theta_r\rangle^2 - 1 \in [-1, 0].$$

Thus, it follows from (3.11) that

$$\mathbb{E}\hat{P}_r = (1+b_r)P_r + T_r. \tag{3.14}$$

We obtain from (3.8) and (3.14) that

$$\hat{P}_r - (1+b_r)P_r = L_r(E) + S_r(E) - \mathbb{E}S_r(E) + T_r. \tag{3.15}$$

Denote

$$\rho_r(u) := \langle (\hat{P}_r - (1+b_r)P_r)\theta_r, u\rangle, u \in \mathbb{H}.$$

As in [KL16], the function $\rho_r(u), u \in \mathbb{H}$ will be used in what follows to control the linear forms $\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle, u \in \mathbb{H}$. First, we need to derive some bounds on $\rho_r(u)$.

The following lemma is an immediate consequence of (3.15), (3.10) and (3.13).

Lemma 3.4.1. *Suppose condition (3.9) holds for some $\gamma \in (0,1)$. Then, for all $u \in \mathbb{H}$ and for all $t \geq 1$ with probability at least $1 - e^{-t}$*

$$|\rho_r(u) - \langle L_r(E)\theta_r, u\rangle| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\sqrt{\frac{t}{n}}\|u\|. \tag{3.16}$$

We will need simple concentration and normal approximation bounds for $\langle L_r(E)\theta_r, u\rangle$ given in the next lemma.

Lemma 3.4.2. *For all $t \geq 1$ with probability at least $1 - e^{-t}$*

$$|\langle L_r(E)\theta_r, u\rangle| \lesssim \sigma_r(\Sigma; u)\left(\sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right). \tag{3.17}$$

Moreover, if $\sigma_r(\Sigma; u) > 0$, then

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big\{\frac{\sqrt{n}\langle L_r(E)\theta_r, u\rangle}{\sigma_r(\Sigma; u)} \leq x\Big\} - \Phi(x)\Big| \lesssim \frac{1}{\sqrt{n}}, \tag{3.18}$$

where Φ is the distribution function of standard normal r.v.

Proof. Without loss of generality, assume that the space $\mathbb{H}$ is finite-dimensional (the general case follows by a simple approximation argument). Since $L_r(E) = P_rEC_r + C_rEP_r$ and $C_r\theta_r = 0$, we have

$$\langle L_r(E)\theta_r, u\rangle = \langle C_rEP_r\theta_r, u\rangle = \langle E\theta_r, C_ru\rangle = \langle E, \theta_r \otimes C_ru\rangle.$$

Since E is self-adjoint, we obtain that

$$\langle L_r(E)\theta_r, u\rangle = \frac{1}{2}\langle E, \theta_r \otimes C_ru + C_ru \otimes \theta_r\rangle.$$

Let $Z, Z_1, \dots, Z_n$ be i.i.d. standard normal vectors in $\mathbb{H}$ such that $X_j = \Sigma^{1/2} Z_j$. Then

$$E = \Sigma^{1/2}\biggl(n^{-1}\sum_{j=1}^n Z_j \otimes Z_j - \mathbb{E}(Z\otimes Z)\biggr)\Sigma^{1/2}.$$

Defining

$$D := \frac{1}{2}\Sigma^{1/2}(\theta_r \otimes C_r u + C_r u \otimes \theta_r)\Sigma^{1/2} = \frac{1}{2}\bigl(\Sigma^{1/2}\theta_r \otimes \Sigma^{1/2} C_r u + \Sigma^{1/2} C_r u \otimes \Sigma^{1/2}\theta_r\bigr),$$

we obtain that

$$\begin{aligned}\langle L_r(E)\theta_r, u\rangle &= \biggl\langle n^{-1}\sum_{j=1}^n Z_j\otimes Z_j - \mathbb{E}(Z\otimes Z), D\biggr\rangle \\ &= n^{-1}\sum_{j=1}^n (\langle DZ_j, Z_j\rangle - \mathbb{E}\langle DZ, Z\rangle).\end{aligned}$$

Clearly, $\langle DZ, Z\rangle \stackrel{d}{=} \sum_k \lambda_k g_k^2$, where $\{\lambda_k\}$ are the eigenvalues of D and $\{g_k\}$ are i.i.d. standard normal r.v. It easily follows that

$$\mathbb{E}\langle DZ, Z\rangle = \operatorname{tr}(D) = 0$$

and

$$\operatorname{Var}(\langle DZ, Z\rangle) = 2\sum_k \lambda_k^2 = 2\|D\|_2^2 = \sigma_r^2(\Sigma; u).$$

We can now represent $\langle L_r(E)\theta_r, u\rangle$ as follows:

$$\langle L_r(E)\theta_r, u\rangle \stackrel{d}{=} n^{-1}\sum_{j=1}^n \sum_k \lambda_k (g_{k,j}^2 - 1),$$

where $\{g_{k,j}\}$ are i.i.d. standard normal r.v. Using standard exponential bounds for sums of independent ψ_1 r.v. (see, e.g., [Ver12], Proposition 5.16 or Theorem 3.1.9 in [GN16]), we obtain that with probability at least $1 - e^{-t}$

$$\bigl|n^{-1}\sum_{j=1}^n \sum_k \lambda_k (g_{k,j}^2 - 1)\bigr| \lesssim \biggl(\sum_k \lambda_k^2\biggr)^{1/2}\sqrt{\frac{t}{n}} \bigvee \sup_k |\lambda_k| \frac{t}{n},$$

which implies that with the same probability

$$|\langle L_r(E)\theta_r, u\rangle| \lesssim \|D\|_2 \sqrt{\frac{t}{n}} \bigvee \|D\| \frac{t}{n}.$$

Since $\|D\| \le \|D\|_2 = \frac{1}{2}\sigma_r^2(\Sigma; u)$, bound (3.17) follows.

To prove (3.18), we use the Berry-Esseen bound that implies

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big\{\frac{\sum_{j=1}^n\sum_k\lambda_k(g_{k,j}^2-1)}{\sqrt{n}(2\sum_k\lambda_k^2)^{1/2}}\le x\Big\}-\Phi(x)\Big|\lesssim\frac{\sum_k|\lambda_k|^3}{(\sum_k\lambda_k^2)^{3/2}}\frac{1}{\sqrt{n}},$$

and therefore

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big\{\frac{\sqrt{n}\langle L_r(E)\theta_r,u\rangle}{\sigma_r(\Sigma;u)}\le x\Big\}-\Phi(x)\Big|\lesssim\frac{\|D\|_3^3}{\|D\|_2^3}\frac{1}{\sqrt{n}}\lesssim\frac{\|D\|}{\|D\|_2}\frac{1}{\sqrt{n}}\lesssim\frac{1}{\sqrt{n}}.$$

□

The following bounds on $\rho_r(u)$ immediately follow from (3.16) and (3.17).

Lemma 3.4.3. *Suppose condition (3.9) holds for some $\gamma\in(0,1)$. Then, for all $u\in\mathbb{H}$ and for all $t\ge 1$ with probability at least $1-e^{-t}$*

$$|\rho_r(u)|\lesssim_\gamma\sigma_r(\Sigma;u)\Big(\sqrt{\tfrac{t}{n}}\bigvee\tfrac{t}{n}\Big)+\tfrac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\tfrac{t}{n}}\Big)\sqrt{\tfrac{t}{n}}\|u\|.\tag{3.19}$$

Moreover, with the same probability

$$|\rho_r(u)|\lesssim_\gamma\tfrac{\|\Sigma\|}{g_r}\sqrt{\tfrac{t}{n}}\|u\|+\tfrac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\tfrac{t}{n}}\Big)\sqrt{\tfrac{t}{n}}\|u\|.\tag{3.20}$$

and, for $u=\theta_r$,

$$|\rho_r(\theta_r)|\lesssim_\gamma\frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\Big)\sqrt{\frac{t}{n}}.\tag{3.21}$$

Note that we dropped the term $\frac{t}{n}$ in some of the expressions on the right hand side of the above bounds (compare with (3.16)). This term is dominated by $\sqrt{\frac{t}{n}}$ for $t\le n$. Moreover, it follows from the definition of $\rho_r(u)$ that it is upper bounded by $2\|u\|$. Since $\frac{\|\Sigma\|}{g_r}\ge 1$, this easily implies that, for $t\ge n$, the right hand side of bound (3.20) (with a proper constant) is larger than $|\rho_r(u)|$. Bound (3.21) follows from (3.16) since $\langle L_r(E)\theta_r,\theta_r\rangle=0$.

To study concentration and normal approximation of the linear form

$$\langle\hat\theta_r-\sqrt{1+b_r}\theta_r,u\rangle,u\in\mathbb{H},$$

it remains to prove that it can be approximated by $\langle L_r(E)\theta_r,u\rangle$.

Lemma 3.4.4. *Suppose that for some $\gamma\in(0,1)$ condition (3.9) holds and, in addition,*

$$1+b_r\ge\gamma.\tag{3.22}$$

Then, for all $u \in \mathbb{H}$ and for all $t \geq 1$, with probability at least $1 - e^{-t}$

$$|\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \langle L_r(E)\theta_r, u\rangle| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right)\sqrt{\frac{t}{n}}\|u\|. \tag{3.23}$$

Proof. We use the following representation obtained in [KL16] (see (6.7) in [KL16]), which holds provided that $\hat{\theta}_r$ and θ_r are properly aligned so that $\langle \hat{\theta}_r, \theta_r\rangle \geq 0$:

$$\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle = \frac{\rho_r(u)}{\sqrt{1+b_r+\rho_r(\theta_r)}} \tag{3.24}$$
$$- \frac{\sqrt{1+b_r}}{\sqrt{1+b_r+\rho_r(\theta_r)}(\sqrt{1+b_r+\rho_r(\theta_r)}+\sqrt{1+b_r})}\rho_r(\theta_r)\langle \theta_r, u\rangle$$

(it is clear from the proof given in [KL16] that $1 + b_r + \rho_r(\theta_r) \geq 0$). Denote

$$\nu_r := \frac{\rho_r(\theta_r)}{1+b_r}.$$

Then, it is easy to see that

$$\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle = \rho_r(u) - \frac{b_r/(1+b_r)+\nu_r}{1+\nu_r+\sqrt{(1+\nu_r)/(1+b_r)}}\rho_r(u) \tag{3.25}$$
$$- \frac{\nu_r\sqrt{1+b_r}}{1+\nu_r+\sqrt{1+\nu_r}}\langle \theta_r, u\rangle.$$

Recall that (3.9) and (3.22) hold for some $\gamma \in (0,1)$. If $|\nu_r| \leq 1/2$, then (3.25) easily implies that

$$|\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \rho_r(u)| \leq \frac{1}{\gamma}(|b_r| + |\nu_r|)|\rho_r(u)| + |\nu_r||\langle \theta_r, u\rangle|. \tag{3.26}$$

It also follows from (3.21) that, under condition (3.22),

$$|\nu_r| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}} \tag{3.27}$$

with probability at least $1 - e^{-t}$. On the other hand, bound (3.12) implies that

$$|b_r| \lesssim \frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n}. \tag{3.28}$$

It follows from (3.27) that for the condition $|\nu_r| \leq 1/2$ to hold with probability at least $1 - e^{-t}$, it is enough to have

$$\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}} \leq c_\gamma \tag{3.29}$$

for a small enough constant $c_\gamma > 0$. Assume that (3.29) holds. Note also that it implies

that $t \lesssim n$ and condition (3.9) and Theorem 3.2.2 imply that $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \lesssim 1$. It follows from (3.26), (3.20), (3.27) and (3.28) that with probability at least $1-3e^{-t}$:

$$
\begin{aligned}
&|\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \rho_r(u)| \\
\lesssim_\gamma &\left[\frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n} + \left(\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\right) \wedge 1/2\right] \\
&\times \left[\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}}\|u\| + \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|\right] \\
&+ \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|. \qquad (3.30)
\end{aligned}
$$

Using the facts that

$$
\frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n} \lesssim \frac{\|\Sigma\|}{g_r}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \lesssim 1,
$$

that

$$
\frac{\|\Sigma\|^2}{g_r^2}\frac{t}{n} \lesssim \frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}} \lesssim 1
$$

and that

$$
\frac{\|\Sigma\|^2}{g_r^2}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\sqrt{\frac{t}{n}} \lesssim \frac{\|\Sigma\|}{g_r}\left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{1/4}\left(\frac{t}{n}\right)^{1/4} \le \frac{\|\Sigma\|}{g_r}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)
$$

(that follow from condition (3.29)), it is easy to conclude that the last term

$$
\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|
$$

in the right hand side of bound (3.30) is dominant. Hence, with probability at least $1-e^{-t}$

$$
|\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \rho_r(u)| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\| \qquad (3.31)
$$

provided that condition (3.29) holds. On the other hand, if

$$
\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}} > c_\gamma,
$$

then

$$
\begin{aligned}
&|\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \rho_r(u)| \le |\langle \hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle| + |\rho_r(u)| \\
&\le (\|\hat{\theta}_r\| + \sqrt{1+b_r}\|\theta_r\|)\|u\| + (\|\hat{P}_r\| + (1+b_r)\|P_r\|)\|\theta_r\|\|u\| \le 4\|u\| \\
&\lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|.
\end{aligned}
$$

Thus, we proved that with probability at least $1-e^{-t}$

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle - \rho_r(u)| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|. \tag{3.32}$$

It remains to combine this with the bound (3.16) to complete the proof.

□

The following result is a slightly improved version of Theorem 6 in [KL16].

Lemma 3.4.5. *Under conditions (3.9) and (3.22) for some $\gamma \in (0,1)$, the following bounds hold for all $t \geq 1$ with probability at least $1-e^{-t}$:*

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle| \lesssim_\gamma \frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}}\|u\| \tag{3.33}$$

and

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, \theta_r\rangle| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}. \tag{3.34}$$

Proof. Indeed, it follows from (3.23) and (3.17) that, for some constants $C, C_\gamma > 0$ with probability at least $1-e^{-t}$

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle| \leq C\sigma_r(\Sigma;u)\left(\sqrt{\frac{t}{n}}\bigvee\frac{t}{n}\right) + C_\gamma\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|.$$

Since $\sigma_r(\Sigma;u) \lesssim \frac{\|\Sigma\|}{g_r}\|u\|$, with the same probability

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle| \leq C\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}}\|u\| + C_\gamma\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|.$$

We dropped the term $\frac{t}{n}$ present in bounds (3.23) and (3.17) since for $t \geq n$ (the only case when it is needed), the right hand side already dominates the left hand side (which is smaller than $2\|u\|$). Note that condition (3.9) and Theorem 3.2.2 imply that $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \leq c_\gamma$ for some constant $c_\gamma > 0$. Assuming that also $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}} \leq c_\gamma$, which implies that $t \lesssim n$, we obtain that for some constant $C_\gamma > 0$ with probability at least $1-e^{-t}$ bound (3.33) holds. On the other hand, if $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}} > c_\gamma$, then

$$|\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle| \leq (\|\hat{\theta}_r\| + \sqrt{1+b_r}\|\theta_r\|)\|u\| \leq 2\|u\| \lesssim_\gamma \frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}}\|u\|,$$

implying again (3.33). For $u = \theta_r$, $\langle L_r(E)\theta_r, u\rangle = 0$ and bound (3.23) implies that with probability at least $1-e^{-t}$ (3.34) holds. □

The following two lemmas will be used to derive normal approximation bounds for $\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle$ from the corresponding bounds for $\langle L_r(E)\theta_r, u\rangle$ as well as to control

the risk for loss functions satisfying Assumption 3.3.1. We state them without proofs (which are elementary).

Lemma 3.4.6. *For random variables ξ, η, denote*

$$\Delta(\xi;\eta) := \sup_{x\in\mathbb{R}} |\mathbb{P}\{\xi \le x\} - \mathbb{P}\{\eta \le x\}|$$

and

$$\delta(\xi;\eta) := \inf\{\delta > 0 : \mathbb{P}\{|\xi-\eta| \ge \delta\} + \delta\}.$$

Then, for a standard normal r.v. Z,

$$\Delta(\xi;Z) \le \Delta(\eta;Z) + \delta(\xi;\eta).$$

Under Assumption 3.3.1, for all $A > 0$

$$|\mathbb{E}\ell(\xi) - \mathbb{E}\ell(\eta)| \le 4\ell(A)\Delta(\xi;\eta) + \mathbb{E}\ell(\xi)I(|\xi| \ge A) + \mathbb{E}\ell(\eta)I(|\eta| \ge A).$$

Lemma 3.4.7. *Let ξ be a random variable such that for some $\tau_1 \ge 0$ and $\tau_2 \ge 0$ and for all $t \ge 1$ with probability at least $1 - e^{-t}$*

$$|\xi| \le \tau_1\sqrt{t} \vee \tau_2 t.$$

Let ℓ be a loss function satisfying Assumption 3.3.1. If $2c_2\tau_2 < 1$, then

$$\mathbb{E}\ell^2(\xi) \le 2e\sqrt{2\pi}c_1^2 e^{2c_2^2\tau_1^2} + \frac{ec_1^2}{1-2c_2\tau_2}. \tag{3.35}$$

Next we prove the normal approximation bounds for linear forms $\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle$.

Lemma 3.4.8. *Suppose that conditions (3.9) and (3.22) hold for some $\gamma \in (0,1)$ and also that $n \ge 2\mathbf{r}(\Sigma)$. Assume that, for some $u \in \mathbb{H}$, $\sigma_r(\Sigma;u) > 0$. Let $\alpha \ge 1$. Then the following bound holds: for some constants $C, C_{\gamma,\alpha} > 0$,*

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big\{\frac{\sqrt{n}\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle}{\sigma_r(\Sigma;u)} \le x\Big\} - \Phi(x)\Big| \\
&\le Cn^{-1/2} + \frac{C_{\gamma,\alpha}}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}\log\frac{n}{\mathbf{r}(\Sigma)}} \bigvee \frac{\log\frac{n}{\mathbf{r}(\Sigma)}}{\sqrt{n}}\right)\|u\| + \left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{\alpha}.
\end{aligned} \tag{3.36}$$

Moreover, under Assumption 3.3.1 on the loss ℓ, there exist constants $C, C_\gamma, C_{\gamma,\alpha} > 0$ such that

$$\Big|\mathbb{E}\ell\Big(\frac{\sqrt{n}\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle}{\sigma_r(\Sigma;u)}\Big) - \mathbb{E}\ell(Z)\Big|$$

$$\leq c_1 e^{c_2 A}\left(Cn^{-1/2} + \frac{C_{\gamma,\alpha}}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}\log\frac{n}{\mathbf{r}(\Sigma)}}\bigvee\frac{\log\frac{n}{\mathbf{r}(\Sigma)}}{\sqrt{n}}\right)\|u\| + \left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{\alpha}\right)$$
$$+ 2e^{3/2}(2\pi)^{1/4}c_1 e^{c_2^2\tau^2}e^{-A^2/2\tau^2} + c_1 e^{c_2^2}e^{-A^2/4}, \tag{3.37}$$

where

$$\tau := C_\gamma\frac{\|\Sigma\|\|u\|}{g_r\sigma_r(\Sigma;u)}.$$

Proof. We will use the first claim of Lemma 3.4.6 with

$$\xi := \frac{\sqrt{n}\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle}{\sigma_r(\Sigma;u)} \text{ and } \eta := \frac{\sqrt{n}\langle L_r(E)\theta_r, u\rangle}{\sigma_r(\Sigma;u)}.$$

It follows from bound (3.23) that, under conditions (3.9) and (3.22), for some $C_\gamma > 0$

$$\delta(\xi;\eta) \leq \inf_{t\geq 1}\left\{\frac{C_\gamma}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\bigvee\sqrt{\frac{t}{n}}\bigvee\frac{t}{n}\right)\sqrt{t}\|u\| + e^{-t}\right\}.$$

Taking $t := \alpha\log\left(\frac{n}{\mathbf{r}(\Sigma)}\right)$ with some $\alpha \geq 1$ easily yields an upper bound

$$\delta(\xi;\eta) \leq \frac{C_{\gamma,\alpha}}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}\log\frac{n}{\mathbf{r}(\Sigma)}}\bigvee\frac{\log\frac{n}{\mathbf{r}(\Sigma)}}{\sqrt{n}}\right)\|u\| + \left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{\alpha}.$$

Using bound (3.18) to control $\Delta(\eta;Z)$, we obtain from Lemma 3.4.6 that bound (3.36) holds with some constants $C, C_{\gamma,\alpha} > 0$. To prove the second statement, we use the second bound of Lemma 3.4.6 with the random variable $\xi := \frac{\sqrt{n}\langle\hat{\theta}_r - \sqrt{1+b_r}\theta_r, u\rangle}{\sigma_r(\Sigma;u)}$ and $\eta = Z$. The following exponential bound on ξ is an easy corollary of bound (3.33): for some constant $C_\gamma > 0$ and for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$|\xi| \leq C_\gamma\frac{\|\Sigma\|}{g_r\sigma_r(\Sigma;u)}\sqrt{t}\|u\| = \tau\sqrt{t}. \tag{3.38}$$

Using bound (3.35) with $\tau_1 = \tau$ and $\tau_2 = 0$, we obtain

$$\mathbb{E}\ell^2(\xi) \leq 2e\sqrt{2\pi}c_1^2 e^{2c_2^2\tau_1^2} + ec_1^2 \leq 4e\sqrt{2\pi}c_1^2 e^{2c_2^2\tau_1^2}$$

Therefore,

$$\mathbb{E}\ell(\xi)I(|\xi|\geq A) \leq \mathbb{E}^{1/2}\ell^2(\xi)\mathbb{P}^{1/2}\{|\xi|\geq A\} \leq 2e^{3/2}(2\pi)^{1/4}c_1 e^{c_2^2\tau^2}e^{-A^2/2\tau^2}.$$

We also have

$$\mathbb{E}\ell(Z)I(|Z|\geq A) \leq c_1 e^{c_2^2}e^{-A^2/4}.$$

Using bound (3.36), we can now deduce bound (3.37) from the second statement of Lemma 3.4.6. □

Lemma 3.4.8 immediately implies Theorem 3.3.2 (by passing to the limit as $n \to \infty$ in (3.36) and as $n \to \infty$ and then $A \to \infty$ in (3.37)).

3.4.1 Proof of Proposition 3.3.5

Denote

$$A_r(\Sigma) := 2\,\mathrm{trace}(P_r \Sigma P_r)\,\mathrm{trace}(C_r \Sigma C_r) = 2\sum_{s \neq r} \frac{\mu_r \mu_s m_s}{(\mu_r - \mu_s)^2}.$$

It was shown in [KL17c] that

$$\mathbb{E}\|L_r(E)\|_2^2 = \frac{A_r(\Sigma)}{n},$$

where $E = \hat{\Sigma} - \Sigma$. Note that

$$\frac{A_r(\Sigma)}{2} \leq \frac{\mu_r}{g_r^2}(\mathrm{trace}(\Sigma) - \mu_r) \leq \frac{\|\Sigma\|^2}{g_r^2}\mathbf{r}(\Sigma) \tag{3.39}$$

and

$$\frac{A_r(\Sigma)}{2} \geq \frac{\mu_1 \mu_r}{(\mu_1 - \mu_r)^2 \vee \mu_r^2}(\mathbf{r}(\Sigma) - 1). \tag{3.40}$$

Lemma 3.4.9. *The following representation holds:*

$$b_r(\Sigma) = -\frac{1}{2}\frac{A_r(\Sigma)}{n} + \beta_r,$$

where

$$|\beta_r| \lesssim \frac{\|\Sigma\|^3}{g_r^3}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n}\right)^3.$$

Proof. Recall representation (3.11) and bound (3.13). Note that

$$b_r = \mathrm{trace}(P_r(\mathbb{E}\hat{P}_r - P_r)P_r)$$

and

$$\mathbb{E}\hat{P}_r - P_r = \mathbb{E}S_r(E).$$

We will use the following representation for $S_r(E)$ (based on perturbation series for $\hat{P}_r$) that easily follows from Lemma 4 in [KL17b]:

$$S_r(E) = P_r E C_r E C_r + C_r E P_r E C_r + C_r E C_r E P_r$$

$$-P_r E P_r E C_r^2 - P_r E C_r^2 E P_r - C_r^2 E P_r E P_r + S_r^{(3)}(E),$$

where

$$\|S_r^{(3)}(E)\| \lesssim \frac{\|E\|^3}{g_r^3}.$$

Since $P_r C_r = C_r P_r = 0$ this implies

$$P_r S_r(E) P_r = -P_r E C_r^2 E P_r + P_r S_r^{(3)}(E) P_r.$$

Therefore we obtain

$$b_r = \mathbb{E}\,\mathrm{trace}(P_r S_r(E) P_r) = -\mathbb{E}\,\mathrm{trace}(P_r E C_r^2 E P_r) + \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r)$$

$$= -\mathbb{E}\|P_r E C_r\|_2^2 + \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r) = -\frac{1}{2}\mathbb{E}\|P_r E C_r + C_r E P_r\|_2^2 + \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r)$$

$$-\frac{1}{2}\mathbb{E}\|L_r(E)\|_2^2 + \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r) = -\frac{1}{2}\frac{A_r(\Sigma)}{n} + \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r).$$

Thus, $\beta_r = \mathbb{E}\,\mathrm{trace}(P_r S_r^{(3)}(E) P_r)$ and, using bound (3.4), we get

$$|\beta_r| \le \mathbb{E}\|S_r^{(3)}(E)\|\|P_r\|_1 \le \mathbb{E}\|S_r^{(3)}(E)\| \lesssim \frac{\mathbb{E}\|E\|^3}{g_r^3}$$

$$\lesssim \frac{\|\Sigma\|^3}{g_r^3}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n}\right)^3,$$

which completes the proof. □

It follows from the lower bound (3.40) on $\frac{A_r(\Sigma)}{2}$ and the bound of Lemma 3.4.9 that, under the assumption $\mathbf{r}(\Sigma) \le n$, with some constant $C > 0$

$$|b_r| \ge \frac{\mu_1\mu_r}{(\mu_1-\mu_r)^2 \vee \mu_r^2}\frac{\mathbf{r}(\Sigma)-1}{n} - C\frac{\|\Sigma\|^3}{g_r^3}\left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{3/2}. \tag{3.41}$$

Next note that

$$|\langle\hat\theta_r - \theta_r, u\rangle| \ge |\sqrt{1+b_r}-1||\langle\theta_r,u\rangle| - |\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle|$$

$$\ge \frac{|b_r|}{1+\sqrt{1+b_r}}|\langle\theta_r,u\rangle| - |\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle|$$

$$\ge \frac{|b_r|}{2}|\langle\theta_r,u\rangle| - |\langle\hat\theta_r - \sqrt{1+b_r}\theta_r, u\rangle|.$$

Using bounds (3.33) and (3.41), we obtain that for all $t \ge 1$ with probability at least $1-e^{-t}$

$$|\langle\hat\theta_r - \theta_r, u\rangle| \ge \frac{1}{2}|\langle\theta_r,u\rangle|\left(\frac{\mu_1\mu_r}{(\mu_1-\mu_r)^2\vee\mu_r^2}\frac{\mathbf{r}(\Sigma)-1}{n} - C\frac{\|\Sigma\|^3}{g_r^3}\left(\frac{\mathbf{r}(\Sigma)}{n}\right)^{3/2}\right) - C_\gamma\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n}}\|u\|. \tag{3.42}$$

We will show that there exists a covariance $\Sigma_0 \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)$ such that $|\langle\theta_r(\Sigma_0), u\rangle| \ge$

$\frac{\|u\|}{2}$,

$$\frac{\mu_1(\Sigma_0)\mu_r(\Sigma_0)}{(\mu_1(\Sigma_0)-\mu_r(\Sigma_0))^2 \vee \mu_r^2(\Sigma_0)} \geq c_1$$

for some constant $c_1 > 0$ that might depend on r, a, σ_0 and $\mathbf{r}(\Sigma_0) - 1 \geq \mathfrak{r}_n/2$. Assuming that such a Σ_0 exists, we choose $t_n \to \infty$, $t_n = o(\frac{\mathfrak{r}_n^2}{n})$ and applying bound (3.42) to $\Sigma = \Sigma_0$, we immediately obtain that

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \mathbb{P}_\Sigma\Big\{|\langle \hat{\theta}_r - \theta_r(\Sigma), u\rangle| \geq \left(\frac{c_1}{8}\frac{\mathfrak{r}_n}{n} - \frac{C}{4}a^3\left(\frac{\mathfrak{r}_n}{n}\right)^{3/2} - C_\gamma a\sqrt{\frac{t_n}{n}}\right)\|u\|\Big\}$$

$$\geq 1 - e^{-t_n} \to 1.$$

Since

$$\left(\frac{c_1}{8}\frac{\mathfrak{r}_n}{n} - \frac{C}{4}a^3\left(\frac{\mathfrak{r}_n}{n}\right)^{3/2} - C_\gamma a\sqrt{\frac{t_n}{n}}\right)\|u\| = \left(\frac{c_1}{8} + o(1)\right)\frac{\mathfrak{r}_n}{n}\|u\|,$$

this implies the claim of Proposition 3.3.5.

It remains to define a Σ_0 with the desired properties. Let

$$\Sigma_0 = \sum_{s=1}^{r+1} \mu_s P_s,$$

where $P_s = \theta_s \otimes \theta_s, s = 1, \dots r$, $\theta_1, \dots, \theta_r$ being arbitrary orthonormal vectors in $\mathbb{H}$ and P_{r+1} is an orthogonal projection on a d-dimensional subspace of $\mathbb{H}$ orthogonal to $\theta_1, \dots, \theta_r$. Let $\mu_s := \mu_1\big(1 - \frac{s-1}{a}\big), s = 1, \dots, r+1$. Then $\bar{g}_r(\Sigma_0) = \frac{\mu_1}{a}$ and the condition $\frac{\|\Sigma_0\|}{\bar{g}_r(\Sigma_0)} \leq a$ is satisfied. For simplicity, assume that $\|u\| = 1$. Moreover, since $\theta_1, \dots, \theta_r$ are arbitrary orthonormal vectors, we can assume without loss of generality that, for $r > 1$, $u := \frac{1}{\sqrt{2}}\theta_1 + \frac{1}{\sqrt{2}}\theta_r$. Then $\langle \theta_r(\Sigma_0), u\rangle = \frac{1}{\sqrt{2}} > \frac{1}{2}\|u\|$ and, by a simple computation,

$$\sigma_r^2(\Sigma_0; u) = \sum_{s \neq r} \frac{\mu_r \mu_s}{(\mu_r - \mu_s)^2}\|P_s u\|^2 = \frac{1}{2}\frac{\mu_1\mu_r}{(\mu_1 - \mu_r)^2} = \frac{1}{2}\left[\frac{a^2}{(r-1)^2} - \frac{a}{r-1}\right].$$

Assuming that $\sigma_0^2 \leq \frac{1}{2}\left[\frac{a^2}{(r-1)^2} - \frac{a}{r-1}\right]$, we conclude that the condition $\sigma_r^2(\Sigma_0; u) \geq \sigma_0^2$ is satisfied. For $r = 1$, we can assume that $u := \frac{1}{\sqrt{2}}\theta_1 + \frac{1}{\sqrt{2}}\theta_2$ with a slight modification of the argument. Finally, we take dimension $d = d_n$ so that

$$\mathbf{r}(\Sigma_0) = \sum_{s=1}^{r}\frac{\mu_s}{\mu_1} + \frac{\mu_{r+1}}{\mu_1}d_n = \sum_{s=1}^{r}\big(1 - \frac{s-1}{a}\big) + \big(1 - \frac{r}{a}\big)d_n \in (\mathfrak{r}_n/2 + 1, \mathfrak{r}_n].$$

Then $\Sigma_0 \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)$. This completes the proof.

3.5 Proof of Theorem 3.3.6

Recall that the estimator $\check{\theta}_r$ is based on empirical eigenvectors $\hat{\theta}_r^{\delta_j,j}, j = 1,2,3$ with parameters $\delta_j = \tau\|\hat{\Sigma}^{(j)}\|$ and with a proper choice of τ (as in Lemma 3.2.5). These eigenvectors are in turn defined in terms of empirical spectral projections $\hat{P}_r^{\delta_j,j}$ of sample covariances $\hat{\Sigma}^{(j)}$ (based on δ_j-clusters of its spectrum $\sigma(\hat{\Sigma}^{(j)})$). We will, however, replace $\check{\theta}_r$ by the estimator $\tilde{\theta}_r$ defined in terms of empirical spectral projections $\hat{P}_r^{(j)}, j = 1,2,3$, $\hat{P}_r^{(j)}$ being the orthogonal projection onto direct sum of eigenspaces of $\hat{\Sigma}^{(j)}$ corresponding to its eigenvalues $\lambda_k(\hat{\Sigma}^{(j)}), k \in \Delta_r$. Since $\mathrm{card}(\Delta_r) = m_r = 1$, $\hat{P}_r^{(j)} = \hat{\theta}_r^{(j)} \otimes \hat{\theta}_r^{(j)}$ and we can define

$$\hat{d}_r := \frac{\langle \hat{\theta}_r^{(1)}, \hat{\theta}_r^{(2)} \rangle}{\langle \hat{\theta}_r^{(2)}, \hat{\theta}_r^{(3)} \rangle^{1/2}}$$

and

$$\tilde{\theta}_r := \frac{\hat{\theta}_r^{(1)}}{\hat{d}_r \vee (1/2)}.$$

The reduction to this case is based on Lemma 3.2.5 (implying that $\hat{P}_r^{\delta_j,j} = \hat{P}_r^{(j)}$ with a high probability) and is straightforward (as in the proof of Theorem 3.3.2).

The rest of the proof is based on several lemmas stated and proved below.

Lemma 3.5.1. *Suppose that for some $\gamma \in (0,1)$ condition (3.9) holds for the sample covariance $\hat{\Sigma}^{(2)}$ based on m observations:*

$$\mathbb{E}\|\hat{\Sigma}^{(2)} - \Sigma\| \le \frac{(1-\gamma)g_r}{2} \tag{3.43}$$

Then, for all $t \ge 1$ with probability at least $1 - e^{-t}$

$$\left|\langle \hat{\theta}_r^{(1)}, \hat{\theta}_r^{(2)} \rangle - \sqrt{1 + b_r^{(n')}}\sqrt{1 + b_r^{(m)}}\right| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}. \tag{3.44}$$

and with the same probability

$$\left|\langle \hat{\theta}_r^{(2)}, \hat{\theta}_r^{(3)} \rangle - (1 + b_r^{(m)})\right| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}. \tag{3.45}$$

Proof. Obviously, condition (3.43) holds also for the sample covariance $\hat{\Sigma}^{(2)}$ (which is based on a sample of the same size m). Moreover, it also holds for the sample covariance $\hat{\Sigma}^{(1)}$ based on $n' \ge m$ observations since the sequence $n \mapsto \mathbb{E}\|\hat{\Sigma}_n - \Sigma\|$ is non-increasing (see, e.g., Lemma 2.4.5 in [vdVW96]).

The following representation is obvious:

$$\begin{aligned}\langle \hat{\theta}_r^{(1)}, \hat{\theta}_r^{(2)} \rangle &= \sqrt{1 + b_r^{(n')}}\sqrt{1 + b_r^{(m)}}\langle \theta_r, \theta_r \rangle \\ &+ \sqrt{1 + b_r^{(m)}}\langle \hat{\theta}_r^{(1)} - \sqrt{1 + b_r^{(n')}}\theta_r, \theta_r \rangle\end{aligned}$$

$$+\sqrt{1+b_r^{(n')}}\langle\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r,\theta_r\rangle$$
$$\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\rangle. \tag{3.46}$$

By bound (3.34), with probability at least $1-e^{-t}$

$$|\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,\theta_r\rangle|\lesssim_\gamma \tfrac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{n'}}\bigvee\sqrt{\tfrac{t}{n'}}\right)\sqrt{\tfrac{t}{n'}} \tag{3.47}$$

Similarly, with probability at least $1-e^{-t}$

$$|\langle\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r,\theta_r\rangle|\lesssim_\gamma \tfrac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\tfrac{t}{m}}\right)\sqrt{\tfrac{t}{m}}. \tag{3.48}$$

To bound the last term in the right hand side of (3.46), we apply bound (3.33) to $\hat\theta_r^{(1)}$ conditionally on the second sample (similarly to the proof of Theorem 6 in [KL16]). This yields that with probability at least $1-e^{-t}$

$$|\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\rangle|\lesssim_\gamma \frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n'}}\|\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\|. \tag{3.49}$$

On the other hand, under the assumption that $\langle\hat\theta_r,\theta_r\rangle\ge 0$,

$$\|\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\|\le\|\hat\theta_r^{(2)}-\theta_r\|+\left|\sqrt{1+b_r^{(m)}}-1\right|$$
$$=\sqrt{2-2\langle\hat\theta_r^{(2)},\theta_r\rangle}+\tfrac{|b_r^{(m)}|}{\sqrt{1+b_r^{(m)}}+1}\le\sqrt{2-2\langle\hat\theta_r^{(2)},\theta_r\rangle^2}+|b_r^{(m)}|$$
$$=\sqrt{2-2\langle\hat P_r^{(2)},P_r\rangle}+|b_r^{(m)}|=\|\hat P_r^{(2)}-P_r\|_2+|b_r^{(m)}|.$$
$$\le\sqrt{2}\|\hat P_r^{(2)}-P_r\|+|b_r^{(m)}|.$$

By a standard perturbation bound (see, e.g., [KL16]),

$$\|\hat P_r^{(2)}-P_r\|\le 4\frac{\|\hat\Sigma^{(2)}-\Sigma\|}{g_r}.$$

Thus,

$$\|\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\|\le 4\sqrt{2}\frac{\|\hat\Sigma^{(2)}-\Sigma\|}{g_r}+|b_r^{(m)}|. \tag{3.50}$$

Using the exponential bound (3.3) on $\|\hat\Sigma^{(2)}-\Sigma\|$ and bound (3.28), we obtain that with probability at least $1-e^{-t}$

$$\left\|\hat\theta_r^{(2)}-\sqrt{1+b_r^{(m)}}\theta_r\right\|\lesssim\frac{\|\Sigma\|}{g_r}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\frac{\mathbf{r}(\Sigma)}{m}\bigvee\sqrt{\frac{t}{m}}\bigvee\frac{t}{m}\right)+\frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{m}. \tag{3.51}$$

Under assumption (3.43), we have $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \lesssim 1$, which implies $\frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{m} \lesssim \frac{\|\Sigma\|}{g_r}\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}$. Thus, the first term in the right hand side of bound (3.51) is dominant. Moreover, we can drop the term $\frac{\mathbf{r}(\Sigma)}{m}$ and, for $t \le m$, we can also drop the term $\frac{\|\Sigma\|}{g_r}\frac{t}{m}$ in the right hand side. Since the left hand side of (3.51) is not larger than 2, for $t > m$, the term $\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{m}}$ is larger (up to a constant) than the left hand side. Thus, the term $\frac{\|\Sigma\|}{g_r}\frac{t}{m}$ can be dropped for all the values of t and the bound (3.51) simplifies as follows

$$\left\|\hat\theta_r^{(2)} - \sqrt{1+b_r^{(m)}}\theta_r\right\| \lesssim \frac{\|\Sigma\|}{g_r}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right) \tag{3.52}$$

and it still holds with probability at least $1-e^{-t}$. It follows from bound (3.49) and (3.52) that for all $t \ge 1$ with probability at least $1-2e^{-t}$

$$|\langle \hat\theta_r^{(1)} - \sqrt{1+b_r^{(n')}}\theta_r, \hat\theta_r^{(2)} - \sqrt{1+b_r^{(m)}}\theta_r\rangle| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{n'}}. \tag{3.53}$$

Taking into account that $n' \ge m$, it easily follows from representation (3.46) and bounds (3.47), (3.48) and (3.53) that with probability at least $1-e^{-t}$

$$\left|\langle \hat\theta_r^{(1)}, \hat\theta_r^{(2)}\rangle - \sqrt{1+b_r^{(n')}}\sqrt{1+b_r^{(m)}}\right| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}},$$

which proves (3.44). The proof of bound (3.45) is similar. □

Define

$$\Delta_1 := \frac{\langle \hat\theta_r^{(1)}, \hat\theta_r^{(2)}\rangle}{\sqrt{1+b_r^{(n')}}\sqrt{1+b_r^{(m)}}} - 1$$

and

$$\Delta_2 := \frac{\langle \hat\theta_r^{(2)}, \hat\theta_r^{(3)}\rangle}{1+b_r^{(m)}} - 1.$$

Assuming that

$$1+b_r^{(n')} \ge (3/4)^2 \quad \text{and} \quad 1+b_r^{(m)} \ge (3/4)^2, \tag{3.54}$$

we obtain that, for some constant $C_\gamma > 0$ and for $t \ge 1$ on an event E of probability at least $1-e^{-t}$

$$|\Delta_1| \vee |\Delta_2| \le C_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}. \tag{3.55}$$

Next we have

$$\hat d_r = \frac{\langle \hat\theta_r^{(1)}, \hat\theta_r^{(2)}\rangle}{\langle \hat\theta_r^{(2)}, \hat\theta_r^{(3)}\rangle^{1/2}} = \frac{\langle \hat\theta_r^{(1)}, \hat\theta_r^{(2)}\rangle/((1+b_r^{(n')})^{1/2}(1+b_r^{(m)})^{1/2})}{\langle \hat\theta_r^{(2)}, \hat\theta_r^{(3)}\rangle^{1/2}/(1+b_r^{(m)})^{1/2}}\sqrt{1+b_r^{(n')}}$$

$$= \frac{1+\Delta_1}{\sqrt{1+\Delta_2}}\sqrt{1+b_r^{(n')}} = \sqrt{1+b_r^{(n')}} + \frac{1+\Delta_1-\sqrt{1+\Delta_2}}{\sqrt{1+\Delta_2}}\sqrt{1+b_r^{(n')}},$$

which implies

$$\begin{aligned}\Big|\hat{d}_r - \sqrt{1+b_r^{(n')}}\Big| &\le \sqrt{1+b_r^{(n')}}\,\frac{\Big|(1+\Delta_1)^2-(1+\Delta_2)\Big|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})} \\ &\le \tfrac{2|\Delta_1|+\Delta_1^2+|\Delta_2|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})}. \end{aligned} \tag{3.56}$$

Under the assumption that

$$\frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\Big)\sqrt{\frac{t}{m}} \le c_\gamma \tag{3.57}$$

for a sufficiently small constant $c_\gamma > 0$, bounds (3.56) and (3.55) imply that on the event E

$$\Big|\frac{\hat{d}_r}{\sqrt{1+b_r^{(n')}}} - 1\Big| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\Big)\sqrt{\frac{t}{m}}. \tag{3.58}$$

Moreover, on the same event E,

$$\begin{aligned}\hat{d}_r &\ge \sqrt{1+b_r^{(n')}} - \tfrac{2|\Delta_1|+\Delta_1^2+|\Delta_2|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})} \\ &\ge \tfrac{3}{4} - \tfrac{2|\Delta_1|+\Delta_1^2+|\Delta_2|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})} \ge \tfrac{1}{2}, \end{aligned} \tag{3.59}$$

$$\Big|\frac{\sqrt{1+b_r^{(n')}}}{\hat{d}_r} - 1\Big| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\Big)\sqrt{\frac{t}{m}} \tag{3.60}$$

and also, using bound (3.28), we obtain that

$$\begin{aligned}|\hat{d}_r - 1| &\le |\sqrt{1+b_r^{(n')}} - 1| + \tfrac{2|\Delta_1|+\Delta_1^2+|\Delta_2|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})} \\ &\le |b_r^{(n')}| + \tfrac{2|\Delta_1|+\Delta_1^2+|\Delta_2|}{\sqrt{1+\Delta_2}(1+\Delta_1+\sqrt{1+\Delta_2})} \\ &\lesssim_\gamma \tfrac{\|\Sigma\|^2}{g_r^2}\tfrac{\mathbf{r}(\Sigma)}{n'} + \tfrac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\tfrac{t}{m}}\Big)\sqrt{\tfrac{t}{m}}. \end{aligned} \tag{3.61}$$

and

$$\Big|\frac{1}{\hat{d}_r} - 1\Big| \lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n'} + \frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\frac{t}{m}}\Big)\sqrt{\frac{t}{m}}. \tag{3.62}$$

The key ingredient of the proof of Theorem 3.3.6 is the following lemma.

Lemma 3.5.2. *Suppose that, for some $\gamma \in (0,1)$, conditions (3.43) and (3.54) hold. Then, for all $t \ge 1$ with probability at least $1 - e^{-t}$*

$$\big|\langle \tilde{\theta}_r - \theta_r, u\rangle - \langle L_r(\hat{\Sigma}^{(1)} - \Sigma)\theta_r, u\rangle\big|$$

$$\lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\frac{t}{m}}\bigvee\frac{t}{m}\right)\sqrt{\frac{t}{m}}\|u\|. \tag{3.63}$$

Proof. We use the following simple representation:

$$\begin{aligned}\langle\tilde\theta_r-\theta_r,u\rangle &= \langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,u\rangle\\ +\left(\frac{1}{\hat d_r}-1\right)\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,u\rangle &+\left(\frac{\sqrt{1+b_r^{(n')}}}{\hat d_r}-1\right)\langle\theta_r,u\rangle\end{aligned} \tag{3.64}$$

that holds on the event E (where $\hat d_r\ge 1/2$). Using bounds (3.60) and (3.62) that both hold under assumption (3.57) on the event E as well as bound (3.33) (applied to $\hat\theta_r^{(1)}$ with $n=n'$), we obtain that with probability at least $1-2e^{-t}$

$$\begin{aligned}&\left|\langle\tilde\theta_r-\theta_r,u\rangle-\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,u\rangle\right|\\ \lesssim_\gamma\ &\frac{\|\Sigma\|^2}{g_r^2}\frac{\mathbf{r}(\Sigma)}{n'}\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n'}}\|u\|+\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}\frac{\|\Sigma\|}{g_r}\sqrt{\frac{t}{n'}}\|u\|\\ &+\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}\|u\|.\end{aligned}$$

It is easy to check that the last term in the right hand side is dominant yielding the simpler bound

$$\begin{aligned}&\left|\langle\tilde\theta_r-\theta_r,u\rangle-\langle\hat\theta_r^{(1)}-\sqrt{1+b_r^{(n')}}\theta_r,u\rangle\right|\\ &\lesssim_\gamma \frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}\|u\|\end{aligned} \tag{3.65}$$

that holds under assumption (3.57) with probability at least $1-e^{-t}$. Since the left hand side is bounded by $5\|u\|$, bound (3.65) also holds trivially when

$$\frac{\|\Sigma\|^2}{g_r^2}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{m}}\bigvee\sqrt{\frac{t}{m}}\right)\sqrt{\frac{t}{m}}>c_\gamma.$$

It remains to combine (3.65) with the bound (3.23) (applied to $\hat\theta_r^{(1)}$) to complete the proof. □

The following statement is an immediate consequence of Lemma 3.5.2 and Lemma 3.4.2. As always, we dropped the terms $\frac{t}{n'}$, $\frac{t}{m}$ from the bounds since the left-hand side is smaller that $3\|u\|$ and, for $t\ge n'$ or $t\ge m$ (the only cases when these terms might be needed), it is dominated by the expression with $\sqrt{\frac{t}{n'}}$, $\sqrt{\frac{t}{m}}$ only.

Corollary 3.5.3. *Suppose that, for some $\gamma\in(0,1)$, conditions (3.43) and (3.54) hold.*

Then, for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$\big|\langle \tilde{\theta}_r - \theta_r, u\rangle\big| \lesssim_\gamma \tfrac{\|\Sigma\|}{g_r}\sqrt{\tfrac{t}{n'}}\|u\| + \tfrac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\tfrac{\mathbf{r}(\Sigma)}{m}} \bigvee \sqrt{\tfrac{t}{m}}\Big)\sqrt{\tfrac{t}{m}}\|u\|. \tag{3.66}$$

Lemma 3.5.2 implies the following statement. This, in turn, implies Theorem 3.3.6.

Lemma 3.5.4. *Suppose that $m^2 \geq 2n\mathbf{r}(\Sigma)$ and conditions (3.43) and (3.54) hold for some $\gamma \in (0,1)$. For a given $u \in \mathbb{H}$, suppose that $\sigma_r(\Sigma; u) > 0$. Let $\alpha \geq 1$. Then the following bounds holds: for some constants $C, C_{\gamma,\alpha} > 0$,*

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\big|\mathbb{P}\big\{\frac{\sqrt{n}\langle\tilde{\theta}_r - \theta_r, u\rangle}{\sigma_r(\Sigma;u)} \leq x\big\} - \Phi(x)\big| \\
&\leq C(n')^{-1/2} + \frac{C_{\gamma,\alpha}}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{n\mathbf{r}(\Sigma)}{m^2}\log\frac{m^2}{n\mathbf{r}(\Sigma)}} \bigvee \sqrt{\frac{n\log^2\frac{m^2}{n\mathbf{r}(\Sigma)}}{m^2}}\Big)\|u\| + \Big(\frac{n\mathbf{r}(\Sigma)}{m^2}\Big)^\alpha.
\end{aligned} \tag{3.67}$$

Moreover, denote

$$\tau_1 := C_\gamma\Big(\frac{\|\Sigma\|}{g_r} \bigvee \frac{\|\Sigma\|^2}{g_r^2}\sqrt{\frac{n\mathbf{r}(\Sigma)}{m^2}}\Big)\|u\|$$

and

$$\tau_2 := C_\gamma\frac{\|\Sigma\|^2}{g_r^2}\sqrt{\frac{n}{m^2}}\|u\|.$$

Suppose that Assumptions 3.3.1 on the loss ℓ holds and $c_2\tau_2 \leq 1/4$. There exist constants $C, C_\gamma, C_{\gamma,\alpha} > 0$ such that

$$\begin{aligned}
&\Big|\mathbb{E}\ell\Big(\frac{\sqrt{n}\langle\tilde{\theta}_r - \theta_r, u\rangle}{\sigma_r(\Sigma;u)}\Big) - \mathbb{E}\ell(Z)\Big| \\
&\leq c_1 e^{c_2 A}\Big(C(n')^{-1/2} + \frac{C_{\gamma,\alpha}}{\sigma_r(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\Big(\sqrt{\frac{n\mathbf{r}(\Sigma)}{m^2}\log\frac{m^2}{n\mathbf{r}(\Sigma)}} \bigvee \sqrt{\frac{n\log^2\frac{m^2}{n\mathbf{r}(\Sigma)}}{m^2}}\Big)\|u\| + \Big(\frac{n\mathbf{r}(\Sigma)}{m^2}\Big)^\alpha\Big) \\
&+ 2e^{3/2}(2\pi)^{1/4}c_1 e^{c_2^2\tau_1^2}(e^{-A^2/2\tau_1^2} \vee e^{-A/2\tau_2}) + c_1 e^{c_2^2}e^{-A^2/4}.
\end{aligned} \tag{3.68}$$

Proof. The proof is similar to that of Lemma 3.4.8. To prove (3.67), we apply the first bound of Lemma 3.4.6 to the random variables

$$\xi := \frac{\sqrt{n}\langle\tilde{\theta}_r - \theta_r, u\rangle}{\sigma_r(\Sigma;u)}, \quad \eta := \frac{\langle L_r(\hat{\Sigma}^{(1)} - \Sigma)\theta_r, u\rangle}{\sigma_r(\Sigma;u)}.$$

and use the bound of Lemma 3.5.2 with $t = \alpha\log\big(\frac{m^2}{n\mathbf{r}(\Sigma)}\big)$ to control $\delta(\xi,\eta)$.

To prove the bound (3.68), observe that, by bound (3.66), for all $t \geq 1$ with probability at least $1 - e^{-t}$

$$|\xi| \leq \tau_1\sqrt{t} \vee \tau_2 t.$$

Under assumption $c_2\tau_2 \le 1/4$, bound (3.35) implies that

$$\mathbb{E}\ell^2(\xi) \le 2e\sqrt{2\pi}c_1^2 e^{2c_2^2\tau_1^2} + \frac{ec_1^2}{1-2c_2\tau_2} \le 4e\sqrt{2\pi}c_1^2 e^{2c_2^2\tau_1^2}.$$

Therefore,

$$\mathbb{E}\ell(\xi)I(|\xi| \ge A) \le \mathbb{E}^{1/2}\ell^2(\xi)\mathbb{P}^{1/2}\{|\xi| \ge A\} \le 2e^{3/2}(2\pi)^{1/4}c_1 e^{c_2^2\tau_1^2}(e^{-A^2/2\tau_1^2} \vee e^{-A/2\tau_2}).$$

It remains to repeat the rest of the proof of the second statement of Lemma 3.4.8. □

3.6 Proof of Corollary 3.3.7

The proof is based on a deterministic bound on $|\sigma_r^2(\tilde{\Sigma};u) - \sigma_r^2(\Sigma;u)|$ for a small perturbation $\tilde{\Sigma}$ of Σ provided by the following lemma.

Lemma 3.6.1. *Let $m_r = 1$. Denote $E := \tilde{\Sigma} - \Sigma$ and suppose that $\|E\| \le g_r/4$. Then*

$$|\sigma_r^2(\tilde{\Sigma};u) - \sigma_r^2(\Sigma;u)| \lesssim \frac{\|\Sigma\|^2}{g_r^2}\frac{\|E\|}{g_r}\|u\|^2. \tag{3.69}$$

and

$$\left|\frac{\sigma_r(\tilde{\Sigma};u)}{\sigma_r(\Sigma;u)} - 1\right| \lesssim \frac{1}{\sigma_r^2(\Sigma;u)}\frac{\|\Sigma\|^2}{g_r^2}\frac{\|E\|}{g_r}\|u\|^2. \tag{3.70}$$

Proof. We use the Riesz representation of the spectral projector $P_r(\tilde{\Sigma})$

$$P_r(\tilde{\Sigma}) = -\frac{1}{2\pi i}\oint_{\gamma_r} R_{\tilde{\Sigma}}(\eta)d\eta,$$

where $R_B(\eta) = (B - \eta I)^{-1}$ denotes the resolvent of operator B and γ_r is the circle in $\mathbb{C}$ with center μ_r and radius $g_r/2$ (and with counterclockwise orientation). Since $\|\tilde{E}\| \le \frac{g_r}{4}$ and $m_r = 1$, it is easy to see that there is only one eigenvalue $\mu_r(\tilde{\Sigma})$ of $\tilde{\Sigma}$ inside γ_r and that $\operatorname{dist}(\eta;\sigma(\tilde{\Sigma})) \ge \frac{g_r}{4}, \eta \in \gamma_r$. Note also that, for all $\eta \in \gamma_r$,

$$\|R_\Sigma(\eta)\| \le \frac{2}{g_r}, \quad \|R_{\tilde{\Sigma}}(\eta)\| \le \frac{4}{g_r} \tag{3.71}$$

and

$$\begin{aligned} R_{\tilde{\Sigma}}(\eta) - R_\Sigma(\eta) &= (\Sigma - \eta I + E)^{-1} - (\Sigma - \eta I)^{-1} \\ &= \left[(I + R_\Sigma(\eta)E)^{-1} - I\right] R_\Sigma(\eta). \end{aligned} \tag{3.72}$$

It follows that, for all $\eta \in \gamma_r$,

$$\|R_{\tilde{\Sigma}}(\eta) - R_\Sigma(\eta)\| \leq \frac{2}{g_r}\| \left(I + R_\Sigma(\eta)E\right)^{-1} - I\| \leq \frac{2}{g_r}\sum_{k=1}^{\infty}\|R_\Sigma(\eta)E\|^k \leq \frac{8\|E\|}{g_r^2}. \tag{3.73}$$

Denote $A(\Sigma) := \theta_r(\Sigma) \otimes u + u \otimes \theta_r(\Sigma)$, $B(\Sigma) := P_r(\Sigma) \otimes C_r(\Sigma) + C_r(\Sigma) \otimes P_r(\Sigma)$ and

$$D(\Sigma) := B(\Sigma)A(\Sigma) = \theta_r(\Sigma) \otimes C_r(\Sigma)u + C_r(\Sigma)u \otimes \theta_r(\Sigma).$$

We have

$$\begin{aligned}\frac{1}{2\pi i}\oint_{\gamma_r} R_\Sigma(\eta) \otimes R_\Sigma(\eta)d\eta &= \sum_{s,s'=1}^{\infty}\frac{1}{2\pi i}\oint_{\gamma_r}\frac{d\eta}{(\mu_s - \eta)(\mu_{s'} - \eta)}P_s \otimes P_{s'} \\ &= \sum_{s\neq r}\frac{1}{\mu_r - \mu_s}(P_r \otimes P_s + P_s \otimes P_r) \\ &= P_r(\Sigma) \otimes C_r(\Sigma) + C_r(\Sigma) \otimes P_r(\Sigma) = B(\Sigma). \end{aligned} \tag{3.74}$$

Hence, using (3.71) and (3.73), we derive the following bound for any bounded operator H :

$$\begin{aligned}&\left\|(B(\tilde{\Sigma}) - B(\Sigma))H\right\| \\ =&\left\|\frac{1}{2\pi i}\oint_{\gamma_r}\left[R_{\tilde{\Sigma}}(\eta)HR_{\tilde{\Sigma}}(\eta) - R_\Sigma(\eta)HR_\Sigma(\eta)\right]d\eta\right\| \\ =&\left\|\frac{1}{2\pi i}\oint_{\gamma_r}\left[(R_{\tilde{\Sigma}}(\eta) - R_\Sigma(\eta))HR_{\tilde{\Sigma}}(\eta) + R_\Sigma(\eta)H(R_{\tilde{\Sigma}}(\eta) - R_\Sigma(\eta))\right]d\eta\right\| \\ \leq&\frac{g_r}{2}\frac{8\|E\|}{g_r^2}\|H\|\left(\frac{4}{g_r} + \frac{2}{g_r}\right) \leq \frac{24\|E\|\|H\|}{g_r^2}. \end{aligned} \tag{3.75}$$

Note also that

$$\|A(\tilde{\Sigma})\| \leq 2\|u\|, \tag{3.76}$$

and, using the bound $\|C_r(\Sigma)\| \leq \frac{1}{g_r}$,

$$\|B(\Sigma)H\| \leq \|P_r(\Sigma)HC_r(\Sigma)\| + \|C_r(\Sigma)HP_r(\Sigma)\| \leq \frac{2}{g_r}\|H\|. \tag{3.77}$$

Finally, observe that, by standard perturbation bounds,

$$\begin{aligned}&\|A(\tilde{\Sigma}) - A(\Sigma)\| \leq 2\|\theta_r(\tilde{\Sigma}) - \theta_r(\Sigma)\|\|u\| \\ &\leq 2\|P_r(\tilde{\Sigma}) - P_r(\Sigma)\|_2\|u\| \leq 2\sqrt{2}\|P_r(\tilde{\Sigma}) - P_r(\Sigma)\|_2\|u\| \\ &\leq \frac{8\sqrt{2}\|E\|\|u\|}{g_r}. \end{aligned} \tag{3.78}$$

It follows from bounds (3.75), (3.76), (3.77) and (3.78) that

$$\begin{aligned}
&\|D(\tilde{\Sigma}) - D(\Sigma)\| \le \|(B(\tilde{\Sigma}) - B(\Sigma))A(\tilde{\Sigma})\| + \|B(\Sigma)(A(\tilde{\Sigma}) - A(\Sigma))\| \\
&\le \frac{24\|E\|\|A(\tilde{\Sigma})\|}{g_r^2} + \frac{2}{g_r}\|A(\tilde{\Sigma}) - A(\Sigma)\| \le \frac{48\|E\|\|u\|}{g_r^2} + \frac{2}{g_r}\frac{8\sqrt{2}\|E\|\|u\|}{g_r} \\
&\le \frac{80\|E\|\|u\|}{g_r^2}.
\end{aligned} \tag{3.79}$$

Now, recall that

$$\begin{aligned}
&\sigma_r^2(\Sigma; u) = \langle \Sigma\theta_r(\Sigma), \theta_r(\Sigma)\rangle\langle \Sigma C_r(\Sigma)u, C_r(\Sigma)u\rangle \\
&= \frac{1}{2}\Big\|\Sigma^{1/2}(\theta_r(\Sigma)\otimes C_r(\Sigma)u + C_r(\Sigma)u\otimes\theta_r)\Sigma^{1/2}\Big\|_2^2 \\
&= \frac{1}{2}\|\Sigma^{1/2}D(\Sigma)\Sigma^{1/2}\|_2^2 = \frac{1}{2}\mathrm{tr}(\Sigma^{1/2}D(\Sigma)\Sigma^{1/2}\Sigma^{1/2}D(\Sigma)\Sigma^{1/2}) \\
&= \frac{1}{2}\mathrm{tr}(\Sigma D(\Sigma)\Sigma D(\Sigma)).
\end{aligned} \tag{3.80}$$

Hence, by the duality between operator and nuclear norms and since $\mathrm{rank}(D(\Sigma)) \le 2, \mathrm{rank}(D(\tilde{\Sigma})) \le 2$, we have that

$$\begin{aligned}
|\sigma_r^2(\tilde{\Sigma}; u) - \sigma_r^2(\Sigma; u)| =& \frac{1}{2}\big|\,\mathrm{trace}(\tilde{\Sigma}D(\tilde{\Sigma})\tilde{\Sigma}D(\tilde{\Sigma})) - \mathrm{trace}(\Sigma D(\Sigma)\Sigma D(\Sigma))\big| \\
=& \frac{1}{2}\big|\,\mathrm{trace}((\tilde{\Sigma}-\Sigma)D(\tilde{\Sigma})\tilde{\Sigma}D(\tilde{\Sigma})) + \mathrm{trace}(\Sigma(D(\tilde{\Sigma}) - D(\Sigma))\tilde{\Sigma}D(\tilde{\Sigma})) \\
&+ \mathrm{trace}(\Sigma D(\Sigma)(\tilde{\Sigma}-\Sigma)D(\tilde{\Sigma})) + \mathrm{trace}(\Sigma D(\Sigma)\Sigma(D(\tilde{\Sigma}) - D(\Sigma)))\big| \\
&\le \frac{1}{2}\|\tilde{\Sigma}-\Sigma\|\big(\|D(\tilde{\Sigma})\tilde{\Sigma}D(\tilde{\Sigma})\|_1 + \|D(\tilde{\Sigma})\Sigma D(\Sigma)\|_1\big) \\
&+ \frac{1}{2}\|D(\tilde{\Sigma}) - D\|\big(\|\tilde{\Sigma}D(\tilde{\Sigma})\Sigma\|_1 + \|\Sigma D(\Sigma)\Sigma\|_1\big) \\
&\le \|\tilde{\Sigma}-\Sigma\|\big(\|D(\tilde{\Sigma})\tilde{\Sigma}D(\tilde{\Sigma})\| + \|D(\tilde{\Sigma})\Sigma D(\Sigma)\|\big) \\
&+ \|D(\tilde{\Sigma}) - D\|\big(\|\tilde{\Sigma}D(\tilde{\Sigma})\Sigma\| + \|\Sigma D(\Sigma)\Sigma\|\big).
\end{aligned} \tag{3.81}$$

It remains to observe that $\|C_r(\Sigma)\| \le \frac{1}{g_r}$, $\|C_r(\tilde{\Sigma})\| \le \frac{2}{g_r}$ and that

$$\|D(\Sigma)\| \le 2\|(\theta_r(\Sigma)\otimes C_r(\Sigma)u)\| \le 2\|C_r(\Sigma)\|\|u\| \le \frac{2\|u\|}{g_r},$$

$$\|D(\tilde{\Sigma})\| \le 2\|(\theta_r(\tilde{\Sigma})\otimes C_r(\tilde{\Sigma})u)\| \le 2\|C_r(\tilde{\Sigma})\|\|u\| \le \frac{4\|u\|}{g_r}$$

and

$$\|\tilde{\Sigma}\| \le \|\Sigma\| + \|E\| \le \|\Sigma\| + \frac{g_r}{4} \le 2\|\Sigma\|,$$

implying the bounds

$$\|D(\tilde\Sigma)\tilde\Sigma D(\tilde\Sigma)\| \le \frac{32\|\Sigma\|\|u\|^2}{g_r^2}, \quad \|D(\tilde\Sigma)\Sigma D(\Sigma)\| \le \frac{8\|\Sigma\|\|u\|^2}{g_r^2},$$
$$\|\tilde\Sigma D(\tilde\Sigma)\Sigma\| \le \frac{8\|\Sigma\|^2\|u\|}{g_r} \quad \text{and} \quad \|\Sigma D(\Sigma)\Sigma\| \le \frac{2\|\Sigma\|^2\|u\|}{g_r}. \tag{3.82}$$

Bound (3.69) now follows from (3.81), (3.79) and (3.82). Bound (3.70) follows from (3.69). □

It remains to apply this lemma to $\tilde\Sigma = \hat\Sigma$ and to use standard bounds on $\|\hat\Sigma - \Sigma\|$ to obtain the following inequalities.

Proposition 3.6.2. *Suppose that condition (3.9) holds for some $\gamma \in (0,1)$. Then, there exists a constant $c_\gamma > 0$ such that for all $t \in [1, c_\gamma n]$ with probability at least $1 - e^{-t}$*

$$|\sigma_r^2(\hat\Sigma; u) - \sigma_r^2(\Sigma; u)| \lesssim \frac{\|\Sigma\|^3}{g_r^3}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\|u\|^2 \tag{3.83}$$

and

$$\left|\frac{\sigma_r(\hat\Sigma; u)}{\sigma_r(\Sigma; u)} - 1\right| \lesssim \frac{1}{\sigma_r^2(\Sigma; u)}\frac{\|\Sigma\|^3}{g_r^3}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\|u\|^2. \tag{3.84}$$

The consistency of estimator $\sigma_r(\hat\Sigma; u)$ immediately follows:

Proposition 3.6.3. *Suppose $\mathfrak{r}_n > 1$, $\mathfrak{r}_n = o(n)$ as $n \to \infty$. For any sequence $\delta_n \to 0$ such that $\frac{\mathfrak{r}_n}{n} = o(\delta_n^2)$ as $n \to \infty$,*

$$\sup_{\Sigma \in \mathcal{S}^{(r)}(\mathfrak{r}_n, a, \sigma_0, u)} \mathbb{P}_\Sigma\Big\{\Big|\tfrac{\sigma_r(\hat\Sigma;u)}{\sigma_r(\Sigma;u)} - 1\Big| \ge \delta_n\Big\} \to 0 \text{ as } n \to \infty.$$

Corollary 3.3.7 can be easily proved using the first statement of Theorem 3.3.6, Proposition 3.6.3 and Lemma 3.4.6.

3.7 Proof of Theorem 3.3.4

Note that the set $\mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ is open in nuclear norm topology. This easily follows from the continuity of functions $\Sigma \mapsto \|\Sigma\|$, $\Sigma \mapsto \bar g_r(\Sigma)$ and $\Sigma \mapsto \sigma_r^2(\Sigma; u)$ with respect to the operator norm (for the last function, see Lemma 3.6.1) and, as a consequence, with respect to the nuclear norm, and of the functions $\Sigma \mapsto \mathrm{tr}(\Sigma)$ and $\Sigma \mapsto \mathbf{r}(\Sigma)$ with respect to the nuclear norm.

Let $\Sigma = \sum_{s=1}^\infty \mu_s P_s \in \mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$. Without loss of generality, assume that Σ is of finite rank. Otherwise, consider $\Sigma_N := \sum_{s=1}^N \mu_s P_s$. Clearly,

$$\mathbf{r}(\Sigma_N) \le \mathbf{r}(\Sigma) < \mathfrak{r}$$

and, for all $N > r$,

$$\frac{\|\Sigma_N\|}{\bar{g}_r(\Sigma_N)} = \frac{\|\Sigma\|}{\bar{g}_r(\Sigma)} < a.$$

Moreover, since $\|\Sigma_N - \Sigma\| \to 0$ as $N \to \infty$, we also have that $\sigma_r^2(\Sigma_N; u) \to \sigma_r^2(\Sigma; u)$ as $N \to \infty$, implying that $\sigma_r^2(\Sigma_N; u) > \sigma_0^2$ for all large enough N. Thus, $\Sigma_N \in \mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ for a sufficiently large N and we can replace Σ by Σ_N. Assuming that $\operatorname{rank}(\Sigma) < \infty$, let $L := \operatorname{Im}(\Sigma)$. We can now restrict Σ to an operator acting from L to L, which is non-singular. In what follows, all the covariance operators from the class $\mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ that are of interest to us will have L as an image and could be viewed as operators from L to L. For simplicity, we just assume that $\mathbb{H} = L$ is a finite-dimensional space. For a fixed Σ, consider the following parametric family of perturbations of Σ :

$$\Sigma_t := \Sigma + \frac{tH}{\sqrt{n}}, |t| \leq c,$$

where H is a self-adjoint operator and $c > 0$ is a constant. Denote

$$\mathcal{S}_{\Sigma,c} := \{\Sigma_t : t \in [-c, c]\}.$$

Since the set $\mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ is open in nuclear norm topology, there exists $\delta > 0$ such that the condition

$$\frac{c\|H\|_1}{\sqrt{n}} < \delta, \tag{3.85}$$

implies that $\mathcal{S}_{\Sigma,c} \subset \mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$. Moreover, we will assume that

$$\delta < \|\Sigma^{-1}\|^{-1} \tag{3.86}$$

and

$$\delta < \frac{1}{4}\bar{g}_r(\Sigma). \tag{3.87}$$

Under these assumptions and condition (3.85), Σ_t is a small enough perturbation of Σ so that Σ_t is non-singular and we can define in a standard way the one-dimensional spectral projection operator $P_t := P_r(\Sigma_t) = \theta_t \otimes \theta_t$, where $\theta_t = \theta_r(\Sigma_t)$ is the corresponding unit eigenvector as well as operators $C_t := C_r(\Sigma_t)$ and

$$L_t(H) := L_r(\Sigma_t)(H) = P_t H C_t + C_t H P_t.$$

It is easy to see that (for a given $c > 0$ and large enough n so that the perturbation is small) one can choose $t \mapsto \theta_t$ in such a way that $\langle \theta_t, \theta_{t'} \rangle \geq 0, t, t' \in [-c, c]$. Based on these definitions, we also define the functions $g(t) := \langle \theta_t, u \rangle$ and $\sigma^2(t) := \sigma_r^2(\Sigma_t; u)$. Concerning the function g, we need the following lemma.

Lemma 3.7.1. *The function g is continuously differentiable in the interval $[-c, c]$ and*

the following statements hold:

i) $g'(t) = \frac{1}{\sqrt{n}}\langle L_t(H)\theta_t, u\rangle, t \in [-c, c]$.

ii) $|g'(t) - g'(0)| \lesssim \frac{|t|\|H\|^2}{g_r^2 n}\|u\|, t \in [-c, c]$.

Proof. Let $\delta \in (-1, 1)$. Similarly to (3.24) (see also (6.6) in [KL16]),

$$\begin{aligned} g(t+\delta) - g(t) &= \langle \theta_{t+\delta} - \theta_t, u\rangle \\ &= \frac{\langle (P_{t+\delta} - P_t)\theta_t, u\rangle - (\sqrt{1 + \langle (P_{t+\delta} - P_t)\theta_t, \theta_t\rangle} - 1)\langle \theta_t, u\rangle}{\sqrt{1 + \langle (P_{t+\delta} - P_t)\theta_t, \theta_t\rangle}}. \end{aligned} \tag{3.88}$$

Applying the first order perturbation expansion (similar to (3.8)) to the spectral projections $P_t, P_{t+\delta}$, we obtain that

$$P_{t+\delta} - P_t = L_t(\delta H/\sqrt{n}) + S_t(\delta H/\sqrt{n}) \tag{3.89}$$

with the remainder term satisfying the bound

$$\|S_t(\delta H/\sqrt{n})\| \lesssim \frac{\delta^2\|H\|^2}{g_r^2 n} = O(\delta^2). \tag{3.90}$$

Moreover, since $C_t\theta_t = 0$,

$$\langle L_t(\delta H/\sqrt{n})\theta_t, \theta_t\rangle = \frac{1}{\sqrt{n}}\langle (P_t H C_t + C_t H P_t)\theta_t, \theta_t\rangle = 0 \tag{3.91}$$

and therefore we have that

$$|\langle (P_{t+\delta} - P_t)\theta_t, \theta_t\rangle| \lesssim \frac{\delta^2\|H\|^2}{g_r^2 n} = O(\delta^2). \tag{3.92}$$

Hence, using again (3.88), (3.90) and (3.92), we have that

$$\frac{g(t+\delta) - g(t)}{\delta} = \frac{1}{\sqrt{n}}\frac{\langle L_t(H)\theta_t, u\rangle}{1 + O(\delta)} + O(\delta). \tag{3.93}$$

Passing to the limit as $\delta \to 0$ implies the first assertion.

We now prove the second claim. First note that

$$\begin{aligned} |g'(t) - g'(0)| &= |\langle L_t(H/\sqrt{n})\theta_t - L_0(H/\sqrt{n})\theta_0, u\rangle| \\ &\le |\langle (L_t(H/\sqrt{n}) - L_0(H/\sqrt{n}))\theta_t, u\rangle| + |\langle L_0(H/\sqrt{n})(\theta_t - \theta_0), u\rangle| \\ &\le \|L_t(H/\sqrt{n}) - L_0(H/\sqrt{n})\|\|u\| + \|L_0(H/\sqrt{n})\|\|\theta_t - \theta_0\|\|u\|. \end{aligned} \tag{3.94}$$

Also,

$$L_t(H/\sqrt{n}) = -\frac{1}{2\pi i}\oint_{\gamma_r} R_{\Sigma_t}(\eta)\frac{H}{\sqrt{n}}R_{\Sigma_t}(\eta)d\eta, \tag{3.95}$$

where γ_r is the circle of radius $g_r/2$ with the center at μ_r and with counterclockwise orientation. Therefore, by a standard argument already used in the proof of Lemma 3.6.1,

$$\|L_t(H/\sqrt{n}) - L_0(H/\sqrt{n})\| \le \frac{g_r}{2}\sup_{\eta\in\gamma_r}\|R_{\Sigma_t}(\eta) - R_\Sigma(\eta)\|(\|R_\Sigma(\eta)\| + \|R_{\Sigma_t}(\eta)\|)\frac{\|H\|}{\sqrt{n}}. \tag{3.96}$$

By (3.71) and (3.73), we have

$$\|R_\Sigma(\eta)\| \le \frac{2}{g_r}, \quad \|R_{\Sigma_t}(\eta)\| \le \frac{4}{g_r}$$

and

$$\|R_{\Sigma_t}(\eta) - R_\Sigma(\eta)\| \le \frac{8}{g_r^2}\frac{|t|\|H\|}{\sqrt{n}}.$$

Therefore, it follows from (3.96) that

$$\|L_t(H/\sqrt{n}) - L_0(H/\sqrt{n})\| \le \frac{24|t|\|H\|^2}{g_r^2 n}. \tag{3.97}$$

It remains to observe that

$$\|L_0(H)\| = \|P_r H C_r + C_r H P_r\| \le \frac{2\|H\|}{g_r}$$

and

$$\|\theta_t - \theta_0\| \le \|P_t - P_0\|_2 \le \frac{4\sqrt{2}|t|\|H\|}{g_r\sqrt{n}}$$

(where we also used the fact that $\mathrm{rank}(P_t - P_0) \le 2$ and $\|P_t - P_0\|_2 \le \sqrt{2}\|P_t - P_0\|$). This implies the bound

$$\|L_0(H/\sqrt{n})\|\|\theta_t - \theta_0\|\|u\| \le \frac{8\sqrt{2}|t|\|H\|^2}{g_r^2 n}\|u\|. \tag{3.98}$$

The second assertion follows from the bounds (3.94), (3.97) and (3.98).

The continuity of the derivative $g'(t)$ easily follows from the continuity of the functions $t \mapsto \theta_t$ and $t \mapsto L_t(H/\sqrt{n})$ (which could be proved using representation (3.95)).

□

We will study the following estimation problem. Let Σ be fixed and let $X_1, \dots, X_n$ be *i.i.d.* random variables in $\mathbb{H}$ sampled from $N(0; \Sigma_t)$, $|t| \le c$, t being an unknown parameter.

The goal is to estimate the function $g(t)$ based on the observations $X_1, \dots, X_n$. We will use van Trees inequality to obtain a minimax lower bound on the risk of estimation of $g(t)$ with respect to quadratic loss. To this end, let π be a smooth probability density on $[-1,1]$, satisfying the boundary conditions $\pi(-1) = \pi(1) = 0$ as well the condition $J_\pi := \int_{-1}^{1} \frac{\pi'(s)^2}{\pi(s)} ds < +\infty$. Let $\pi_c(t) := \frac{1}{c}\pi\big(\frac{t}{c}\big), t \in [-c,c]$ be a prior on $[-c,c]$. Then (see e.g. [GL95]), for any estimator $T_n = T_n(X_1, \dots, X_n)$ of $g(t)$ the following bound holds

$$\begin{aligned} \sup_{|t|\le c} n\mathbb{E}_t(T_n - g(t))^2 &\ge n\int_{-c}^{c} \mathbb{E}_t(T_n - g(t))^2 \pi_c(t)dt \\ &\ge \frac{n\big(\int_{-c}^{c} g'(t)\pi_c(t)dt\big)^2}{\int_{-c}^{c} \mathbb{I}_n(t)\pi_c(t)dt + J_{\pi_c}}, \end{aligned} \tag{3.99}$$

where $\mathbb{I}_n(t) = n\mathbb{I}(t)$ denotes the Fisher information for the model

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} N(0, \Sigma_t),$$

$t \in [-c,c]$. Let $\mathbb{I}(t) := \mathbb{I}_1(t)$. It is well known that the Fisher information for the model $X \sim N(0;\Sigma)$ with non-singular covariance matrix Σ is $\mathbb{I}(\Sigma) = \frac{1}{2}(\Sigma^{-1} \otimes \Sigma^{-1})$ (see, e.g., [Eat83]). Thus,

$$\mathbb{I}_n(t) = n\mathbb{I}(t) = n\Big\langle \mathbb{I}(\Sigma_t)\frac{d\Sigma_t}{dt}, \frac{d\Sigma_t}{dt}\Big\rangle = \frac{n}{2}\Big\langle (\Sigma_t^{-1}\otimes\Sigma_t^{-1})\frac{H}{\sqrt{n}}, \frac{H}{\sqrt{n}}\Big\rangle$$

$$= \frac{1}{2}\langle \Sigma_t^{-1} H \Sigma_t^{-1}, H\rangle = \frac{1}{2}\mathrm{tr}(\Sigma_t^{-1} H \Sigma_t^{-1} H).$$

We will now bound the numerator of the expression in the right hand side of inequality (3.99) from below and its denominator from above.

Bound on the numerator. We use Lemma 3.7.1 to obtain that for some constant $B_1 > 0$

$$\begin{aligned} \Big(\int_{-c}^{c} g'(t)\pi_c(t)dt\Big)^2 &= \Big(\int_{-c}^{c} [g'(0) + (g'(t) - g'(0))]\pi(t/c)dt/c\Big)^2 \\ &\ge g'(0)^2 + 2g'(0)\int_{-c}^{c} (g'(t) - g'(0))\pi(t/c)dt/c \\ &\ge g'(0)^2 - 2|g'(0)|\int_{-c}^{c} |g'(t) - g'(0)|\pi(t/c)dt/c \\ &\ge g'(0)^2 - B_1 c|g'(0)|\int_{-1}^{1} |t|\pi(t)dt \frac{\|H\|^2}{g_r^2 n}\|u\| \\ &= g'(0)^2 - B_1 c|g'(0)|\frac{\|H\|^2}{g_r^2 n}\|u\| \\ &= \frac{\langle L_r(H)\theta_r, u\rangle^2}{n} - |\langle L_r(H)\theta_r, u\rangle|\frac{B_1 c\|H\|^2}{g_r^2 n^{3/2}}\|u\|. \end{aligned} \tag{3.100}$$

Bound on the denominator. First note that, by a simple computation,

$$J_{\pi_c} = J_\pi / c^2. \tag{3.101}$$

Then, we need to bound $\mathbb{I}_n(t) = \frac{1}{2}\mathrm{tr}(\Sigma_t^{-1} H \Sigma_t^{-1} H)$ in terms of $\mathbb{I}_n(0) = \frac{1}{2}\mathrm{tr}(\Sigma^{-1} H \Sigma^{-1} H)$. Assume that

$$\frac{c\|\Sigma^{-1}H\|}{\sqrt{n}} \leq \frac{1}{2}. \tag{3.102}$$

Arguing as in the proof of Lemma 3.6.1, we easily get that

$$\Sigma_t^{-1} = \Sigma^{-1} + \underbrace{\left[\left(I + \frac{t\Sigma^{-1}H}{\sqrt{n}}\right)^{-1} - I\right]}_{=:D} \Sigma^{-1}, \tag{3.103}$$

where

$$\|D\| \leq 2|t|\frac{\|\Sigma^{-1}H\|}{\sqrt{n}} \leq 1.$$

Furthermore, note that

$$\mathrm{trace}\left(\Sigma_t^{-1} H \Sigma_t^{-1} H\right) = \mathrm{trace}(\Sigma^{-1} H \Sigma^{-1} H) + 2\,\mathrm{trace}(D\Sigma^{-1} H \Sigma^{-1} H) + \mathrm{trace}(D\Sigma^{-1} H D \Sigma^{-1} H).$$

and thus we have that

$$\begin{aligned}\mathbb{I}_n(t) &\leq \mathbb{I}_n(0) + \|D\|\|\Sigma^{-1}H\Sigma^{-1}H\|_1 + \frac{\|D\Sigma^{-1}H\|_2\|H\Sigma^{-1}D\|_2}{2} \\ &\leq \mathbb{I}_n(0) + \left(\|D\| + \frac{\|D\|^2}{2}\right)\|\Sigma^{-1}H\|_2^2 \leq \mathbb{I}_n(0) + 3\frac{|t|\|\Sigma^{-1}H\|_2^3}{\sqrt{n}}.\end{aligned} \tag{3.104}$$

Using (3.104), we obtain the following bound:

$$\begin{aligned}\int_{-c}^{c} \mathbb{I}_n(t)\pi_c(t)dt &\leq \mathbb{I}_n(0) + 3\frac{\|\Sigma^{-1}H\|_2^3}{\sqrt{n}}\int_{-c}^{c}|t|\pi(t/c)dt/c \\ &\leq \mathbb{I}_n(0) + \frac{3c\|\Sigma^{-1}H\|_2^3}{\sqrt{n}}.\end{aligned} \tag{3.105}$$

Substituting (3.100), (3.105) and (3.101) into van Trees inequality (3.99) and taking into account that

$$\mathbb{I}_n(0) = \frac{1}{2}\mathrm{tr}(\Sigma^{-1}H\Sigma^{-1}H) = \frac{1}{2}\|\Sigma^{-1/2}H\Sigma^{-1/2}\|_2^2$$

and

$$\langle L_r(H)\theta_r, u\rangle = \langle (P_r H C_r + C_r H P_r)\theta_r, u\rangle = \langle H\theta_r, C_r u\rangle$$

$$= \frac{1}{2}\langle H, \theta_r \otimes C_r u + C_r u \otimes \theta_r\rangle = \langle \Sigma^{-1/2}H\Sigma^{-1/2}, \Sigma^{-1/2}B\Sigma^{-1/2}\rangle,$$

where

$$B := \frac{1}{2}(\Sigma\theta_r \otimes \Sigma C_r u + \Sigma C_r u \otimes \Sigma\theta_r),$$

we obtain that

$$\begin{aligned}
&\sup_{|t|\le c} n\mathbb{E}_t(T_n - g(t))^2 \\
&\ge \frac{\langle \Sigma^{-1/2}H\Sigma^{-1/2}, \Sigma^{-1/2}B\Sigma^{-1/2}\rangle^2 - |\langle \Sigma^{-1/2}H\Sigma^{-1/2}, \Sigma^{-1/2}B\Sigma^{-1/2}\rangle| \frac{B_1 c\|H\|^2}{g_r^2\sqrt{n}}\|u\|}{\frac{1}{2}\|\Sigma^{-1/2}H\Sigma^{-1/2}\|_2^2 + \frac{3c\|\Sigma^{-1}H\|_2^3}{\sqrt{n}} + J_\pi/c^2}.
\end{aligned} \tag{3.106}$$

In what follows, we set $H := B$. Note that with this choice of H

$$2\|\Sigma^{-1/2}B\Sigma^{-1/2}\|_2^2 = \frac{1}{2}\|\Sigma^{1/2}\theta_r \otimes \Sigma^{1/2}C_r u + \Sigma^{1/2}C_r u \otimes \Sigma^{1/2}\theta_r\|_2^2$$

$$= \frac{1}{2}\Big(\|\Sigma^{1/2}\theta_r \otimes \Sigma^{1/2}C_r u\|_2^2 + \|\Sigma^{1/2}C_r u \otimes \Sigma^{1/2}\theta_r\|_2^2\Big) = \|\Sigma^{1/2}\theta_r\|^2\|\Sigma^{1/2}C_r u\|^2 = \sigma_r^2(\Sigma; u).$$

Also, by a simple computation (using that $\operatorname{rank}(B) = 2$), we have that

$$\|B\| \le \|B\|_2 \le \frac{1}{\sqrt{2}}\frac{\|\Sigma\|^2}{g_r}\|u\|, \quad \|B\|_1 \le \frac{\|\Sigma\|^2}{g_r}\|u\| \tag{3.107}$$

and that

$$\|\Sigma^{-1}B\| \le \|\Sigma^{-1}B\|_2 \le \frac{1}{\sqrt{2}}\frac{\|\Sigma\|}{g_r}\|u\|. \tag{3.108}$$

These bounds imply that, for any given $c > 0$ and for all n large enough, $H = B$ satisfies condition (3.85) for a small enough δ such that $\mathcal{S}_{\Sigma,c} \subset \mathring{\mathcal{S}}^{(r)}(\mathfrak{r}, a, \sigma_0, u)$ and conditions (3.86), (3.87) hold. Also, $H = B$ satisfies condition (3.102) (for any given $c > 0$ and all large enough n).

For $H = B$, inequality (3.106) becomes

$$\begin{aligned}
&\sup_{|t|\le c} n\mathbb{E}_t(T_n - g(t))^2 \\
&\ge \frac{\|\Sigma^{-1/2}B\Sigma^{-1/2}\|_2^4 - \|\Sigma^{-1/2}B\Sigma^{-1/2}\|_2^2 \frac{B_1 c\|B\|^2}{g_r^2\sqrt{n}}\|u\|}{\frac{1}{2}\|\Sigma^{-1/2}B\Sigma^{-1/2}\|_2^2 + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2} \\
&\ge \sigma_r^2(\Sigma; u)\left(1 - \frac{\frac{B_1 c\|B\|^2}{2g_r^2\sqrt{n}}\|u\| + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}{\frac{1}{4}\sigma_r^2(\Sigma; u) + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}\right).
\end{aligned} \tag{3.109}$$

It remains to replace $\sigma_r^2(\Sigma; u)$ with $\sigma^2(t) = \sigma_r^2(\Sigma_t; u)$. To this end, we use the bound (3.69)

to obtain that for some constant $D_1 > 0$

$$\sup_{t\in[-c,c]} \frac{\sigma^2(t)}{\sigma_r^2(\Sigma;u)} \le 1 + \frac{D_1}{\sigma_r^2(\Sigma;u)} \frac{\|\Sigma\|^2}{g_r^3} \frac{c\|B\|}{\sqrt{n}} \|u\|^2. \tag{3.110}$$

It follows from (3.109) that

$$\sup_{t\in[-c,c]} \frac{\sigma^2(t)}{\sigma_r^2(\Sigma;u)} \sup_{|t|\le c} \frac{n\mathbb{E}_t(T_n - g(t))^2}{\sigma^2(t)} \ge 1 - \frac{\frac{B_1 c\|B\|^2}{2g_r^2\sqrt{n}}\|u\| + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}{\frac{1}{4}\sigma_r^2(\Sigma;u) + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}. \tag{3.111}$$

Suppose

$$\frac{D_1}{\sigma_r^2(\Sigma;u)} \frac{\|\Sigma\|^4}{g_r^4} \frac{c}{\sqrt{n}} \|u\|^3 \le 1, \tag{3.112}$$

which holds for any given $c > 0$ and all large enough n and which, in view of bounds (3.107), implies that

$$\frac{D_1}{\sigma_r^2(\Sigma;u)} \frac{\|\Sigma\|^2}{g_r^3} \frac{c\|B\|}{\sqrt{n}} \|u\|^2 \le 1.$$

Under condition (3.112), bounds (3.111) and (3.110) (and also bounds (3.107) and (3.108)) imply that

$$\begin{aligned}
&\sup_{\Sigma\in\mathring{\mathcal{S}}^{(r)}(\mathfrak{r},a,\sigma_0,u)} \frac{\mathbb{E}_\Sigma(T_n - \langle\theta_r(\Sigma),u\rangle)^2}{\sigma_r^2(\Sigma;u)} \ge \sup_{|t|\le c} \frac{n\mathbb{E}_t(T_n - g(t))^2}{\sigma^2(t)} \\
&\ge \left(1 - \frac{\frac{B_1 c\|B\|^2}{2g_r^2\sqrt{n}}\|u\| + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}{\frac{1}{4}\sigma_r^2(\Sigma;u) + \frac{3c\|\Sigma^{-1}B\|_2^3}{\sqrt{n}} + J_\pi/c^2}\right)\left(1 - \frac{D_1}{\sigma_r^2(\Sigma;u)} \frac{\|\Sigma\|^2}{g_r^3} \frac{c\|B\|}{\sqrt{n}} \|u\|^2\right) \\
&\ge \left(1 - \frac{B_1 a^4\|u\|^3\frac{c}{\sqrt{n}} + 3a^3\|u\|^3\frac{c}{\sqrt{n}} + J_\pi/c^2}{\frac{\sigma_0^2}{4} + 3a^3\|u\|^3\frac{c}{\sqrt{n}} + J_\pi/c^2}\right)\left(1 - \frac{D_1}{\sigma_0^2} a^4\|u\|^3 \frac{c}{\sqrt{n}}\right).
\end{aligned} \tag{3.113}$$

It remains to pass to the limit in inequality (3.113) first as $n \to \infty$ and then as $c \to \infty$ to complete the proof.

A local version of the theorem easily follows from the above arguments since, for all $\varepsilon > 0, c > 0$ and for all large enough n, $\mathcal{S}_{\Sigma_0,c} \subset \{\Sigma : \|\Sigma - \Sigma_0\|_1 \le \varepsilon\}$.

Chapter 4

Spectral thresholding for the estimation of Markov chain transition operators

4.1 Introduction

We consider an aperiodic and irreducible Markov chain $(X_i)_{i\in\mathbb{N}}$ with the d-dimensional torus $\mathbb{T}^d$ as state space. The dynamics of this chain are described by its transition operator,

$$Pf(x) = \mathbb{E}[f(X_1)|X_0 = x] = \int_{\mathbb{T}^d} f(y)p(x,y)\mathrm{d}y,$$

where $f \in L^2 = L^2(\mathbb{T}^d)$. We are interested in nonparametric estimation of the transition density $p(\cdot,\cdot)$ and thus the transition operator P, too.
Nonparametric estimation of p when assuming smoothness of p has been thoroughly studied, e.g. [AL11, Bir13, Clé00, Lac07, Sar14]. If $p \in H^s$, where H^s denotes the L^2-Sobolev space of smoothness s, the L^2-minimax rates for estimating p are

$$n^{-\frac{s}{2s+2d}}.$$

Here we use the additional information provided by assuming that P has an approximately *low rank* structure to improve these rates. Precisely, since P is a compact operator on $L^2(\mu)$ it has functional singular value decomposition

$$Pf = \sum_{k\geq 0} \lambda_k \langle u_k, f\rangle_\mu v_k \qquad f \in L^2(\mu),$$

and we assume that the singular values λ_k decay exponentially fast, in the sense that for constants $c, C > 0$.

$$\lambda_k \leq C \exp\left(-ck^{\frac{2}{d}}\right).$$

This assumption is motivated by discrete, low frequency observations of periodised, reversible diffusion processes for which it is fulfilled by virtue of Weyl's law [Gar53, Hör79, Ivr00, Ivr16, Wey11]. Indeed, for a 1-periodic Lipschitz continuous vector field $b(x) = (b_1(x), \ldots, b_d(x))$ and a scalar 1-periodic $\sigma(x)$ define the multi-dimensional diffusion process

$$dY_t = b(Y_t)dt + \sigma(Y_t)dW_t, \qquad t \geq 0,$$

and consider its periodised version

$$X_t = Y_t \quad \text{modulo} \quad \mathbb{Z}^d, \quad t \geq 0.$$

Then $P = P_1$ is one instance of the Feller semigroup $(P_t)_{t\in\mathbb{R}_+}$ with infinitesimal generator $L : H^2 \to L^2$, and one obtains that $P = \exp(L)$ where L is given by

$$L = \frac{\sigma^2(x)}{2} \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} + \sum_{i=1}^{d} b_i(x) \frac{\partial}{\partial x_i}.$$

L is an elliptic operator and moreover, since the diffusion is assumed to be reversible, L is self-adjoint with respect to the invariant measure μ. Hence, Weyl's law[Ivr00] applies and states that its k-th eigenvalue is of order $-k^{\frac{2}{d}}$. This implies the exponential decay of the eigenvalues, and thus the singular values, of P.

Such a decay of the singular values is also observed empirically in applications such as molecular dynamics (see e.g. [RZMC11]). This has prompted practitioners and applied mathematicians to estimate only the first few eigenpairs of P and discard the rest in their analysis [CSP+07, CKL+08, KWNS18, RZMC11, Sch98, SMP14, SHWP15]. However, often no theoretical guarantees are provided and it is not clear whether their procedures are optimal from a statistical point of view.

Low rank assumptions for Markov chains have only recently began to be considered in the statistical literature, primarily in the finite state case [LWZ18, ZW18]. In these works it is assumed that the transition matrix has a low rank structure and they show nearly optimal rates for their algorithms. Moreover, Zhang and Wang [ZW18] extend their result to continuous state space Markov chains. By contrast, they assume that P has fixed (constant) rank whereas we assume decay of the eigenvalues. This leads to a more difficult analysis in our setting as one has to take bias due to discarding eigenvalues into account.

We investigate a modified version of one popular method from molecular dynamics for the estimation of P, where the number of eigenpairs kept is chosen in a data driven way. Considering a Galerkin-type estimator [GHR04, Sch98, SMP14] we use techniques from low rank matrix estimation [CP11, Klo14, KLT11]. Particularly we show that hard thresholding eigenvalues yields minimax optimal L^2-rates

$$n^{-\frac{s}{2s+d}} \log(n)^{\frac{d}{2}\frac{s}{2s+d}}$$

over the class of Markov chains with exponentially decaying eigenvalues. This improves the dependance on the dimension d from $2d$ to almost d compared to the case without eigenvalue decay. Moreover, our analysis reveals that our algorithm keeps at most $C\log(n)^{\frac{d}{2}}$ eigenpairs of the estimated transition operator, thus justifying the commonly used approach to discard most of them. Simulations complement our theoretical results and show the improved performance when thresholding eigenvalues.

4.2 Main results

4.2.1 Basic Notation

Let $\mathbb{T}^d$ denote the d-dimensional torus, isomorphic to the unit cube $[0,1]^d$ when opposite points are identified, equipped with Lebesgue measure λ. By $L^2 = L^2(\mathbb{T}^d, \lambda)$ we denote the space of square integrable functions (with respect to λ) on $\mathbb{T}^d$ equipped with Euclidean inner product $\langle \cdot, \cdot \rangle$ and corresponding L^2 norm $\|\cdot\|_{L^2}$. We also denote the Euclidean inner product for any finite dimensional vector space by $\langle \cdot, \cdot \rangle$ and the corresponding norm by $\|\cdot\|_2$. For any probability measure μ on $\mathbb{T}^d$ if μ has a density with respect to the Lebesgue measure, we denote it in slight abuse of notation by μ, too. Moreover, when considering functions in $L^2(\mu) = L^2(\mathbb{T}^d, \mu)$, we use the canonical scalar product and denote it by $\langle \cdot, \cdot \rangle_\mu$ with corresponding norm $\|\cdot\|_{L^2(\mu)}$. $\|\cdot\|_{L^\infty}$ denotes the L^∞ norm. $\|\cdot\|_F$ denotes the Hilbert–Schmidt (Frobenius) norms of operators on L^2, while $\|\cdot\|_\infty$ denotes the spectral norm for the λ scalar product, respectively.
For $s \in \mathbb{N}$ we define the Sobolev space of smoothness s as

$$H^s := \{f \in L^2 : \|f\|_{H^s} := \sum_{|i| \leq s} \|D^i f\|_{L^2} < \infty\}.$$

For $s \notin \mathbb{N}$, H^s is defined through interpolation or equivalently through Fourier methods (see Chapter I.9 in [LM72] or Section 7 in [AF03]). For $s > 0$ we will also use the Hölder spaces C^s equipped with Hölder norm $\|\cdot\|_{C^s}$. We also employ the same notation for vector fields $f = (f_1, \ldots, f_d)$. For example $f \in C^s$ means that $\|f\|_{C^s} := \sum_i \|f_i\|_{C^s} < \infty$. We will sometimes use the notation $a \lesssim b$, meaning that $a \leq Cb$ for some universal constant $C > 0$ which does not depend on n.

4.2.2 Assumptions on the model

We assume that we observe a Markov chain $(X_i)_{0\leq i\leq n}$ with state space $\mathbb{T}^d$ and we introduce a set of Markov chains with smoothness index s denoted by $\mathcal{M}(s) = \mathcal{M}(s, C_\mu, c_\mu, C_1, C_2, \ldots, C_6)$ fulfilling the following assumptions:

A1: $(X_i)_{i\in\mathbb{N}_0}$ is irreducible, aperiodic and has invariant measure μ which has a density which we will also denote by μ

A2: The invariant measure μ is bounded away from 0 and ∞, i.e. for constants $C_\mu > c_\mu > 0$, $c_\mu \leq \mu \leq C_\mu$.

A3: For a $s \geq d$, $\mu \in H^s$ and $\|\mu\|_{H^s} \leq C_1$ for some constant $C_1 > 0$.

Note that assumption **A2** implies that $L^2 = L^2(\mu)$ and that the pairs of norms $\|\cdot\|_{L^2}$ and $\|\cdot\|_{L^2(\mu)}$are equivalent. We assume that $X_0 \sim \mu$. Recall, that the transition operator P is defined on $L^2(\mu)$ by

$$Pf(x) = \mathbb{E}\left[f(X_1) \mid X_0 = x\right].$$

We assume that P is a compact integral operator with kernel $p(x, y)$, the transition density.

A4: $C_2 > p(x, y) > 0$ for all $x, y \in \mathbb{T}^d$ and for a constant $C_2 > 1$.

As P is compact in $L^2(\mu)$ it has a functional singular value decomposition: there exists two orthonormal bases $(u_k)_{k\in\mathbb{N}}$ and $(v_k)_{k\in\mathbb{N}}$ of $L^2(\mu)$ and a non-negative decreasing sequence $(\lambda_k)_{k\in\mathbb{N}}$ such that,

$$Pf = \sum_k \lambda_k \langle u_k, f\rangle_\mu v_k, \quad f \in L^2(\mu), \tag{4.1}$$

$$p(x, y) = \sum_k \lambda_k u_k(y)\mu(y)v_k(x). \tag{4.2}$$

Having obtained the representation (4.1) it is thus natural to formulate the remaining assumptions on the singular values and left and right singular functions. We assume that P has an approximately low rank structure with exponential decay of the singular values and that the left and right singular functions obey a certain degree of smoothness.

A4: The k-th singular value (counting multiplicity) is bounded by $C_3 \exp\left(-C_4 k^{\frac{2}{d}}\right)$ for positive constants C_3 and C_4.

A5: The absolute spectral gap $\gamma := 1 - \lambda_1 = 1 - \sup_{f\in L^2(\mu),\langle f,1\rangle_\mu=0,\|f\|_{L^2(\mu)}=1} \|Pf\|_{L^2(\mu)}$ is bounded away from zero by some constant $C_5 > 0$.

A6: The singular functions (u_k, v_k) fulfill $\sum_k \lambda_k^2(\|u_k\|_{H^s}^2 + \|v_k\|_{H^s}^2) \leq C_6$ for some constant $C_6 > 0$ for the same $s \geq d$ as in **A3**.

When considering the class $\mathcal{M}(s) = \mathcal{M}(s, C_\mu, c_\mu, C_1, \ldots, C_6)$ we will suppress the dependence on all parameters except s, since they, treating them as constants, do not change the minimax rate as long $\mathcal{M}(s)$ has non-empty interior. We will also write that $p \in \mathcal{M}(s)$ or $P \in \mathcal{M}(s)$ if it is the transition density or the transition operator of a Markov chain in $\mathcal{M}(s)$, respectively.

Periodised diffusion processes (which have also been considered in [Abr18, NR19, vWvZ16]) fulfill these assumptions under mild conditions on σ and b detailed in the Lemma below. This includes for example periodised versions of the Langevin processes considered by Roberts and Tweedie [RT96]. The proof follows after an application of Weyl's law for operators with non-smooth coefficients due to Ivrii [Ivr00] and by using p.d.e. results for elliptic operators from a recent article by Nickl and Ray [NR19].

Lemma 4.2.1. *For a vector field $b(x) = (b_1(x), \ldots, b_d(x)$ and a scalar $\sigma(x)$ consider the diffusion process $dY_t = b(Y_t)dt + \sigma(Y_t)dW_t$, $t \geq 0$, and its periodised version $X_t = Y_t$ modulo $\mathbb{Z}^d$. Assume that we observe the chain $(X_i)_{i\in\mathbb{N}_0}$. Moreover, assume that $\sigma(m + x) = \sigma(x)$ and $b(x + m) = b(x)$ for all $m \in \mathbb{Z}^d$ and that $\sigma^{-2}b = \nabla B$ for some $B \in C^2$. If $\|\sigma^{-2}\|_{C^{s-1}}$, $\|\sigma^2\|_{C^{s-1}}$ and $\|b\|_{C^{s-1}}$ are bounded by a constant $C > 0$ for some $s > 2$, then $p \in \mathcal{M}(s)$.*

4.2.3 Construction of the estimator

Here we describe how to obtain estimators for p and P given observations $(X_i)_{0\leq i\leq n}$, using a Galerkin approach. This method has also been employed for estimating the drift and volatility functions in a scalar diffusion model in the seminal paper by Gobet et al. [GHR04] and the first part of our construction is closely related.
Instead of estimating p in the functional space, the Galerkin approach estimates the action of P on a suitable approximation space and we obtain plug-in estimators for p and P.

Lemma 4.2.2. *For any non-negative integral operator P whose kernel p satisfies assumption **A4** and for any orthonormal basis $(f_k)_{k\in\mathbb{Z}^d}$ of L^2 we have that*

$$p(x, y) = \sum_{k,k'} \langle f_k, P f_{k'} \rangle f_k(x) f_{k'}(y)$$

in L^2. In particular this defines an isometry between P and p.

Working with P instead of p is advantageous because we can fully use its low-rank nature. We construct our estimator as a modified version of the estimator described by Gobet et al. [GHR04], adjusted to the non-reversible case:
Let $\{\Psi_{jk},\ j \in \mathbb{N} \cup \{-1\},\ k \in \mathbb{Z}^d\}$ be a tensorized and sufficiently smooth (with regularity greater than s) periodic wavelet basis of $\mathbb{T}^d$. For convenience, we denote this basis $\{\Psi_\lambda\}$

where $\lambda = (j, k_1, \ldots, k_d)$ is a multi index. We define V_J as the linear span of wavelets up to resolution level J,

$$V_J := \operatorname{span}\{\Psi_\lambda,\ |\lambda| = |(j,k)| := j \leq J\},$$

and denote by $\mathbf{V}_J$ the corresponding space of wavelet coefficients. The dimension of V_J is bounded by $C2^{Jd}$. One can find the construction of such a wavelet basis for instance in chapters 4.3.4 and 4.3.6 in [GN16].

Remark 4.2.1 (Other basis functions). The proof of Theorem 4.2.1, our main result, requires the Jackson and Bernstein inequalities and the bound $\|v\|_{L^\infty} \leq C\sqrt{\dim(V_J)}$ for any $v \in V_J$ satisfying $\|v\|_{L^2} \leq 1$. Thus, arguing as in Remark 5 in Chorowski and Trabs [CT16] the conclusions of Theorem 4.2.1 remain valid for the trigonometric and the B-spline basis if one strengthens the assumptions **A3** and **A7** to $\|\mu\|_{C^s} \leq c$ and $\sum \lambda_k^2(\|u_k\|_{C^s}^2 + \|v_k\|_{C^s}^2) \leq C$ for some constants $c, C > 0$.

As in Gobet et al. [GHR04], we will use bold letters for the coefficient expansions in the wavelet basis (Ψ_λ) of functions and operators in and on L^2. These denote vector and matrix like elements. The corresponding functions and operators - which do not depend on the basis - are in italic. In the case of vectors or matrix elements whose coefficients are only defined for $|\lambda| \leq J$, such as $\hat{\mathbf{R}}_J$, we will sometimes consider them as elements in the whole sequence space. This is done through setting the undefined coefficients to 0.
Let now J be a resolution level which we will choose later. Following Gobet et al. [GHR04] we construct a first estimator $\hat{\mathbf{R}}_J$ with coefficients :

$$\left(\hat{\mathbf{R}}_J\right)_{\lambda,\lambda'} = \frac{1}{n}\sum_{i=0}^{n-1} \Psi_\lambda(X_i)\,\Psi_{\lambda'}(X_{i+1}) \quad \text{for } |\lambda| \leq J,\ |\lambda'| \leq J.$$

The ergodic theorem implies that each of these coefficients converges almost surely to its expectation,

$$\mathbb{E}\left[\Psi_\lambda(X_0)\,\Psi_{\lambda'}(X_1)\right] = \langle \Psi_\lambda, P\Psi_{\lambda'}\rangle_\mu.$$

We thus also introduce $\mathbf{R}_J$ which is defined as the expectation of $\hat{\mathbf{R}}_J$, i.e.

$$(\mathbf{R}_J)_{\lambda,\lambda} = \langle \Psi_\lambda, P\Psi_{\lambda'}\rangle_\mu \quad \text{for } |\lambda| \leq J,\ |\lambda'| \leq J.$$

As $\overline{\cup_{J\in\mathbb{N}} V_J} = L^2$, we can define $\mathbf{R}$, the limit of $\mathbf{R}_J$ (with respect to the Hilbert–Schmidt norm). Note that $\mathbf{R}$ is defined through the $L^2(\mu)$-inner product and therefore

$$\mathbf{R} \neq \mathbf{P} := (\langle \Psi_\lambda, P\Psi_{\lambda'}\rangle)_{\lambda,\lambda'}.$$

We need to match the scalar products to estimate P. Let G be the Gram operator with corresponding sequence representation $\mathbf{G} = (\langle \Psi_\lambda, G\Psi_{\lambda'}\rangle)_{\lambda,\lambda'}$. G is such that $\forall u, v \in L^2$

$\langle u, Gv\rangle = \langle u, v\rangle_\mu$. Therefore, defining $\mathbf{u} = (\langle u, \Psi_\lambda\rangle)_\lambda$ (and $\mathbf{v}$ similarly), we have that

$$\langle \mathbf{u}, \mathbf{R}\mathbf{v}\rangle = \langle u, Pv\rangle_\mu = \langle u, GPv\rangle = \langle \mathbf{u}, \mathbf{G}\mathbf{P}\mathbf{v}\rangle .$$

If we estimate $\mathbf{G}^{-1}$ we are thus able to estimate $\mathbf{P}$. Following Gobet et al. [GHR04], we define

$$(\mathbf{G}_J)_{\lambda,\lambda'} := \langle \Psi_\lambda, \Psi_{\lambda'}\rangle_\mu \quad \text{for} \quad |\lambda| \le J, |\lambda'| \le J$$

and $\hat{\mathbf{G}}_J$ as:

$$\left(\hat{\mathbf{G}}_J\right)_{\lambda,\lambda'} = \frac{1}{n+1}\sum_{i=0}^{n} \Psi_\lambda(X_i)\,\Psi_{\lambda'}(X_i) \quad \text{for } |\lambda| \le J,\ |\lambda'| \le J.$$

From here on, our approach differs from that in Gobet et al. [GHR04]. In their (self-adjoint) setting, recovering the first non-trivial eigenpair is sufficient, as the drift and volatility functions are identified in terms of this eigenpair and the invariant measure. Since our objective is to estimate p and P we have to consider *all* singular triples instead. By assumption **A5** P is approximately low rank and hence $\mathbf{R}_J$, the matrix of projected coefficients of GP, is an approximately low rank matrix. For this reason we use the usual scheme for estimating low rank matrices, see for instance [YELM07, Klo11, BSW11, KLT11] and hard threshold the singular values of $\hat{\mathbf{R}}_J$. This yields which singular triples should be discarded in a data driven way.
We denote the SVD of $\hat{\mathbf{R}}_J$ by

$$\hat{\mathbf{R}}_J = \sum \hat{\lambda}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T,$$

where $\hat{\lambda}_k$ denotes the k-th eigenvalue of $\hat{\mathbf{R}}_J$ and $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{v}}_k$ the corresponding singular vectors. We define the spectral hard threshold estimator at level α, $\tilde{\mathbf{R}}_J = \tilde{\mathbf{R}}_J(\alpha)$ as,

$$\tilde{\mathbf{R}}_J := \sum \hat{\lambda}_k \mathbf{1}\left(|\hat{\lambda}_k| > \alpha\right) \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T. \tag{4.3}$$

Finally, we define the estimator for the action of P on V_J as

$$\tilde{\mathbf{P}}_J := \hat{\mathbf{G}}_J^{-1} \tilde{\mathbf{R}}_J. \tag{4.4}$$

We have the relation

$$Pf(x) = \sum_\lambda (\mathbf{P}\mathbf{f})_\lambda \Psi_\lambda(x),$$

and hence we estimate P by $\tilde{P}$ which we define as

$$\tilde{P}f(x) := \sum_{|\lambda|\le J} (\tilde{\mathbf{P}}\mathbf{f})_\lambda \Psi_\lambda(x). \tag{4.5}$$

This also yields an estimator for p by plug-in, given by

$$\tilde{p}(x,y) := \sum_{|\lambda|\leq J,\ |\lambda'|\leq J} \left(\tilde{\mathbf{P}}_J\right)_{\lambda,\lambda'} \Psi_\lambda(x)\Psi_{\lambda'}(y). \tag{4.6}$$

We finally choose for a constant $C > 0$ and for $\lceil\cdot\rceil$ denoting the ceiling function,

$$J = \left\lceil \log_2(n^{\frac{1}{2s+d}} \log(n)^{-\frac{d}{4s+2d}}) \right\rceil \quad \text{and} \quad \alpha = C\sqrt{\frac{2^{Jd}}{n}}, \tag{4.7}$$

to obtain the theoretical results in Theorem 4.2.1 in the next section.

Remark 4.2.2 (From P to P_τ). In molecular dynamics it is often desired to obtain an estimate for the transition operator,

$$P_\tau f(x) := \mathbb{E}[f(X_\tau)|X_0 = x], \qquad f \in L^2(\mu),$$

and its transition density p_τ, $\tau > 1$, for example for simulating or visualizing the Markov chain at a coarser timescale.
Given the estimator $\tilde{\mathbf{P}}$ in (4.4) and $\tau \in \mathbb{N}$ it is possible to obtain an estimator for p_τ as follows: if $\tau \leq c\log(n)$ we use the plug-in estimator $(\tilde{\mathbf{P}})^\tau$ and the induced estimator for p_τ in (4.6) and we are able to obtain similar theoretical results as in our main result, Theorem 4.2.1 (up to logarithmic factors).
If $\tau > C\log(n)$ it suffices to estimate the invariant density μ as in this case all singular values of P_τ except the first one are of smaller order than $1/n$.

Remark 4.2.3 (Adaptivity). The correct choice of J depends on the smoothness parameter s. In practice s is unknown, but one can use for instance Lepski's method to adapt to s. The proof that this works is a straightforward adaptation of results of Chorowski and Trabs [CT16].

4.2.4 Convergence rates

We now give our main theoretical result for the estimator $\tilde{p}$ of the transition density p constructed in (4.6). The upper bounds attained in L^2-loss for estimating p match the lower bounds and are therefore minimax optimal, showing that the logarithmic factors are inherent in the information-geometric structure of the problem. Heuristically this can be explained by the need to estimate approximately $\log(n)^{\frac{d}{2}}$ singular triples with d-dimensional rate for each triple.
Comparing our result to the standard Markov chain case without singular value decay where the L^2 minimax rates are $n^{-\frac{s}{2s+2d}}$ (e.g. [Clé00, Lac08]), we see that the effect of the dimension on the rate improves, up to the logarithmic factor, from $2d$ to d.

Theorem 4.2.1. *Suppose that we observe $(X_i)_{0\leq i\leq n}$ drawn from a stationary Markov Chain with $p \in \mathcal{M}(s)$ for some $s \geq d$. Then, for the estimator $\tilde{p}$ defined in (4.6) and a constant $C > 0$ we have, for n sufficiently large enough, with probability at least $1 - 12\exp\left(-n^{\frac{d}{2s+d}}\log(n)^{-\frac{d^2}{4s+2d}}\right)$ that*

$$\|p - \tilde{p}\|_{L^2} \leq C\log(n)^{\frac{d}{2}\frac{s}{2s+d}}\, n^{-\frac{s}{2s+d}}. \tag{4.8}$$

Moreover, the following minimax lower bound holds: for constants $c, p_0 > 0$,

$$\inf_{\hat{p}} \sup_{p\in\mathcal{M}(s)} \mathbb{P}_p\left(\|p - \hat{p}\|_{L^2} \geq c\log(n)^{\frac{d}{2}\frac{s}{2s+d}}\, n^{-\frac{s}{2s+d}}\right) \geq p_0 > 0. \tag{4.9}$$

In addition, by isometry this implies the same upper and lower bounds for estimating P. The proof of the upper bounds for $\tilde{p}$ in (4.8) is based on an application of concentration inequalities for Markov chains by Jiang et al. [JSF18] , combined with an ϵ-net argument to obtain tight bounds for the spectral norm rate of $\hat{\mathbf{R}}_J$ and an application of the general theory for rank penalized estimators by Klopp[Klo11].
The lower bound (4.9) requires different arguments compared to the case without decay. There an application of Assouad's Lemma and flipping coefficients suffices [Clé00]. Instead, here we adapt an idea by Koltchinskii an Xia [KX15] to our nonparametric setting by using projection matrices to infuse the low rank structure of P.
Additionally, the rank of $\tilde{\mathbf{P}}$ in (4.4) is bounded by approximately $\log(n)^{\frac{d}{2}}$, implying the same low rank structure for $\tilde{P}$. This justifies the approach of practitioners such as [CSP+07, CKL+08, KWNS18, SMP14] to dismiss most eigenpairs in their analysis.

Lemma 4.2.3. *Under the conditions of Theorem 4.2.1, we have for the estimator $\tilde{P}$ given in (4.5), for some constant $C > 0$, that, on the same event of probability at least $1 - 12\exp\left(-n^{\frac{d}{2s+d}}\log(n)^{-\frac{d^2}{4s+2d}}\right)$ on which (4.8) holds,*

$$\operatorname{rank}(\tilde{P}) \leq C\log(n)^{\frac{d}{2}}. \tag{4.10}$$

4.2.5 Numerical Experiments

In this section we illustrate our theoretical findings with simulated data from two diffusion processes. We consider one-dimensional, real valued Ornstein-Uhlenbeck and Cox-Ingersoll-Ross (CIR) processes.
Our theoretical findings are constrained to Markov chains with compact state space and thus, strictly speaking, do not apply for those. However, due to their drift pushing both of these processes close to the origin, all of our simulated observations where in fact bounded by 1.5 and 2.5 for the Ornstein-Uhlenbeck and CIR processes respectively, effectively con-

fining them to a compact set. Therefore we believe that the use of our methodology is justified here.
The Ornstein-Uhlenbeck process is given by

$$dX_t = -\theta X_t dt + \sigma dW_t, \quad t \geq 0 \tag{4.11}$$

and the CIR process by,

$$dX_t = -\theta(X_t - \mu)dt + \sigma\sqrt{X_t}dW_t, \quad t \geq 0. \tag{4.12}$$

In each case we generated observations at discrete time steps $X_0, X_1, \ldots, X_n$. For the Ornstein-Uhlenbeck process we simulated $X_0, X_1, \ldots, X_n$ exactly whereas we used the Euler-Maruyama scheme with step size 0.005 to generate the CIR process. The transition density of the Ornstein-Uhlenbeck process is the density of a Gaussian random variable and given by,

$$p(x,y) = \frac{1}{\sqrt{\pi\sigma^2(1-e^{-2\theta})/\theta}} e^{\frac{\theta(y-xe^{-\theta})^2}{\sigma^2(1-e^{-2\theta})}},$$

whereas for the CIR process the transition density is the density of a non-central χ^2-distribution and can be expressed as,

$$p(x,y) = \frac{\beta(\frac{y}{x})^{\frac{\nu}{2}} e^{\frac{\theta\nu}{2} - \beta y} e^{\frac{-\beta(x+y)}{e^\theta - 1}} I_\nu\left(\frac{\beta\sqrt{xy}}{\sinh(\theta/2)}\right)}{\Gamma(\beta\mu)(1-e^{-\theta})},$$

where $\beta = 2\theta\sigma^{-2}$, $\nu = \beta\mu - 1$, Γ denotes the Gamma function and I_ν the modified Bessel function of first kind with index ν.

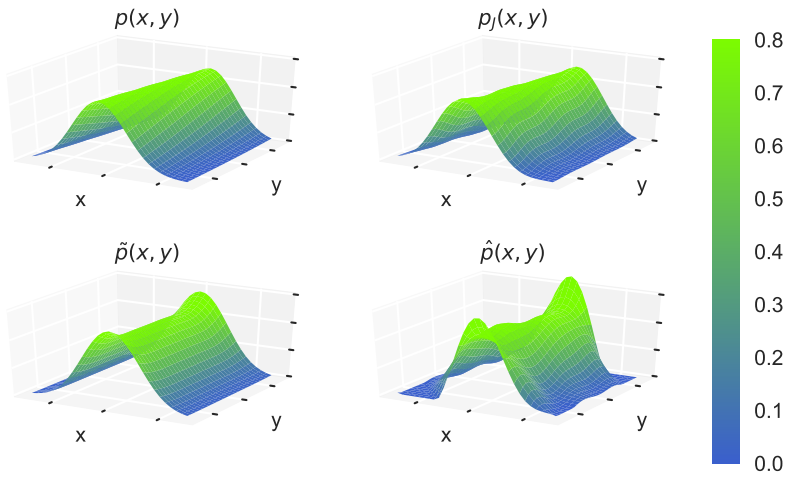


Figure 4.1: In clockwise order starting in the upper left corner: Transition-density $p(x, y)$ for the Ornstein-Uhlenbeck process (4.11) with parameters $\theta = 2$, $\sigma = 1$ and plotted in the region $[-1.5, 1.5]^2$; transition density projected on the approximation space of the first $J = 7$ trigonometric basis functions in each direction; non-thresholded estimator (i.e. $\tilde{p}$ with $\alpha = 0$) $\hat{p}$ for $n = 500$, $X_0 = 0.5$ and $J = 7$; thresholded estimator $\tilde{p}$ with the same settings and threshold level $\alpha = 0.1$.

As basis functions, following Remark 1, we use the trigonometric basis on the interval $[c - b, c + b]$, given by

$$\Psi_k(x) = \begin{cases} \frac{1}{\sqrt{2b}} & k = 0 \\ \frac{1}{\sqrt{b}} \cos\left(\frac{\pi(x-c)k}{2b}\right) & k = 2i,\ i \in \mathbb{N} \\ \frac{1}{\sqrt{b}} \sin\left(\frac{\pi(x-c)(k+1)}{2b}\right) & k = 2i-1,\ i \in \mathbb{N}. \end{cases}$$

For the Ornstein-Uhlenbeck process we choose $c = 0$ and $b = 2$ and for the CIR-process $c = b = 2$. Moreover, we symmetrize the estimator $\hat{\mathbf{R}}_J$ as originally proposed by Gobet et al. [GHR04], since one-dimensional diffusion processes are always reversible.
In the plots one can see that spectral hard thresholding eigenvalues reduces the noise level and smoothes the estimated transition density. This allows to use a larger resolution level than would be optimal for the non-thresholded estimator and thus to estimate finer details of the transition densities.

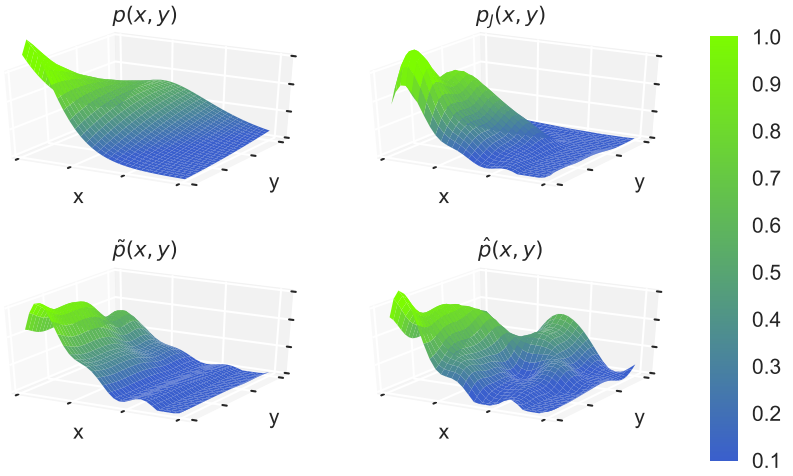


Figure 4.2: In clockwise order starting in the upper left corner: Transition-density $p(x, y)$ for the CIR process (4.12) with parameters $\theta = 1.2$, $\sigma = 1.1$, $\mu = 0.8$ and plotted in the region $[0.1, 3.1]^2$; transition density projected on the approximation space of the first $J = 8$ trigonometric basis functions in each direction; non-thresholded estimator (i.e. $\tilde{p}$ with $\alpha = 0$) $\hat{p}$ for $n = 1000$, $X_0 = 1$ and $J = 8$; thresholded estimator $\tilde{p}$ with the same settings and threshold level $\alpha = 0.08$.

4.3 Proofs

Throughout the results and proofs, the constants involved will be denoted by C and c; we will not always keep track of them and they may change from equation to equation. However one can check that they can be bounded by functions of constants defining the model in **A1**-**A7**.

4.3.1 Upper bounds - proof of (4.8)

Decomposing the error term

We first decompose the error term and then bound each term separately. We have that

$$\begin{aligned}
\|\tilde{\mathbf{P}}_J - \mathbf{P}\|_F \leq & \|\hat{\mathbf{G}}_J^{-1}(\tilde{\mathbf{R}}_J - \mathbf{R}_{r,J}\|_F + \|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_{r,J}\|_F \\
& + \|\mathbf{G}_J^{-1}(\mathbf{R}_{r,J} - \mathbf{R}_J)\|_F + \|\mathbf{G}_J^{-1}\mathbf{R}_J - P\|_F \\
\leq & \|\hat{\mathbf{G}}_J^{-1}\|_\infty \|\tilde{\mathbf{R}}_J - \mathbf{R}_J\|_F + (\|\mathbf{G}_J^{-1}\|_\infty + \|\hat{\mathbf{G}}_J^{-1}\|_\infty)\|\mathbf{R}_{r,J} - \mathbf{R}_J)\|_F \\
& + r^{1/2}\|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_{r,J}\|_\infty + \|\mathbf{G}_J^{-1}\mathbf{R}_J - \mathbf{P}\|_F \\
=: & I + II + III + IV, \qquad (4.13)
\end{aligned}$$

where $\mathbf{R}_{r,J}$ denotes a rank-r approximation of $\mathbf{R}_J$ with r to be chosen. We therefore have to take care of 4 terms: Variance bounds in Frobenius norm (I), rank-r approximation error (II), correction of the scalar product in spectral norm (III), and smoothness approximation error (IV).

Bounding I - variance bounds in spectral and Frobenius norm

In this section we bound the first term $\|\hat{\mathbf{G}}_J^{-1}\|_\infty \|\tilde{\mathbf{R}}_J - \mathbf{R}_J\|_F$. We will first obtain a bound for $\|\tilde{\mathbf{R}}_J - \mathbf{R}_J\|_F$. In our proof, we follow the usual line of arguments from the low rank literature [Klo11, KLT11] and bound the spectral norm of $\hat{\mathbf{R}}_J - \mathbf{R}_J$. Moreover, we also prove spectral norm bounds for $\hat{\mathbf{G}}_J - \mathbf{G}_J$ and $(\hat{\mathbf{G}}_J - \mathbf{G}_J)\mathbf{P}_J$, where $\mathbf{P}_J$ denotes the restriction of $\mathbf{P}$ to $\mathbf{V}_J$.

Lemma 4.3.1. *Assume $2^{3Jd} \leq cn$ for some small enough constant $c > 0$. Then for constants $C, C', C'' > 0$ we have that*

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{R}}_J - \boldsymbol{R}_J\right\|_\infty \leq C\sqrt{\frac{2^{Jd}}{n}}\right) \geq 1 - 4\exp\left(-2^{Jd}\right), \tag{4.14}$$

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{G}}_J - \boldsymbol{G}_J\right\|_\infty \leq C'\sqrt{\frac{2^{2Jd}}{n}}\right) \geq 1 - 4\exp\left(-2^{Jd}\right), \tag{4.15}$$

$$\mathbb{P}\left(\left\|(\hat{\boldsymbol{G}}_J - \boldsymbol{G}_J)\boldsymbol{P}_J\right\|_\infty \leq C''\sqrt{\frac{2^{Jd}}{n}}\right) \geq 1 - 4\exp\left(-2^{Jd}\right). \tag{4.16}$$

Proof. We only prove (4.14) as the two other bounds follow from the same argument, appealing to the bounds (4.34) and (4.35), respectively, instead.
We use an ϵ-net argument, arguing exactly as in the proof of Lemma 1.1 in Candès and Plan [CP11]. Indeed, arguing as in [CP11] we have, since $\mathbf{V}_J$ has dimension $C2^{Jd}$, that there exists a $\frac{1}{4}$-net $D_{\frac{1}{4}}$ of the unit sphere in $\mathbf{V}_J$ for Euclidean distance of cardinality less than $9^{C2^{Jd}}$.
Now let $\mathbf{v}$ and $\mathbf{u}$ with $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$ such that $\left\|\hat{\mathbf{R}}_J - \mathbf{R}_J\right\|_\infty = \mathbf{v}^T(\hat{\mathbf{R}}_J - \mathbf{R}_J)\mathbf{u}$ and $\mathbf{u}_0$ and $\mathbf{v}_0$ contained in $D_{\frac{1}{4}}$ such that $\|\mathbf{u} - \mathbf{u}_0\|_2 \leq 1/4$, $\|\mathbf{v} - \mathbf{v}_0\|_2 \leq 1/4$. We obtain that

$$\begin{aligned}
&\left\|\hat{\mathbf{R}}_J - \mathbf{R}_J\right\|_\infty \\
=&\langle \mathbf{v}_0, (\hat{\mathbf{R}}_J - \mathbf{R}_J)\mathbf{u}_0\rangle + \langle \mathbf{v} - \mathbf{v}_0, (\hat{\mathbf{R}}_J - \mathbf{R}_J)\mathbf{u}\rangle + \langle \mathbf{v}_0, (\hat{\mathbf{R}}_J - \mathbf{R}_J)(\mathbf{u} - \mathbf{u}_0)\rangle \\
\leq&\frac{1}{2}\left\|\hat{\mathbf{R}}_J - \mathbf{R}_J\right\|_\infty + \langle \mathbf{v}_0, (\hat{\mathbf{R}}_J - \mathbf{R}_J)\mathbf{u}_0\rangle
\end{aligned}$$

and hence it suffices to bound $\mathbf{v}^T\left(\hat{\mathbf{R}}_J - \mathbf{R}_J\right)\mathbf{u}$ on $D_{\frac{1}{4}}$. Generalizing formula (24) in Lemma 19 in Nickl and Söhl [NS17] by using a Bernstein inequality for non reversible Markov chains by Jiang et al. [JSF18] we obtain Lemma 4.4.1 which can be found in the

appendix. Applying Lemma 4.4.1 and using a union bound we obtain,

$$\begin{aligned}&\mathbb{P}\left(\left\|\left(\hat{\mathbf{R}}_J-\mathbf{R}_J\right)\right\|_\infty>2C\sqrt{\frac{2^{Jd}}{n}}x\right)\\ \leq&\mathbb{P}\left(\max_{\mathbf{u}_0,\mathbf{v}_0\in D_{\frac{1}{4}}}\langle\mathbf{v}_0,(\hat{\mathbf{R}}_J-\mathbf{R}_J)\mathbf{u}_0\rangle\ >C\sqrt{\frac{2^{Jd}}{n}}x\right)\leq 4\cdot 9^{C2^{Jd}}\left(e^{-2^{Jd}x}\right).\end{aligned}$$

Applying this with $x=1+C\ln(9)$ finishes the proof of the Lemma. □

Throughout the rest of the proof we will work on the event

$$\begin{aligned}\Omega:=\Bigg\{&\left\|\hat{\mathbf{R}}_J-\mathbf{R}_J\right\|_\infty\leq C\sqrt{\frac{2^{Jd}}{n}},\ \left\|\hat{\mathbf{G}}_J-\mathbf{G}_J\right\|_\infty\leq C'\sqrt{\frac{2^{2Jd}}{n}},\\ &\left\|(\hat{\mathbf{G}}_J-\mathbf{G}_J)\mathbf{P}_J\right\|_\infty\leq C''\sqrt{\frac{2^{Jd}}{n}}\Bigg\}\end{aligned}\tag{4.17}$$

which happens by Lemma 4.3.1 with probability at least $1-12e^{-2^{Jd}}$.
We now prove Frobenius norm bounds by applying Theorem 2 (iii) by Klopp [Klo11]. For completeness, we briefly present her proof below.
As noted by Bunea et. al. [BSW11] the hard threshold estimator (4.3) is the solution of the rank penalized problem

$$\tilde{\mathbf{R}}_J=\operatorname*{arg\,min}_{\mathbf{S}\in\mathbf{V}_J\times\mathbf{V}_J}\|\hat{\mathbf{R}}_J-\mathbf{S}\|_F^2+\alpha^2\mathrm{rank}(\mathbf{S}).\tag{4.18}$$

We suppose that α is such that $\alpha\geq 2C\sqrt{2^{Jd}/n}$. Since $\tilde{\mathbf{R}}_J$ is the minimizer of (4.18) the first inequality holds for any $\mathbf{S}$ and afterwards we use that $\langle A,B\rangle\leq\mathrm{rank}(A)\|A\|_F\|B\|_\infty$ and that $2ab\leq a^2+b^2$ to obtain that

$$\begin{aligned}&\|\tilde{\mathbf{R}}_J-\mathbf{R}_J\|_F^2\leq\|\mathbf{S}-\mathbf{R}_J\|_F^2+2\langle\hat{\mathbf{R}}_J-\mathbf{R}_J,\tilde{\mathbf{R}}_J-\mathbf{S}\rangle+\alpha^2(\mathrm{rank}(\mathbf{S})-\mathrm{rank}(\tilde{\mathbf{R}}_J))\\ \leq&\|\mathbf{S}-\mathbf{R}_J\|_F^2+\alpha\sqrt{\mathrm{rank}(\tilde{\mathbf{R}}_J)+\mathrm{rank}(\mathbf{S})}\|\tilde{\mathbf{R}}_J-\mathbf{S}\|_F+\alpha^2(\mathrm{rank}(\mathbf{S})-\mathrm{rank}(\tilde{\mathbf{R}}_J))\\ \leq&\|\mathbf{S}-\mathbf{R}_J\|_F^2+\alpha\sqrt{\mathrm{rank}(\tilde{\mathbf{R}}_J)+\mathrm{rank}(\mathbf{S})}\|\tilde{\mathbf{R}}_J-\mathbf{R}_J\|_F\\ &+\|\mathbf{S}-\mathbf{R}_J\|_F^2+\alpha\sqrt{\mathrm{rank}(\tilde{\mathbf{R}}_J)+\mathrm{rank}(\mathbf{S})}\|\mathbf{R}_J-\mathbf{S}\|_F+\alpha^2(\mathrm{rank}(\mathbf{S})-\mathrm{rank}(\tilde{\mathbf{R}}_J))\\ \leq&\frac{3}{2}\|\mathbf{S}-\mathbf{R}_J\|_F^2+\frac{1}{2}\|\tilde{\mathbf{R}}_J-\mathbf{R}_J\|_F^2+4\alpha^2\mathrm{rank}(\mathbf{S}).\end{aligned}$$

Summarizing, rearranging terms, we have that on Ω

$$\begin{aligned}I=\|\hat{\mathbf{G}}_J^{-1}\|_\infty\left\|\check{\mathbf{R}}_J-\mathbf{R}_J\right\|_F^2&\lesssim\|\hat{\mathbf{G}}_J^{-1}\|_\infty\inf_{\mathbf{S}\in\mathbf{V}_J\times\mathbf{V}_J}\left(\|\mathbf{S}-\mathbf{R}_J\|_F^2+\alpha^2\mathrm{rank}\left(\mathbf{S}\right)\right)\\ &\leq\|\hat{\mathbf{G}}_J^{-1}\|_\infty\left(\|\mathbf{R}_{r,J}-\mathbf{R}_J\|_F^2+r\alpha^2\right)\end{aligned}\tag{4.19}$$

We now find the adequate $\mathbf{R}_{r,J}$ in (4.19) .

Bounding II - low rank approximation error

By construction of the extension of operators on V_J as operators in the sequence space we have that, $\mathbf{R}_J = \pi_J^\lambda \mathbf{G}\mathbf{P}\pi_J^\lambda$, where π_J^λ is the orthogonal projection on $\mathbf{V}_J$ with respect to the Euclidean scalar product.
For any rank r approximation P_r of P, $\mathbf{R}_{r,J} := \pi_J^\lambda \mathbf{G}\mathbf{P}_r\pi_J^\lambda$ is a rank r approximation of $\mathbf{R}_J$ and fulfills $\|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F \lesssim \|P_r - P\|_F$. We define the rank r approximation of P as follows:

$$P_r f := \sum_{k=0}^{r-1} \lambda_k \langle u_k, f\rangle_\mu v_k \quad \text{for } f \in L^2(\mu). \tag{4.20}$$

This provides a sequence of approximations $\mathbf{R}_{r,J}$ of $\mathbf{R}_J$ satisfying

$$\|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F^2 \lesssim \sum_{k\geq r} \lambda_k^2. \tag{4.21}$$

We recall that by assumption **A5** $\lambda_k \leq C_3 \exp\left(-C_4 k^{2/d}\right)$. Denote by $\lceil\cdot\rceil$ the ceiling function and set

$$r := \left\lceil C \log\left(\frac{1}{\alpha}\right)^{\frac{d}{2}} \right\rceil + 2 \tag{4.22}$$

for $C > 0$ large enough. With this choice we obtain that

$$\|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F^2 \lesssim \sum_{k\geq r} \lambda_k^2 \lesssim \int_{2\sqrt{\log\frac{1}{\alpha}}}^{\infty} x^{d-1} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x. \tag{4.23}$$

If $d \geq 3$, we use integration by parts

$$\begin{aligned} F_d(y) := \int_y^\infty x^{d-1} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x &= y^{d-2} \exp\left(-\frac{y^2}{2}\right) + (d-1)\int_y^\infty x^{d-3} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \\ &= y^{d-2} \exp\left(-\frac{y^2}{2}\right) + (d-1) F_{d-2}(y), \end{aligned}$$

and it remains to bound F_d for $d = 2$ and $d = 1$. For $y \geq 1$ we have that $F_1(y) \leq F_2(y) = \exp(-y^2/2)$ and therefore we obtain overall that

$$\|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F^2 \lesssim \left(\log\frac{1}{\alpha}\right)^{\frac{d}{2}} \alpha^2. \tag{4.24}$$

Since $\operatorname{rank}(\mathbf{R}_{r,J}) = r$, (4.19) implies that on Ω

$$\left\|\tilde{\mathbf{R}}_J - \mathbf{R}_J\right\|_F^2 \lesssim \alpha^2 \left(\log\frac{1}{\alpha}\right)^{\frac{d}{2}} \lesssim \frac{2^{Jd}}{n} (\log n)^{\frac{d}{2}}. \tag{4.25}$$

Bounding III - correction of the scalar product

In this section we bound the third term, $r^{1/2}\|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_{r,J}\|_\infty$, in the decomposition (4.13). Moreover, we first prove that $\|\hat{\mathbf{G}}_J^{-1}\|_\infty \lesssim 1$ on Ω.
The fact that the invariant density is bounded away from 0 implies that $\inf_{\|\mathbf{u}\|_2=1} \|\mathbf{G}_J\mathbf{u}\|_2 \geq c$, which proves that $\mathbf{G}_J$ has bounded inverse in spectral norm. On the event

$$\left\{\left\|\hat{\mathbf{G}}_J - \mathbf{G}_J\right\|_\infty \leq c/2\right\} \supset \Omega,$$

we have by Lidski's inequality that

$$\forall \mathbf{u} \in \mathbf{V}_J, \quad \left\|\hat{\mathbf{G}}_J\mathbf{u}\right\|_2 \geq \|\mathbf{G}_J\mathbf{u}\|_2 - \left\|\mathbf{G}_J - \hat{\mathbf{G}}_J\right\|_\infty \|\mathbf{u}\|_2 .$$

Therefore, for any $\mathbf{u} \in \mathbf{V}_J$ we have that

$$\left\|\hat{\mathbf{G}}_J\mathbf{u}\right\|_2 \geq \frac{c}{2}\|\mathbf{u}\|_2 ,$$

implying that on Ω

$$\|\hat{\mathbf{G}}_J^{-1}\|_\infty \lesssim 1. \tag{4.26}$$

Moreover, due to the bounds (4.16) and (4.25), we obtain that

$$\begin{aligned}
&\|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_{r,J}\|_\infty \\
\lesssim &\|\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1}\|_\infty \|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F + \|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_J\|_\infty \\
\lesssim &\|\hat{\mathbf{G}}_J - \mathbf{G}_J\|_\infty \|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F + \|(\hat{\mathbf{G}}_J^{-1} - \mathbf{G}_J^{-1})\mathbf{R}_J\|_\infty \\
\lesssim &\|\hat{\mathbf{G}}_J - \mathbf{G}_J\|_\infty \left(\|\mathbf{R}_{r,J} - \mathbf{R}_J\|_F + \|\mathbf{G}_J^{-1}\mathbf{R}_J - \mathbf{P}_J\|_F\right) + \|(\hat{\mathbf{G}}_J - \mathbf{G}_J)\mathbf{P}_J\|_\infty \\
\lesssim &\sqrt{\frac{2^{2Jd}}{n}} \cdot \left(\sqrt{\frac{2^{Jd}}{n}}\log(n)^{\frac{d}{4}} + IV\right) + \sqrt{\frac{2^{Jd}}{n}}.
\end{aligned}$$

Bounding IV - bias bounds

It is left to bound the term IV in (4.13).
Let π_J^λ and π_J^μ be the orthogonal projectors on V^J for the λ and μ scalar products respectively. [GHR04] remarks that the non-zero eigenpairs of $\pi_J^\mu \mathbf{P} \pi_J^\lambda$ and $\mathbf{G}_J^{-1}\mathbf{R}_J$ are identical. We quickly prove this here for completeness.

Lemma 4.3.2. *We have the equality*

$$\pi_J^\mu = (\pi_J^\lambda G \pi_J^\lambda)^{-1}\pi_J^\lambda G$$

which implies that

$$G_J^{-1}R_J = \left(\pi_J^\lambda G \pi_J^\lambda\right)^{-1}\pi_J^\lambda R \pi_J^\lambda = \pi_J^\mu P \pi_J^\lambda. \tag{4.27}$$

Indeed, π_J^μ minimizes

$$\|G^{1/2}(I - \pi_J^\mu \pi_J^\lambda)\|_F^2,$$

leading to the normal equation

$$\begin{aligned} \pi_J^\lambda G(I - \pi_J^\lambda \pi_J^\mu) = 0 \quad &\implies \pi_J^\lambda G \pi_J^\lambda \pi_J^\mu = \pi_J^\lambda G \\ &\implies \pi_J^\mu = (\pi_J^\lambda G \pi_J^\lambda)^{-1} \pi_J^\lambda G. \end{aligned}$$

□

Using this identity, we establish the bias bounds.

Lemma 4.3.3. *The bias satisfies :*

$$\left\| \boldsymbol{G}_J^{-1} \boldsymbol{R}_J - \boldsymbol{P} \right\|_F \lesssim 2^{-Js}. \tag{4.28}$$

Proof. Note that $(I - \pi_J^\mu) = (I - \pi_J^\mu)(I - \pi_J^\lambda)$ and that $\|I - \pi_J^\mu\|_\infty \lesssim 1$ arguing as in the proof of Lemma 4.4. in [GHR04]: indeed, by our assumptions on μ we have that

$$\begin{aligned} \|\pi_J^\mu\|_\infty &= \sup_{f,g \in L^2:\ \|f\|_{L^2} = \|g\|_{L^2} = 1} \langle f, \pi_J^\mu g \rangle \\ &\leq \|\mu^{-1}\|_{L^\infty} \|\mu\|_{L^\infty} \sup_{f,g \in L^2(\mu)\ \|f\|_{L^2(\mu)} = \|g\|_{L^2(\mu)} = 1} \langle f, \pi_J^\mu g \rangle_\mu \lesssim 1. \end{aligned}$$

Therefore, we obtain that

$$\begin{aligned} \left\| \mathbf{G}_J^{-1} \mathbf{R}_J - \mathbf{P} \right\|_F = \|\pi_J^\mu P \pi_J^\lambda - P\|_F &\leq \|(I - \pi_J^\mu)P\|_F + \|P(I - \pi_J^\lambda)\|_F \\ &\lesssim \|(I - \pi_J^\lambda)P\|_F + \|P(I - \pi_J^\lambda)\|_F. \end{aligned}$$

Since the $L^2(\mu)$ and the L^2 norm are equivalent due to μ being bounded from above and below the induced Hilbert-Schmidt (Frobenius) norms are also equivalent. Particularly, this implies that for any operator A acting on $L^2(\mu)$ the inequality $\|A\|_F^2 \lesssim \sum_k \|Au_k\|_{L^2(\mu)}^2$ holds, where $\{u_k\}$ is the basis of $L^2(\mu)$-orthonormal right singular functions of P. Hence, we obtain that

$$\begin{aligned} \|(I - \pi_J^\lambda)P\|_F^2 &\lesssim \sum_k \|(I - \pi_J^\lambda)Pu_k\|_{L^2(\mu)}^2 = \sum_k \lambda_k^2 \|(I - \pi_J^\lambda)v_k\|_{L^2(\mu)}^2 \\ &\lesssim \sum_k \lambda_k^2 \|v_k\|_{H^s}^2 2^{-2Js} \lesssim 2^{-2Js}, \end{aligned}$$

where we used Jackson's inequality. Finally, we bound

$$\|P(I - \pi_J^\lambda)\|_F^2 = \sum_{\Psi_\lambda \notin V_J} \|P\Psi_\lambda\|_{L^2}^2$$

$$\lesssim \sum_{\Psi_\lambda \notin V_J} \| \sum_k \lambda_k v_k \langle u_k, \Psi_\lambda \rangle_\mu \|^2_{L^2(\mu)}$$
$$= \sum_k \lambda_k^2 \|(I - \pi_J^\lambda)(\mu u_k)\|^2_{L^2} \lesssim 2^{-2Js}$$

where we used Jackson's inequality and that for $s > d/2$ H^s is a Banach algebra. □

Rates of convergence for $\tilde{\mathbf{P}}$

Taking all the above bounds together we obtain on the event Ω which happens with probability at least $1 - 12e^{-2^{Jd}}$ that

$$\|\tilde{\mathbf{P}} - \mathbf{P}\|_F \lesssim \left(\log(n)^{\frac{d}{4}} \sqrt{\frac{2^{Jd}}{n}} + 2^{-Js} \right) \left(1 + \sqrt{\frac{2^{2Jd}}{n}} \log(n)^{\frac{d}{4}} \right) \tag{4.29}$$

and hence, choosing the optimal resolution level from (4.7) and since $s > d/2$, we obtain that with probability at least $1 - 12 \exp\left(-n^{\frac{d}{2s+d}} \log(n)^{-\frac{d^2}{4s+2d}} \right)$

$$\left\| \mathbf{P} - \tilde{\mathbf{P}} \right\|_F \lesssim \log(n)^{\frac{d}{2}\frac{s}{2s+d}} n^{-\frac{s}{2s+d}}. \tag{4.30}$$

The identification between P and $\mathbf{P}$ is isometric, and therefore, this proves the rates for estimation of P. By Lemma 4.2.2 the correspondence between P and p is also isometric, and thus the estimator $\tilde{p}$ achieves the same L^2-rates as in (4.30) on the same high probability event. This ends the proof of (4.8) in Theorem 4.2.1. □

4.3.2 Proof of Lemma 4.2.3

We have that $\operatorname{rank}(\tilde{P}) = \operatorname{rank}(\tilde{\mathbf{P}}) \leq \operatorname{rank}(\tilde{\mathbf{R}}_J) =: \tilde{r}$. Moreover, since $\tilde{\mathbf{R}}_J$ is a hard thresholding estimator, we have, by Lidski's inequality and denoting by $\lambda_k(\mathbf{R}_J)$ the k-th singular value of $\mathbf{R}_J$ that

$$\tilde{r} \geq k \implies \lambda_k(\mathbf{R}_J) + \|\mathbf{R}_J - \hat{\mathbf{R}}_J\|_\infty > 2C\sqrt{\frac{2^{Jd}}{n}}.$$

On the other hand, on Ω we have that $\|\mathbf{R}_J - \hat{\mathbf{R}}_J\|_\infty \leq C\sqrt{2^{Jd}/n}$. Finally, note that as in (4.23) we have that for some small enough $c > 0$

$$\lambda_k(\mathbf{R}_J)^2 \lesssim \sum_{l \geq k} \lambda_l^2 \lesssim \exp(-ck^{\frac{2}{d}}).$$

Hence, for $k = C' \log(n)^{\frac{d}{2}}$ for some $C' > 0$ large enough, we have that

$$\lambda_k(\mathbf{R}_J) \leq C\sqrt{\frac{2^{Jd}}{n}}.$$

Thus, $\lambda_k(\hat{\mathbf{R}}_J)$ and the preceding singular values are set to 0 by the hard thresholding procedure, implying that $\tilde{r} \lesssim \log(n)^{\frac{d}{2}}$. □

4.3.3 Lower bounds - proof of (4.9)

In this section, we prove the minimax lower bounds showing that the rates attained by our estimator are optimal.

We first construct a sufficiently rich sub-set $M \subset \mathcal{M}(s)$ of transition densities. Let π_0 be the λ-orthogonal projector onto constants. Let $(\Psi_\lambda)_\lambda$ be a s-regular orthonormal periodic wavelet family with at least one vanishing moment and compactly supported. Let (N_J) be for each J a maximal subset of wavelets of resolution J such that two different wavelets in R_J have disjoint support. We have that $|N_J| \geq c2^{Jd}$. Let $W_J = \operatorname{span}(\Psi \in N_J)$.
Let $\mathcal{G}_{k,J}$ denote the set of all k-dimensional subspaces of W_J. For every element $S \in \mathcal{G}_{k,J}$, we denote π_S the orthogonal projector from L^2 to S, and define $P_S = \pi_0 + \eta\varepsilon_n\pi_S$, with

$$\varepsilon_n = (\log n)^{-\frac{d}{4}\frac{d}{2s+d}} n^{-\frac{s}{2s+d}}$$

and for $\eta > 0$ a constant. The following lemma shows that these P_S are contained in $\mathcal{M}(s)$ for an appropriate choice of k and J:

Lemma 4.3.4. *Choose k and J such that*

$$\frac{c_k}{2}(-\log\varepsilon_n)^{\frac{d}{2}} \leq k \leq c_k(-\log\varepsilon_n)^{\frac{d}{2}}$$
$$\frac{c_J}{2}\log(n)^{-\frac{d}{2}\frac{1}{2s+d}} n^{1/(2s+d)} \leq 2^J \leq c_J \log(n)^{-\frac{d}{2}\frac{1}{2s+d}} n^{1/(2s+d)}.$$

Then for any choice of constants defining $\mathcal{M}(s)$ such that $\mathcal{M}(s) \neq \emptyset$, we can choose positive constants c_ε, c_k and c_J, such that for n large enough $\forall S \in \mathcal{G}_{k,J}$ P_S is contained in $\mathcal{M}(s)$.

Proof. We carefully check that **A1**-**A7** are fulfilled.
We first check **A1**-**A4** together. Let $b = (f_i)_{1\leq i\leq k}$ be an orthonormal basis of S. Complete it into $\overline{b} = (f_i)_{1\leq i\leq |N_J|}$ an orthonormal basis of W_J and let $\mathbf{f}_{i,\lambda} = \langle f_i, \Psi_\lambda\rangle$ be the change of coordinate matrix between $(\Psi_\lambda)_{\lambda\in R_J}$ and $\overline{b}$. Then

$$p_S(x,y) = 1 + \varepsilon_n\eta\sum_{i=1}^{k}\sum_{\lambda\in R_J}\sum_{\lambda'\in R_J}\mathbf{f}_{i,\lambda}\Psi_\lambda(x)\,\mathbf{f}_{i,\lambda'}\Psi_{\lambda'}(y)$$

Note that this formula implies that λ is the invariant measure and thus $\mathbf{A1}-\mathbf{A3}$ once we have proved that p_S defines a probability density. Since the Ψ_λ have disjoint support,

$$1 - C\eta 2^{Jd}\varepsilon_n \leq p_S(x,y) \leq 1 + C\eta 2^{Jd}\varepsilon_n.$$

Since $s \geq d\ 2^{Jd}\varepsilon_n$ goes to 0 as n grows, implying that for any $c > 0$, for n large enough, $0 < 1-c \leq p_S(x,y) \leq 1+c$. Moreover, p integrates to 1 and hence p is indeed a probability density and **A1**-**A4** follow. Moreover, by definition of P_S the first eigenvalue is 1, the next k eigenvalues are $\eta\varepsilon_n$ and the remaining eigenvalues are zero. With our choices of k and ε_n we thus obtain **A5**. Likewise **A6** is fulfilled as the spectral gap is precisely $1-\eta\varepsilon_n$ which can be made arbitrary close to one. Finally, by the relation $\|f_i\|_{H^s} \leq C2^{Js}\|f_i\|_{L^2}$ which holds for arbitrary $f_i \in W_J$ (see equation 4.166 and following in chapter 4.3.6 in [GN16]) we obtain that

$$\sum_k \lambda_k^2 \|e_k\|_{H_s}^2 \leq 1 + Ck\eta^2\varepsilon_n^2 2^{2Js} \leq C$$

for n large enough and thus **A7** holds. □

We now choose a maximal subset M of $\mathcal{G}_{k,J}$ such that for any two projections in M, denoted by S_1 and S_2 we have that,

$$\|p_{S_1} - p_{S_2}\|_{L^2} = \|P_{S_1} - P_{S_2}\|_F \geq c_0\varepsilon_n\sqrt{k} \tag{4.31}$$

for a constant $c_0 > 0$. By Proposition 8 in [Paj98] we have for some universal constants $c, C > 0$ that,

$$\left(\frac{c}{c_0}\right)^{k(|R_J|-k)} \leq |M| \leq \left(\frac{C}{c_0}\right)^{k(|R_J|-k)}. \tag{4.32}$$

We finally add the element $p_0 = 1$ to M.

We now apply Theorem 2.5 in [Tsy08] and check that its conditions are fulfilled for our choices of k and ε_n. For $p_S \in M$ denote by $\mathcal{P}_S^n$ the probability measure for the Markov chain $(X_0, \ldots, X_n)$ with transition density p_S and invariant measure 1. We first show that we can control the Kullback–Leibler divergence $K(\mathcal{P}_S^n, \mathcal{P}_0^n)$ defined for two probability measures $\mathcal{P}$ and $\mathcal{Q}$ with densities $\mathrm{d}\mathcal{P}$ and $\mathrm{d}\mathcal{Q}$ respectively as,

$$K(\mathcal{P},\mathcal{Q}) := \begin{cases} \int_{\mathbb{T}^d} \log\left(\frac{\mathrm{d}\mathcal{P}(x)}{\mathrm{d}\mathcal{Q}(x)}\right) \mathrm{d}\mathcal{P}(x) & \mathcal{P} \text{ is absolutely continous with respect to } \mathcal{Q} \\ \infty & \text{else} \end{cases}$$

by the squared L^2 norm of $p_S - p_0$

$$K\left(\mathcal{P}_S^n, \mathcal{P}_0^n\right) \leq n\left\|p_S - p_0\right\|_{L^2}^2.$$

Indeed,

$$K\left(\mathcal{P}_S^n, \mathcal{P}_0^n\right) = \mathbb{E}_{\mathcal{P}_S^n}\left[\log\left(\frac{\mathrm{d}\mathcal{P}_S^n\left(X_0, X_1, \ldots, X_n\right)}{\mathrm{d}\mathcal{P}_0^n\left(X_0, X_1, \ldots, X_n\right)}\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{P}_S^n}\left[\log\left(\frac{p_S(X_0,X_1)\dots p_S(X_{n-1},X_n)}{p_0(X_0,X_1)\dots p_0(X_{n-1},X_n)}\right)\right] \\
&= n\mathbb{E}_{\mathcal{P}_S^1}\left[\log\left(\frac{p_S(X_0,X_1)}{p_0(X_0,X_1)}\right)\right].
\end{aligned}
$$

Further evaluating the last equation we find,

$$
\mathbb{E}_{\mathcal{P}_S^1}\left[\log\left(\frac{p_S(X_0,X_1)}{p_0(X_0,X_1)}\right)\right] = \int_x\int_y \log\left(p_S(x,y)\right) p_S(x,y)\,\mathrm{d}x\mathrm{d}y.
$$

We can decompose $p_S = 1 + \varepsilon_n H_b$. Then, since $\log(1+\varepsilon_n H_b) \le \varepsilon_n H_b$, we have that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{P}_S^1}\left[\log\left(\frac{p_S(X_0,X_1)}{p_0(X_0,X_1)}\right)\right] &\le \int_x\int_y \varepsilon_n H_b(x,y)\left(1+\varepsilon_n H_b(x,y)\right)\mathrm{d}x\mathrm{d}y \\
&= \int_x\int_y \varepsilon_n^2 H_b(x,y)^2\,\mathrm{d}x\mathrm{d}y \\
&= \|p_0 - p_S\|_{L^2}^2 = \eta^2\varepsilon_n^2\|\pi_S\|_F^2 = \eta^2\varepsilon_n^2 k
\end{aligned}
$$

Thus, ordering the elements $p_S \in M$ from 0 to $|M|$ with $p_0 = 1$ and denoting by $\mathcal{P}_i^n$ the respective probability measure for the chain $(X_0,\dots,X_n)$, we obtain that

$$
\frac{1}{|M|}\sum_{j=1}^{|M|} \mathrm{K}\left(\mathcal{P}_j^n, \mathcal{P}_0^n\right) \le n\eta^2\varepsilon_n^2 k.
$$

The bound (4.32) on $|M|$ and our choices of k and J described in Lemma 4.3.4 then imply

$$
n\eta^2\varepsilon_n^2 k \le \alpha C k 2^{Jd} \le k\left(2^{Jd} - k\right)\log(\frac{c}{c_0}) \le \log|M|,
$$

by choosing η small enough. Thus, using also (4.31), all conditions of Theorem 2.5 in [Tsy08] are met and we obtain (4.9). Moreover, by isometry the same lower bound holds for P. □

4.4 Appendix

4.4.1 Proof of Lemma 4.2.1

The condition $\sigma^{-2}b = \nabla B$ for some $B \in C^2$ implies, by Theorem 4.2 in [Ken78], that the chain X_t is reversible with invariant measure satisfying $\mu \propto e^B$. This identity and the bounds on the C^{s-1} norms of b and σ^{-2} imply $\mu \in H^s$ and that $c \le \mu \le C$ for constants $c, C > 0$. Moreover, irreducibility and aperiodicity follow by the upper and lower bounds on p below and thus **A1** – **A3** are fulfilled. Assumption **A4** follows by estimates for the heat kernel, see e.g. Theorem 1.1 in [Nor97] and by noting that $\sum_{x'=x+\mathbb{Z}^d} Ce^{-c\|x'-y\|_2^2}$ is summable for every $x, y \in \mathbb{T}^d$. Also note that these estimates yield $p(x,y) > c > 0$

uniformly for $x, y \in \mathbb{T}^d$.
Assumption **A5** is implied by Weyl's law for elliptic operators with non-smooth coefficients on closed manifolds, Theorem 3.1. in [Ivr00]. Particularly, **A5** follows by inverting formula (3.4) in [Ivr00] applied to the operator $\tilde{L} = G^{-1/2} L G^{1/2}$ where L is the infinitesimal generator L

$$L = \frac{\sigma(x)}{2} \sum_{i=1}^{d} \frac{\partial^2}{\partial^2 x_i} + \sum_{i=1}^{d} b_i(x) \frac{\partial}{\partial x_i}$$

(with $m = 1$ there) and by noting that the $L^2(\mu)$-eigenvalues of L equal the $L^2(\lambda)$-eigenvalues of $\tilde{L}$ and that the $L^2(\mu)$-eigenvalues of P equal the exponentiated $L^2(\mu)$-eigenvalues of L.
A6 follows from arguing as [Abr18] in the proof of Theorem 6, using exercise 7 on p. 493 in [BW09] instead of the cited Lemma 2.3 there and the lower bound on p from above.
We now show that assumption **A7** is fulfilled. Adapting Lemma 11 in [NR19] to our situation with non-constant but scalar σ is straightforward and we obtain that there exists a $C = C(\|\sigma^{-2}\|_{C^{s-1}}, \|b\|_{C^{s-2}}) > 0$ such that for all $f \in L^2$ with $\mathbb{E}[f(X_0)] = 0$ we have for $t \leq s$ that

$$\left\|L^{-1}(f)\right\|_{H^t} \leq C \|f\|_{H^{t-2}}, \qquad \left\|(L^*)^{-1}(f)\right\|_{H^t} \leq C \|f\|_{H^{t-2}}$$

where $L^{-1}(f)$ denotes the solution u to the inhomogeneous p.d.e. $Lu = f$. Since P and L are self-adjoint the left and right singular functions coincide, are called eigenfunctions, and we denote them by e_k. Since $\langle e_k, 1\rangle_\mu = 0$ for $k > 0$ we can use this repeatedly for the eigenfunctions e_k which fulfill $L^{-1} e_k = \log(\lambda_k) e_k$. This implies that

$$\|e_k\|_{H^s} \lesssim |\log \lambda_k|^{\lceil s/2 \rceil} \|e_k\|_{L^2} \lesssim |\log \lambda_k|^{\lceil s/2 \rceil} \lesssim k^{\frac{s+2}{d}},$$

where the last inequality follows by using Weyl's law again. Therefore we obtain that,

$$\sum_k \lambda_k^2 \|e_k\|_{H^s}^2 \lesssim \sum_k k^{\frac{2s+4}{d}} e^{-ck^{\frac{2}{d}}} \lesssim 1$$

and **A7** follows. □

4.4.2 Lemma 4.4.1

Lemma 4.4.1. *Assume $2^{3Jd} \leq n$ and that $\kappa n 2^{-3Jd} \geq x \geq 1$ for some constant $\kappa > 1$. Then for constants $C = C(\kappa), C' = C'(\kappa)$ and $C'' = C''(\kappa)$ and $\forall u, v \in V_J$ with $\|u\|_{L^2} = \|v\|_{L^2} = 1$ the three following bounds hold:*

$$\mathbb{P}\left(\boldsymbol{v}^T \left(\hat{\boldsymbol{R}}_J - \boldsymbol{R}_J \right) \boldsymbol{u} > C \sqrt{\frac{2^{Jd}}{n}} x \right) \leq 2 e^{-2^{Jd} x} \tag{4.33}$$

$$\mathbb{P}\left(\boldsymbol{v}^T\left(\hat{\boldsymbol{G}}_J-\boldsymbol{G}_J\right)\boldsymbol{u}>C'\sqrt{\frac{2^{2Jd}}{n}x}\right)\leq 2e^{-2^{Jd}x} \tag{4.34}$$

$$\mathbb{P}\left(\boldsymbol{v}^T\left(\hat{\boldsymbol{G}}_J-\boldsymbol{G}_J\right)\boldsymbol{P}_J\boldsymbol{u}>C''\sqrt{\frac{2^{Jd}}{n}x}\right)\leq 2e^{-2^{Jd}x}. \tag{4.35}$$

In each case the proof is an application of the Bernstein type inequality in Theorem 1.1 by Jiang et al. [JSF18]. Also note that the proof is similar to the proof of Lemma 19 in Nickl and Söhl [NS17] but that they use a different concentration inequality. We prove (4.33) carefully and only sketch the proofs of the remaining two inequalities as they follow along the same line of argumentation.
Without loss of generality assume that n is even. We use the identity

$$\begin{aligned}\mathbf{v}^T\left(\hat{\mathbf{R}}_J-\mathbf{R}_J\right)\mathbf{u}&=\frac{1}{n}\sum_{i=0}^{n-1}\left(v\left(X_i\right)u\left(X_{i+1}\right)-\mathbb{E}\left[v\left(X_0\right)u\left(X_1\right)\right]\right)\\ =&\frac{1}{n}\sum_{i=0}^{n/2-1}\left(v\left(X_{2i}\right)u\left(X_{2i+1}\right)-\mathbb{E}\left[v\left(X_0\right)u\left(X_1\right)\right]\right)\\ &+\frac{1}{n}\sum_{i=0}^{n/2-2}\left(v\left(X_{2i+1}\right)u\left(X_{2i+2}\right)-\mathbb{E}\left[v\left(X_0\right)u\left(X_1\right)\right]\right)\end{aligned}$$

where $v(x)=\sum_\lambda \mathbf{v}_\lambda\Psi_\lambda(x)$ and u is defined likewise. We only treat the first term in the equation above as the second one can be bounded with the same arguments. By Lemma 24 in [NS17] the invariant density of the chain $(X_{2i},X_{2i+1})_{i\in\mathbb{N}_0}$ is

$$\mu_2(x_1,x_2)=\mu(x)p(x,y).$$

Moreover, denoting by P_2 the transition operator of $(X_{2i},X_{2i+1})_{i\in\mathbb{N}_0}$, we can bound its absolute spectral gap by the absolute spectral gap of the original chain $(X_i)_{i\in\mathbb{N}_0}$ by applying Lemma 24 in [NS17], i.e. for any $f\in L^2(\mu_2)$, $\langle f,1\rangle_{\mu_2}=0$, we have that

$$\|P_2f\|_{L^2(\mu_2)}\leq\lambda_1\|f\|_{L^2(\mu_2)}.$$

We upper bound the variance

$$\begin{aligned}V_{v,u}:&=\left\|v\left(x\right)u\left(y\right)-\mathbb{E}\left[v\left(X_0\right)u\left(X_1\right)\right]\right\|^2_{L^2(\mu_2)}\\ &\leq\int v(x)^2u(y)^2\mu(x)p(x,y)dxdy\leq\|\mu\|_{L^\infty}\|p\|_{L^\infty}\leq C\end{aligned}$$

for some constant $C>0$. Next we bound

$$\left\|v\left(x\right)u\left(y\right)-\mathbb{E}\left[v\left(X_0\right)u\left(X_1\right)\right]\right\|_{L^\infty}\leq 2\left\|v\left(x\right)u\left(y\right)\right\|_{L^\infty}$$

$$\leq C' 2^{Jd}.$$

We now apply Theorem 1.1 by [JSF18] (with $\epsilon = x\sqrt{\frac{2^{Jd}}{n}}$ for some constant $\sqrt{n}2^{-3Jd/2} \geq x \geq 1$, $\sigma^2 \leq C$ and $c = C'2^{Jd}$ there) to obtain overall that for some constants $\tau, \tau' > 0$

$$\mathbb{P}\left(\mathbf{v}^T\left(\hat{\mathbf{R}}_J - \mathbf{R}_J\right)\mathbf{u} > x\sqrt{\frac{2^{Jd}}{n}}\right) \leq \exp\left(\frac{-x^2 2^{Jd}}{\tau + \tau' \frac{x 2^{3Jd/2}}{\sqrt{n}}}\right).$$

Using also the assumption $2^{3Jd} \leq n$ this yields for another constant $\tau'' > 0$ that

$$\mathbb{P}\left(\mathbf{v}^T\left(\hat{\mathbf{R}}_J - \mathbf{R}_J\right)\mathbf{u} > x\tau''\sqrt{\frac{2^{Jd}}{n}}\right) \leq \exp\left(-x2^{Jd}\right).$$

For the proof of (4.34) note that we have the equality

$$\mathbf{v}^T\left(\hat{\mathbf{G}}_J - \mathbf{G}_J\right)\mathbf{u} = \frac{1}{n+1}\sum_{i=0}^{n} v(X_i)u(X_i) - \mathbb{E}v(X_0)u(X_0).$$

Hence, it remains to bound the variance and obtain a pointwise bound. We have that

$$\|v(x)u(x) - \mathbb{E}v(X_0)u(X_0)\|^2_{L^2(\mu)} \leq \int v(x)^2 u(x)^2 \mu(x)\mathrm{d}x \lesssim 2^{Jd}$$

and

$$\|v(x)u(x) - \mathbb{E}v(X_0)u(X_0)\|_{L^\infty} \lesssim 2^{Jd}.$$

For the proof of (4.35) we argue as before, this time working with the equality

$$\mathbf{v}^T\left(\hat{\mathbf{G}}_J - \mathbf{G}_J\right)\mathbf{P}_J\mathbf{u} = \frac{1}{n+1}\sum_{i=0}^{n} v(X_i)\tilde{u}(X_i) - \mathbb{E}v(X_0)\tilde{u}(X_0),$$

where $\tilde{u} = P_J u$. As above it remains to bound the variance and obtain a pointwise bound. In this case we have that

$$\|v(x)\tilde{u}(x) - \mathbb{E}v(X_0)\tilde{u}(X_0)\|^2_{L^2(\mu)} \leq \int v(x)^2\tilde{u}(x)^2\mu(x)\mathrm{d}x \lesssim \|\tilde{u}(x)\|_{L^\infty}.$$

Moreover, denoting by p_J the L^2-projection of p to $V_J \times V_J$, we have by Young's convolution inequality

$$\|\tilde{u}\|_{L^\infty} = \|\int p_J(\cdot, y)u(y)dy\|_{L^\infty} \leq \|u\|_{L^2}\|p_J\|_{L^2} \lesssim 1$$

Thus, we obtain that

$$\|v(X_0)\tilde{u}(X_0) - \mathbb{E}v(X_0)\tilde{u}(X_0)\|_{L^\infty} \lesssim 2^{Jd/2}.$$

□

4.4.3 Proof of Lemma 4.2.2

We split this proof in two parts: we first show the mapping is an isometry from Hilbert–Schmidt operators to its image endowed with the $L^2 \times L^2$ norm, and then prove that its sends the transition operators into their respective transition probabilities.

Lemma 4.4.2. *The following function*

$$T : A_{\lambda,\lambda'} \to \left(x, y \to \sum_{\lambda,\lambda'} \langle \Psi_\lambda, A\Psi_{\lambda'} \rangle_{L^2(\lambda)} \Psi_\lambda(x) \Psi_{\lambda'}(y) \right)$$

is an isometry from the space of Hilbert–Schmidt operators on L^2 endowed with the Hilbert–Schmidt norm to a subset of $L^2 \times L^2$.

Proof. We first assume that a finite number of $A_{\lambda,\lambda'}$ are non-zero; this ensures that we work with proper functions and justifies exchanging summation and integration. The general case follows using a density argument. For A with a finite number of non-zero coefficients we have that,

$$\begin{aligned} \|T(A)\|_{L^2}^2 &= \int \sum_{\lambda_1,\lambda_2,\lambda_3,\lambda_4} A_{\lambda_1,\lambda_2} A_{\lambda_3,\lambda_4} \Psi_{\lambda_1}(x) \Psi_{\lambda_3}(x) \Psi_{\lambda_2}(y) \Psi_{\lambda_4}(y) \, \mathrm{d}x\mathrm{d}y \\ &= \sum_{\lambda_1,\lambda_2,\lambda_3,\lambda_4} A_{\lambda_1,\lambda_2} A_{\lambda_3,\lambda_4} \delta_{\lambda_1,\lambda_3} \delta_{\lambda_2,\lambda_4} = \sum_{\lambda_1,\lambda_2} A_{\lambda_1,\lambda_2}^2 = \|A\|_F^2 . \end{aligned}$$

□

Lemma 4.4.3. *The function*

$$T : A_{\lambda,\lambda'} \to \left(x, y \to \sum_{\lambda,\lambda'} A_{\lambda,\lambda'} \Psi_\lambda(x) \Psi_{\lambda'}(y) \right)$$

sends a transition operator P into its transition density p.

Proof. We show that P_{T}, the transition operator for the kernel $\mathrm{T}(P)$, equals P. We have

for any $f, g \in L^2$ that

$$\begin{aligned}\langle g, P_{\mathrm{T}} f\rangle &= \int_x g(x) \left(P_{\mathrm{T}} f\right)(x) \mathrm{d}x \\ &= \int_x \int_y \sum_{\lambda,\lambda'} g(x) \Psi_\lambda(x) P_{\lambda,\lambda'} \Psi_{\lambda'}(y) f(y) \mathrm{d}y \mathrm{d}x \\ &= \sum_{\lambda,\lambda'} \langle \Psi_\lambda, g\rangle P_{\lambda,\lambda'} \langle \Psi_{\lambda'}, f\rangle = \langle g, Pf\rangle .\end{aligned}$$

□

Bibliography

[Abr18] K. Abraham. Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data. *Bernoulli, to appear*, 2018. (Cited on pages 103 and 120).

[ABDJ06] F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006. (Cited on page 7).

[AF03] R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Elsevier/Academic Press, 2003. (Cited on page 101).

[ANW12] A. Agarwal, S. Negahban, and M.J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 2012. (Cited on page 3).

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In P.N. Petrov and F. Csaki, editors, *Proceedings 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973. (Cited on page 11).

[AL11] N. Akakpo and C. Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electron. J. Statist.*, 5:1618–1653, 2011. (Cited on page 99).

[And51] T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.*, 22(327-351), 1951. (Cited on page 3).

[And63] T.W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963. (Cited on page 57).

[AG57] T.W. Anderson and L.A. Goodman. Statistical Inference about Markov Chains. *Ann. Math. Statist.*, 28(1):89–110, 1957. (Cited on page 4).

[ACCP11] E. Arias-Castro, E.J. Candès, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Statist.*, 39(5):2533–2556, 2011. (Cited on page 20).

[Arl19] S. Arlot. Minimal penalties and the slope heuristics: a survey. *arXiv preprint*, 2019. (Cited on page 13).

[Ass83] P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math*, 296:1021–1024, 1983. (Cited on page 17).

[ATW19] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Ann. Statist.*, 47(2):1148–1178, 2019. (Cited on page 8).

[BB18] N. Baldin and Q. Berthet. Optimal Link Prediction with Matrix Logistic Regression. *arXiv preprint*, 2018. (Cited on page 12).

[BvH16] A.S. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016. (Cited on pages 21, 53, and 55).

[Bar04] Y. Baraud. Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2):528–551, 2004. (Cited on page 26).

[BBM99] A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113:301–413, 1999. (Cited on pages 10 and 11).

[Bar51] M.S. Bartlett. The frequency goodness of fit test for probability chains. *Proc. Cambridge Philos. Soc.*, 47:86–95, 1951. (Cited on page 4).

[BK19] B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, 47(4):2261–2285, 2019. (Cited on page 8).

[BH05] F. Bauer and T. Hohage. A Lepskij-type stopping rule for regularized Newton methods. *Inverse Problems*, 21(6):1975–1991, 2005. (Cited on page 9).

[BG03] E. Belitser and S. Ghosal. Adaptive bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–5598, 2003. (Cited on page 9).

[BC13] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. (Cited on pages 12 and 13).

[BCCW18] A. Belloni, V. Chernozhukov, D. Chetverikov, and Y. Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Ann. Statist.*, 46(6B):3643–3675, 2018. (Cited on page 15).

[BCK15] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015. (Cited on page 16).

[BCW11] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. (Cited on page 13).

[BCW14] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 2014. (Cited on page 13).

[BL07] J. Bennett and S. Lanning. The Netflix Prize. *Proceedings of KDD Cup and Workshop*, 2007. (Cited on pages 2, 20, 22, 25, 26, 27, 33, and 38).

[BR13] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, 41(4):1780–1815, 2013. (Cited on pages 12, 20, and 57).

[BW09] R.N. Bhattacharya and E.C. Waymire. *Stochastic Processes with Applications*. SIAM, 2009. (Cited on page 120).

[BR88] J.P. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A*, 50:381–393, 1988. (Cited on page 14).

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. (Cited on page 12).

[BL08a] P.J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008. (Cited on page 3).

[BL08b] P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008. (Cited on page 4).

[BR03] P.J. Bickel and Y. Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4):1033–1053, 2003. (Cited on page 13).

[BR73] P.J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *Ann. Statist.*, 1(6):1071–1095, 1973. (Cited on page 16).

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. für Wahrscheinlichkeitstheorie und Verw. Geb.*, 65(2):181–237, 1983. (Cited on page 18).

[Bir13] L. Birgé. Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 47–64. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2013. (Cited on page 99).

[BM93] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993. (Cited on pages 7, 8, and 18).

[BM98] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998. (Cited on page 8).

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001. (Cited on pages 11 and 13).

[BM07] L. Birgé and P. Massart. Minimal Penalties for Gaussian Model Selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. (Cited on pages 10 and 13).

[BLWY06] P. Biswas, T. Liang, T. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, 2(2):188–220, 2006. (Cited on page 27).

[BBZ07] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Mach Learn*, 66(2-3):259–294, 2007. (Cited on pages 22 and 57).

[BHR18a] G. Blanchard, M. Hoffmann, and M. Reiß. Early stopping for statistical inverse problems via truncated SVD estimation. *Electron. J. Stat.*, 12(2):3204–3231, 2018. (Cited on page 9).

[BHR18b] G. Blanchard, M. Hoffmann, and M. Reiß. Optimal adaptation for early stopping in statistical inverse problems. *SIAM-ASA J UNCERTAIN*, 6(3):1043–1075, 2018. (Cited on page 9).

[BKYY16] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Probab. Theory Related Fields*, 164(1-2):459–552, 2016. (Cited on page 68).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, 2013. (Cited on page 48).

[BT15] S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. *Electron. J. Statist.*, 9:2751–2792, 2015. (Cited on page 10).

[Bre01] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. (Cited on page 8).

[BCLZ02] L.D. Brown, T.T. Cai, M.G. Low, and C-H. Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.*, 30(3):688–707, 2002. (Cited on page 5).

[BCLZ04] L.D. Brown, A.V. Carter, M.G. Low, and C-H. Zhang. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.*, 32(5):2074–2097, 2004. (Cited on page 5).

[BL96] L.D. Brown and M.G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996. (Cited on pages 5 and 19).

[Büh06] P. Bühlmann. Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2):559–583, 2006. (Cited on page 9).

[BvdG11] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011. (Cited on pages 2, 3, and 7).

[BN13] A.D. Bull and R. Nickl. Adaptive confidence sets in L^2. *Probab. Theory Related Fields*, 156(3):889–919, 2013. (Cited on pages 21, 26, and 35).

[BSW11] F. Bunea, Y. She, and M.H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 39(2):1282–1309, 2011. (Cited on pages 3, 12, 105, and 112).

[BSW12] F. Bunea, Y. She, and M.H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388, 2012. (Cited on page 3).

[BTW07] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007. (Cited on page 9).

[BI13] C. Butucea and Y.I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013. (Cited on page 3).

[CLL12] T. Cai, W. Liu, and X. Luo. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *J Am Stat Assoc.*, 106(494):594–607, 2012. (Cited on page 4).

[CZ16] T. T. Cai and W. Zhou. Matrix completion via max-norm constrained optimization. *Electron. J. Statist.*, 10(1):1493–1525, 2016. (Cited on pages 25, 26, and 31).

[CG17] T.T. Cai and Z. Guo. Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity. *Ann. Statist.*, 45(2):615–646, 2017. (Cited on pages 15, 20, 23, and 64).

[CG18a] T.T. Cai and Z. Guo. Accuracy assessment for high-dimensional linear regression. *Ann. Statist.*, 46(4):1807–1836, 2018. (Cited on pages 17 and 26).

[CG18b] T.T. Cai and Z. Guo. Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications. *arXiv preprint*, 2018. (Cited on page 15).

[CLR16] T.T. Cai, T. Liang, and A. Rakhlin. Geometric inference for general high-dimensional linear inverse problems. *Ann. Statist.*, 44(4):1536–1563, 2016. (Cited on page 12).

[CLR17] T.T. Cai, T. Liang, and A. Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Ann. Statist.*, 45(4):1404–1430, 2017. (Cited on pages 3 and 12).

[CLZ16] T.T. Cai, W. Liu, and H.H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488, 2016. (Cited on pages 4 and 18).

[CL04a] T.T. Cai and M. Low. Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.*, 32(2):552–576, 2004. (Cited on page 19).

[CL04b] T.T. Cai and M.G. Low. An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, 32(5):1805–1840, 2004. (Cited on page 26).

[CL05] T.T. Cai and M.G. Low. On adaptive estimation of linear functionals. *Ann. Statist.*, 33(5):2311–2343, 2005. (Cited on pages 15 and 19).

[CMW13] T.T. Cai, Z. Ma, and Y. Wu. Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013. (Cited on page 57).

[CRZ13] T.T. Cai, Z. Ren, and H.H. Zhou. Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields*, 156:101–143, 2013. (Cited on page 4).

[CW19] T.T. Cai and Y. Wu. Statistical and Computational Limits for Sparse Matrix Detection. *Ann. Statist., to appear*, 2019. (Cited on pages 3 and 12).

[CZZ10] T.T. Cai, C-H. Zhang, and H.H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010. (Cited on pages 4 and 18).

[CZ12] T.T. Cai and H.H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, 40(5):2389–2420, 2012. (Cited on page 18).

[CP10] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. (Cited on pages 12, 30, and 31).

[CT10] E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. (Cited on page 26).

[CP11] E.J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011. (Cited on pages 12, 26, 31, 101, and 111).

[CR09] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. (Cited on page 25).

[Car13] A. Carpentier. Honest and adaptive confidence sets in L^p. *Electron. J. Statist.*, 7:2875–2923, 2013. (Cited on page 30).

[CEGN15] A. Carpentier, J. Eisert, D. Gross, and R. Nickl. Uncertainty Quantification for Matrix Compressed Sensing and Quantum Tomography Problems. *arXiv preprint*, 2015. (Cited on pages 21, 26, 27, 32, and 33).

[CK18] A. Carpentier and A.K.H. Kim. An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution. *Stat. Sin.*, 28:1371–1393, 2018. (Cited on page 15).

[CKLN18] A. Carpentier, O. Klopp, M. Löffler, and R. Nickl. Adaptive confidence sets for matrix completion. *Bernoulli*, 24(4A):2429–2460, 2018. (Cited on page ix).

[CN15] A. Carpentier and R. Nickl. On signal detection and confidence sets for low rank inference problems. *Electron. J. Statist.*, 9(2):2675–2688, 2015. (Cited on pages 20, 30, 43, and 46).

[CvdV12] I. Castillo and A. van der Vaart. Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101, 2012. (Cited on page 9).

[Cen62] N.N. Cencov. Evaluation of an unknown distribution density from observations. *Soviet. Math.*, 3:1559–1562, 1962. (Cited on page 8).

[Cha15] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015. (Cited on pages 25, 26, and 31).

[CCK13] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013. (Cited on pages 13 and 16).

[CCK14a] V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.*, 42(5):1787–1818, 2014. (Cited on pages 10 and 16).

[CCK14b] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597, 2014. (Cited on page 16).

[CZC+13] E.C. Chi, H. Zhou, G.K. Chen, D.O. Del Vecchyo, and K. Lange. Genotype imputation via matrix completion. *Genome Res.*, 23(3):509–518, 2013. (Cited on pages 20 and 25).

[CSP+07] J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, and W.C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics*, 126(15):155101, 2007. (Cited on pages 24, 100, and 107).

[CT16] J. Chorowski and M. Trabs. Spectral estimation for diffusions with random sampling times. *Stochastic Process. Appl.*, 126(10):2976–3008, 2016. (Cited on pages 104 and 106).

[CvK03] G. Claeskens and I. van Keilegom. Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.*, 31(6):1852–1884, 2003. (Cited on page 16).

[Clé00] S. Clémençon. *Méthodes d'ondelettes pour la statistique non paramétrique des chaines de Markov*. PhD thesis, Université Denis Diderot Paris 7, 2000. (Cited on pages 23, 99, 106, and 107).

[CKL+08] R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008. (Cited on pages 24, 100, and 107).

[CCT17] O. Collier, L. Comminges, and A.B. Tsybakov. Minimax estimation of linear and quadratic functionals under sparsity constraints. *Ann. Statist.*, 45(3):923–958, 2017. (Cited on page 18).

[CCTV18] O. Collier, L. Comminges, A.B. Tsybakov, and N. Verzelen. Optimal adaptive estimation of linear functionals under sparsity. *Ann. Statist.*, 46(6A):3130–3150, 2018. (Cited on pages 15 and 19).

[Cox88] D.D. Cox. Approximation of Least Squares Regression on Nested Subspaces. *Ann. Statist.*, 16(2):713–732, 1988. (Cited on page 8).

[CW79] P. Craven and G. Wahba. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1979. (Cited on page 9).

[DT08] A. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach Learn*, 72:39–61, 2008. (Cited on page 9).

[DGT06] A.S. Dalalyan, G.K. Golubev, and A.B. Tsybakov. Penalized maximum likelihood and semiparametric second-order efficiency. *Ann. Statist.*, 34(1):169–201, 2006. (Cited on page 19).

[DGP18] A.S. Dalalyan, E. Grappin, and Q. Paris. On the Exponentially Weighted Aggregate with the Laplace Prior. *Ann. Statist.*, 46(5):2452–2478, 2018. (Cited on page 9).

[DR06] A.S. Dalalyan and M. Reiß. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields*, 134(2):248–282, 2006. (Cited on page 5).

[DPR82] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, 1982. (Cited on page 57).

[DT82] P.T. Davies and M.K-S. Tso. Procedures for Reduced-Rank Regression. *J R Stat Soc Ser C Appl Stat*, 31(3):244–255, 1982. (Cited on page 3).

[Dem72] A.P. Dempster. Covariance Selection. *Biometrics*, 28(1), 1972. (Cited on page 4).

[DBZ17] R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, 2017. (Cited on page 16).

[Don97] D.L. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997. (Cited on page 8).

[DJ94a] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 84(3):425–455, 1994. (Cited on page 12).

[DJ94b] D.L. Donoho and I.M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields*, 99:277–303, 1994. (Cited on pages 7 and 12).

[DJKP96] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996. (Cited on page 17).

[DL91a] D.L. Donoho and R.C. Liu. Geometrizing rates of convergence, ii. *Ann. Statist.*, 19(2):633–667, 1991. (Cited on page 19).

[DL91b] D.L. Donoho and R.C. Liu. Geometrizing rates of convergence, iii. *Ann. Statist.*, 19(2):668–701, 1991. (Cited on page 19).

[Eat83] M.L. Eaton. *Multivariate Statistics: A Vector Space Approach*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1983. (Cited on pages 3, 64, and 95).

[Efr08] S. Efromovich. Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.*, 36(3):1127–1155, 2008. (Cited on page 9).

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Ann. Statist.*, 32(2):407–451, 2004. (Cited on page 12).

[EK08] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756, 2008. (Cited on page 3).

[FRW15] J. Fan, P. Rigollet, and W. Wang. Estimation of functionals of sparse covariance matrices. *Ann. Statist.*, 43(6):2706–2737, 2015. (Cited on pages 15 and 18).

[Fan61] R.M. Fano. *Transmission Of Information: A Statistical Theory Of Communication*. Cambridge, Mass: MIT Press, 1961. (Cited on page 17).

[GMZ17] C. Gao, Z. Ma, and H.H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.*, 45(5):2074–2101, 2017. (Cited on page 12).

[GvdVZ15] C. Gao, A.W. van der Vaart, and H.H. Zhou. A General Framework for Bayes Structured Linear Models. *arXiv preprint*, 2015. (Cited on page 9).

[GZ15] C. Gao and H.H. Zhou. Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.*, 43(2):785–818, 2015. (Cited on page 57).

[GZ16] C. Gao and H.H. Zhou. Bernstein-von mises theorems for functionals of the covariance matrix. *Electron. J. Statist.*, 10(2):1751–1806, 2016. (Cited on pages 23 and 58).

[Gar53] L. Garding. On the asymptotic distribution of the eigenvalues and eigenfunctions of elliptic differential operators. *MATHEMATICA SCANDINAVICA*, 1:237–255, 1953. (Cited on page 100).

[Gei75] S. Geisser. The Predictive Sample Reuse Method with Applications. *J. Amer. Statist. Assoc.*, 70(350):320–328, 1975. (Cited on page 9).

[GH82] S. Geman and C-R. Hwang. Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *Ann. Statist.*, 10(2):401–414, 1982. (Cited on page 8).

[GvdV17] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge:Cambridge University Press, 2017. (Cited on page 8).

[GL95] R.D. Gill and B.Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995. (Cited on pages 19, 58, and 95).

[GN08a] E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47–61, 2008. (Cited on page 14).

[GN08b] E. Giné and R. Nickl. Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields*, 141(3-4):333–387, 2008. (Cited on page 13).

[GN09a] E. Giné and R. Nickl. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields*, 143(3-4):569–596, 2009. (Cited on page 13).

[GN09b] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Prob.*, 37(4):1605–1646, 2009. (Cited on page 13).

[GN10] E. Giné and R. Nickl. Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170, 2010. (Cited on pages 16 and 26).

[GN16] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Methods*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016. (Cited on pages 3, 4, 8, 10, 16, 17, 19, 26, 27, 30, 35, 43, 48, 51, 53, 72, 104, and 118).

[GHR04] E. Gobet, M. Hoffmann, and M. Reiss. Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.*, 32(5):2223–2253, 2004. (Cited on pages 18, 24, 101, 103, 104, 105, 109, 114, and 115).

[GNOT92] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992. (Cited on pages 20, 25, 26, 27, and 33).

[GL08] A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008. (Cited on page 10).

[GL11] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011. (Cited on page 10).

[GL14] A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probab. Theory Related Fields*, 159(3-4):479–543, 2014. (Cited on page 10).

[GNZ10] G.K Golubev, M. Nussbaum, and H.H. Zhou. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.*, 38(1):181–214, 2010. (Cited on page 5).

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. (Cited on page 8).

[Gre81] U. Grenander. *Abstract Inference*. Wiley, New York, 1981. (Cited on page 8).

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011. (Cited on page 26).

[Gun11] A. Guntuboyina. *Minimax Lower Bounds*. PhD thesis, Yale University, 2011. (Cited on page 17).

[GWCL19] Z. Guo, W. Wang, T.T. Cai, and H. Li. Optimal Estimation of Genetic Relatedness in High-dimensional Linear Models. *J. Amer. Statist. Assoc.*, 114(525):358–369, 2019. (Cited on page 15).

[HWX15] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. In *Proceedings of the 28th Conference on Learning Theory, vol. 40 of Proceedings of Machine Learning Research*. 2015. (Cited on page 12).

[Hal91] P. Hall. On convergence rates of suprema. *Probab. Theory Related Fields*, 89(4):447–455, 1991. (Cited on page 16).

[HH13] P. Hall and J. Horowitz. A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.*, 41(4):1892–1921, 2013. (Cited on page 16).

[HM85] W. Härdle and J.S. Marron. Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *Ann. Statist.*, 13(4):1465–1481, 1985. (Cited on page 9).

[HN11] M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *Ann. Statist.*, 39(5):2383–2409, 2011. (Cited on pages 17 and 26).

[Hör79] L. Hörmander. The Weyl Calculus of Pseudo-Differential Operators. *Comm. Pure Appl. Math.*, 32:359–443, 1979. (Cited on page 100).

[HMZ08] J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high- dimensional regression models. *Statist. Sinica*, 18:1603–1618, 2008. (Cited on page 12).

[Hua04] T-M. Huang. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593, 2004. (Cited on page 9).

[Hub67] P.J Huber. The behavior of maximum likelihood estimates under non-standard conditions. In *Proc. 5th Berkeley Symp. Math. Stat. Probab., vol. 1*, pages 221–233. Berkeley, CA: University of California Press, 1967. (Cited on page 7).

[Hub73] P.J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Statist.*, 1(5):799–821, 1973. (Cited on pages 14 and 23).

[IK82] I.A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation. Asymptotic Theory.* Springer, New York, 1982. (Cited on page 17).

[IS03] Y.I. Ingster and I.A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer New York, 2003. (Cited on page 19).

[ITV10] Y.I. Ingster, A.B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electron. J. Statist.*, 4:1476–1526, 2010. (Cited on pages 20 and 22).

[Ivr00] V. Ivrii. Sharp spectral asymptotics for operators with irregular coefficients. *Int. Math. Res. Notices*, 2000(22):1155–1166, 2000. (Cited on pages 24, 100, 103, and 120).

[Ivr16] V. Ivrii. 100 years of Weyl's law. *Bull. Math. Sci.*, 6(3):379–452, 2016. (Cited on page 100).

[Ize75] A.J. Izenman. Reduced-Rank Regression for the Multivariate Linear Model. *J. Multivariate Anal.*, 5:248–264, 1975. (Cited on page 3).

[JvdG15] J. Janková and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Statist.*, 9(1):1205–1229, 2015. (Cited on page 15).

[Jv18] J. Janková and S. van de Geer. De-biased sparse PCA: Inference and testing for eigenstructure of large covariance matrices. *arXiv preprint*, 2018. (Cited on page 15).

[JvdG18] J. Janková and S. van de Geer. Semi-parametric efficiency bounds for high-dimensional models. *Ann. Statist.*, 46(5):2336–2359, 2018. (Cited on pages 19 and 58).

[JM14] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014. (Cited on pages 14 and 64).

[JM18] A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *Ann. Statist.*, 46(6A):2593–2622, 2018. (Cited on pages 15 and 23).

[JSF18] B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general markov chains. *arXiv preprint*, 2018. (Cited on pages 24, 107, 111, 121, and 122).

[JHW18] J. Jiao, Y. Han, and T. Weissman. Bias Correction with Jackknife, Bootstrap, and Taylor Series. *arXiv preprint*, 2018. (Cited on page 14).

[Joh01] I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001. (Cited on page 57).

[JL09] I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, 104(486):682–693, 2009. (Cited on page 3).

[JS04] I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, 2004. (Cited on pages 7 and 9).

[JLL04] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Math. Methods Statist.*, 12(4):410–428, 2004. (Cited on pages 17 and 26).

[Ken78] J. Kent. Time-Reversible Diffusions. *Adv. Appl. Probab.*, 10:819–835, 1978. (Cited on page 119).

[KP92] G. Kerkyacharian and D. Picard. Density estimation in besov spaces. *Statistics & Probability Letters*, 13(1):15–24, 1992. (Cited on page 17).

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010. (Cited on pages 12, 26, 27, and 31).

[Klo11] O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electron. J. Statist.*, 5:1161–1183, 2011. (Cited on pages 24, 105, 107, 111, and 112).

[Klo14] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014. (Cited on pages 13, 25, 26, 31, 32, 33, and 101).

[Klo15] O. Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electron. J. Statist.*, 9(2):2348–2369, 2015. (Cited on pages 25 and 31).

[KLT17] O. Klopp, K. Lounici, and A.B. Tsybakov. Robust matrix completion. *Probab. Theory Related Fields*, 169(1-2):523–564, 2017. (Cited on page 3).

[KT15] O. Klopp and A.B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015. (Cited on page 3).

[KSvdVvZ16] B.T. Knapik, B.T. Szabó, A.W. van der Vaart, and J.H. van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields*, 164(3):771–813, 2016. (Cited on page 9).

[KWNS18] P. Koltai, H. Wu, F. Noé, and C. Schütte. Optimal Data-Driven Estimation of Generalized Markov State Models for Non-Equilibrium Dynamics. *Computation*, 6(1), 2018. (Cited on pages 100 and 107).

[Kol11] V. Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.*, 39(6):2936–2973, 2011. (Cited on page 3).

[Kol17] V. Koltchinskii. Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators. *arXiv preprint*, 2017. (Cited on pages 14 and 68).

[KLN19] V. Koltchinskii, M. Löffler, and R. Nickl. Efficient Estimation of Linear Functionals of Principal Components. *Ann. Statist., to appear*, 2019. (Cited on page ix).

[KL16] V. Koltchinskii and K. Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(4):1976–2013, 2016. (Cited on pages 23, 58, 61, 63, 64, 66, 67, 70, 71, 74, 76, 83, and 93).

[KL17a] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. (Cited on pages 22, 58, and 60).

[KL17b] V. Koltchinskii and K. Lounici. New asymptotic results in principal component analysis. *Sankhya A*, 79(2):254–297, 2017. (Cited on pages 58 and 79).

[KL17c] V. Koltchinskii and K. Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.*, 45(1), 2017. (Cited on pages 58 and 79).

[KLT11] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy Low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. (Cited on pages 3, 12, 18, 21, 25, 26, 27, 30, 31, 36, 55, 101, 105, and 111).

[KX15] V. Koltchinskii and D. Xia. Optimal Estimation of Low Rank Density Matrices. *J Mach Learn Res.*, 16:1757–1792, 2015. (Cited on pages 24 and 107).

[KZ18] V. Koltchinskii and M. Zhilova. Efficient Estimation of Smooth Functionals in Gaussian Shift Models. *arXiv preprint*, 2018. (Cited on page 14).

[KT93] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction.* Lectures Notes in Statistics 82. Springer, New York, 1993. (Cited on page 17).

[Lac07] C. Lacour. Adaptative estimation of the transition density of a Markov Chain. *Ann. Henri Poincaré*, 43(5):571–597, 2007. (Cited on pages 23 and 99).

[Lac08] C. Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Process. Appl.*, 118(2):232–260, 2008. (Cited on page 106).

[LM15] C. Lacour and P. Massart. Minimal penalty for goldenshluger-lepski method. *Stochastic Process. Appl.*, 126(12):3774–3789, 2015. (Cited on page 10).

[LeC73] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1(1):38–53, 1973. (Cited on page 17).

[LSHM10] M. Lee, H. Shen, J.Z. Huang, and J.S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010. (Cited on page 3).

[LMS97] O.V. Lepski, E. Mammen, and V.G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997. (Cited on page 9).

[Lep90] O.V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466, 1990. (Cited on pages 9, 15, and 19).

[Lep91] O.V. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36(4):682–697, 1991. (Cited on page 9).

[Lep92] O.V. Lepskii. Asymptotically minimax adaptive estimation II: Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.*, 37(3):433–448, 1992. (Cited on pages 9, 15, and 19).

[LWZ18] X. Li, M. Wang, and A. Zhang. Estimation of Markov chain via rank-constrained likelihood. *Proceedings of the 35th International Conference on Machine Learning*, 2018. (Cited on pages 4 and 100).

[LAS16] E. Lila, J.A.D. Aston, and L.M. Sangalli. Smooth Principal Component Analysis over two-dimensional manifolds with an application to Neuroimaging. *Ann. Appl. Stat.*, 10(4):1854–1879, 2016. (Cited on page 57).

[LM72] J. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications*. Springer-Verlag, 1972. (Cited on page 101).

[LG18] H. Liu and C. Gao. Density Estimation with Contaminated Data: Minimax Rates and Theory of Adaptation. *arXiv preprint*, 2018. (Cited on pages 9 and 15).

[LP19] M. Löffler and A. Picard. Spectral thresholding for the estimation of Markov chain transition operators. *arXiv preprint*, 2019. (Cited on page ix).

[Lou14] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014. (Cited on page 4).

[LPVDGT11] K. Lounici, M. Pontil, S. Van De Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011. (Cited on page 3).

[Low97] M. G. Low. On nonparametric confidence intervals. *Ann. Statist.*, 25(6):2547–2554, 1997. (Cited on page 26).

[Ma13] Z. Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801, 2013. (Cited on page 3).

[MMS14] Z. Ma, Z. Ma, and T. Sun. Adaptive estimation in two-way sparse reduced-rank regression. *arXiv preprint*, 2014. (Cited on page 3).

[MW15] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, 2015. (Cited on pages 12 and 18).

[Mal73] C.L. Mallows. Some comments on c_p. *Technometrics*, 15(4):661–675, 1973. (Cited on page 11).

[Mam89] E. Mammen. Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap. *Ann. Statist.*, 17(1):382–400, 1989. (Cited on page 14).

[Mat06] Peter Mathé. The Lepskii principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006. (Cited on page 9).

[MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34(3):1436–1462, 2006. (Cited on page 4).

[MB10] N. Meinshausen and P. Bühlmann. Stability Selection. *J R Stat Soc Series B Stat Methodol*, 72(4):417–473, 2010. (Cited on page 13).

[MGS18] J. Mourtada, S. Gaïffas, and E. Scornet. Minimax optimal rates for mondrian trees and forests. *arXiv preprint*, 2018. (Cited on page 8).

[MPL15] R. Mukherjee, N.S. Pillai, and X. Lin. Hypothesis testing for high-dimensional sparse binary regression. *Ann. Statist.*, 43(1):352–381, 2015. (Cited on page 20).

[Nad08] B. Nadler. Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann. Statist.*, 36(6):2791–2817, 2008. (Cited on page 57).

[NSU18] A.A. Naumov, V.G. Spokoiny, and V.V. Ulyanov. Confidence sets for Spectral Projectors of Covariance Matrices. *Dokl. Math.*, 2018. (Cited on page 58).

[NW12] S. Negahban and M.J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012. (Cited on pages 25, 26, and 31).

[Nic07] R. Nickl. Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields*, 138:411–449, 2007. (Cited on page 13).

[NR19] R. Nickl and K. Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusion processes. *Ann. Statist., to appear*, 2019. (Cited on pages 24, 103, and 120).

[NS17] R. Nickl and J. Söhl. Nonparametric bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, 45(4):1664–1693, 2017. (Cited on pages 111 and 121).

[NS16] R. Nickl and B. Szabó. A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.*, 126(12):3913–3934, 2016. (Cited on page 26).

[Nv13] R. Nickl and S. van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013. (Cited on pages 17, 20, 21, 22, 26, 35, and 43).

[NL17] Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, 45(1):158–195, 2017. (Cited on pages 15 and 58).

[Nor97] J.R. Norris. Long-time behaviour of heat flow: global estimates and exact asymptotics. *Arch. Rational Mech. Anal.*, 140:161–195, 1997. (Cited on page 119).

[NJB+08] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, and C.D. Bustamante. Genes mirror geography within europe. *Nature*, 456:98–101, 2008. (Cited on page 22).

[Nus96] M. Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996. (Cited on page 5).

[Owe07] A.B. Owen. A robust hybrid of lasso and ridge regression. In *Prediction and discovery*, volume 443, pages 59–71. Amer. Math. Soc., Providence, RI, 2007. (Cited on page 13).

[Paj98] A. Pajor. Metric Entropy of the Grassmann Manifold. *Convex Geometric Analysis*, 34:181–188, 1998. (Cited on page 118).

[Pau07] D. Paul. Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. *Statist. Sinica*, 17(4):1617–1642, 2007. (Cited on pages 57 and 68).

[Pfa69] J. Pfanzagl. On the measurability and consistency of minimum contrast estimates. *Metrika*, 14:249–272, 1969. (Cited on page 7).

[Pin80] M.S. Pinsker. Optimal Filtering of Square-Integrable Signals in Gaussian Noise. *Probl. Peredachi Inf.*, 16(2):52–68, 1980. (Cited on page 6).

[Por84] S. Portnoy. Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I.Consistency. *Ann. Statist.*, 12(4):1298–1309, 1984. (Cited on page 14).

[Por85] S. Portnoy. Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. II.Asymptotic normality. *Ann. Statist.*, 13(4):403–1417, 1985. (Cited on page 14).

[RS05] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2005. (Cited on pages 22 and 57).

[RWY11] G. Raskutti, M.J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10), 2011. (Cited on page 18).

[RSH18] K. Ray and J. Schmidt-Hieber. The Le Cam distance between density estimation, Poisson processes and Gaussian white noise. *Mathematical Statistics and Learning*, 1(2), 2018. (Cited on page 5).

[Rec11] B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res*, 12:3413–3430, 2011. (Cited on pages 26 and 31).

[RV98] G. Reinsel and R. Velu. *Multivariate reduced-rank regression: Theory and applications*. New York: Springer, 1998. (Cited on page 3).

[Rei08] M. Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982, 2008. (Cited on page 5).

[RW16] M. Reiss and M. Wahl. Non-asymptotic upper bounds for the reconstruction error of PCA. *Ann. Statist., to appear*, 2016. (Cited on page 58).

[RSZZ15] Z. Ren, T. Sun, C.H. Zhang, and H.H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *Ann. Statist.*, 43(3):991–1026, 2015. (Cited on page 58).

[RT11] P. Rigollet and A.B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011. (Cited on page 18).

[RT12] P. Rigollet and A.B. Tsybakov. Comment: ?minimax estimation of large covariance matrices under ℓ_1-norm? *Statist. Sinica*, 22:1358–1367, 2012. (Cited on page 18).

[RT96] G.O. Roberts and R.L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. (Cited on page 103).

[RVDV07] J. Robins and A.W. Van Der Vaart. Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253, 2007. (Cited on pages 17, 21, 26, and 27).

[RT11] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39 (2):887–930, 2011. (Cited on pages 3, 12, and 25).

[RZMC11] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134:124116, 2011. (Cited on page 100).

[Rou69] G.G. Roussas. Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.*, 21:73–87, 1969. (Cited on page 4).

[RS17] J. Rousseau and B. Szabo. Asymptotic behaviour of the empirical bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.*, 45(2):833–865, 2017. (Cited on page 9).

[Sar14] M. Sart. Estimation of the transition density of a Markov chain. *Ann. Henri Poincaré*, 50(3):1028–1068, 2014. (Cited on page 99).

[SH19] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Ann. Statist., to appear*, 2019. (Cited on page 8).

[Sch98] C. Schütte. Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules. *Habilitation Thesis*, 1998. (Cited on pages 100 and 101).

[SMP14] C. R. Schwantes, R. T. McGibbon, and V. S. Pande. Perspective: Markov models for long-timescale biomolecular dynamics. *The Journal of Chemical Physics*, 141(9):090901, 2014. (Cited on pages 24, 100, 101, and 107).

[Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2), 1978. (Cited on page 11).

[SW94] X. Shen and W.H. Wong. Convergence Rate of Sieve Estimates. *Ann. Statist.*, 22(2):580–615, 1994. (Cited on page 8).

[SHWP15] D. Shukla, C.X. Hernández, J.K. Weber, and V.S. Pande. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Accounts of Chemical Research*, 48(2):414–422, 2015. (Cited on page 100).

[Sin08] A. Singer. A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. U.S.A.*, 105(28):9507–9511, 2008. (Cited on pages 20, 25, and 27).

[Sto80] C.J. Stone. Optimal Rates of Convergence for Nonparametric Estimators. *Ann. Statist.*, 8(6):1348–1360, 1980. (Cited on page 8).

[Sto74] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *J R Stat Soc Series B Stat Methodol*, 36(2):111–147, 1974. (Cited on page 9).

[SC16] W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016. (Cited on page 7).

[SvdVvZ15] B. Szabó, A. van der Vaart, and H. van Zanten. Frequentist coverage of adaptive nonparametric bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428, 2015. (Cited on page 26).

[Tal96] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996. (Cited on pages 21, 52, and 53).

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, 58(1):267–288, 1996. (Cited on pages 11, 12, and 13).

[Tsy98] A.B. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the sobolev classes. *Ann. Statist.*, 26(6):2420–2469, 1998. (Cited on page 6).

[Tsy08] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008. (Cited on pages 3, 17, 118, and 119).

[Tsy14] A.B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians (Seoul, August 2014)*, volume 3, pages 225–246. 2014. (Cited on page 9).

[vdG90] S. van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990. (Cited on pages 8 and 18).

[vdG00] S. van de Geer. *Empirical Processes in M-Estimation*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000. (Cited on page 8).

[vdG01] S. van de Geer. Least squares estimation with complexity penalties. *Maths. Methods Statist*, 10:335–374, 2001. (Cited on page 10).

[vdG16] S. van de Geer. Worst possible sub-directions in high-dimensional models. *J. Multivariate Anal.*, 146:248–260, 2016. (Cited on page 12).

[vdGBRD14] S. van de Geer, P.B. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014. (Cited on pages 14, 19, 23, 58, and 64).

[vdG08] S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2):614–645, 2008. (Cited on page 12).

[vdV98] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, 1998. (Cited on pages 6, 13, 16, 19, and 58).

[vdVvZ09] A.W. van der Vaart and J.H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009. (Cited on page 9).

[vdVW96] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996. (Cited on page 82).

[vT68] H.L. van Trees. *Detection, Estimation and Modulation Theory.* New York: Wiley, 1968. (Cited on pages 19 and 23).

[vWvZ16] J. van Waaij and H. van Zanten. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electron. J. Statist.*, 10(1):628–645, 2016. (Cited on page 103).

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*, pages 210–268. Cambridge University Press, 2012. (Cited on pages 58 and 72).

[VG18] N. Verzelen and E. Gassiat. Adaptive estimation of High-Dimensional Signal-to-Noise Ratios. *Bernoulli*, 24(4B):3683–3710, 2018. (Cited on page 15).

[VL13] V. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41(6):2905–2947, 2013. (Cited on pages 18 and 57).

[WW75] G. Wahba and S. Wold. A completely automatic french curve: fitting spline functions by cross validation. *Comm. Statist. Theory Methods*, 4(1):1–17, 1975. (Cited on page 9).

[WBS16] T. Wang, Q. Berthet, and R.J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, 44(5):1896–1930, 2016. (Cited on page 57).

[WF17] W. Wang and J. Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.*, 45(3):1342–1374, 2017. (Cited on pages 57 and 65).

[Wan13] Y. Wang. Asymptotic equivalence of quantum state tomography and noisy matrix completion. *Ann. Statist.*, 41(5), 2013. (Cited on page 5).

[Wey11] H. Weyl. Über die Asymptotische Verteilung der Eigenwerte. *Nachr. Königl. Ges. Wiss. Göttingen*, pages 110–117, 1911. (Cited on pages 24 and 100).

[WS95] W.H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *Ann. Statist.*, 23(2):339–362, 1995. (Cited on page 8).

[WP03] W.B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003. (Cited on page 4).

[WZ13] Z. Wu and H.H. Zhou. Model selection and sharp asymptotic minimaxity. *Probab. Theory Related Fields*, 156(1-2):165–191, 2013. (Cited on page 7).

[YMB16] D. Yang, Z. Ma, and A. Buja. Rate Optimal Denoising of Simultaneously Sparse and Low Rank Matrices. *J. Mach. Learn. Res*, 17:1–27, 2016. (Cited on page 3).

[YB99] Y. Yang and A.B. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999. (Cited on page 18).

[YM79] V.J. Yohai and R.A. Maronna. Asymptotic Behavior of M-Estimators for the Linear Model. *Ann. Statist.*, 7(2):258–268, 1979. (Cited on page 14).

[Yu97] B. Yu. Assouad, fano, and lecam. In D. Pollard, E. Torgersen, and G.L. Yang, editors, *Festschrift for Lucien Le Cam*. Springer, New York, NY, 1997. (Cited on page 17).

[YELM07] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression. *J R Stat Soc Series B Stat Methodol*, 69(3):329–346, 2007. (Cited on pages 3, 12, and 105).

[YL07] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. (Cited on page 4).

[ZW18] A. Zhang and M. Wang. State Compression of Markov Processes via Empirical Low-Rank Estimation. *ArXiv preprint*, 2018. (Cited on pages 4 and 100).

[Zha05] C.-H. Zhang. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.*, 33(1):54–100, 2005. (Cited on page 9).

[ZZ14] C.H. Zhang and S.S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014. (Cited on pages 14 and 64).

[Zha13] L. Zhang. Nearly optimal minimax estimator for high-dimensional sparse linear regression. *Ann. Statist.*, 41(4):2149–2175, 2013. (Cited on page 9).

[Zha09] T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009. (Cited on page 12).

[ZY05] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Ann. Statist.*, 33(4):1538–1579, 2005. (Cited on page 9).

[ZC17] X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.*, 112(518):757–768, 2017. (Cited on page 16).

[Zou07] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2007. (Cited on page 12).

[ZHT07] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007. (Cited on page 13).