Generalization bounds for random samples in Hilbert spaces

Ilaria Giulini

ENS

RDMath IdF
Domaine d'Intérêt Majeur (DIM) en Mathématiques

île de France

Thèse de doctorat

En vue de l'obtention du grade de

DOCTEUR
DE L'ÉCOLE NORMALE SUPÉRIEURE

École doctorale 386 de sciences mathématiques de Paris-Centre

Spécialité : Mathématiques

Estimation statistique dans les espaces de Hilbert

Generalization bounds for random samples in Hilbert spaces

---

par

Ilaria GIULINI

Présentée et soutenue le 24 septembre 2015 devant le jury composé de:

M. Pierre ALQUIER	*Examinateur*
M. Gérard BIAU	*Examinateur*
M. Stéphane BOUCHERON	*Examinateur*
M. Olivier CATONI	*Directeur de thèse*
M. Nicolò CESA-BIANCHI	*Rapporteur*
M. Stéphane MALLAT	*Examinateur*
M. Pascal MASSART	*Rapporteur*
M. Alexandre TSYBAKOV	*Examinateur*

Contents

Introduction

This thesis focuses on obtaining generalization bounds for random samples in reproducing kernel Hilbert spaces. The approach consists in first obtaining non-asymptotic dimension-free bounds in finite-dimensional spaces using some PAC-Bayesian inequalities and then in generalizing the results to separable Hilbert spaces. We will investigate questions arising in the field of unsupervised classification and more specifically spectral clustering. We will also make some preliminary experiments on the possible relevance of some new spectral clustering algorithms in image analysis.

The goal of image analysis is to recognize and classify patterns that may have been transformed by some set of transformations, such as translations, rotations or scaling, and that more generally are affected by the conditions in which the images have been taken. The challenge is to find a representation in which those transformations of a same pattern are grouped together. However this objective is difficult to achieve even in the case of translations, that represent the simplest kind of transformations. Indeed, translated copies of a single pattern span a high dimensional vector space, so that it is hard to delineate clusters.

What we envision in this thesis is to use kernel methods to characterize clusters as a set of directions in some potentially infinite-dimensional Hilbert space.

Kernel-based methods are a class of statistical learning methods used to detect general types of relations between data. Detecting linear structures in the data has been the subject of much research in statistics and machine learning, and standard techniques, such as linear support vector machines (SVMs) and principal component analysis (PCA [13]), are efficient and well understood. However real-world data analysis problems often have a non-linear structure. Kernel methods provide a workaround. Indeed, they provide a way to deal with non-linear structures in the data with the efficiency of linear algorithms and they have been successfully used for supervised learning (e.g. SVMs) and for unsupervised learning (e.g. kernel-PCA [27]).

An early survey on support vector machines for pattern recognition can be found in Burges [4]. For an introduction to support vector machines and kernel methods we refer to Cristianini and Shawe-Taylor [9] and to Schölkopf and Smola [26]. For a survey on kernel methods we refer to Hofmann, Schölkopf and Smola [12] and to Shawe-Taylor and Cristianini [29].

The basic idea of kernel methods is that any type of data can be implicitly embedded, via a (non-linear) function called a feature map, in a higher-dimensional space, the feature space, in which traditional linear methods become efficient. In other words, they simplify the geometry of the problem but increase the dimension of the space which may now be extremely large (or even infinite).

Kernel-based methods consist of two parts: firstly, the embedding in the feature space is computed and secondly, a learning algorithm is designed to detect the linear structure of the embedded data. The interesting point is that the embedding can be done without ex-

plicit knowledge of the feature map but by simply computing the inner products between the images of the points in the feature space, using a kernel function. This approach is called the *kernel trick*.

In these kernel-based methods an important role is played by kernel random matrices which are matrices of the form $k(X_i, X_j)$ where k is a positive semi-definite kernel and $X_1, \ldots, X_n$ is a random sample.

We will mainly investigate the statistical framework where $X_1, \ldots, X_n$ is a sample of independent and identically distributed (i.i.d.) vectors drawn according to an unknown probability distribution P. The goal is to estimate the integral operator

$$L_k f(x) = \int k(x, y) f(y) \, \mathrm{dP}(y)$$

from the matrix $k(X_i, X_j)$, under suitable assumptions.

This integral operator is related to the Gram operator defined in the feature space, that is a generalization, to infinite dimension, of the Gram matrix. We now present the Gram matrix in finite dimension, since most of the computations made in this thesis are done in finite dimension first, and then generalized to infinite dimension. To be able to go from finite dimension to infinite dimension, we will establish dimension free inequalities.

Let $X \in \mathbb{R}^d$ be a random (column) vector distributed according to P. The Gram matrix G of X is the symmetric positive semi-definite $d \times d$ matrix

$$G = \mathbb{E}(XX^\top),$$

where $\mathbb{E}$ is the expectation with respect to P.

The study of the spectral properties of the Gram matrix is of interest in the case of a non-centered criterion and it coincides, in the case of centered data (i.e. $\mathbb{E}[X] = 0$), with the study of the covariance matrix

$$\Sigma = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^\top\Big].$$

The empirical Gram matrix is defined as

$$\bar{G} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

and it is obtained by replacing the distribution P in the definition of G with the sample distribution $\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, also called the empirical measure.

It is known, by the law of large numbers, that the empirical Gram matrix converges to the true one as the sample size grows to infinity. Instead of the asymptotic regime, we consider a non-asymptotic approach, that provides, for any fixed sample size n, non asymptotic deviation inequalities holding true with probability close to one.

Results on the estimation of the covariance and Gram matrices follow from random matrix theory and the accuracy of the estimation is usually evaluated in terms of the operator norm.

Many theoretical results have been proposed for the study of spectral properties of covariance matrices. Rudelson [25] uses the non-commutative Khintchine inequality to obtain bounds on the sample covariance matrix of a bounded random vector. Some non-asymptotic results are obtained in Vershynin [35] as a consequence of the analysis of

random matrices with independent rows. More precisely he observes that the empirical Gram matrix can be written as $\frac{1}{n}A^\top A$, where A is the matrix whose i-th row is the vector X_i and hence, by construction, A is a matrix with independent rows. A different approach is proposed by Tropp [34]. He observes that the empirical Gram matrix can be expressed as a sum of independent random matrices and proves that the Bernstein inequality extends to matrices, providing an exponential concentration inequality for the operator norm of a sum of independent random matrices.

The empirical Gram matrix becomes less efficient when the data have a long tail distribution, to improve on this, we construct in chapter 1 a more robust estimator and provide non-asymptotic dimension-free bounds of its estimation error. We then extend these bounds to any separable Hilbert space, taking advantage of the fact that they are independent of the dimension. In chapter 2 we use such an estimator to deduce some dimension-independent results for the classical empirical estimator.

Bounds on the deviations of the empirical Gram operator from the true Gram operator in separable Hilbert spaces can be found in Koltchinskii and Lounici [15] in the case of Gaussian random vectors. Similarly to our results, these bounds are dimension-free and they are characterized in terms of the operator norm of the Gram operator and of its trace (more precisely in terms on the effective rank which is the ratio between the trace of the Gram operator and its operator norm).

Many learning algorithms rely on the spectral properties of the covariance matrix, for example principal component analysis (PCA).

Principal component analysis is a classical dimensionality reduction method that transforms the original coordinates into a new reduced set of variables, called principal components, that correspond to the directions where the variance is maximal. Since this set of directions lies in the space generated by the eigenvectors associated with the largest eigenvalues of the covariance matrix of the sample, the dimensionality reduction is achieved by projecting the dataset into the space spanned by these eigenvectors, that in the following we call *largest eigenvectors*.

Asymptotic results regarding the PCA projectors are provided in Biau and Mas [2]. Results on PCA in Hilbert spaces can be found in Koltchinskii and Lounici [16], [17]. The authors study the problem of estimating the spectral projectors of the covariance operator by their empirical counterparts in the case of Gaussian centered random vectors, based on the bounds obtained in [15], and in the setting where both the sample size n and the trace of the covariance operator are large.

In [30], [31], Shawe-Taylor, Williams, Cristianini and Kandola present concentration bounds for sums of eigenvalues of a kernel random matrix and use these results to provide a bound on the performance of (kernel-)PCA. These studies are continued in Zwald, Bousquet and Blanchard [37]. The main idea of these works is to view the kernel random matrix as the empirical version of an underlying integral operator. A first study on the relationship between the spectral properties of a kernel matrix and the corresponding integral operator is done in Koltchinski and Giné [14] for the case of a symmetric square integrable kernel k. They prove that the ordered spectrum of the kernel matrix $K_{ij} = \frac{1}{n}k(X_i, X_j)$ converges to the ordered spectrum of the kernel integral operator L_k.

In [24], Rosasco, Belkin and De Vito study the connection between the spectral properties of the empirical kernel matrix K_{ij} and those of the corresponding integral operator L_k by introducing two extension operators on the reproducing kernel Hilbert space defined by k, that have the same spectrum (and related eigenfunctions) as K and L_k respectively. The

introduction of these extension operators defined on the same Hilbert space overcomes the difficulty of dealing with objects (K and L_k) operating in different spaces.

Several methods have been proposed in the literature in order to provide a more robust version of PCA, e.g. [5], [21]. In [5], Candès, Li, Ma and Wright show that it is possible to recover the principal components of a data matrix in the case where the observations are contained in a low-dimensional space but arbitrarily corrupted by additive noise. An alternative approach is suggested by Minsker in [21] where a robust estimator of the covariance matrix, based on the geometric median, is used to provide non-asymptotic dimension-independent results concerning PCA.

Chapter 3 of this thesis deals with PCA. More precisely, we study how the results presented in the previous chapters lead to a robust PCA without assuming any geometric structure on the data.

The problem of knowing spectral properties of kernel random matrices also arises in the study of spectral clustering. In von Luxburg, Belkin and Bousquet [36] methods similar to the ones described above are used to prove consistency of spectral clustering and to show *the superiority of normalized spectral clustering over unnormalized spectral clustering from a statistical point of view* [36].

Clustering is the task of grouping a set of objects into classes, called clusters, in such a way that objects in the same group are more similar to each other than to those in other groups. Spectral clustering techniques use the spectrum of some data-dependent matrices to perform clustering. These matrices can be either the affinity (or similarity) matrix [10] or the Laplacian matrix [11].

Given $X_1, \ldots, X_n$ a set of points to cluster, the affinity matrix measures the similarity, in terms of the distance, between each pair of points and a common definition is

$$A_{i,j} = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \tag{1}$$

where σ is a free scale parameter. The Laplacian matrix is obtained by rescaling the affinity matrix by its row sums.

Many spectral clustering algorithms have been proposed, e.g. [32], [20], [22].
Shi and Malik [32] introduce the *normalized cut criterion*, in the context of image segmentation, which consists in separating the dataset into two groups by thresholding the second smallest eigenvector of the unnormalized Laplacian matrix $D - A$, where D is the diagonal matrix whose i-th entry is $D_{ii} = \sum_j A_{ij}$. In order to partitioning the dataset into more than two classes, the method has to be applied recursively. To obtain better performances, Meila and Shi [20] propose to use the first c largest eigenvectors of the stochastic matrix $D^{-1}A$, that has the same eigenvectors as the normalized Laplacian matrix $\mathrm{I} - D^{-1}A$, to compute a partition in c classes, where c is assumed to be known. A different algorithm that uses the c largest eigenvectors of the Laplacian matrix $D^{-1/2}AD^{-1/2}$ simultaneously is proposed by Ng, Jordan and Weiss [22]. It consists of two steps: first the dataset is embedded in a space in which clusters are more evident and then clusters are separated using a standard algorithm, e.g. the k-means algorithm. Also in this case the number of clusters is assumed to be known in advance.

Among all the definitions of the Laplacian matrix, we choose

$$L = D^{-1/2}AD^{-1/2}$$

and we view L as the empirical version of the integral operator with kernel

$$\bar{K}(x,y) = \frac{K(x,y)}{\left(\int K(x,z)\mathrm{dP}(z)\right)^{1/2}\left(\int K(y,z)\mathrm{dP}(z)\right)^{1/2}}.$$

Remark that, in the case when the affinity matrix A has the form described in equation (1), the kernel K is the Gaussian kernel

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right). \tag{2}$$

Since connections between empirical operators and their continuous counterparts have been studied in many works, e.g. [24], [36], the study of spectral clustering can start from the study of the kernel $\bar{K}$.

Our generalization bounds hold when spectral clustering is performed in the feature space defined by a reproducing kernel, opening the possibility of a kernel spectral clustering algorithm as described in chapter 4.

Before we give a short overview of the topics and results covered in this thesis, we explain the meaning of *PAC-Bayesian*.

The PAC-Bayesian approach was first introduced by McAllester [19] and then formalized in Seeger [28] and Catoni [6]. It provides deviation bounds (Probably Approximately Correct bounds) for the generalization error, using Bayesian prior and posterior parameter distributions. Those bounds are frequentist, they do not require any Bayesian hypotheses on the unknown sample distribution. The role of prior and posterior parameter distributions here is purely technical, they serve as tools to obtain inequalities holding true uniformly with respect to the parameter, but are not involved in describing the statistical assumptions made about the sample distribution.

Chapter 1 and Chapter 2: Estimation of the Gram matrix

In order to estimate the Gram matrix, we consider the related problem of estimating the quadratic form

$$\theta^\top G\theta = \mathbb{E}[\langle\theta, X\rangle^2], \quad \theta \in \mathbb{R}^d,$$

that computes the energy in the direction θ. Remark that to recover G it is sufficient to use the polarization identity

$$G_{i,j} = e_i^\top G e_j = \frac{1}{4}\left[(e_i+e_j)^\top G(e_i+e_j) - (e_i-e_j)^\top G(e_i-e_j)\right],$$

where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$.

As shown in Catoni [7], if the distribution of $\langle\theta, X\rangle^2$ has a heavy tail for some values of θ, the quality of the approximation provided by the classical empirical estimator can be improved, using some M-estimator with a suitable influence function and a scale parameter depending on the sample size.

We present a robust estimator of the Gram matrix G and we provide non-asymptotic dimension-free bounds on the approximation error using some PAC-Bayesian inequalities related to Gaussian perturbations.

To reduce the influence of the tail of the distribution of $\langle \theta, X\rangle^2$ we truncate the empirical estimator via a continuous bounded function ψ (that is close to the identity in a neighborhood of zero) and we center it by a parameter $\lambda > 0$ so that we get

$$r_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \psi\left(\langle \theta, X_i\rangle^2 - \lambda\right).$$

Let

$$\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \le 0\}$$

and remark that if ψ is the identity, then the empirical estimator is equal to $\lambda/\widehat{\alpha}(\theta)^2$. Therefore it is natural to consider as an estimator of $\theta^\top G\theta$ a quantity related to $\lambda/\widehat{\alpha}(\theta)^2$, for a suitable value of λ.

We first estimate a confidence region for $\theta^\top G\theta$. Starting from the empirical criterion r_λ we perturb the true parameter θ with a Gaussian perturbation centered in θ and with covariance matrix $\beta^{-1}\mathrm{I}$, where $\beta > 0$ is a free parameter. The use of a PAC-Bayesian inequality leads to upper and lower bounds for r_λ that hold, with high probability, uniformly in θ (i.e. for any $\theta \in \mathbb{R}^d$) and for any choice of β. Choosing the perturbation that optimizes the result, we construct a uniform confidence interval for $\theta^\top G\theta$ that we denote by

$$B_-(\theta) := \sup_{(\lambda,\beta)} \Phi_-\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \le \theta^\top G\theta \le \inf_{(\lambda,\beta)} \Phi_+^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) =: B_+(\theta), \tag{3}$$

where Φ_- and Φ_+^{-1} are non-decreasing functions depending on λ, β and on the kurtosis

$$\kappa = \sup_{\substack{\theta\in\mathbb{R}^d \\ \mathbb{E}(\langle\theta,X\rangle^2)>0}} \frac{\mathbb{E}(\langle\theta,X\rangle^4)}{\mathbb{E}\big(\langle\theta,X\rangle^2\big)^2}.$$

We define as an estimator of G the minimal (in terms of the Frobenius norm) symmetric matrix Q whose corresponding quadratic form belongs to the optimal confidence interval (3) for any θ in a finite δ-net of $\mathbb{S}_d$ (the unit sphere of $\mathbb{R}^d$). To be sure that Q is non-negative, we consider as an estimator its positive part Q_+.

The quadratic estimator $\theta^\top Q_+\theta$ is, by construction, more stable than the empirical estimator $\theta^\top \bar{G}\theta$ and it can be explicitly computed using a convex algorithm. In chapter 1 we provide a bound on the approximation error $|\theta^\top G\theta - \theta^\top Q_+\theta|$ that holds, uniformly in θ, under weak moment assumptions and that does not depend explicitly on the dimension d of the ambient space. The result states that, with probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{S}_d$,

$$|\theta^\top G\theta - \theta^\top Q_+\theta| \le 2\max\{\theta^\top G\theta, \sigma\}\mu\left(\max\{\theta^\top G\theta, \sigma\}\right) + 7\delta\sqrt{\mathbf{Tr}(G^2)} + \sigma$$

with $\sigma > 0$ a suitable small threshold and

$$\mu(t) \simeq \sqrt{\frac{2.032(\kappa-1)}{n}\left(\frac{0.73\ \mathbf{Tr}(G)}{t} + 4.35 + \log(\epsilon^{-1})\right)} + \sqrt{\frac{98.5\,\kappa\,\mathbf{Tr}(G)}{nt}} \tag{4}$$

where $\mathbf{Tr}(G) = \mathbb{E}(\|X\|^2)$ denotes the trace of the Gram matrix and $n \le 10^{20}$ (this bound on n is only used to replace $\log(\log(n))$ with a more informative and more striking numerical constant). The threshold σ can be chosen going to zero as the sample size grows

to infinity.

We observe that the quantity $\kappa - 1$ that appears in the first factor of μ corresponds to the variance term, and more precisely to $\mathbf{Var}(\langle\theta, X\rangle^2)/(\theta^\top G\theta)^2$ (at least for the values of θ where κ is reached).

Remark that if we use a known upper bound instead of the exact value of κ, the result also holds for this upper bound. Remark also that $\kappa \leq 3$ in the case when the distribution P is Gaussian. The bound presented above does not depend explicitly on the dimension d, that has been replaced by the entropy term $\mathbf{Tr}(G)/\max\{\theta^\top G\theta, \sigma\}$.

As a consequence, since the bound is dimension-independent, it can be generalized to any infinite-dimensional Hilbert space. More precisely, let $\mathcal{H}$ be a separable Hilbert space and let P be an unknown probability distribution on $\mathcal{H}$. In the infinite-dimensional setting, the analogous to the Gram matrix is the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{G}\theta = \int \langle\theta, v\rangle_{\mathcal{H}}\ v\ \mathrm{dP}(v).$$

We consider an increasing sequence of finite-dimensional subspaces $(\mathcal{H}_k)_k$ such that $\mathcal{H} = \overline{\bigcup_k \mathcal{H}_k}$ and using a continuity argument we define an optimal confidence region for $\langle\mathcal{G}\theta, \theta\rangle_{\mathcal{H}}$ as in equation (3). Given $X_1, \ldots, X_n \in \mathcal{H}$ an i.i.d. sample drawn according to P, we define

$$V_k = \mathbf{span}\{\Pi_k X_1, \ldots, \Pi_k X_n\},$$

where Π_k is the orthogonal projector on $\mathcal{H}_k$, and we consider the operator

$$\mathcal{Q} = \widehat{\mathcal{G}}_k \circ \Pi_{V_k} \tag{5}$$

where Π_{V_k} is the orthogonal projector on V_k and $\widehat{\mathcal{G}}_k$ is a linear operator on V_k such that $\langle\widehat{\mathcal{G}}_k\theta, \theta\rangle_{\mathcal{H}}$ belongs to the optimal confidence region for any θ in a finite δ-net of $\mathbb{S}_{\mathcal{H}} \cap V_k$. Denoting by $\mathcal{Q}_+$ the positive part of $\mathcal{Q}$, we bound the approximation error

$$|\langle\mathcal{G}\theta, \theta\rangle_{\mathcal{H}} - \langle\mathcal{Q}_+\theta, \theta\rangle_{\mathcal{H}}|,$$

uniformly on the unit sphere of $\mathcal{H}$, with the only additional assumption that the trace of the Gram operator $\mathbf{Tr}(\mathcal{G})$ is finite.

At the end of chapter 1 we generalize these results to estimate the expectation of a symmetric random matrix and we consider the problem of estimating the covariance matrix in the case where the expectation of X is unknown. We show that we do not need to estimate $\mathbb{E}[X]$ but that we can divide instead the data points into groups of a given size q and view the dataset as a sample of size n/q of q-tuples. We use the PAC-Bayesian approach to construct a robust estimator of the covariance matrix and to provide non-asymptotic dimension-free bounds on the approximation error. Finally we present an alternative (robust) estimator, related to the estimated matrix Q, which is easier to implement and we provide some empirical results that show the stability of such an estimator with respect to the empirical one.

In chapter 2 we investigate the problem of estimating the Gram matrix via the classical empirical estimator and we provide non-asymptotic dimension-free bounds for the approximation error $|\theta^\top G\theta - \theta^\top \bar{G}\theta|$ using as a tool the robust estimator introduced before.

The results presented above allow us to characterize the stability of principal component analysis independently of the dimension d.

Chapter 3: Principal Component Analysis

We recall that the study of PCA is linked with the study of the eigenvectors corresponding to the largest eigenvalues of the covariance matrix (or of the Gram matrix). In the following we consider the case of the Gram matrix.

Denoting by $p_1, \dots, p_d$ an orthonormal basis of eigenvectors of G, we observe that the i-th eigenvalue of the Gram matrix is $\lambda_i = p_i^\top G p_i$ (and the same holds for the robust estimator Q_+). Starting from this remark, we are able to prove that each eigenvalue in the ordered spectrum of the Gram matrix is well approximated by the corresponding eigenvalue in the ordered spectrum of Q_+, under weak moment assumptions.

From now on we assume that the eigenvectors are ranked according to the decreasing order of their eigenvalues, so that λ_1 is the largest eigenvalue and p_1 the corresponding eigenvector. A method to determine the number of relevant components (which correspond to the largest eigenvectors of G) is based on the difference in magnitude between successive eigenvalues. Let us assume that a significant gap in the spectrum of the Gram matrix is present at $\lambda_r - \lambda_{r+1}$.

Denoting by Π_r (respectively $\widehat{\Pi}_r$) the orthogonal projector on the r largest eigenvectors of G (respectively Q_+), we provide a bound on the approximation error $\|\Pi_r - \widehat{\Pi}_r\|_\infty$, in terms of the operator norm, which only depends on the trace of the Gram matrix, on its largest eigenvalue λ_1 and on the inverse of the size of the eigengap, but does not depend on the dimension d of the ambient space. In particular, the result relates the behavior of the estimator $\widehat{\Pi}_r$ to the size of the spectral gap $\lambda_r - \lambda_{r+1}$ and a good quality of approximation is determined by a large eigengap.

In order to avoid this kind of requirement, we propose to replace the projection on the r largest eigenvectors of the Gram matrix by a smooth cut-off of its eigenvalues. More precisely we aim at estimating $f(G)$ by $f(Q_+)$, where f is a Lipschitz function which is one on the largest eigenvalues and is zero on the smallest ones. In the case where there exists a (sufficiently large) gap in the spectrum of the Gram matrix, the usual projection on the first eigenvectors coincide with a cut-off with a Lipschitz constant exactly equal to the inverse of the size of the eigengap.

The worst case reformulation of our bound on the approximation error in terms of the operator norm depends, as before, on the trace of the Gram matrix and on its largest eigenvalue λ_1 and it replaces the size of the eigengap by the Lipschitz constant. The result states that, with probability at least $1 - 2\epsilon$, for any $1/L$-Lipschitz function f,

$$\|f(G) - f(Q_+)\|_\infty \leq L^{-1}\left[B(\lambda_1) + \sqrt{6B(\lambda_1)\ \mathbf{Tr}(G) + 4B(\lambda_1)^2}\right]$$

where

$$B(\lambda_1) \simeq 4\lambda_1\left[\sqrt{\frac{2.032(\kappa-1)}{n}\left(\frac{0.73\ \mathbf{Tr}(G)}{\lambda_1} + 4.35 + \log(\epsilon^{-1})\right)} + \sqrt{\frac{98.5\,\kappa\,\mathbf{Tr}(G)}{n\lambda_1}}\right].$$

We also provide a similar bound in terms of the Frobenius norm by slightly changing the definition of the estimator Q_+.

Chapter 4: Spectral Clustering

In this last chapter we present a new algorithm for spectral clustering. The idea is to couple spectral clustering with some preliminary change of representation in a reproducing kernel

Hilbert space in order to bring down the representation of classes to a lower-dimensional space.

In order to better explain our approach we briefly describe one of the most successful spectral clustering algorithms [22]. Given a set of points to cluster and the number c of classes, the algorithm introduced by Ng, Jordan, Weiss in [22] computes the c largest eigenvectors of the Laplacian matrix L and put them in columns to form a new matrix X. After renormalization, each row of X is treated as a vector of $\mathbb{R}^c$ and points are clustered according to this new representation using classical methods as k-means.

Let P be an unknown probability distribution on a compact subset of some separable Hilbert space (of possibly infinite dimension). Our approach relies on viewing the Laplacian matrix as the empirical version of the integral operator with kernel

$$\bar{K}(x,y) = \frac{K(x,y)}{\left(\int K(x,z)\mathrm{dP}(z)\right)^{1/2}\left(\int K(y,z)\mathrm{dP}(z)\right)^{1/2}}$$

and on replacing the mere projection on the c largest eigenvectors of L (computed in [22]) by a power of the kernel operator defined by $\bar{K}$. This iteration, justified by the analysis of spectral clustering in terms of Markov chains, performs a smooth truncation of its eigenvalues that leads to a natural dimensionality reduction. Therefore it makes it possible to propose an algorithm that automatically estimates the number of classes (when it is not known in advance).

The algorithm can be described as a change of representation induced by a change of kernel followed by a (greedy) classification. Let us now describe in more detail the change of representation underlying our proposal for spectral clustering.

We use the Laplacian kernel $\bar{K}$ to build the new kernel

$$\bar{K}_{2m}(x,y) = \int \bar{K}(y,z_1)\bar{K}(z_1,z_2)\ldots\bar{K}(z_{2m-1},x)\ \mathrm{dP}^{\otimes(2m-1)}(z_1,\ldots,z_{2m-1}), \quad m>0.$$

Denoting by $\mathcal{H}$ the reproducing kernel Hilbert space defined by $\bar{K}$ and by ϕ the corresponding feature map, the kernel $\bar{K}_{2m}$ defines a new reproducing kernel Hilbert space with feature map $\mathcal{G}^{m-1/2}\circ\phi$, where $\mathcal{G}$ is the Gram operator on $\mathcal{H}$, so that

$$\bar{K}_{2m}(x,y) = \left\langle \mathcal{G}^{m-1/2}\phi(x), \mathcal{G}^{m-1/2}\phi(y)\right\rangle_{\mathcal{H}}.$$

In order to stabilize the representation induced by $\bar{K}_{2m}$ we project this feature map on the unit sphere. This is equivalent to consider the normalized kernel

$$K_m(x,y) = \bar{K}_{2m}(x,x)^{-1/2}\bar{K}_{2m}(x,y)\bar{K}_{2m}(y,y)^{-1/2}.$$

The change of kernel described above is *ideal* since it requires the knowledge of the law P. In order to obtain a practical algorithm we construct, from a random sample of points in $\mathcal{H}$, an estimated version of this ideal change of representation and we prove the convergence of this empirical algorithm by non-asymptotic bounds that are deduced from the bounds obtained for Gram operators in Hilbert spaces.

We first introduce a robust estimator $\widehat{\bar{K}}$ of $\bar{K}$ that can be written as

$$\widehat{\bar{K}}(x,y) = \langle\widehat{\phi}(x),\widehat{\phi}(y)\rangle_{\mathcal{H}},$$

for a suitable (observable) feature map $\widehat{\phi}$, and then we replace it in the definition of $\bar{K}_{2m}$ to obtain the new kernel

$$\bar{H}_{2m}(x,y) = \left\langle \widehat{\mathcal{G}}^{2m-1}\widehat{\phi}(x),\widehat{\phi}(y)\right\rangle_{\mathcal{H}}$$

where $\widehat{\mathcal{G}}$ is the Gram operator defined by

$$\widehat{\mathcal{G}}v = \int \langle v, \widehat{\phi}(z)\rangle_{\mathcal{H}}\ \widehat{\phi}(z)\ \mathrm{dP}(z), \quad v \in \mathcal{H}.$$

Remark that the Gram operator $\widehat{\mathcal{G}}$ is still not observable since it depends on P. According to the results presented for the estimation of Gram operators in Hilbert spaces, we consider as an estimator of $\widehat{\mathcal{G}}$ the operator $\widehat{\mathcal{Q}}$ defined as the positive part of the operator introduced in equation (5), so that we get the estimated kernel

$$\widehat{H}_{2m}(x,y) = \left\langle \widehat{\mathcal{Q}}^{2m-1}\widehat{\phi}(x), \widehat{\phi}(y)\right\rangle_{\mathcal{H}}.$$

The accuracy of this estimation is given in terms of two suitable quantities $\delta_1, \delta_2 > 0$, that are linked with the function μ defined in equation (4) and depend on the sample size n as $1/\sqrt{n}$. More precisely, with probability at least $1 - 4\epsilon$, for any $x, y \in \mathrm{supp}(\mathrm{P})$,

$$|\widehat{H}_{2m}(x,y) - \bar{K}_{2m}(x,y)| \leq 2m\ (2\delta_1 + \delta_2)\frac{(1 + 2\delta_1 + \delta_2)^{2m-2}}{(1-\delta_1)_+^{4m}}\ \|\phi(x)\|_{\mathcal{H}}\|\phi(y)\|_{\mathcal{H}}.$$

Coupling spectral clustering with a preliminary change of representation in a reproducing kernel Hilbert space can be seen as the analogous for unsupervised learning of the support vector machine (SVM) approach to supervised classification. Indeed, first, the kernel trick increases the dimension of the representation to simplify its geometry and then, spectral clustering decreases the size of the representation again. While SVMs compute a separating hyperplane that can be seen as a classification rule bearing on a representation of dimension one (the normal direction to the hyperplane), we show on some examples that the new representation induced by the kernel K_m sends clusters to the neighborhood of an orthonormal set of c vectors (where c is the number of classes), that are therefore the vertices of a (regular) simplex, making subsequent classification a trivial task.

We apply this strategy to image analysis and we suggest with a small example that it is possible to learn transformation invariant representations from datasets that contain small successive transformations of the same pattern.

Chapter 1

The Gram Matrix

Our first goal is to estimate the Gram matrix from an i.i.d. sample. Based on some PAC-Bayesian inequalities we introduce a robust estimator leading to dimension free bounds on the estimation error. In particular, the dimension of the ambient space is replaced by an entropy term that depends on the trace of the Gram matrix.

1.1 Introduction

Let $X \in \mathbb{R}^d$ be a random vector distributed according to the unknown probability measure $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$. The Gram matrix of X

$$G = \mathbb{E}(XX^\top) = \int x\, x^\top \mathrm{dP}(x)$$

is a symmetric positive semi-definite $d \times d$ random matrix. The aim is to estimate the Gram matrix from a given sample of n independent and identically distributed (i.i.d.) vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to P. In relation to this problem, we consider the quadratic form

$$N(\theta) = \int \langle \theta, x \rangle^2 \mathrm{dP}(x), \quad \theta \in \mathbb{R}^d,$$

which computes the energy in the direction θ, where $\langle \theta, x \rangle$ denotes the standard inner product in $\mathbb{R}^d$. It can be seen as the quadratic form associated with the Gram matrix G. Indeed, according to the polarization identity

$$\xi^\top G \theta = \frac{1}{4} \left[N(\xi + \theta) - N(\xi - \theta) \right],$$

we can recover the Gram matrix from the quadratic form N using the formula

$$G_{i,j} = e_i^\top G e_j = \frac{1}{4} \left[N(e_i + e_j) - N(e_i - e_j) \right],$$

where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$.

The classical empirical estimator

$$\bar{N}(\theta) = \frac{1}{n} \sum_{i=1}^n \langle \theta, X_i \rangle^2, \quad \theta \in \mathbb{R}^d, \tag{1.1}$$

is obtained by replacing, in the definition of N, the distribution P by the sample distribution $\frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and it can be seen as the quadratic form associated with the empirical

Gram matrix

$$\bar{G} = \frac{1}{n}\sum_{i=1}^{n} X_i \, X_i^\top .$$

By the law of large numbers, the empirical Gram matrix converges to G almost surely as n goes to infinity.

However, if $\langle \theta, X\rangle^2$ has a heavy tail distribution for at least some values of θ, the quality of approximation provided by the empirical mean estimator can be improved [7].

Using a PAC-Bayesian approach we introduce in section 1.2 a new robust estimator of the Gram matrix and we provide non-asymptotic dimension-free bounds on the approximation error under weak moment assumptions. Since this result does not depend explicitly on the dimension d of the ambient space, it can be generalized to any infinite-dimensional Hilbert space, with the only assumption that the trace of the Gram matrix is finite, as shown in section 1.3.

In section 1.4 we generalize the results to estimate the expectation of a symmetric random matrix, while in section 1.5 we consider the problem of estimating the covariance matrix in the case of unknown expectation. Finally in section 1.6 we propose some empirical results that compare the performance of our robust estimator to that of the classical empirical one.

1.2 Estimate of the Gram matrix

1.2.1 Preliminaries

Our goal is to estimate, for any $\theta \in \mathbb{R}^d$, the quadratic form

$$N(\theta) = \int \langle \theta, x\rangle^2 \mathrm{dP}(x)$$

from an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to the unknown probability distribution $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$.

In order to construct a robust estimator of N we consider a truncated version of the empirical estimator $\bar{N}$. For any $\lambda > 0$ and for any fixed $\theta \in \mathbb{R}^d$, we define

$$r_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\langle \theta, X_i\rangle^2 - \lambda\right),$$

where λ is a centering parameter and where $\psi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\psi(t) = \begin{cases} \log(2) & \text{if } t \geq 1 \\ -\log\left(1 - t + \frac{t^2}{2}\right) & \text{if } 0 \leq t \leq 1 \\ -\psi(-t) & \text{if } t \leq 0. \end{cases} \tag{1.2}$$

The function ψ is symmetric non-decreasing and bounded, it satisfies

$$-\log\left(1 - t + \frac{t^2}{2}\right) \leq \psi(t) \leq \log\left(1 + t + \frac{t^2}{2}\right), \qquad t \in \mathbb{R},$$

and its role is to reduce the influence of the tail of the distribution of $\langle \theta, X\rangle^2$. The introduction of the function ψ is similar to the use of an influence function in robust

statistics, although here we do not use ψ to modify the definition of the mean and we use a scale parameter depending on the sample size. The scale parameter is somehow hidden in our formulation, in fact its role is played by the norm $\|\theta\|$. We do not introduce an explicit scale parameter because it would have been redundant with the ability to change $\|\theta\|$ and because we will obtain results holding uniformly for any $\theta \in \mathbb{R}^d$.

We consider the vector $\widehat{\alpha}(\theta)\theta$, in the direction of θ, where the multiplicative factor $\widehat{\alpha}(\theta)$ is defined by

$$\widehat{\alpha}(\theta) = \sup\left\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \leq 0\right\}. \tag{1.3}$$

Since the function $\alpha \mapsto r_\lambda(\alpha\theta)$ is continuous, $r_\lambda(\widehat{\alpha}(\theta)\theta) = 0$ as soon as $\widehat{\alpha}(\theta) < +\infty$. Considering the fact that the influence function ψ is close to the identity in a neighborhood of zero, we observe that

$$\begin{aligned} 0 = r_\lambda(\widehat{\alpha}(\theta)\theta) &= \frac{1}{n}\sum_{i=1}^n \psi\left(\widehat{\alpha}(\theta)^2\langle\theta, X_i\rangle^2 - \lambda\right) \\ &\simeq \frac{1}{n}\sum_{i=1}^n \widehat{\alpha}(\theta)^2\langle\theta, X_i\rangle^2 - \lambda \\ &= \widehat{\alpha}(\theta)^2\bar{N}(\theta) - \lambda. \end{aligned}$$

This suggests to consider as an estimator of the quadratic form $N(\theta)$ the quantity $\widehat{N}(\theta) = \lambda/\widehat{\alpha}(\theta)^2$, for a suitable value $\lambda > 0$.

In section 1.2.2 we present a non-asymptotic dimension-free bound on the approximation error that holds under weak moment assumptions (Proposition 1.19). However, since this estimator $\widehat{N}$ is (unfortunately) no more a quadratic form, in section 1.2.3 we construct a quadratic estimator for N.

1.2.2 A PAC-Bayesian approach

We use a PAC-Bayesian approach based on Gaussian perturbations of the parameter θ to first construct a confidence region for $N(\theta)$ that is uniform in θ and then to define and study the robust estimator $\widehat{N}$.

Starting from the empirical criterion r_λ, for any $\theta \in \mathbb{R}^d$, we perturb the parameter θ with the Gaussian perturbation $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathrm{I})$ of mean θ and covariance matrix $\beta^{-1}\mathrm{I}$. The parameter $\beta > 0$ is a free real parameter that can be seen as the inverse of the variance of the perturbation in each direction and will be determined later. The introduction of the perturbation π_θ is a technical tool to analyze the average value of N in a neighborhood of θ.

We first introduce some technical lemmas that are useful to derive an upper bound for the empirical criterion r_λ.

Lemma 1.1. *We have*

$$\int \langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta') = \langle\theta, x\rangle^2 + \frac{\|x\|^2}{\beta}.$$

Proof. Let $W \in \mathbb{R}^d$ be a random variable distributed according to π_θ. Since W is a Gaussian random variable with mean θ and covariance matrix $\beta^{-1}\mathrm{I}$, for any $x \in \mathbb{R}^d$, the random variable $\langle W, x\rangle$ is a one-dimensional Gaussian random variable with mean $\langle\theta, x\rangle$ and variance $x^\top(\beta^{-1}\mathrm{I})x = \frac{\|x\|^2}{\beta}$. Consequently, we have

$$\int \langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta') = \mathbb{E}(\langle W, x\rangle^2) = \mathbb{E}(\langle W, x\rangle)^2 + \mathbf{Var}(\langle W, x\rangle) = \langle\theta, x\rangle^2 + \frac{\|x\|^2}{\beta},$$

which concludes the proof. □

As a consequence, we get

$$\psi(\langle\theta, x\rangle^2 - \lambda) = \psi\Big[\int\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta')\Big].$$

What we would like to do is to pull the expectation with respect to π_θ out of the influence function ψ, with a minimal loss of accuracy. To this purpose, we introduce the function

$$\chi(z) = \begin{cases} \psi(z) & z \le z_1 \\ \psi(z_1) + p_1(z - z_1) - (z - z_1)^2/8 & z_1 \le z \le z_1 + 4p_1 \\ \psi(z_1) + 2p_1^2 & z \ge z_1 + 4p_1 \end{cases} \tag{1.4}$$

where $z_1 \in [0, 1]$ is such that $\psi''(z_1) = -1/4$ and p_1 is defined by the condition $p_1 = \psi'(z_1)$. Using the explicit expression of the first and second derivative of ψ,

$$\psi'(z) = \frac{1 - z}{1 - z + z^2/2} \qquad \text{and} \qquad \psi''(z) = \frac{-z + z^2/2}{(1 - z + z^2/2)^2},$$

we can compute z_1 and p_1. We find

$$\begin{aligned} z_1 &= 1 - \sqrt{4\sqrt{2} - 5}, \\ \psi(z_1) &= -\log[2(\sqrt{2} - 1)], \end{aligned}$$

which implies

$$\begin{aligned} p_1 = \psi'(z_1) &= \frac{\sqrt{4\sqrt{2} - 5}}{2(\sqrt{2} - 1)}, \\ \text{and } \sup\chi = \psi(z_1) + 2p_1^2 &= \frac{1 + 2\sqrt{2}}{2} - \log[2(\sqrt{2} - 1)]. \end{aligned}$$

Further we observe that, for any $z \in \mathbb{R}$,

$$\psi(z) \le \chi(z).$$

Indeed, the inequality is trivial for $z \le z_1$, since $\chi(z) = \psi(z)$. For $z \in [z_1, z_1 + 4p_1]$, performing a Taylor expansion at z_1, we obtain that

$$\begin{aligned} \psi(z) &= \psi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 + \int_{z_1}^{z} \frac{\psi'''(u)}{2}(z - u)^2\,\mathrm{d}u \\ &\le \psi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 = \chi(z), \end{aligned}$$

since $\psi'''(u) \le 0$ for $u \in [0, 1[$. Finally we observe that, for any $z \ge z_1 + 4p_1$,

$$\chi(z) = \psi(z_1) + 2p_1^2 > \log(2) \ge \psi(z).$$

On the other hand, for any $z \in \mathbb{R}$,

$$\chi(z) \le \log(1 + z + z^2/2). \tag{1.5}$$

Indeed, for $z \leq z_1$, we have already seen that the inequality is satisfied since $\chi(z) = \psi(z)$. Moreover we observe that the function

$$f(z) = \log(1 + z + z^2/2)$$

is such that $f(z_1) \geq \chi(z_1)$ and also $f'(z_1) \geq \chi'(z_1)$. Performing a Taylor expansion at z_1, we get

$$\begin{aligned} f(z) &= f(z_1) + f'(z_1)(z - z_1) + \int_{z_1}^{z} f''(u)(z-u)^2 \mathrm{d}u \\ &\geq \chi(z_1) + \chi'(z_1)(z - z_1) + \inf f'' \ \frac{(z - z_1)^2}{2}. \end{aligned}$$

Since for any $t \in [z_1, z_1 + 4p_1]$,

$$\inf f'' = f''(\sqrt{3} - 1) = -1/4 = \chi''(t),$$

we deduce that

$$f(z) \geq \chi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 = \chi(z).$$

In particular, $f(z_1 + 4p_1) \geq \chi(z_1 + 4p_1)$. Recalling that f is an increasing function while χ is constant on the interval $[z_1 + 4p_1, +\infty[$, we conclude that, for all $z \in \mathbb{R}$, equation (1.5) is satisfied.

We are now going to prove that the quantity $\psi(\langle \theta, x \rangle^2 - \lambda)$ can be bounded by the expectation of the function χ up to a loss term. Before, we introduce a lemma that allows us to pull the expectation out of the function χ.

Lemma 1.2. *Let Θ be a measurable space. For any $\rho \in \mathcal{M}_+^1(\Theta)$ and any $h \in L_\rho^1(\Theta)$,*

$$\chi\Big(\int h \,\mathrm{d}\rho\Big) \leq \int \chi(h) \,\mathrm{d}\rho + \frac{1}{8} \mathbf{Var}(h \,\mathrm{d}\rho),$$

where by definition

$$\mathbf{Var}(h \,\mathrm{d}\rho) = \int \Big(h(\theta) - \int h \,\mathrm{d}\rho\Big)^2 \mathrm{d}\rho(\theta) \quad \in \mathbb{R} \cup \{+\infty\}.$$

Proof. Performing a Taylor expansion of the function χ at $z = \int h \,\mathrm{d}\rho$ we see that

$$\begin{aligned} \chi[h(\theta)] &= \chi(z) + (h(\theta) - z)\chi'(z) + \int_z^{h(\theta)} [h(\theta) - u]\chi''(u) \,\mathrm{d}u \\ &\geq \chi(z) + (h(\theta) - z)\chi'(z) + \inf \chi'' \ \frac{(h(\theta) - z)^2}{2}. \end{aligned}$$

Remarking that $\inf \chi'' = -1/4$ and integrating the inequality with respect to ρ we obtain

$$\begin{aligned} \int \chi[h(\theta)] \,\mathrm{d}\rho(\theta) &\geq \chi\Big(\int h \,\mathrm{d}\rho\Big) - \frac{1}{8} \int \Big(h(\theta) - \int h(\theta) \mathrm{d}\rho\Big)^2 \mathrm{d}\rho(\theta) \\ &= \chi\Big(\int h \,\mathrm{d}\rho\Big) - \frac{1}{8} \mathbf{Var}(h \,\mathrm{d}\rho), \end{aligned}$$

which concludes the proof. □

Lemma 1.3. *Let Θ be a measurable space. For any $\rho \in \mathcal{M}_+^1(\Theta)$ and any $h \in L_\rho^1(\Theta)$,*

$$\psi\Big(\int h \,\mathrm{d}\rho\Big) \leq \int \chi(h) \,\mathrm{d}\rho + \min\Big\{\log(4), \frac{1}{8}\,\mathbf{Var}(h \,\mathrm{d}\rho)\Big\}.$$

Proof. We observe that

$$\psi\Big(\int h \,\mathrm{d}\rho\Big) - \int \chi(h) \,\mathrm{d}\rho \leq \sup\psi - \inf\chi \leq \log(4),$$

since $\sup\psi \leq \log(2)$ and $\inf\chi \geq \inf\psi \geq -\log(2)$. Moreover, according to Lemma 1.2, we get

$$\begin{aligned}\psi\Big(\int h \,\mathrm{d}\rho\Big) &\leq \chi\Big(\int h \,\mathrm{d}\rho\Big) \\ &\leq \int \chi(h) \,\mathrm{d}\rho + \frac{1}{8}\,\mathbf{Var}(h \,\mathrm{d}\rho).\end{aligned}$$

We conclude by taking the minimum of these two bounds. □

If we apply Lemma 1.3 to our problem we obtain that

$$\begin{aligned}\psi(\langle\theta, x\rangle^2 - \lambda) &= \psi\Big[\int \Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big) \,\mathrm{d}\pi_\theta(\theta')\Big] \\ &\leq \int \chi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big) \,\mathrm{d}\pi_\theta(\theta') + \min\Big\{\log(4), \frac{1}{8}\,\mathbf{Var}\big[\langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta')\big]\Big\}.\end{aligned}$$

Defining $m = \langle\theta, x\rangle$, $\sigma = \frac{\|x\|}{\sqrt{\beta}}$ and $W \sim \mathcal{N}(0, \sigma^2)$ a centered Gaussian random variable, we observe that the variance of the random variable $(m+W)^2$ is equal to $\mathbf{Var}\big[\langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta')\big]$. Hence

$$\begin{aligned}\mathbf{Var}\big[\langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta')\big] &= \mathbf{Var}\big[(m+W)^2\big] \\ &= \mathbf{Var}(W^2 + 2mW) = \mathbb{E}\big[(W^2 + 2mW)^2\big] - \mathbb{E}\big[W^2 + 2mW\big]^2 \\ &= \mathbb{E}(W^4) + 4m^2\sigma^2 - \sigma^4 = 4m^2\sigma^2 + 2\sigma^4\end{aligned}$$

where in the last line we have used the fact that $\mathbb{E}$
$bigl(W^4) = 3\sigma^4$. We conclude that

$$\mathbf{Var}\big[\langle\theta', x\rangle^2 \,\mathrm{d}\pi_\theta(\theta')\big] = \frac{4\langle\theta, x\rangle^2\|x\|^2}{\beta} + \frac{2\|x\|^4}{\beta^2}.$$

We have thus proved that

$$\psi(\langle\theta, x\rangle^2 - \lambda) \leq \int \chi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big) \,\mathrm{d}\pi_\theta(\theta') + \min\Big\{\log(4), \frac{\langle\theta, x\rangle^2\|x\|^2}{2\beta} + \frac{\|x\|^4}{4\beta^2}\Big\}.$$

Lemma 1.4. *For any positive real numbers a, b and c, and any Gaussian random variable $W \sim \mathcal{N}(0, \sigma^2)$,*

$$\min\{a, bm^2 + c\} \leq \mathbb{E}\big[\min\{2a, 2b(m+W)^2 + 2c\}\big].$$

Proof. We first claim that for any positive real numbers a, b, c,

$$\min\{a, bm^2 + c\} \le \min\{a, b(m+W)^2 + c\} + \min\{a, b(m-W)^2 + c\},$$

since $bm^2 + c \le \max\{b(m+W)^2 + c, b(m-W)^2 + c\}$.
Taking the expectation with respect to W of this inequality and remarking that W and $-W$ have the same probability distribution we conclude that

$$\min\{a, bm^2 + c\} \le 2\mathbb{E}\Big[\min\{a, b(m+W)^2 + c\}\Big] = \mathbb{E}\Big[\min\{2a, 2b(m+W)^2 + 2c\}\Big].$$

□

Applying this result to our problem with $a = \log(4)$, $b = \|x\|^2/(2\beta)$, $c = \|x\|^4/(4\beta^2)$, $m = \langle\theta, x\rangle$, and $m + W \sim \langle\theta', x\rangle$ when $\theta' \sim \pi_\theta$, we get

$$\psi(\langle\theta, x\rangle^2 - \lambda) \le \int \chi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta') + \int \min\Big\{4\log(2), \frac{\langle\theta', x\rangle^2\|x\|^2}{\beta} + \frac{\|x\|^4}{2\beta^2}\Big\}\,\mathrm{d}\pi_\theta(\theta').$$

Lemma 1.5. *For any positive constants b, y and any $z \in \mathbb{R}$,*

$$\chi(z) + \min\{b, y\} \le \log\left(1 + z + \frac{z^2}{2} + y\exp(\sup\chi)\frac{\exp(b) - 1}{b}\right).$$

Proof. For any positive real constants a, b, y,

$$\log(a) + \min\{b, y\} = \log(a\exp(\min\{b, y\})) \le \log\Big(a + a\min\{b, y\}\frac{(\exp(b) - 1)}{b}\Big),$$

since the function $x \mapsto \frac{e^x - 1}{x}$ is non-decreasing for $x \ge 0$. It follows that

$$\log(a) + \min\{b, y\} \le \log\big[a + ya\big(\exp(b) - 1\big)/b\big].$$

Applying this inequality to $a = \exp\big[\chi(z)\big]$ and reminding that $\chi(z) \le \log\big(1 + z + z^2/2\big)$, we conclude the proof. □

According to Lemma 1.5, choosing $b = 4\log(2)$, $z = \langle\theta', x\rangle^2 - \|x\|^2/\beta - \lambda$ and $y = \langle\theta', x\rangle^2\|x\|^2/\beta + \|x\|^4/2\beta^2$, we get

$$\psi(\langle\theta, x\rangle^2 - \lambda) \le \int \log\Bigg[1 + \langle\theta' x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda + \frac{1}{2}\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)^2 + \frac{c\|x\|^2}{\beta}\Big(\langle\theta', x\rangle^2 + \frac{\|x\|^2}{2\beta}\Big)\Bigg]\,\mathrm{d}\pi_\theta(\theta'),$$

where

$$c = \frac{15}{4\log(2)}\exp\big(\sup\chi\big) = \frac{15}{8\log(2)(\sqrt{2} - 1)}\exp\Big(\frac{1 + 2\sqrt{2}}{2}\Big) \le 44.3. \tag{1.6}$$

In terms of the empirical criterion r_λ we have proved the following result.

Proposition 1.6. *With the above notation,*

$$r_\lambda(\theta) \leq \frac{1}{n}\sum_{i=1}^{n}\int \log\left[1+\langle\theta',X_i\rangle^2-\frac{\|X_i\|^2}{\beta}-\lambda+\frac{1}{2}\left(\langle\theta',X_i\rangle^2-\frac{\|X_i\|^2}{\beta}-\lambda\right)^2 + \frac{c\|X_i\|^2}{\beta}\left(\langle\theta',X_i\rangle^2+\frac{\|X_i\|^2}{2\beta}\right)\right]\mathrm{d}\pi_\theta(\theta').$$

We are now ready to use the following general purpose PAC-Bayesian inequality.

Proposition 1.7. *Let $\nu \in \mathcal{M}_+^1(\mathbb{R}^d)$ be a prior probability distribution on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \times \mathbb{R}^d \to [a,+\infty]$ be a measurable function with $a > -1$. For any $\epsilon \in]0,1[$, with probability at least $1-\epsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$,*

$$\int \frac{1}{n}\sum_{i=1}^{n} f(X_i,\theta')\,\mathrm{d}\rho(\theta') \leq \int \log \mathbb{E}\big[\exp\big(f(X,\theta')\big)\big]\,\mathrm{d}\rho(\theta') + \frac{\mathcal{K}(\rho,\nu)+\log\big(\epsilon^{-1}\big)}{n} \tag{1.7}$$

where

$$\mathcal{K}(\rho,\nu) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}\right)\mathrm{d}\rho & \text{if } \rho \ll \nu \\ +\infty & \text{otherwise} \end{cases}$$

is the Kullback divergence of ρ with respect to ν. By convention, a non measurable event is said to happen with probability at least $1-\epsilon$ when it includes a measurable event of probability non-smaller than $1-\epsilon$.

Before proving the result we make some comments. Intuitively the above PAC-Bayesian inequality replaces the empirical mean with the expectation with respect to P. Doing this we loose in accuracy and this lost is evaluated by an entropy term which depends on the Kullbach divergence and on the confidence level set by the parameter ϵ.

In the following we consider as prior distribution $\nu = \pi_0 \sim \mathcal{N}(0,\beta^{-1}\mathbf{I})$. Since the result holds for any $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$, if we consider as a family of posterior distributions the Gaussian perturbations $\left\{\pi_\theta \sim \mathcal{N}(\theta,\beta^{-1}\mathbf{I}) \mid \theta \in \mathbb{R}^d,\ \beta > 0\right\}$ we get a bound that is uniform in θ and we can choose β to optimize the bound.

To conclude we observe that with the above choice of prior and posteriors the Kullback divergence is

$$\mathcal{K}(\pi_\theta,\pi_0) = \frac{\beta\|\theta\|^2}{2}.$$

Proof. We divide the proof into two parts.

Step 1. We prove that given $\nu \in \mathcal{M}_+^1(\mathbb{R}^d)$ a prior probability distribution and $f : \mathbb{R}^d \times \mathbb{R}^d \to [a,+\infty]$ a measurable function, there exists a measurable function $F(X_1,\dots,X_n)$ such that

$$\exp\left(\sup_{\substack{\rho\in\mathcal{M}_+^1(\mathbb{R}^d)\\ \mathcal{K}(\rho,\nu)<+\infty}}\left\{\int n\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i,\theta') - \log\mathbb{E}\left[\exp(f(X,\theta'))\right]\right)\times \mathbb{1}\Big[\mathbb{E}\big[\exp\big(f(X,\theta')\big)\big] < +\infty\Big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho,\nu)\right\}\right) \leq F(X_1,\dots,X_n)$$

and $\mathbb{E}\big[F(X_1,\dots,X_n)\big] \le 1$.

To give a meaning to this statement, we define for any measurable function $h : \mathbb{R}^d \to \mathbb{R}$

$$\int h(\theta')\,\mathrm{d}\rho(\theta') = \begin{cases} -\infty & \text{when } \int \min\{h,0\}\,\mathrm{d}\rho = -\infty \\ +\infty & \text{when } \int \min\{h,0\}\,\mathrm{d}\rho > -\infty \text{ and } \int \max\{h,0\}\,\mathrm{d}\rho = +\infty \\ \displaystyle\int h\,\mathrm{d}\rho & \text{when } h \in L^1_\rho(\mathbb{R}^d). \end{cases}$$

We also introduce, for $\theta' \in \mathbb{R}^d$,

$$M(\theta') = \frac{1}{n}\sum_{i=1}^n f(X_i,\theta') \quad \text{and} \quad B(\theta') = \log \mathbb{E}\left[\exp(f(X,\theta'))\right].$$

For any $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$ such that $\mathcal{K}(\rho,\nu) < +\infty$, by definition of the Kullbach divergence,

$$\exp\left(\int n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho,\nu)\right)$$
$$= \exp\left(\int \left[n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big] - \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta')\right)\right]\,\mathrm{d}\rho(\theta')\right).$$

We put

$$h_\rho = n(M-B)\mathbb{1}[B < +\infty] - \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}\right).$$

In the case when $h_\rho \in L^1_\rho(\mathbb{R}^d)$, by the Jensen inequality,

$$\begin{aligned}
&\exp\left(\int n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho,\nu)\right) \\
&\quad\le \int \exp\left(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big] - \log\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta')\Big)\right)\mathrm{d}\rho(\theta') \\
&\quad= \int \exp\left(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big] - \log\Big(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta')\Big)\right)\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta')\,\mathbb{1}\left[\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta') > 0\right]\mathrm{d}\nu(\theta') \\
&\quad= \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\mathbb{1}\left[\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta') > 0\right]\mathrm{d}\nu(\theta') \\
&\quad\le \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\,\mathrm{d}\nu(\theta').
\end{aligned}$$

Else, when $\int \min\{h_\rho,0\}\,\mathrm{d}\rho = -\infty$, then, with our conventions about ill-defined integrals,

$$\exp\Big(\int h_\rho\,\mathrm{d}\rho\Big) = 0 \le \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\,\mathrm{d}\nu(\theta') \in [0,+\infty].$$

Finally, when $\displaystyle\int \min\{h_\rho,0\}\,\mathrm{d}\rho > -\infty$ and $\displaystyle\int \max\{h_\rho,0\}\,\mathrm{d}\rho = +\infty$, from the Lebesgue monotone convergence theorem and the Jensen inequality,

$$\begin{aligned}
+\infty = \exp\Big(\int h_\rho\,\mathrm{d}\rho\Big) &= \sup_{k\in\mathbb{N}} \exp\Big(\int \min\{h_\rho,k\}\,\mathrm{d}\rho\Big) \\
&\le \sup_{k\in\mathbb{N}} \int \exp[\min\{h_\rho,k\}]\,\mathrm{d}\rho = \int \exp(h_\rho)\,\mathrm{d}\rho.
\end{aligned}$$

By using the definition of h_ρ we conclude that

$$\begin{aligned}+\infty &= \exp\Big(\int h_\rho \,\mathrm{d}\rho\Big) \\ &= \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\mathbb{1}\Big[\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta') > 0\Big]\,\mathrm{d}\nu(\theta') \\ &\leq \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\,\mathrm{d}\nu(\theta').\end{aligned}$$

We have then proved that

$$\exp\Big(\int h_\rho \,\mathrm{d}\rho\Big) \leq \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\,\mathrm{d}\nu(\theta').$$

Hence, taking the supremum in ρ, we obtain that

$$\begin{aligned}\exp\Bigg(\sup_{\substack{\rho\in\mathcal{M}_+^1(\mathbb{R}^d)\\ \mathcal{K}(\rho,\nu)<+\infty}} \Big\{\int n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho,\nu)\Big\}\Bigg) \\ \leq \int \exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\,\mathrm{d}\nu(\theta') \stackrel{\text{def}}{=} F(X_1,\ldots,X_n).\end{aligned}$$

The function F defined above is measurable with respect to $(X_1,\ldots,X_n)$ and its expectation satisfies, from the Fubini theorem for non-negative functions,

$$\mathbb{E}\big[F(X_1,\ldots,X_n)\big] = \int \mathbb{E}\Big[\exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\Big]\,\mathrm{d}\nu(\theta').$$

To conclude we need to prove that $\mathbb{E}\Big[\exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\Big] = 1$. When $B(\theta') = +\infty$, this is obvious. When $B(\theta') < +\infty$, this means, by definition of $B(\theta')$, that $\mathbb{E}\Big[\exp\big(f(X,\theta')\big)\Big] < +\infty$. In this case,

$$\begin{aligned}\mathbb{E}\Big[\exp\big[nM(\theta')\big]\Big] &= \mathbb{E}\left[\exp\left(\sum_{i=1}^n f(X_i,\theta')\right)\right] \\ &= \prod_{i=1}^n \mathbb{E}\Big[\exp\big[f(X_i,\theta')\big]\Big] = \exp\big[nB(\theta')\big],\end{aligned}$$

since $X_1,\ldots,X_n$ is an i.i.d. sample, so that

$$\mathbb{E}\Big[\exp\Big(n\big[M(\theta') - B(\theta')\big]\mathbb{1}\big[B(\theta') < +\infty\big]\Big)\Big] = 1.$$

Step 2. By the Markov inequality, since $\mathbb{E}\big[F(X_1,\ldots,X_n)\big] \leq 1$, we have

$$\mathbb{P}\big[F(X_1,\ldots,X_n) \geq \epsilon^{-1}\big] \leq \epsilon.$$

Hence, with probability at least $1-\epsilon$,

$$\log\big(F(X_1,\ldots,X_n)\big) \leq \log(\epsilon^{-1}).$$

The event $\log\big(F(X_1,\dots,X_n)\big) \le \log(\epsilon^{-1})$ is a measurable event included in the, non necessarily measurable, event

$$\mathcal{A} = \left\{ \sup_{\substack{\rho\in\mathcal{M}_+^1(\mathbb{R}^d),\\ \mathcal{K}(\rho,\nu)<+\infty}} \left[\int n\big[M(\theta') - B(\theta')\big] \mathbb{1}\big[B(\theta')<+\infty\big]\,\mathrm{d}\rho(\theta') - \mathcal{K}(\rho,\nu)\right] \le \log(\epsilon^{-1}) \right\}.$$

Let us consider the event studied in the proposition, defined as

$$\mathcal{B} = \Bigg\{ \forall\, \rho \in \mathcal{M}_+^1(\mathbb{R}^d) \text{ s.t. } \mathcal{K}(\rho,\nu) < +\infty, \\ \int M(\theta')\,\mathrm{d}\rho(\theta') \le \int B(\theta')\,\mathrm{d}\rho(\theta') + \frac{\mathcal{K}(\rho,\nu)+\log(\epsilon^{-1})}{n} \Bigg\}.$$

We observe that the inequality imposed on ρ in event $\mathcal{B}$ is trivially satisfied when $\int B(\theta')\,\mathrm{d}\rho(\theta') = +\infty$. Thus, we have to check the inequality only when $\int B(\theta')\,\mathrm{d}\rho(\theta') < +\infty$.

When $\mathcal{A}$ is satisfied, then for any $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$ such that $\mathcal{K}(\rho,\nu)$ is finite, $\int B(\theta')\,\mathrm{d}\rho(\theta') < +\infty$ and $\int \mathbb{1}\big[B(\theta')=\infty\big]\,\mathrm{d}\rho(\theta') = 0$, we have

$$\begin{aligned}\int M(\theta')\,\mathrm{d}\rho(\theta') &= \int M(\theta')\mathbb{1}\big[B(\theta')<+\infty\big]\,\mathrm{d}\rho(\theta') \\ &\le \int B(\theta')\mathbb{1}\big[B(\theta')<+\infty\big]\,\mathrm{d}\rho(\theta') + \frac{\mathcal{K}(\rho,\nu)+\log(\epsilon^{-1})}{n} \\ &= \int B(\theta')\,\mathrm{d}\rho(\theta') + \frac{\mathcal{K}(\rho,\nu)+\log(\epsilon^{-1})}{n},\end{aligned}$$

proving that event $\mathcal{B}$ is also satisfied. In summary, we have proved that

$$\Big\{\log\big(F(X_1,\dots,X_n)\big) \le \log(\epsilon^{-1})\Big\} \subset \mathcal{A} \subset \mathcal{B}.$$

Moreover, since the event $\Big\{\log\big(F(X_1,\dots,X_n)\big) \le \log(\epsilon^{-1})\Big\}$ is measurable and is realized with probability at least $1-\epsilon$, we conclude that, with probability at least $1-\epsilon$, the event $\mathcal{B}$ is also realized, which proves the proposition. □

We now come back to our problem and we apply the PAC-Bayesian inequality (1.7) to Proposition 1.6 in order to obtain a uniform bound on the empirical criterion. As already said, the prior is $\nu = \pi_0 \sim \mathcal{N}(0, \beta^{-1}\mathbf{I})$ and the family of posterior distributions is

$$\Big\{\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathbf{I}) \mid \theta \in \mathbb{R}^d,\ \beta > 0\Big\}.$$

In what follows, the parameter ϵ can take any value in the interval $]0,1[$.

Proposition 1.8. *With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$r_\lambda(\theta) \le \int \mathbb{E}\left[t(X,\theta') + \frac{1}{2}t(X,\theta')^2 + \frac{c\|X\|^2}{\beta}\left(\langle\theta',X\rangle^2 + \frac{\|X\|^2}{2\beta}\right)\right]\mathrm{d}\pi_\theta(\theta') \\ + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n},$$

where $t(x,\theta') = \langle\theta',x\rangle^2 - \dfrac{\|x\|^2}{\beta} - \lambda$ and $c = \dfrac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\dfrac{1+2\sqrt{2}}{2}\right)$.

Proof. We apply the PAC-Bayesian inequality (1.7) to Proposition 1.6 with

$$f(X_i, \theta') = \log\left[1 + t(X_i, \theta') + \frac{1}{2}t(X_i, \theta')^2 + \frac{c\|X_i\|^2}{\beta}\left(\langle \theta', X_i\rangle + \frac{\|X_i\|^2}{2\beta}\right)\right]. \tag{1.8}$$

We obtain that, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned} r_\lambda(\theta) &\leq \frac{1}{n}\sum_{i=1}^n \int f(X_i, \theta')\mathrm{d}\pi_\theta(\theta') \\ &\leq \int \log \mathbb{E}\left[\exp(f(X, \theta'))\right] \mathrm{d}\pi_\theta(\theta') + \frac{\mathcal{K}(\pi_\theta, \pi_0) + \log(\epsilon^{-1})}{n}, \end{aligned}$$

where $\mathcal{K}(\pi_\theta, \pi_0) = \frac{\beta\|\theta\|^2}{2}$. Observing that

$$\begin{aligned} \log \mathbb{E}\big[\exp(f(X, \theta'))\big] &= \log\left[1 + \mathbb{E}\left[t(X, \theta') + \frac{1}{2}t(X, \theta')^2 + \frac{c\,\|X\|^2}{\beta}\left(\langle \theta', X\rangle^2 + \frac{\|X\|^2}{2\beta}\right)\right]\right] \\ &\leq \mathbb{E}\left[t(X, \theta') + \frac{1}{2}t(X, \theta') + \frac{c\|X\|^2}{\beta}\left(\langle \theta', X\rangle^2 + \frac{\|X\|^2}{2\beta}\right)\right], \end{aligned}$$

since $\log(1+t) \leq t$, we conclude the proof. □

In the following we introduce a sequence of results that provide more explicit bounds on the empirical criterion.

We first integrate explicitly with respect to the Gaussian distribution π_θ.

Proposition 1.9. *With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$\begin{aligned} r_\lambda(\theta) \leq \mathbb{E}\Bigg[\langle \theta, X\rangle^2 - \lambda + \frac{1}{2}\Bigg(\Big(\langle \theta, X\rangle^2 - \lambda\Big)^2 + \frac{4\langle \theta, X\rangle^2\|X\|^2}{\beta} + \frac{2\|X\|^4}{\beta^2}\Bigg) \\ + \frac{c\|X\|^2}{\beta}\left(\langle \theta, X\rangle^2 + \frac{3\|X\|^2}{2\beta}\right)\Bigg] + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}. \end{aligned}$$

Proof. The proof is based on the identities

$$\begin{aligned} \int \langle \theta', X\rangle^2 \,\mathrm{d}\pi_\theta(\theta') &= \langle \theta, X\rangle^2 + \frac{\|X\|^2}{\beta}, \\ \mathbf{Var}\big[\langle \theta', X\rangle^2 \,\mathrm{d}\pi_\theta(\theta')\big] &= \frac{4\langle \theta, X\rangle^2\|X\|^2}{\beta} + \frac{2\|X\|^4}{\beta}, \end{aligned}$$

that we already proved before.

□

Let us introduce

$$s_4 = \mathbb{E}\big(\|X\|^4\big)^{1/4}, \tag{1.9}$$

$$\kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}(\langle \theta, X\rangle^2) > 0}} \frac{\mathbb{E}\big(\langle \theta, X\rangle^4\big)}{\mathbb{E}\big(\langle \theta, X\rangle^2\big)^2}, \tag{1.10}$$

assuming that these two quantities are finite. Using the Cauchy-Schwarz inequality we rewrite the bound in Proposition 1.9 as follows.

Proposition 1.10. *With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$r_\lambda(\theta) \le \frac{\kappa}{2}\big[N(\theta)-\lambda\big]^2 + \left[1+(\kappa-1)\lambda + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta}\right]\big[N(\theta)-\lambda\big] + \frac{(\kappa-1)\lambda^2}{2} + \frac{(2+c)\kappa^{1/2}s_4^2\lambda}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2} + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}. \quad (1.11)$$

Proof. The computation is based on the two inequalities

$$\mathbb{E}\big(\langle X,\theta\rangle^4\big) \le \kappa N(\theta)^2,$$
$$\mathbb{E}\big(\langle\theta,X\rangle^2\|X\|^2\big) \le \kappa^{1/2}s_4^2 N(\theta),$$

and some elementary grouping of factors.

□

We now consider a finite set $\Lambda \subset \big(\mathbb{R}_+\setminus\{0\}\big)^2$ of possible values of the couple of parameters (λ,β) that will be determined later. Let $|\Lambda|$ denote the number of elements of Λ.

Proposition 1.11. *For any choice of $(\lambda,\beta)\in\Lambda$, we put*

$$\begin{aligned}
\xi &= \frac{\kappa\lambda}{2},\\
\mu &= \lambda(\kappa-1) + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta},\\
\gamma &= \frac{\lambda}{2}(\kappa-1) + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2\lambda} + \frac{\log(|\Lambda|/\epsilon)}{n\lambda},\\
\delta &= \frac{\beta}{2n\lambda},
\end{aligned} \quad (1.12)$$

where $c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right)$, s_4 is defined by equation (1.9), and κ is defined by equation (1.10).

With probability at least $1-\epsilon$, for any $\theta\in\mathbb{R}^d$, any $(\lambda,\beta)\in\Lambda$,

$$\frac{r_\lambda(\theta)}{\lambda} \le \xi\left(\frac{N(\theta)}{\lambda}-1\right)^2 + (1+\mu)\left(\frac{N(\theta)}{\lambda}-1\right) + \gamma + \delta\|\theta\|^2. \quad (1.13)$$

Moreover, with probability at least $1-\epsilon$, for any $\theta\in\mathbb{R}^d$, any $(\lambda,\beta)\in\Lambda$,

$$\frac{r_\lambda(\theta)}{\lambda} \ge -\xi\left(\frac{N(\theta)}{\lambda}-1\right)^2 + (1-\mu)\left(\frac{N(\theta)}{\lambda}-1\right) - \gamma - \delta\|\theta\|^2. \quad (1.14)$$

Proof. The first inequality is just a matter of rewriting the previous proposition with more compact notation and taking a union bound with respect to the allowed values of $(\lambda,\beta)\in\Lambda$. To obtain the second inequality, we observe that

$$-r_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \psi(\lambda - \langle\theta,X_i\rangle^2).$$

Proceeding as previously done, with the necessary signs updates, we obtain that with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,

$$-r_\lambda(\theta) \le \int \mathbb{E}\Bigg[-t(X,\theta') + \frac{1}{2}t(X,\theta')^2 + \frac{c\|X\|^2}{\beta}\Big(\langle \theta', X\rangle^2 + \frac{\|X\|^2}{2\beta}\Big)\Bigg] \mathrm{d}\pi_\theta(\theta') + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(|\Lambda|/\epsilon)}{n}.$$

Explicitly integrating with the respect to the Gaussian distribution π_θ and rewriting the bound with compact notation, we conclude the proof. □

From the above Proposition 1.11 we deduce a confidence interval for the quadratic form $N(\theta)$ in terms of $\lambda/\widehat{\alpha}(\theta)^2$, where $\widehat{\alpha}(\theta)$ has been defined in equation (1.3) on page 17. More precisely the following proposition holds.

Proposition 1.12. *With probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,*

$$N(\theta) \le \Phi^{-1}_{\theta,+}\Big(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\Big), \tag{1.15}$$

where $\Phi_{\theta,+}$ is the non-decreasing function defined as

$$\Phi_{\theta,+}(t) = t\left(1 + \frac{\gamma + \delta\lambda\|\theta\|^2/t}{1-\mu-\gamma-2\delta\lambda\|\theta\|^2/t}\right)^{-1} \mathbb{1}\Big[\xi+\mu+\gamma+2\delta\lambda\|\theta\|^2/t < 1\Big].$$

Moreover, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,

$$\Phi_{\theta,-}\Big(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\Big) \le N(\theta), \tag{1.16}$$

where

$$\begin{aligned}\Phi_{\theta,-}(t) &= t\left(1 - \frac{\gamma + \delta\lambda\|\theta\|^2/t}{1+\mu-\gamma-\delta\lambda\|\theta\|^2/t}\right)_+ \mathbb{1}\Big[\xi-\mu+\gamma+\delta\lambda\|\theta\|^2/t < 1\Big] \\ &= t\left(1 - \frac{\gamma + \delta\lambda\|\theta\|^2/t}{1+\mu-\gamma-\delta\lambda\|\theta\|^2/t}\right) \mathbb{1}\Big[\xi-\mu+2\gamma+2\delta\lambda\|\theta\|^2/t < 1\Big]\end{aligned}$$

is an non-decreasing function.

Proof. We first observe that the polynomial

$$-\xi\left(\frac{N(\theta)}{\lambda} - 1\right)^2 + (1-\mu)\left(\frac{N(\theta)}{\lambda} - 1\right) - \gamma - \delta\|\theta\|^2$$

can be written as

$$-\xi\left(\frac{N(\theta)}{\lambda} - 1\right)^2 + \left(1 - \mu - \frac{\lambda\delta\|\theta\|^2}{N(\theta)}\right)\left(\frac{N(\theta)}{\lambda} - 1\right) - \gamma - \frac{\lambda\delta\|\theta\|^2}{N(\theta)}.$$

In order to simplify notation, we introduce the functions $\tau(\theta) = \dfrac{\lambda\delta\|\theta\|^2}{N(\theta)}$ and

$$p_\theta(z) = -\xi z^2 + \big[1-\mu-\tau(\theta)\big]\, z - \gamma - \tau(\theta), \qquad z \in \mathbb{R}.$$

We observe that $\tau(\alpha\theta) = \tau(\theta)$, and consequently $p_{\alpha\theta}(z) = p_\theta(z)$ for any $\alpha \in \mathbb{R}_+$. We consider the case when $p_\theta(1) > 0$, meaning that

$$\xi + \mu + \gamma + 2\tau(\theta) < 1. \tag{1.17}$$

In this case, the second degree polynomial p_θ has two distinct real roots, z_{-1} and z_{+1}, where

$$z_\sigma = \frac{1 - \mu - \tau(\theta) + \sigma\sqrt{\big[1 - \mu - \tau(\theta)\big]^2 - 4\xi\big[\gamma + \tau(\theta)\big]}}{2\xi}, \qquad \sigma \in \{1, -1\}.$$

We claim that

$$z_{-1} < \frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)} < z_{+1},$$

which directly follows once we have shown that

$$p_\theta\bigg(\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg) > 0. \tag{1.18}$$

Observing that equation (1.17) can also be written as $-\xi > -\big[1 - \mu - \gamma - 2\tau(\theta)\big]$, we get

$$\begin{aligned}
&p_\theta\bigg(\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg) \\
&\qquad = -\xi\bigg(\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg)^2 + \big[1 - \mu - \tau(\theta)\big]\bigg(\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg) - \gamma - \tau(\theta) \\
&\qquad > \frac{-\big[\gamma + \tau(\theta)\big]^2}{1 - \mu - \gamma - 2\tau(\theta)} + \big[1 - \mu - \tau(\theta)\big]\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)} - \gamma - \tau(\theta) = 0,
\end{aligned}$$

proving hence equation (1.18). We recall that by definition of $\widehat{\alpha}$ given in equation (1.3) on page 17, $r_\lambda(\alpha\theta) \le 0$, for any $\alpha \le \widehat{\alpha}(\theta)$. Thus, according to equation (1.14), with probability at least $1 - \epsilon$, for any $\alpha \in [0, \widehat{\alpha}(\theta)]$,

$$p_\theta\left(\frac{\alpha^2 N(\theta)}{\lambda} - 1\right) \le \frac{r_\lambda(\alpha\theta)}{\lambda} \le 0.$$

This means that

$$\left[-1, \frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1\right] \cap \,]z_{-1}, z_{+1}[\, = \varnothing.$$

Since $p_\theta(1) > 0 > p_\theta(0)$, then $z_{-1} \ge 0 > -1$. It follows that $\widehat{\alpha}(\theta)^2 N(\theta)/\lambda - 1 \le z_{-1}$. This proves that, with probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$ satisfying equation (1.17),

$$N(\theta) \le \frac{\lambda}{\widehat{\alpha}(\theta)^2}(1 + z_{-1}) \le \frac{\lambda}{\widehat{\alpha}(\theta)^2}\bigg(1 + \frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg).$$

We observe that the last inequality

$$N(\theta) \le \frac{\lambda}{\widehat{\alpha}(\theta)^2}\bigg(1 + \frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\bigg)$$

can be written, when equation (1.17) is satisfied, as

$$\Phi_{\theta,+}\big[N(\theta)\big] \le \frac{\lambda}{\widehat{\alpha}(\theta)^2}.$$

Moreover, this inequality is trivially true when condition (1.17) is not satisfied, because its left-hand side is equal to zero and its right-hand side is non-negative. Thus, we have proved that with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\Phi_{\theta,+}\big[N(\theta)\big] \leq \frac{\lambda}{\widehat{\alpha}(\theta)^2}.$$

Proving the second part of the proposition requires a new argument and not a mere update of signs in the proof of the first part. Although it may seen at first sight that we are just aiming at a reverse inequality, the situation is more subtle than that.
Let us first remark that in the case when

$$\xi - \mu + \gamma + \delta\widehat{\alpha}(\theta)^2\|\theta\|^2 < 1, \tag{1.19}$$

is not satisfied, equation (1.16) is trivially satisfied because $\Phi_{\theta,-}\left(\dfrac{\lambda}{\widehat{\alpha}(\theta)^2}\right) = 0$. Now, in the case when equation (1.19) is true, $\widehat{\alpha}(\theta) < +\infty$, so that $r_\lambda\big[\widehat{\alpha}(\theta)\theta\big] = 0$, and therefore, according to equation (1.13), with probability at least $1-\epsilon$,

$$0 \leq q_{\widehat{\alpha}(\theta)\theta}\left(\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1\right),$$

where

$$q_\theta(z) = \xi z^2 + (1+\mu)z + \gamma + \delta\|\theta\|^2.$$

Since condition (1.19) can also be written as $q_{\widehat{\alpha}(\theta)\theta}(-1) < 0$, it implies that the second order polynomial $q_{\widehat{\alpha}(\theta)\theta}$ has two roots and that $\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1$ is on the right of its largest root, which is larger than -1. On the other hand, we observe that, under condition (1.19), putting $\widehat{\tau}(\theta) = \delta\widehat{\alpha}(\theta)^2\|\theta\|^2$, we get

$$\begin{aligned} q_{\widehat{\alpha}(\theta)\theta}\left(-\frac{\gamma+\widehat{\tau}(\theta)}{1+\mu-\gamma-\widehat{\tau}(\theta)}\right) &= \xi\left(\frac{\gamma+\widehat{\tau}(\theta)}{1+\mu-\gamma-\widehat{\tau}(\theta)}\right)^2 - (1+\mu)\left(\frac{\gamma+\widehat{\tau}(\theta)}{1+\mu-\gamma-\widehat{\tau}(\theta)}\right) + \gamma + \widehat{\tau}(\theta) \\ &< \frac{\big(\gamma+\widehat{\tau}(\theta)\big)^2}{1+\mu-\gamma-\widehat{\tau}(\theta)} - \frac{(1+\mu)\big[\gamma+\widehat{\tau}(\theta)\big]}{1+\mu-\gamma+\widehat{\tau}(\theta)} + \gamma + \widehat{\tau}(\theta) = 0. \end{aligned}$$

Therefore, with probability at least $1-\epsilon$, if condition (1.19) is satisfied,

$$\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1 \geq -\frac{\gamma+\widehat{\tau}(\theta)}{1+\mu-\gamma-\widehat{\tau}(\theta)},$$

which rewrites as equation (1.16) under condition (1.19). Hence, we conclude that with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$, equation (1.16) holds. □

To simplify notation, we omit the dependence on θ of the functions defined above, so that, from now on, we write Φ_- and Φ_+ instead of $\Phi_{\theta,-}$ and $\Phi_{\theta,+}$.

Now that we have an observable non-asymptotic confidence interval for $N(\theta)$, we are going to define and study in the following an estimator for $N(\theta)$. Let $\Lambda \subset \big(\mathbb{R}_+ \setminus \{0\}\big)^2$ be a finite set as before. We consider, for some energy level $\sigma \in \mathbb{R}_+$ to be chosen later, the set

$$\Gamma = \left\{(\lambda,\beta,t) \in \Lambda \times \mathbb{R}_+ \mid \xi + \mu + \gamma + 2\frac{\delta\lambda}{\max\{t,\sigma\}} < 1\right\}$$

and the bound

$$B_{\lambda,\beta}(t) = \begin{cases} \dfrac{\gamma + \lambda\delta/\max\{t,\sigma\}}{1-\mu-\gamma-2\lambda\delta/\max\{t,\sigma\}} & (\lambda,\beta,t)\in\Gamma \\ +\infty & \text{otherwise.} \end{cases} \tag{1.20}$$

We consider the family of estimators

$$\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2}, \quad \theta\in\mathbb{R}^d, \tag{1.21}$$

and we observe that, since $\widehat{\alpha}$ is homogeneous (of degree -1) in θ,

$$\widetilde{N}_\lambda(\theta) = \|\theta\|^2 \widetilde{N}_\lambda(\theta/\|\theta\|).$$

Before defining our estimator and presenting some uniform bounds on the estimation error, we introduce a technical lemma.

Lemma 1.13. *With probability at least $1-\epsilon$, for any $\theta\in\mathbb{R}^d$, any $(\lambda,\beta)\in\Lambda$,*

$$\max\{N(\theta),\sigma\|\theta\|^2\}\Big[1+B_{\lambda,\beta}\Big(\|\theta\|^{-2}N(\theta)\Big)\Big]^{-1} \le \max\{\widetilde{N}_\lambda(\theta),\sigma\|\theta\|^2\}. \tag{1.22}$$

With probability at least $1-\epsilon$, for any $\theta\in\mathbb{R}^d$, any $(\lambda,\beta)\in\Lambda$,

$$\max\{\widetilde{N}_\lambda(\theta),\sigma\|\theta\|^2\}\Big[1-B_{\lambda,\beta}\Big(\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\Big)\Big] \le \max\{N(\theta),\sigma\|\theta\|^2\}.$$

With probability at least $1-\epsilon$, for any $\theta\in\mathbb{R}^d$, any $(\lambda,\beta)\in\Lambda$,

$$\max\{N(\theta),\sigma\|\theta\|^2\} \le \max\{\widetilde{N}_\lambda(\theta),\sigma\|\theta\|^2\}\Big[1+B_{\lambda,\beta}\Big(\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\Big)\Big].$$

Proof. We may assume, without loss of generality, that $\|\theta\|=1$. We observe that, for any $z,t,\sigma\in\mathbb{R}_+$, if $\Phi_+(z)\le t$ then

$$\Phi_+(z)\le\max\{t,\sigma\}.$$

Moreover, since by definition of Φ_+, we have $\Phi_+(\sigma)\le\sigma$, in particular

$$\Phi_+(\sigma)\le\max\{t,\sigma\}.$$

We have then proved that, if $\Phi_+(z)\le t$, then

$$\Phi_+\big(\max\{z,\sigma\}\big)\le\max\{t,\sigma\}.$$

The same reasoning applies to Φ_- proving that, if $\Phi_-(z)\le t$, then

$$\Phi_-\big(\max\{z,\sigma\}\big)\le\max\{t,\sigma\}.$$

We also observe that, by definition of $B_{\lambda,\beta}$, for any $(\lambda,\beta)\in\Lambda$,

$$\Phi_+\big(\max\{z,\sigma\}\big) = \max\{z,\sigma\}\Big(1+B_{\lambda,\beta}(z)\Big)^{-1} \tag{1.23}$$

$$\Phi_-\big(\max\{z,\sigma\}\big) \ge \max\{z,\sigma\}\Big(1-B_{\lambda,\beta}(z)\Big). \tag{1.24}$$

Applying these results to Proposition 1.12 we obtain the first two inequalities.

To obtain the third inequality, we recall that, by definition of inverse function,

$$\Phi_+^{-1}(t) = \sup\{y \mid \Phi_+(y) \leq t\}.$$

Denoting $\bar{y} = \sup\{y \mid \Phi_+(y) \leq t\}$, we get

$$\Phi_+^{-1}\big(\max\{t,\sigma\}\big) = \max\{\bar{y},\sigma\} \leq \max\{t,\sigma\}\Big(1 + B_{\lambda,\beta}(t)\Big), \tag{1.25}$$

since $t \leq \bar{y}$ and $B_{\lambda,\beta}$ is a non-increasing function. To conclude, it is sufficient to apply equation (1.25) to inequality

$$N(\theta) \leq \Phi_+^{-1}\Big(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\Big).$$

□

Let us put

$$(\widehat{\lambda},\widehat{\beta}) = \arg\min_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\Big[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\Big].$$

We define our robust estimator $\widehat{N}$ as

$$\widehat{N}(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta) \tag{1.26}$$

where $\widetilde{N}_\lambda$ is defined in equation (1.21).

Next proposition provides a (uniform) bound on the approximation error.

Proposition 1.14. *With probability at least* $1-2\epsilon$*, for any* $\theta \in \mathbb{R}^d$*,*

$$\left|\frac{\max\{N(\theta),\sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta),\sigma\|\theta\|^2\}} - 1\right| \leq \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\Bigg(\frac{\|\theta\|^{-2}N(\theta)}{1 + B_{\lambda,\beta}\Big[\|\theta\|^{-2}N(\theta)\Big]}\Bigg).$$

Proof. We may assume that $\|\theta\| = 1$ without loss of generality, from homogeneity considerations. By Lemma 1.13 with the choice of parameters $(\widehat{\lambda},\widehat{\beta})$, we get that with probability at least $1-2\epsilon$,

$$\left|\frac{\max\{N(\theta),\sigma\}}{\max\{\widehat{N}(\theta),\sigma\}} - 1\right| \leq B_{\widehat{\lambda},\widehat{\beta}}\left(\widehat{N}(\theta)\right).$$

Since, by definition, $(\widehat{\lambda},\widehat{\beta})$ are the values which minimize $B_{\lambda,\beta}\Big[\widetilde{N}_\lambda(\theta)\Big]$,

$$\begin{aligned} B_{\widehat{\lambda},\widehat{\beta}}\left(\widehat{N}(\theta)\right) &= \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left(\widetilde{N}_\lambda(\theta)\right) \\ &= \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left(\max\{\widetilde{N}_\lambda(\theta),\sigma\}\right). \end{aligned}$$

Equation (1.22) concludes the proof. Indeed,

$$\begin{aligned} \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left(\max\{\widetilde{N}_\lambda(\theta),\sigma\}\right) &\leq \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left(\max\{N(\theta),\sigma\}\left(1 + B_{\lambda,\beta}\left(N(\theta)\right)\right)^{-1}\right) \\ &\leq \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left(N(\theta)\left(1 + B_{\lambda,\beta}\left(N(\theta)\right)\right)^{-1}\right). \end{aligned}$$

□

In the following we are going to present some simplified and more explicit bounds on the approximation error.

We introduce the subset Γ' of Γ defined as

$$\Gamma' = \Big\{(\lambda,\beta,t) \in \Lambda \times \mathbb{R}_+ \mid \begin{array}{r} \xi + \mu + \gamma + 4\delta\lambda/\max\{t,\sigma\} < 1, \\ \mu + \gamma + 2\delta\lambda/\max\{t,\sigma\} \leq 1/2, \\ \text{and } 2\gamma + \delta\lambda/\max\{t,\sigma\} \leq 1/2 \end{array} \Big\}$$

and the function

$$\widetilde{B}_{\lambda,\beta}(t) = \begin{cases} \dfrac{\gamma + \lambda\delta/\max\{t,\sigma\}}{1 - \mu - 2\gamma - 4\lambda\delta/\max\{t,\sigma\}} & (\lambda,\beta,t) \in \Gamma' \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 1.15. *For any $(\lambda,\beta) \in \Lambda$ and any $t \in \mathbb{R}_+$,*

$$B_{\lambda,\beta}\left(\frac{t}{1+B_{\lambda,\beta}(t)}\right) \leq \widetilde{B}_{\lambda,\beta}(t),$$
$$\frac{B_{\lambda,\beta}(t)}{1-B_{\lambda,\beta}(t)} \leq \widetilde{B}_{\lambda,\beta}(t).$$

Proof. We first observe that when $(\lambda,\beta,t) \notin \Gamma'$ then $\widetilde{B}_{\lambda,\beta}(t) = +\infty$ and hence the two inequalities are trivial. We now assume $(\lambda,\beta,t) \in \Gamma'$ and we put $\tau = \lambda\delta/\max\{t,\sigma\}$. We prove the second inequality first. Since $\Gamma' \subset \Gamma$, we have

$$\frac{B_{\lambda,\beta}(t)}{1-B_{\lambda,\beta}(t)} = \frac{\gamma+\tau}{1-\mu-2\gamma-3\tau} \leq \widetilde{B}_{\lambda,\beta}(t).$$

In order to prove the first inequality, we first check that $\left(\lambda,\beta,\frac{t}{1+B_{\lambda,\beta}(t)}\right) \in \Gamma$. We first observe that, since

$$\max\{t/[1+B_{\lambda,\beta}(t)],\sigma\} \geq \max\{t,\sigma\}/[1+B_{\lambda,\beta}(t)],$$

then

$$\begin{aligned}\xi + \mu + \gamma + 2\delta\lambda/\max\left\{\frac{t}{1+B_{\lambda,\beta}(t)},\sigma\right\} &\leq \xi + \mu + \gamma + 2[1+B_{\lambda,\beta}(t)]\frac{\delta\lambda}{\max\{t,\sigma\}} \\ &= \xi + \mu + \gamma + 2\tau + 2\tau B_{\lambda,\beta}(t).\end{aligned}$$

Moreover, as $(\lambda,\beta,t) \in \Gamma'$, we get

$$B_{\lambda,\beta}(t) = \frac{\gamma+\tau}{1-\mu-\gamma-2\tau} \leq 1,$$

so that

$$\xi + \mu + \gamma + 2\delta\lambda/\max\{t/[1+B_{\lambda,\beta}(t)],\sigma\} \leq \xi + \mu + \gamma + 4\tau < 1,$$

which proves that, indeed, $\left(\lambda,\beta,\frac{t}{1+B_{\lambda,\beta}(t)}\right) \in \Gamma$. Therefore

$$\begin{aligned}B_{\lambda,\beta}\left(\frac{t}{1+B_{\lambda,\beta}(t)}\right) &\leq \frac{\gamma+\tau\,(1+B_{\lambda,\beta}(t))}{1-\mu-\gamma-2\tau[1+\tau B_{\lambda,\beta}(t)]} \\ &= \frac{(\gamma+\tau)\,(1+\tau/(1-\mu-\gamma-2\tau))}{1-\mu-\gamma-2\tau-2\tau B_{\lambda,\beta}(t)},\end{aligned}$$

where in the last line we have used the definition of $B_{\lambda,\beta}$. Observing that

$$1-\mu-\gamma-2\tau-2\tau B_{\lambda,\beta}(t)=\frac{(1-\mu-\gamma-2\tau)^2-2\tau\,(\gamma+\tau)}{1-\mu-\gamma-2\tau},$$

we obtain

$$\begin{aligned}B_{\lambda,\beta}\left(\frac{t}{1+B_{\lambda,\beta}(t)}\right)&\leq\frac{(\gamma+\tau)\,(1-\mu-\gamma-\tau)}{(1-\mu-\gamma-2\tau)^2-2\tau\,(\gamma+\tau)}\\&=\frac{(\gamma+\tau)\,(1-\mu-\gamma-\tau)}{(1-\mu-\gamma-\tau)^2+\tau^2-2\tau(1-\mu-\gamma-\tau)-2\tau^2-2\gamma\tau}\\&=\frac{\gamma+\tau}{1-\mu-\gamma-\tau-2\tau-\left(\tau^2+2\gamma\tau\right)/\left(1-\mu-\gamma-\tau\right)}.\end{aligned}$$

Considering that

$$\left(\tau^2+2\gamma\tau\right)/\left(1-\mu-\gamma-\tau\right)\leq\tau,$$

since when $(\lambda,\beta,t)\in\Gamma'$, it is true that $1-\mu-\gamma-\tau\geq 1/2$ and $2\gamma+\tau\leq 1/2$, we conclude that

$$B_{\lambda,\beta}\left(\frac{t}{1+B_{\lambda,\beta}(t)}\right)\leq\frac{\gamma+\tau}{1-\mu-\gamma-4\tau}=\widetilde{B}_{\lambda,\beta}(t).$$

□

Combining the above lemma with Proposition 1.14, we obtain the following result.

Proposition 1.16. *With probability at least* $1-2\epsilon$*, for any* $\theta\in\mathbb{R}^d$*,*

$$\left|\frac{\max\{N(\theta),\sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta),\sigma\|\theta\|^2\}}-1\right|\leq\inf_{(\lambda,\beta)\in\Lambda}\widetilde{B}_{\lambda,\beta}\big[\|\theta\|^{-2}N(\theta)\big].$$

We now compute an explicit upper bound for the right-hand side of the inequality presented in the above proposition.
We first observe that, for any $\theta\in\mathbb{R}^d$,

$$\|\theta\|^{-2}N(\theta)=\mathbb{E}\Big[\|\theta\|^{-2}\langle\theta,X\rangle^2\Big]\leq\mathbb{E}\big[\|X\|^2\big]\leq\mathbb{E}\big[\|X\|^4\big]^{1/2}=s_4^2,$$

where s_4 is defined in equation (1.9) on page 26. Let $a>0$ and

$$K=1+\left\lceil a^{-1}\log\Big(\frac{n}{72(2+c)\kappa^{1/2}}\Big)\right\rceil,$$

where we recall that $c=\frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right)$ and κ is defined in equation (1.10). The choice of K will be justified by equation (1.29). We fix a geometric grid for $\|\theta\|^{-2}N(\theta)/s_4^2$ of the form

$$\exp\left(-ja\right),\quad 0\leq j<K,$$

and we define the finite set Λ of all possible values of the couple (λ,β) as

$$\Lambda=\{(\lambda_j,\beta_j)\mid 0\leq j<K\},\tag{1.27}$$

where

$$\lambda_j = \sqrt{\frac{2}{n(\kappa-1)}\left(\frac{(2+3c)}{4(2+c)\kappa^{1/2}\exp(-ja)} + \log(K/\epsilon)\right)}$$
$$\beta_j = \sqrt{2(2+c)\kappa^{1/2}s_4^4 n \exp\big[-(j-1/2)a\big]}.$$

Proposition 1.17. *Let $\sigma > 0$ be a threshold such that $\sigma \leq s_4^2$. We define the explicit bound*

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{t}}\cosh(a/2)$$

and

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & \big[6 + (\kappa-1)^{-1}\big]\zeta(\max\{t,\sigma\}) \leq \sqrt{n} \\ +\infty & \textit{otherwise.} \end{cases}$$

For any $t \in \mathbb{R}_+$,

$$\inf_{(\lambda,\beta)\in\Lambda} \widetilde{B}_{\lambda,\beta}(t) \leq B_*(\min\{t, s_4^2\}). \tag{1.28}$$

As a consequence, with probability at least $1-2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \leq B_*\Big[\|\theta\|^{-2}N(\theta)\Big].$$

Before proving the result we recall explicitly some definitions previously introduced. We have defined the set

$$\Gamma = \Big\{(\lambda,\beta,t) \in \Lambda \times \mathbb{R}_+ \mid \\ \frac{(4\kappa-3)\lambda}{2} + \frac{2(2+c)\kappa^{1/2}s_4^2}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2\lambda} + \frac{\log(K/\epsilon)}{n\lambda} + \frac{\beta}{n\max\{t,\sigma\}} < 1\Big\},$$

and the bound

$$\widetilde{B}_{\lambda,\beta}(t) = \frac{\dfrac{(\kappa-1)\lambda}{2} + \dfrac{(2+c)\kappa^{1/2}s_4^2}{\beta} + \dfrac{(2+3c)s_4^4}{2\beta^2\lambda} + \dfrac{\log(K/\epsilon)}{n\lambda} + \dfrac{\beta}{2n\max\{t,\sigma\}}}{1 - \dfrac{3(\kappa-1)\lambda}{2} - \dfrac{2(2+c)\kappa^{1/2}s_4^2}{\beta} - \dfrac{(2+3c)s_4^4}{2\beta^2\lambda} - \dfrac{\log(K/\epsilon)}{n\lambda} - \dfrac{\beta}{n\max\{t,\sigma\}}}$$

when $(\lambda,\beta,t) \in \Gamma$, and $\widetilde{B}_{\lambda,\beta}(t) = +\infty$ otherwise. The definition of $\widehat{N}$ is given in equation (1.26).

We remark that the function ζ is non-increasing and observable, so that in the case when the threshold σ is defined by the equation

$$\big[6 + (1-\kappa)^{-1}\big]\zeta(\sigma) = \sqrt{n},$$

then the condition in the definition of B_* is satisfied for all $t \in \mathbb{R}_+$. Moreover the function $B_* : \mathbb{R}_+ \to \mathbb{R}_+$ is non-increasing and in this case

$$B_*(t) \leq B_*(0) \leq \frac{1}{2}, \qquad t \in \mathbb{R}_+.$$

We also observe that it is not restrictive to assume $\sigma \leq s_4^2$, since we know that $N(\theta) \leq s_4^2\|\theta\|^2$. Moreover, we will see later that if we choose the threshold σ such that

$$8\zeta(\sigma) = \sqrt{n},$$

then it decays to zero at speed $1/n$ as the sample size grows to infinity.

Proof. We recall that the function $\widetilde{B}_{\lambda,\beta}$ is non-increasing so that $\widetilde{B}_{\lambda,\beta}(t) \leq \widetilde{B}_{\lambda,\beta}(\min\{t, s_4^2\})$. Moreover, since

$$\max\{\min\{t, s_4^2\}, \sigma\} = \min\{\max\{t, \sigma\}, s_4^2\},$$

it is sufficient to prove equation (1.28) for $\max\{t, \sigma\} \in [0, s_4^2]$.

As equation (1.28) is trivial when $B_*(t) = +\infty$, we may assume that $B_*(t) < +\infty$, so that $6\zeta(\max\{t, \sigma\}) \leq \sqrt{n}$. In particular, by considering only the second factor in the definition of ζ, we obtain that

$$\sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{\max\{t,\sigma\}}} \leq \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{\max\{t,\sigma\}}}\cosh(a/2) \leq \frac{\sqrt{n}}{6},$$

which implies

$$\frac{\max\{t,\sigma\}}{s_4^2} \geq \frac{72(2+c)\kappa^{1/2}}{n} \geq \exp\left(-a(K-1)\right).$$

Therefore we have

$$\log\left(\frac{\max\{t,\sigma\}}{s_4^2}\right) \in \left[-a(K-1), 0\right] \tag{1.29}$$

which justifies the definition of K. Because of equation (1.29), there exists $\widehat{\jmath} \in \{0, \dots, K-1\}$, for which

$$\left|\log\left(\frac{\max\{t,\sigma\}}{s_4^2}\right) + \widehat{\jmath}a\right| \leq a/2.$$

Equivalently, it means that

$$\exp\left(-\left(\widehat{\jmath}+\frac{1}{2}\right)a\right) \leq \frac{\max\{t,\sigma\}}{s_4^2} \leq \exp\left(-\left(\widehat{\jmath}-\frac{1}{2}\right)a\right). \tag{1.30}$$

We now recall that by equation (1.12) on page 27

$$\gamma + \delta\lambda/\max\{t,\sigma\} = \frac{\lambda}{2}(\kappa-1) + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2\lambda} + \frac{\log(K/\epsilon)}{n\lambda} + \frac{\beta}{2n\max\{t,\sigma\}}$$

and we observe that (λ_*, β_*) defined as

$$\lambda_* = \lambda_*(t) = \sqrt{\frac{2}{n(\kappa-1)}\left(\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}\max\{t,\sigma\}} + \log(K/\epsilon)\right)} \tag{1.31}$$

$$\beta_* = \sqrt{2(2+c)\kappa^{1/2}s_4^2\max\{t,\sigma\}n} \tag{1.32}$$

are the desired values that optimize $\gamma + \delta\lambda/\max\{t,\sigma\}$. We also remark that, by equation (1.30),

$$\beta_{\widehat{\jmath}}\exp(-a/2) \leq \beta_* \leq \beta_{\widehat{\jmath}} \tag{1.33}$$

$$\lambda_{\widehat{\jmath}}\exp(-a/4) \leq \lambda_* \leq \lambda_{\widehat{\jmath}}\exp(a/4). \tag{1.34}$$

Thus, evaluating $\gamma + \delta\lambda/\max\{t,\sigma\}$ in $(\lambda_{\widehat{\jmath}}, \beta_{\widehat{\jmath}}) \in \Lambda$, we obtain that

$$\begin{aligned}
&\gamma_{\widehat{\jmath}} + \delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\} \\
&= \frac{\lambda_*(\kappa-1)}{2}\frac{\lambda_{\widehat{\jmath}}}{\lambda_*} + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta_*}\frac{\beta_*}{\beta_{\widehat{\jmath}}} + \frac{(2+3c)s_4^4}{2\,\beta_{\widehat{\jmath}}^2\,\lambda_*}\frac{\lambda_*}{\lambda_{\widehat{\jmath}}} + \frac{\log(K/\epsilon)}{n\lambda_*}\frac{\lambda_*}{\lambda_{\widehat{\jmath}}} + \frac{\beta_*}{2n\max\{t,\sigma\}}\frac{\beta_{\widehat{\jmath}}}{\beta_*} \\
&\qquad \le \frac{\lambda_*(\kappa-1)}{2}\frac{\lambda_{\widehat{\jmath}}}{\lambda_*} + \frac{1}{n\lambda_*}\left[\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}\max\{t,\sigma\}} + \log(K/\epsilon)\right]\frac{\lambda_*}{\lambda_{\widehat{\jmath}}} \\
&\qquad\qquad + \sqrt{\frac{(2+c)\kappa^{1/2}s_4^2}{2n\max\{t,\sigma\}}}\left(\frac{\beta_*}{\beta_{\widehat{\jmath}}} + \frac{\beta_{\widehat{\jmath}}}{\beta_*}\right) \\
&\qquad \le \sqrt{\frac{2(\kappa-1)}{n}\left[\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}\max\{t,\sigma\}} + \log(K/\epsilon)\right]}\cosh\log\left(\frac{\lambda_{\widehat{\jmath}}}{\lambda_*}\right) \\
&\qquad\qquad + \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{n\max\{t,\sigma\}}}\cosh\log\left(\frac{\beta_{\widehat{\jmath}}}{\beta_*}\right).
\end{aligned}$$

Using equation (1.33) and equation (1.34), we get

$$\begin{aligned}
&\gamma_{\widehat{\jmath}} + \delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\} \\
&\qquad \le \sqrt{\frac{2(\kappa-1)}{n}\left[\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}\max\{t,\sigma\}} + \log(K/\epsilon)\right]}\cosh\left(\frac{a}{4}\right) \\
&\qquad\qquad + \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{n\max\{t,\sigma\}}}\cosh\left(\frac{a}{2}\right).
\end{aligned}$$

We also observe that

$$\mu_{\widehat{\jmath}} + \gamma_{\widehat{\jmath}} + 4\delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\} \le 4\big[\gamma_{\widehat{\jmath}} + \delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\}\big] \le 4n^{-1/2}\zeta(t),$$

since by definition $\mu_{\widehat{\jmath}} \le 2\gamma_{\widehat{\jmath}}$. In the same way, observing that

$$\xi_{\widehat{\jmath}} = \frac{\kappa\lambda_{\widehat{\jmath}}}{2} \le \gamma_{\widehat{\jmath}}\left(1 + \frac{1}{\kappa-1}\right)$$

we obtain

$$\begin{aligned}
\xi_{\widehat{\jmath}} + \mu_{\widehat{\jmath}} + \gamma_{\widehat{\jmath}} + 4\delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\} &< \big[4 + (\kappa-1)^{-1}\big]\gamma_{\widehat{\jmath}} + 4\delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\} \\
&\le \big[6 + (\kappa-1)^{-1}\big]n^{-1/2}\zeta(\max\{t,\sigma\})
\end{aligned}$$

and similarly,

$$\begin{aligned}
2\big[\mu_{\widehat{\jmath}} + \gamma_{\widehat{\jmath}} + 2\delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\}\big] &\le 6n^{-1/2}\zeta(\max\{t,\sigma\}), \\
2\big[2\gamma_{\widehat{\jmath}} + \delta_{\widehat{\jmath}}\lambda_{\widehat{\jmath}}/\max\{t,\sigma\}\big] &\le 4n^{-1/2}\zeta(\max\{t,\sigma\}).
\end{aligned}$$

This implies that, whenever $B_*(t) < +\infty$, then $(\lambda_{\widehat{\jmath}}, \beta_{\widehat{\jmath}}, t) \in \Gamma'$. We have then proved that

$$\inf_{(\lambda,\beta)\in\Lambda} \widetilde{B}_{\lambda,\beta}(t) \le \widetilde{B}_{\lambda_{\widehat{\jmath}},\,\beta_{\widehat{\jmath}}}(t) \le \frac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4n^{-1/2}\zeta(\max\{t,\sigma\})} = B_*(t).$$

We now need to prove the last part of the proposition. We first apply equation (1.28) to Proposition 1.16 and we obtain that, with probability at least $1-2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1 \right| \leq B_*\big[\min\{\|\theta\|^{-2}N(\theta), s_4^2\}\big].$$

Since $\|\theta\|^{-2}N(\theta) \leq s_4^2$, we conclude the proof. □

Before introducing some comments on the bound presented in Proposition 1.17 we make more explicit its formulation by replacing the quantity s_4^2 by the trace of the Gram matrix.

Lemma 1.18. *We have*

$$s_4^2 \leq \sqrt{\kappa}\ \mathbf{Tr}(G),$$

where $\mathbf{Tr}(G) = \int \|x\|^2 \,\mathrm{dP}(x)$ *denotes the trace of* G.

Proof. By definition

$$s_4^4 = \int \|x\|^4 \mathrm{dP}(x) = \sum_{\substack{1\leq j\leq d\\ 1\leq k\leq d}} \int x_j^2 x_k^2 \mathrm{dP}(x)$$

and by the Cauchy-Schwarz inequality we obtain

$$s_4^4 \leq \sum_{\substack{1\leq j\leq d\\ 1\leq k\leq d}} \left(\int x_j^4 \mathrm{dP}(x)\right)^{\frac{1}{2}} \left(\int x_k^4 \mathrm{dP}(x)\right)^{\frac{1}{2}}.$$

Recalling that, by definition, $\kappa = \sup_{\theta\in\mathbb{R}^d} \frac{\int\langle\theta,x\rangle^4 \mathrm{dP}(x)}{\left(\int\langle\theta,x\rangle^2 \mathrm{dP}(x)\right)^2}$, we get

$$\begin{aligned} s_4^4 &\leq \kappa \sum_{\substack{1\leq j\leq d\\ 1\leq k\leq d}} \left(\int x_j^2 \mathrm{dP}(x)\right)\left(\int x_k^2 \mathrm{dP}(x)\right) \\ &= \kappa \left(\int \sum_{j=1}^d x_j^2 \mathrm{dP}(x)\right)^2 \\ &= \kappa \left(\int \|x\|^2 \mathrm{dP}(x)\right)^2, \end{aligned}$$

which concludes the proof. □

We now apply the above lemma to Proposition 1.17. We recall that, given $a > 0$,

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil$$

where κ is defined in equation (1.10) on page 26 and $c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2}\right)$.

Proposition 1.19. *Let us fix a threshold* $\sigma \in \mathbb{R}_+$ *such that* $\sigma \leq s_4^2$. *With probability at least* $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1 \right| \leq B_*\Big[\|\theta\|^{-2}N(\theta)\Big],$$

where

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & [6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \leq \sqrt{n} \\ +\infty & \textit{otherwise} \end{cases}$$

and

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\ \mathbf{Tr}(G)}{4(2+c)t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa\ \mathbf{Tr}(G)}{t}}\cosh(a/2).$$

Let us provide some numerical evaluation of the constants. Let us choose $a = 1/2$. We can bound the logarithmic factor $\log\log(n)$ hidden in $\log(K)$ with a relatively small constant assuming any reasonable upper bound for the number n of observations. For instance, if we assume that $n \leq 10^{20}$, and use the fact that $\kappa \geq 1$ we can bound K with the constant 77, so that $\log(K) \leq 4.35$. In this case, ζ is bounded by

$$\zeta(t) \leq \sqrt{2.032(\kappa-1)\left(\frac{0.73\ \mathbf{Tr}(G)}{t} + 4.35 + \log(\epsilon^{-1})\right)} + \sqrt{\frac{98.5\,\kappa\,\mathbf{Tr}(G)}{t}}, \quad t \in \mathbb{R}_+.$$

We observe that the quantity $\kappa - 1$ corresponds to a variance factor, and more precisely to $\mathbf{Var}[\langle\theta, X\rangle^2]\ N(\theta)^{-2}$.

In the special case when X is a Gaussian random vector we have $\kappa = 3$. Moreover, the bound B_* does not depend explicitly on the dimension d and in particular the dimension is replaced by the entropy term $\mathbf{Tr}(G)/\max\{\|\theta\|^{-2}N(\theta), \sigma\}$. The remarkable fact is that, since the result is dimension-independent, it can be generalized to any infinite-dimensional Hilbert space, with the only assumption that the trace of the Gram matrix $\mathbf{Tr}(G)$ is finite. For further details we refer to section 1.3.

Let us also make clear that we do not need to know the exact values of κ and $\mathbf{Tr}(G)$ to compute the estimator and evaluate the bound. We only need to know upper bounds for these two quantities. If we use those upper bounds in the definition of the estimator, the estimation error will hold true with κ and $\mathbf{Tr}(G)$ replaced with their upper bounds.

In order to have a meaningful (finite) bound we can choose for example the threshold σ such that

$$8\,\zeta(\sigma) = \sqrt{n},$$

so that $B_*(t) < +\infty$ for any $t \in \mathbb{R}_+$, assuming that we work with $\kappa \geq 3/2$. With this choice the threshold decays to zero as the sample size grows to infinity. More precisely, using the inequality $(\sqrt{a} + \sqrt{b})^2 \leq 2(a+b)$, we get in this case that

$$\sigma \leq \frac{100\,\kappa\,\mathbf{Tr}(G)}{n/128 - 4.35 - \log(\epsilon^{-1})}. \tag{1.35}$$

1.2.3 A quadratic estimator of N

We have already said that our estimator $\widehat{N}$ defined in equation (1.26) is unfortunately not a quadratic form. In this section we are going to deduce a quadratic estimator of the quadratic form N.

Let $\Lambda \subset (\mathbb{R}_+ \setminus \{0\})^2$ be the finite set defined by equation (1.27). Proposition 1.12 on page 28 provides a confidence region for $N(\theta)$. Define

$$B_+(\theta) = \min_{(\lambda,\beta)\in\Lambda} \Phi_{\theta,+}^{-1}\Big(\widetilde{N}_\lambda(\theta)\Big) \qquad \text{and} \qquad B_-(\theta) = \max_{(\lambda,\beta)\in\Lambda} \Phi_{\theta,-}\Big(\widetilde{N}_\lambda(\theta)\Big) \tag{1.36}$$

where we recall that $\widetilde{N}_\lambda(\theta) = \lambda/\widehat{\alpha}(\theta)^2$ and $\widehat{\alpha}(\theta)$ is defined in equation (1.3) on page 17. According to Proposition 1.12 we get the following result.

Proposition 1.20. *With probability at least* $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$B_-(\theta) \le N(\theta) \le B_+(\theta).$$

In the following we are going to study the properties of a symmetric matrix Q that satisfies $\mathbf{Tr}(Q^2) \le \mathbf{Tr}(G^2)$ and

$$B_-(\theta) \le \theta^\top Q\theta \le B_+(\theta), \qquad \theta \in \Theta_\delta, \tag{1.37}$$

where Θ_δ is a finite δ-net of the unit sphere $\mathbb{S}_d = \{\theta \in \mathbb{R}^d, \|\theta\| = 1\}$, meaning that

$$\sup_{\theta\in\mathbb{S}_d} \min_{\xi\in\Theta_\delta} \|\theta - \xi\| \le \delta.$$

Such a quadratic form exists with probability at least $1-2\epsilon$ according to Proposition 1.20 and it can be computed using a convex optimization algorithm described in section 1.2.4. Since, for any $\theta \in \mathbb{S}_d$, there is $\xi \in \Theta_\delta$ such that $\|\theta - \xi\| \le \delta$, we have

$$\begin{aligned} |\theta^\top Q\theta - \xi^\top Q\xi| &= (\theta+\xi)^\top Q(\theta-\xi) \\ &\le \|\theta+\xi\|\|Q\|_\infty\|\theta-\xi\| \le 2\delta\sqrt{\mathbf{Tr}(Q^2)} \le 2\delta\sqrt{\mathbf{Tr}(G^2)}. \end{aligned} \tag{1.38}$$

Let us put $\eta = 2\delta\sqrt{\mathbf{Tr}(G^2)}$.

Lemma 1.21. *With probability at least* $1-2\epsilon$, *for any* $\theta \in \mathbb{S}_d$ *and any* $(\lambda,\beta) \in \Lambda$,

$$\begin{aligned} \Phi_- \circ \Phi_+(\theta^\top Q\theta - \eta) &\le N(\theta) + \eta, \\ \Phi_- \circ \Phi_+(N(\theta) - \eta) &\le \theta^\top Q\theta + \eta, \end{aligned}$$

where Φ_- *and* Φ_+ *are the values taken by* $\Phi_{\theta,-}$ *and* $\Phi_{\theta,+}$, *when* $\theta \in \mathbb{S}_d$.

We recall that the functions $\Phi_{\theta,-}$ and $\Phi_{\theta,+}$ depend on θ through $\|\theta\|$ only.

Proof. According to Proposition 1.20, with probability at least $1-2\epsilon$, for any $(\lambda,\beta) \in \Lambda$,

$$\theta^\top Q\theta \le \Phi_+^{-1}(\widetilde{N}_\lambda(\xi)) + \eta \le \Phi_+^{-1} \circ \Phi_-^{-1}(N(\xi)) + \eta.$$

Since equation (1.38) holds true also for N, we conclude that

$$\theta^\top Q\theta \le \Phi_+^{-1} \circ \Phi_-^{-1}(N(\theta) + \eta) + \eta.$$

In the same way

$$\theta^\top Q\theta \ge \Phi_-(\widetilde{N}_\lambda(\xi)) - \eta \ge \Phi_- \circ \Phi_+(N(\xi)) - \eta \ge \Phi_- \circ \Phi_+(N(\theta) - \eta) - \eta.$$

□

We need some technical lemmas to take advantage of this result.

Lemma 1.22. *Let us put*

$$F(t) = \max\{t, \sigma\} B_*(\min\{t, s_4^2\}),$$

where B_* is defined in Proposition 1.17. In the case when

$$\kappa \geq 3/2 \quad \text{and} \quad 8\zeta(\sigma) \leq \sqrt{n},$$

the function $t \mapsto F(t)$ is non-decreasing for any $t \in \mathbb{R}_+$.

Proof. If $\sigma \geq s_4^2$, then $B_*(\min\{t, s_4^2\}) = B_*(\sigma)$, so that $F(t) = \max\{t, \sigma\} B_*(\sigma)$ is obviously non-decreasing. Otherwise, $\sigma \leq s_4^2$, so that

$$\zeta\left(\max\left\{\min\{t, s_4^2\}, \sigma\right\}\right) = \zeta\left(\min\left\{\max\{t, \sigma\}, s_4^2\right\}\right).$$

Therefore the function F is of the form

$$F(t) = c\frac{ug(u)}{(1 - g(u))},$$

where $u = \max\{t, \sigma\}$,

$$g(u) = \sqrt{a_1/u + a_2} + \sqrt{a_3/u},$$

$g(\sigma) \leq 1/2$, and the constants c, a_1, a_2, and a_3 are positive. Let $h(u) = \sqrt{a_1/u} + \sqrt{a_3/u}$ and observe that

$$g'(u) = -\frac{1}{2u}\left(\frac{a_1/u}{(a_1/u + a_2)^{1/2}} + \sqrt{a_3/u}\right) \geq -\frac{1}{2u}\left(\sqrt{a_1/u} + \sqrt{a_3/u}\right) = h'(u)$$

and that $g(u) \geq h(u)$. Therefore $h(u) \leq g(u) \leq 1/2$, for any $u \geq \sigma$, and

$$\frac{\partial}{\partial u}\log\left(\frac{ug(u)}{1 - g(u)}\right) = \frac{1}{u} + \frac{g'(u)}{g(u)(1 - g(u))} \geq \frac{1}{u} + \frac{h'(u)}{h(u)(1 - h(u))} = \frac{1}{u} - \frac{1}{2u(1 - h(u))} \geq 0,$$

showing that F is non-decreasing. □

In the following we always assume that $\kappa \geq 3/2$.

Lemma 1.23. *For any $(a, b) \in \mathbb{R}^2$ such that, for any $(\lambda, \beta) \in \Lambda$,*

$$\Phi_- \circ \Phi_+(a - \eta) \leq b + \eta, \quad \text{and} \quad \Phi_- \circ \Phi_+(b - \eta) \leq a + \eta,$$

and any threshold $\sigma \in \mathbb{R}_+$ such that $8\zeta(\sigma) \leq \sqrt{n}$ and $\sigma \leq s_4^2$, we have

$$\left|\max\{a, \sigma\} - \max\{b, \sigma\}\right| \leq 2\max\{a + \eta, \sigma\} B_*(\min\{a + \eta, s_4^2\}) + 2\eta \tag{1.39}$$

$$\text{and } \left|\max\{a, \sigma\} - \max\{b, \sigma\}\right| \leq 2\max\{b + \eta, \sigma\} B_*(\min\{b + \eta, s_4^2\}) + 2\eta. \tag{1.40}$$

Proof. By symmetry of a and b, equation (1.40) is a consequence of equation (1.39).

Step 1. We will prove that

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\}\left(1 + 2\widetilde{B}_{\lambda,\beta}(a + \eta)\right), \tag{1.41}$$

where $\widetilde{B}_{\lambda,\beta}$ is defined as in Lemma 1.15 on page 33.

Case 1. Assume that

$$\max\{\Phi_+(b-\eta),\sigma\} \le \max\{a+\eta,\sigma\},$$

and remark that, since Φ_+ is non-decreasing and $\Phi_+(\sigma) \le \sigma$,

$$\begin{aligned}\max\{\Phi_+(b-\eta),\sigma\} &\ge \max\{\Phi_+(b-\eta),\Phi_+(\sigma)\}\\ &= \Phi_+\big(\max\{(b-\eta),\sigma\}\big) = \frac{\max\{b-\eta,\sigma\}}{1+B_{\lambda,\beta}(b-\eta)},\end{aligned}$$

according to equation (1.23) on page 31, where $B_{\lambda,\beta}$ is defined in equation (1.20). Therefore in this case,

$$\max\{b-\eta,\sigma\} \le \max\{a+\eta,\sigma\}\Big(1+B_{\lambda,\beta}(b-\eta)\Big), \tag{1.42}$$

but when $\max\{b-\eta,\sigma\} > \max\{a+\eta,\sigma\}$,

$$B_{\lambda,\beta}(b-\eta) \le B_{\lambda,\beta}(a+\eta)$$

because $B_{\lambda,\beta}(t)$ is a non-increasing function of $\max\{t,\sigma\}$, thus equation (1.42) implies that

$$\max\{b-\eta,\sigma\} \le \max\{a+\eta,\sigma\}\Big(1+B_{\lambda,\beta}(a+\eta)\Big).$$

Since $B_{\lambda,\beta} \le \widetilde{B}_{\lambda,\beta}$, equation (1.41) holds true.

Case 2. Assume now that we are not in **Case 1**, implying that

$$\max\{b-\eta,\sigma\} \ge \max\{\Phi_+(b-\eta),\sigma\} > \max\{a+\eta,\sigma\}.$$

In this case

$$\begin{aligned}\max\{a+\eta,\sigma\} &\ge \max\{\Phi_-\circ\Phi_+(b-\eta),\sigma\} \ge \max\{\Phi_-\circ\Phi_+(b-\eta),\Phi_-(\sigma)\}\\ &\ge \Phi_-\big(\max\{\Phi_+(b-\eta),\sigma\}\big) \ge \max\{\Phi_+(b-\eta),\sigma\}\Big[1-B_{\lambda,\beta}\Big(\max\{\Phi_+(b-\eta),\sigma\}\Big)\Big]\end{aligned}$$

according to equation (1.24). Moreover, continuing the above chain of inequalities

$$\begin{aligned}\max\{a+\eta,\sigma\} &\ge \max\{\Phi_+(b-\eta),\Phi_+(\sigma)\}\Big[1-B_{\lambda,\beta}(\max\{a+\eta,\sigma\})\Big]\\ &= \Phi_+\big(\max\{b-\eta,\sigma\}\big)\Big[1-B_{\lambda,\beta}(a+\eta)\Big]\\ &\ge \max\{b-\eta,\sigma\}\frac{1-B_{\lambda,\beta}(a+\eta)}{1+B_{\lambda,\beta}(\max\{b-\eta,\sigma\})}\\ &\ge \max\{b-\eta,\sigma\}\frac{1-B_{\lambda,\beta}(a+\eta)}{1+B_{\lambda,\beta}(a+\eta)}.\end{aligned}$$

Therefore

$$\begin{aligned}\max\{b-\eta,\sigma\} &\le \max\{a+\eta,\sigma\}\frac{1+B_{\lambda,\beta}(a+\eta)}{1-B_{\lambda,\beta}(a+\eta)}\\ &= \max\{a+\eta,\sigma\}\left(1+\frac{2B_{\lambda,\beta}(a+\eta)}{1-B_{\lambda,\beta}(a+\eta)}\right) \le \max\{a+\eta,\sigma\}(1+2\widetilde{B}_{\lambda,\beta}(a+\eta))\end{aligned}$$

according to Lemma 1.15. This concludes the proof of **Step 1**.

Step 2 Taking the infimum in $(\lambda,\beta)\in\Lambda$ in equation (1.41), according to equation (1.28), we obtain that

$$\max\{b-\eta,\sigma\}\leq\max\{a+\eta,\sigma\}\Big(1+2B_*\big(\min\{a+\eta,s_4^2\}\big)\Big).$$

We can then use the fact that $t\mapsto\max\{t,\sigma\}B_*(\min\{t,s_4^2\})$ is non-decreasing (proved in Lemma 1.22) to deduce that

$$\max\{b-\eta,\sigma\}\leq\max\{a+\eta,\sigma\}+2\max\{b+\eta,\sigma\}B_*(\min\{b+\sigma,s_4^2\}),$$

since there is nothing to prove when already $\max\{b+\eta,\sigma\}\leq\max\{a+\eta,\sigma\}$. Remark that $\max\{a+\eta,\sigma\}\leq\max\{a+\eta,\sigma+\eta\}\leq\max\{a,\sigma\}+\eta$ and that in the same way $\max\{b-\eta,\sigma\}\geq\max\{b,\sigma\}-\eta$. This proves that

$$\begin{aligned}\max\{b,\sigma\}-\max\{a,\sigma\}&\leq 2\max\{a+\eta,\sigma\}B_*\big(\min\{a+\eta,s_4^2\}\big)+2\eta\\ \text{and }\max\{b,\sigma\}-\max\{a,\sigma\}&\leq 2\max\{b+\eta,\sigma\}B_*\big(\min\{b+\eta,s_4^2\}\big)+2\eta.\end{aligned}$$

By symmetry, we can then exchange a and b to prove the same bounds for $\max\{a,\sigma\}-\max\{b,\sigma\}$, and therefore also for the absolute value of this quantity, which ends the proof of the lemma. □

Lemma 1.24. *For any $(a,b)\in\mathbb{R}^2$ such that, for any $(\lambda,\beta)\in\Lambda$,*

$$\Phi_-\circ\Phi_+(a-\eta)\leq b+\eta,\qquad\text{and}\qquad\Phi_-\circ\Phi_+(b-\eta)\leq a+\eta,$$

and any threshold $\sigma\in\mathbb{R}_+$ such that $8\zeta(\sigma)\leq\sqrt{n}$ and $\sigma\leq s_4^2$, we have

$$\begin{aligned}\big|\max\{a,\sigma\}-\max\{b,\sigma\}\big|&\leq 2\max\{a,\sigma\}B_*\big(\min\{a,s_4^2\}\big)+5\eta/2\\ \big|\max\{a,\sigma\}-\max\{b,\sigma\}\big|&\leq 2\max\{b,\sigma\}B_*\big(\min\{b,s_4^2\}\big)+5\eta/2.\end{aligned}$$

Proof. This is a consequence of the previous lemma, of the fact that $B_*(\min\{t,s_4^2\})\leq 1/4$, and of the fact that $\max\{a+\eta,\sigma\}\leq\max\{a,\sigma\}+\eta$. □

We deduce from Lemma 1.21 and Lemma 1.24 the analogous for the quadratic estimator $\theta^\top Q\theta$ of the dimension-free bound presented in Proposition 1.17 for $\widehat{N}(\theta)$.

Lemma 1.25. *Let us assume that $8\zeta(\sigma)\leq\sqrt{n}$, $\sigma\leq s_4^2$ and that $\kappa\geq 3/2$. With probability at least $1-2\epsilon$, for any $\theta\in\mathbb{S}_d$,*

$$\begin{aligned}\Big|\max\{\theta^\top Q\theta,\sigma\}-\max\{N(\theta),\sigma\}\Big|&\leq 2\max\{N(\theta),\sigma\}B_*(N(\theta))+5\delta\sqrt{\mathbf{Tr}(G^2)},\\ \Big|\max\{\theta^\top Q\theta,\sigma\}-\max\{N(\theta),\sigma\}\Big|&\leq 2\max\{\theta^\top Q\theta,\sigma\}B_*\big(\min\{\theta^\top Q\theta,s_4^2\}\big)+5\delta\sqrt{\mathbf{Tr}(G^2)}.\end{aligned}$$

There is still some improvement to bring: at this stage we are not sure that Q is non-negative. Nevertheless we know that, for any $\theta\in\mathbb{S}_d$, there is $\xi\in\Theta_\delta$ such that $\|\theta-\xi\|\leq\delta$, so that

$$\theta^\top Q\theta\geq\xi^\top Q\xi-\eta\geq B_-(\xi)-\eta\geq-\eta.$$

Decomposing Q into its positive and negative parts and writing $Q=Q_+-Q_-$, we deduce that

$$\|Q_-\|_\infty=\sup_{\theta\in\mathbb{S}_d}\theta^\top Q_-\theta=-\inf_{\theta\in\mathbb{S}_d}\theta^\top Q\theta\leq\eta,$$

where we recall that $\eta=2\delta\sqrt{\mathbf{Tr}(G^2)}$. Therefore, for any $\theta\in\mathbb{S}_d$,

$$\Big|\max\{\theta^\top Q\theta,\sigma\}-\max\{\theta^\top Q_+\theta,\sigma\}\Big|\leq\Big|\theta^\top Q\theta-\theta^\top Q_+\theta\Big|=\theta^\top Q_-\theta\leq\eta.$$

We have proved the following result.

Proposition 1.26. *Let us assume that* $8\zeta(\sigma) \leq \sqrt{n}$, $\sigma \leq s_4^2$ *and that* $\kappa \geq 3/2$. *With probability at least* $1 - 2\epsilon$, *for any* $\theta \in \mathbb{S}_d$,

$$\left|\max\{\theta^\top Q_+ \theta, \sigma\} - \max\{N(\theta), \sigma\}\right| \leq 2\max\{N(\theta), \sigma\} B_*(N(\theta)) + 7\delta\sqrt{\mathbf{Tr}(G^2)},$$

$$\left|\max\{\theta^\top Q_+ \theta, \sigma\} - \max\{N(\theta), \sigma\}\right| \leq 2\max\{\theta^\top Q_+ \theta, \sigma\} B_*(\min\{\theta^\top Q_+ \theta, s_4^2\}) + 7\delta\sqrt{\mathbf{Tr}(G^2)},$$

where B_* *is defined in Proposition 1.17.*

Let us recall that we can obtain a bound stated in terms of $\mathbf{Tr}(G)$, using the inequality $s_4^2 \leq \kappa^{1/2}\,\mathbf{Tr}(G)$ (proved in Lemma 1.18) that implies

$$\zeta(t) \leq \sqrt{2(\kappa-1)\left(\frac{(2+3c)\ \mathbf{Tr}(G)}{4(2+c)t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa\ \mathbf{Tr}(G)}{t}}\cosh(a/2).$$

Remark that, for any $a, b \in \mathbb{R}_+$,

$$a - b \leq \max\{a, \sigma\} - \max\{b, \sigma\} + \sigma,$$

so that

$$|a - b| \leq \left|\max\{a, \sigma\} - \max\{b, \sigma\}\right| + \sigma.$$

Corollary 1.27. *Assume that* $8\zeta(\sigma) \leq \sqrt{n}$, $\sigma \leq s_4^2$ *and that* $\kappa \geq 3/2$. *With probability at least* $1 - 2\epsilon$, *for any* $\theta \in \mathbb{S}_d$,

$$|\theta^\top Q_+ \theta - N(\theta)| \leq 2\max\{N(\theta), \sigma\} B_*(N(\theta)) + 7\delta\sqrt{\mathbf{Tr}(G^2)} + \sigma$$

$$|\theta^\top Q_+ \theta - N(\theta)| \leq 2\max\{\theta^\top Q_+ \theta, \sigma\} B_*(\min\{\theta^\top Q_+ \theta, s_4^2\}) + 7\delta\sqrt{\mathbf{Tr}(G^2)} + \sigma.$$

We conclude the section by providing two more bounds on the approximation error $|N(\theta) - \theta^\top Q_+ \theta|$ that depend on the largest eigenvalue of N and on the largest eigenvalue of Q_+ respectively. Applying Lemma 1.22 to Corollary 1.27 we obtain the following result.

Corollary 1.28. *Assume that* $8\zeta(\sigma) \leq \sqrt{n}$, $\sigma \leq s_4^2$ *and that* $\kappa \geq 3/2$. *With probability at least* $1 - 2\epsilon$, *for any* $\theta \in \mathbb{S}_d$,

$$|N(\theta) - \theta^\top Q_+ \theta| \leq 2\max\{\|G\|_\infty, \sigma\} B_*(\|G\|_\infty) + 7\delta\|G\|_F + \sigma,$$

$$|N(\theta) - \theta^\top Q_+ \theta| \leq 2\max\{\|Q_+\|_\infty, \sigma\} B_*(\min\{\|Q_+\|_\infty, s_4^2\}) + 7\delta\|G\|_F + \sigma,$$

where $\|G\|_F = \sqrt{\mathbf{Tr}(G^2)}$ *is the Frobenius norm of the Gram matrix.*

1.2.4 How to compute Q

In this section we provide an explicit construction of the quadratic form Q defined in equation (1.37) on page 40, inspired by [8]. Later, in section 1.6 we will present a simplified construction that leads to an estimator which is not exactly the same as Q but presents the same kind of behavior.

From a theoretical point of view, to obtain a quadratic estimator of N it is sufficient to choose the quadratic form associated with any non-negative symmetric matrix Q satisfying

$$B_-(\theta) \leq \theta^\top Q \theta \leq B_+(\theta), \quad \theta \in \mathbb{S}_d, \tag{1.43}$$

and we know, by Proposition 1.20, that such a quadratic form exists with probability at least $1 - 2\epsilon$. From a practical point of view, we prefer to satisfy the constraints on a finite set, as in equation (1.37).

Lemma 1.29. *With probability at least $1-2\epsilon$, for any finite set $\Theta \subset \mathbb{R}^d$, the symmetric matrix*

$$Q = \sum_{\theta\in\Theta} \left[\widehat{\xi}_+(\theta) - \widehat{\xi}_-(\theta)\right] \theta\theta^\top,$$

where

$$\left(\widehat{\xi}_+, \widehat{\xi}_-\right) = \arg \max_{\xi_+,\xi_-\in\left(\mathbb{R}^2_+\right)^\Theta} \left\{ -\frac{1}{2} \sum_{(\theta,\theta')\in\Theta^2} \left[\xi_+(\theta) - \xi_-(\theta)\right] \left[\xi_+(\theta') - \xi_-(\theta')\right] \langle\theta,\theta'\rangle^2 \right.$$
$$\left. + \sum_{\theta\in\Theta} \left[\xi_+(\theta)B_-(\theta) - \xi_-(\theta)B_+(\theta)\right] \right\},$$

is such that

$$Q = \arg\min \left\{ \|H\|_F \mid H^\top = H,\ B_-(\theta) \le \theta^\top H\theta \le B_+(\theta),\ \theta\in\Theta \right\}.$$

Moreover $\|Q\|_F \le \|G\|_F \le \mathbf{Tr}(G)$, and

$$|\theta_2^\top Q\theta_2 - \theta_1^\top Q\theta_1| \le \|\theta_1+\theta_2\|\|G\|_F\|\theta_1-\theta_2\|, \quad \theta_1,\theta_2\in\mathbb{R}^d. \tag{1.44}$$

Proof. We define

$$\widehat{H} = \arg\min \left\{ \|H\|_F \mid H^\top = H,\ B_-(\theta) \le \theta^\top H\theta \le B_+(\theta),\ \theta\in\Theta \right\}$$

and we observe that it is the solution of the (minimax) optimization problem

$$V = \inf_{H^\top=H} \sup_{\xi_+,\xi_-} V(H,\xi_+,\xi_-),$$

where

$$V(H,\xi_+,\xi_-) = \frac{1}{2}\|H\|_F^2 + \sum_{\theta\in\Theta} \left[\xi_+(\theta)\left(B_-(\theta) - \theta^\top H\theta\right) + \xi_-(\theta)\left(\theta^\top H\theta - B_+(\theta)\right)\right].$$

On the event of probability $1-2\epsilon$ described in Proposition 1.20 where

$$B_-(\theta) \le \theta^\top G\theta \le B_+(\theta),$$

the constraints are satisfied when $H=G$ and therefore, since the constraints are linear, by the Slater condition,

$$\inf_H \sup_{\xi_+,\xi_-} V(H,\xi_+,\xi_-) = \sup_{\xi_+,\xi_-} \inf_H V(H,\xi_+,\xi_-).$$

We can compute explicitly the solution of

$$\inf_H V(H,\xi_+,\xi_-)$$

and we obtain

$$\inf_H V(H,\xi_+,\xi_-) = V(Q,\xi_+,\xi_-).$$

It follows that $\widehat{H} = Q$, since $\sup_{\xi_+,\xi_-} \inf_H V(H,\xi_+,\xi_-) = V(Q,\widehat{\xi}_+,\widehat{\xi}_-)$.

Moreover, for any $\theta_1,\theta_2\in\mathbb{R}^d$, we have

$$|\theta_2^\top Q\theta_2 - \theta_1^\top Q\theta_1| \le (\theta_1+\theta_2)^\top Q\,(\theta_2-\theta_1) \le \|\theta_1+\theta_2\|\|Q\|_\infty\|\theta_2-\theta_1\|.$$

We conclude observing that $\|Q\|_\infty \le \|Q\|_F \le \|G\|_F$. The inequality $\|G\|_F \le \mathbf{Tr}(G)$ is a consequence of the fact that the eigenvalues of G are non-negative, so that

$$\|G\|_F^2 = \mathbf{Tr}(G^2) \le \mathbf{Tr}(G)^2.$$

□

1.3 Gram operators in Hilbert spaces

We have already underlined the fact that the results presented in the previous section do not explicitly depend on the dimension d of the ambient space and that, for this reason, they can be generalized to any separable Hilbert space. In this section we study how these results extend to infinite-dimensional Hilbert spaces.

Let $\mathcal{H}$ be a separable Hilbert space and let $\mathrm{P} \in \mathcal{M}_+^1(\mathcal{H})$ be an unknown probability distribution on $\mathcal{H}$. We consider the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{G}\theta = \int \langle \theta, v \rangle_{\mathcal{H}} \; v \; \mathrm{dP}(v)$$

and we assume $\mathbf{Tr}(\mathcal{G}) = \mathbb{E}(\|X\|_{\mathcal{H}}^2) < +\infty$, where $X \in \mathcal{H}$ denotes a random vector of law P. In analogy to the previous section we denote by N the quadratic form associated with $\mathcal{G}$ so that

$$N(\theta) = \langle \mathcal{G}\theta, \theta \rangle_{\mathcal{H}} = \int \langle \theta, v \rangle_{\mathcal{H}}^2 \; \mathrm{dP}(v), \quad \theta \in \mathcal{H}.$$

We consider $(\mathcal{H}_k)_k$ an increasing sequence of subspaces of $\mathcal{H}$ of finite dimension such that $\overline{\cup_k \mathcal{H}_k} = \mathcal{H}$ and we endow each space $\mathcal{H}_k$ with the probability P_k which arises from the disintegration of P. We denote by N_k the quadratic form in $\mathcal{H}_k$ associated with the probability distribution P_k and we observe that, denoting by Π_k the orthogonal projector on $\mathcal{H}_k$, for any $\theta \in \mathcal{H}$, we have

$$N_k\left(\Pi_k \theta\right) = N\left(\Pi_k \theta\right).$$

In the following we consider i.i.d. samples of size n in $\mathcal{H}$ drawn according to P. According to Proposition 1.12 on page 28, the event

$$\mathcal{A}_k = \left\{ \forall \theta \in \mathcal{H}_k, \; \forall (\lambda, \beta) \in \Lambda, \; \Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \leq \Phi_{\theta,+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \right\}$$

is such that $\mathrm{P}^{\otimes n}(\mathcal{A}_k) \geq 1 - 2\epsilon$. Since $\mathcal{A}_{k+1} \subset \mathcal{A}_k$, by the continuity of measure,

$$\mathrm{P}^{\otimes n}\left(\bigcap_{k \in \mathbb{N}} \mathcal{A}_k \right) \geq 1 - 2\epsilon.$$

This means that with probability at least $1 - 2\epsilon$, for any $\theta \in \bigcup_k \mathcal{H}_k$ and any $(\lambda, \beta) \in \Lambda$,

$$\Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \leq \Phi_{\theta,+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right).$$

Consequently, since $N(\theta) = \lim_{k \to +\infty} N(\Pi_k(\theta))$, for any $\theta \in \mathcal{H}$, the following result holds.

Proposition 1.30. *With probability at least* $1 - 2\epsilon$*, for any* $\theta \in \mathcal{H}$*,*

$$B_-(\theta) \leq N(\theta) \leq B_+(\theta)$$

where

$$B_-(\theta) = \lim \sup_{k \to +\infty} \sup_{(\lambda,\beta) \in \Lambda} \Phi_{\Pi_k(\theta),-}\left(\frac{\lambda}{\widehat{\alpha}(\Pi_k(\theta))^2}\right),$$

$$B_+(\theta) = \lim \inf_{k \to +\infty} \inf_{(\lambda,\beta) \in \Lambda} \Phi_{\Pi_k(\theta),+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\Pi_k(\theta))^2}\right).$$

If we do not want to go to the limit, we can use the explicit bound

$$\begin{aligned}|N(\theta) - N(\Pi_k(\theta))| &= |\langle \theta + \Pi_k(\theta), \mathcal{G}(\theta - \Pi_k(\theta))\rangle_{\mathcal{H}}| \\ &\leq 2\|\theta\|_{\mathcal{H}}\|\mathcal{G}\|_\infty\|\theta - \Pi_k(\theta)\|_{\mathcal{H}} \leq 2\|\theta\|_{\mathcal{H}}\,\mathbf{Tr}(\mathcal{G})\|\theta - \Pi_k(\theta)\|_{\mathcal{H}} \\ &= 2\|\theta\|_{\mathcal{H}}\mathbb{E}(\|X\|_{\mathcal{H}}^2)\|\theta - \Pi_k(\theta)\|_{\mathcal{H}}.\end{aligned}$$

This bound depends on $\|\theta - \Pi_k\theta\|_{\mathcal{H}}$. We will see in the following another bound that goes uniformly to zero for any $\theta \in \mathbb{S}_{\mathcal{H}}$ when k tends to infinity. In the same way, proceeding as already done in section 1.2 we state the analogous of Proposition 1.17.

Proposition 1.31. *Let*

$$\kappa \geq \sup_{\substack{\theta\in\mathcal{H} \\ \mathbb{E}(\langle\theta,X\rangle_{\mathcal{H}}^2)>0}} \frac{\mathbb{E}(\langle\theta, X\rangle_{\mathcal{H}}^4)}{\mathbb{E}\big(\langle\theta, X\rangle_{\mathcal{H}}^2\big)^2} \qquad \textit{and} \qquad s_4 \geq \mathbb{E}(\|X\|_{\mathcal{H}}^4)^{1/4}$$

be known constants. Let $a > 0$ and

$$K = 1 + \left\lceil a^{-1}\log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil$$

with $c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right)$. Define

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{t}}\cosh(a/2)$$

and consider, according to equation (1.26) on page 32, the estimator

$$\widehat{N}(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta), \qquad \theta \in \bigcup_k \mathcal{H}_k.$$

For any $\theta \in \mathcal{H}$, define $\widehat{N}(\theta)$ by choosing any accumulation point of the sequence $\widehat{N}(\Pi_k(\theta))$. Define the bound

$$B_*(t) = \frac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})},$$

where $\sigma \in]0, s_4^2]$ is some energy level such that

$$[6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \leq \sqrt{n}.$$

With probability at least $1 - 2\epsilon$, for any $\theta \in \mathcal{H}$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|_{\mathcal{H}}^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|_{\mathcal{H}}^2\}} - 1\right| \leq B_*\Big[\|\theta\|_{\mathcal{H}}^{-2}N(\theta)\Big].$$

Proof. This is a consequence of the fact that $\lim_{k\to+\infty} N(\Pi_k(\theta)) = N(\theta)$ and of the continuity of B_*. □

We recall that if we choose the threshold σ such that $8\zeta(\sigma) = \sqrt{n}$, then σ goes to zero as the sample size grows to infinity as shown in equation (1.35) on page 39.

Let $X_1, \ldots, X_n \in \mathcal{H}$ be an i.i.d. sample drawn according to P. Define $V = \mathbf{span}\{X_1, \ldots, X_n\}$ and

$$V_k = \mathbf{span}\{\Pi_k X_1, \ldots, \Pi_k X_n\} = \Pi_k(V).$$

Let Θ_δ be a δ-net of $\mathbb{S}_{\mathcal{H}} \cap V_k$ (where $\mathbb{S}_{\mathcal{H}}$ denotes the unit sphere in $\mathcal{H}$). Remark that Θ_δ is finite because $\dim(V_k) \leq n < +\infty$. We can compute as in Lemma 1.29 on page 45 a linear operator $\widehat{\mathcal{G}}_k : V_k \to V_k$ such that $\mathbf{Tr}(\widehat{\mathcal{G}}_k^2) \leq \mathbf{Tr}(\mathcal{G}^2)$ and

$$\Phi_-\big(\widetilde{N}_\lambda(\theta)\big) \leq \langle \widehat{\mathcal{G}}_k\theta, \theta\rangle_{\mathcal{H}} \leq \Phi_+^{-1}\big(\widetilde{N}_\lambda(\theta)\big), \qquad \theta \in \Theta_\delta.$$

Consider the estimator of $\mathcal{G}$ defined as

$$\mathcal{Q} = \widehat{\mathcal{G}}_k \circ \Pi_{V_k}, \tag{1.45}$$

where Π_{V_k} is the orthogonal projector on V_k. For any $\theta \in \mathbb{S}_{\mathcal{H}}$,

$$\langle \theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} = \langle \Pi_{V_k}\theta, \mathcal{Q}\Pi_{V_k}\theta\rangle_{\mathcal{H}} \leq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2(\langle \xi, \mathcal{Q}\xi\rangle_{\mathcal{H}} + \eta),$$

where $\xi \in \Theta_\delta$ is the closest point in Θ_δ to $\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\Pi_{V_k}\theta$ and where $\eta = 2\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)}$. Since $\xi \in \mathcal{H}_k$, with probability at least $1-\epsilon$, for any $(\lambda, \beta) \in \Lambda$,

$$\langle \xi, \mathcal{Q}\xi\rangle_{\mathcal{H}} \leq \Phi_+^{-1}(\widetilde{N}_\lambda(\xi)) = \Phi_+^{-1}\Big[\widetilde{N}_\lambda\Big(\xi + \|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}(\Pi_k - \Pi_{V_k})\theta\Big)\Big].$$

Let us now remark that for any $a \in [0,1]$, we have $\Phi_+(at) \leq a\Phi_+(t)$, so that $a\Phi_+^{-1}(t) \leq \Phi_+^{-1}(at)$. Let us also remark that $\widetilde{N}_\lambda(a\theta) = a^2\widetilde{N}_\lambda(\theta)$. Therefore

$$\begin{aligned}
\langle \theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} &\leq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\ \Phi_+^{-1}\Big\{\widetilde{N}_\lambda\Big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\Big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big)\Big]\Big\} + \eta \\
&\leq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\ \Phi_+^{-1} \circ \Phi_-^{-1}\Big\{N\Big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\Big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big)\Big]\Big\} + \eta \\
&\leq \Phi_+^{-1} \circ \Phi_-^{-1}\Big\{N\Big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big)\Big]\Big\} + \eta \\
&\leq \Phi_+^{-1} \circ \Phi_-^{-1}\Big(N(\Pi_k\theta) + \eta\Big) + \eta.
\end{aligned}$$

Indeed,

$$\Big\|\Big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big) - \Pi_k\theta\Big\| \leq \delta,$$

and this is a difference of two vectors belonging to the unit ball. In the same way

$$\begin{aligned}
\langle \theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} &\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2(\langle \xi, \mathcal{Q}\xi\rangle_{\mathcal{H}} - \eta) \\
&\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\ \Phi_-\Big\{\widetilde{N}_\lambda\Big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\Big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big)\Big]\Big\} - \eta \\
&\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\ \Phi_- \circ \Phi_+\Big\{N\Big[\|\Pi_k\theta\|_{\mathcal{H}}^{-1}\Big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\Big)\Big]\Big\} - \eta \\
&\geq \Phi_- \circ \Phi_+\Big(N(\Pi_k\theta) - \eta\Big) - \eta.
\end{aligned}$$

Let us decompose $\mathcal{Q}$ in its positive and negative parts and write $\mathcal{Q} = \mathcal{Q}_+ - \mathcal{Q}_-$. Using Lemma 1.24, we deduce the analogous of Proposition 1.26.

Proposition 1.32. *For any threshold $\sigma \in \mathbb{R}_+$ such that $\sigma \leq s_4^2$ and $8\zeta(\sigma) \leq \sqrt{n}$, in the case when $\kappa \geq 3/2$, with probability at least $1-2\epsilon$, for any $\theta \in \mathbb{S}_{\mathcal{H}}$, for any k,*

$$\begin{aligned}
\big|\max\{\langle \theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle \Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}, \sigma\}\big| &\leq 2\max\{\langle \Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}, \sigma\}B_*(\langle \Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}) \\
&\quad + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)} \\
\big|\max\{\langle \theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle \Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}, \sigma\}\big| &\leq 2\max\{\langle \theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\}B_*(\min\{\langle \theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, s_4^2\}) \\
&\quad + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)}.
\end{aligned}$$

Let us consider $\{p_i\}_{i=1}^{+\infty}$ an orthonormal basis of eigenvectors of $\mathcal{G}$ such that the corresponding sequence of eigenvalues $\{\lambda_i, i = 1, \ldots, +\infty\}$ is non-increasing. For any $\theta \in \mathbb{S}_{\mathcal{H}}$, we have

$$\begin{aligned}\big|\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}} - \langle\Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}\big| &= \Big|\sum_{i=1}^{+\infty}(\langle\Pi_k\theta, p_i\rangle_{\mathcal{H}}^2 - \langle\theta, p_i\rangle_{\mathcal{H}}^2)\lambda_i\Big| \\ &= \Big|\sum_{i=1}^{+\infty}(\langle\theta, \Pi_k p_i\rangle_{\mathcal{H}}^2 - \langle\theta, p_i\rangle_{\mathcal{H}}^2)\lambda_i\Big| = \Big|\sum_{i=1}^{+\infty}\langle\theta, \Pi_k p_i + p_i\rangle_{\mathcal{H}}\langle\theta, p_i - \Pi_k p_i\rangle_{\mathcal{H}}\lambda_i\Big| \\ &\leq \inf_{m=1,\ldots,+\infty}\left(\sum_{i=1}^{m-1} 2\lambda_i\|p_i - \Pi_k p_i\|_{\mathcal{H}} + \lambda_m\right).\end{aligned}$$

Indeed,

$$\sum_{i=m}^{+\infty}\langle\Pi_k\theta, p_i\rangle_{\mathcal{H}}^2 \leq 1,$$

so that

$$\sum_{i=m}^{+\infty}\langle\Pi_k\theta, p_i\rangle_{\mathcal{H}}^2\lambda_i \leq \sup_{i=m,\ldots,+\infty}\lambda_i = \lambda_m,$$

and in the same way

$$\sum_{i=m}^{+\infty}\langle\theta, p_i\rangle_{\mathcal{H}}^2\lambda_i \leq \lambda_m.$$

We deduce the following result.

Proposition 1.33. *Consider some threshold $\sigma \in \mathbb{R}_+$ such that $\sigma \leq s_4^2$ and $8\zeta(\sigma) \leq \sqrt{n}$. Assume that $\kappa \geq 3/2$. With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{S}_{\mathcal{H}}$, for any k,*

$$\begin{aligned}\big|\max\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}}, \sigma\}\big| &\leq 2\max\{\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}}, \sigma\}B_*(\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}}) \\ &\quad + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)} + 3\nu_k \\ \big|\max\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}}, \sigma\}\big| &\leq 2\max\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\}B_*(\min\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, s_4^2\}) \\ &\quad + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)} + 2\nu_k,\end{aligned}$$

where B_* is defined in Proposition 1.31 and

$$\begin{aligned}\nu_k &= \inf_{m=1,\ldots,+\infty}\left(\sum_{i=1}^{m-1}\lambda_i\|p_i - \Pi_k p_i\|_{\mathcal{H}} + \lambda_m/2\right) \\ &\leq \inf_{m=1,\ldots,+\infty}\left(\sum_{i=1}^{m-1}\lambda_i\|p_i - \Pi_k p_i\|_{\mathcal{H}} + \mathbf{Tr}(\mathcal{G})/(2m)\right) \underset{k\to+\infty}{\longrightarrow} 0.\end{aligned}$$

Remark that we can use this result to bound $\big|\langle\theta, (\mathcal{G} - \mathcal{Q}_+)\theta\rangle_{\mathcal{H}}\big|$, using the inequality

$$\big|\langle\theta, (\mathcal{G} - \mathcal{Q}_+)\theta\rangle_{\mathcal{H}}\big| \leq \big|\max\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle\theta, \mathcal{G}\theta\rangle_{\mathcal{H}}, \sigma\}\big| + \sigma.$$

Let us mention to conclude this section another way to extend the results from finite dimension to separable Hilbert spaces. This second method consists in defining directly a Gaussian perturbation of the parameter in the Hilbert space. Indeed, taking a basis, we can identify any separable Hilbert space $\mathcal{H}$ with the sequence space $\ell^2 \subset \mathbb{R}^{\mathbb{N}}$. We can take

as our prior parameter distribution π_0 the law on $\mathbb{R}^{\mathbb{N}}$ of a sequence of independent one-dimensional Gaussian random variables with mean 0 and variance $1/\beta$. For any $\theta \in \ell^2$, we can then define the posterior π_θ as $\bigotimes_{i\in\mathbb{N}} \mathcal{N}(\theta_i, 1/\beta) \in \mathcal{M}_+^1(\mathbb{R}^{\mathbb{N}})$. One can prove that

$$\mathcal{K}(\pi_\theta, \pi_0) = \sum_{i=1}^{+\infty} \mathcal{K}\Big(\mathcal{N}(\langle\theta_i, \beta^{-1}), \mathcal{N}(0, \beta^{-1})\Big) = \sum_{i=1}^{+\infty} \frac{\beta}{2}\theta_i^2 = \frac{\beta\|\theta\|_{\mathcal{H}}^2}{2}.$$

Moreover, when $\theta' \sim \pi_\theta$,

$$W(x) = \lim_{n\to+\infty} \sum_{i=0}^{n} \theta_i' x_i, \qquad x \in \ell^2,$$

exists π_θ-almost surely and is a Gaussian process indexed by $\ell^2 \simeq \mathcal{H}$, with mean

$$\mathbb{E}[W(x)] = \langle\theta, x\rangle_{\mathcal{H}}, \qquad x \in \mathcal{H},$$

and covariance

$$\mathbb{E}[W(x)W(y)] - \mathbb{E}[W(x)]\mathbb{E}[W(y)] = \frac{1}{\beta}\langle x, y\rangle_{\mathcal{H}}, \qquad x, y \in \mathcal{H}.$$

Using these properties, we can directly generalize to $\ell^2 \simeq \mathcal{H}$ the PAC-Bayesian bounds stated in $\mathbb{R}^d$ in the previous section.

1.4 Symmetric random matrices

1.4.1 Preliminaries

In this section we generalize the previous results to estimate the expectation of a symmetric random matrix. Indeed the Gram matrix can be viewed as the expectation of the symmetric positive semi-definite random matrix $XX^\top$, where $X \in \mathbb{R}^d$ is a random vector.

Let $A \in M_d(\mathbb{R})$ be a symmetric random matrix of size d. As already observed for the Gram matrix, the expectation of A can be recovered via the polarization identity from the quadratic form

$$N_A(\theta) = \mathbb{E}(\theta^\top A\theta), \qquad \theta \in \mathbb{R}^d,$$

where the expectation is taken with respect to the unknown probability distribution of A on the space of symmetric matrices of size d. We decompose $A = UDU^\top$ in its positive and negative parts

$$A = A_+ - A_-,$$

where A_+ and A_- are symmetric positive semi-definite random matrices defined by

$$A_+ = UD_+U^\top \quad \text{and} \quad A_- = UD_-U^\top,$$

and D_+ (respectively D_-) is the diagonal matrix whose entries are the positive (respectively negative) parts of those of D. According to the previous decomposition, the quadratic form N_A rewrites as

$$\begin{aligned} N_A(\theta) &= \mathbb{E}(\theta^\top A\theta) \\ &= \mathbb{E}(\theta^\top A_+\theta) - \mathbb{E}(\theta^\top A_-\theta) = N_{A_+}(\theta) - N_{A_-}(\theta). \end{aligned}$$

Thus in the following we will consider the case where $A \in M_d(\mathbb{R})$ is a symmetric positive semi-definite random matrix of size d.

1.4.2 Symmetric positive semi-definite random matrices

Let $A \in M_d(\mathbb{R})$ be a symmetric positive semi-definite random matrix of size d and let P be the (unknown) probability distribution of A on the space of symmetric positive semi-definite matrices of size d. Our goal is to estimate the quadratic form

$$N(\theta) = \mathbb{E}(\theta^\top A\theta), \qquad \theta \in \mathbb{R}^d,$$

where $\mathbb{E}$ is the expectation with respect to P, from an i.i.d. sample $A_1, \dots, A_n \in M_d(\mathbb{R})$ of symmetric positive semi-definite matrices drawn according to P.

Since A can be decomposed as

$$A = UDU^\top = \left(UD^{1/2}U^\top\right)^2,$$

where $D \in M_d(\mathbb{R})$ is the diagonal matrix of the eigenvalues of A and $U \in M_d(\mathbb{R})$ is the orthogonal matrix of its eigenvectors, the quadratic form N rewrites as

$$N(\theta) = \mathbb{E}\left(\|A^{1/2}\theta\|^2\right)$$

where $A^{1/2} = UD^{1/2}U^\top$ is the square root of A.

The construction of the robust estimator of the quadratic form N follows the one already done in the case of the Gram matrix with the necessary adjustments.

For any $\lambda > 0$ and for any $\theta \in \mathbb{R}^d$, we define

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi \left(\|A_i^{1/2}\theta\|^2 - \lambda \right),$$

where the influence function ψ is defined as in equation (1.2) on page 16.

We use a PAC-Bayesian approach linked with Gaussian perturbations to define and study a robust estimator of N.

We perturb θ with the Gaussian perturbation $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathrm{I})$ of mean θ and covariance matrix $\beta^{-1}\mathrm{I}$, where $\beta > 0$ is a free real parameter to be determined. The following result holds.

Lemma 1.34. *We have*

$$\int \|A^{1/2}\theta'\|^2 \, \mathrm{d}\pi_\theta(\theta') = \|A^{1/2}\theta\|^2 + \frac{\mathbf{Tr}(A)}{\beta}.$$

Proof. Let $W \in \mathbb{R}^d$ be a Gaussian random variable with mean $A^{1/2}\theta$ and covariance matrix $\beta^{-1}A$. We have

$$\begin{aligned} \mathbb{E}\Big[\|A^{1/2}\theta'\|^2 \mathrm{d}\pi_\theta(\theta')\Big] &= \mathbb{E}(\|W\|^2) \\ &= \sum_{i=1}^{d} \mathbb{E}(\langle W, e_i\rangle^2) \end{aligned}$$

where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. Since $\langle W, e_i\rangle$ is a one-dimensional Gaussian random variable with mean $\langle A^{1/2}\theta, e_i\rangle$ and variance $\beta^{-1}e_i^\top A e_i$, we conclude that

$$\begin{aligned} \mathbb{E}\Big[\|A^{1/2}\theta'\|^2 \mathrm{d}\pi_\theta(\theta')\Big] &= \sum_{i=1}^{d} \Big[\mathbf{Var}(\langle W, e_i\rangle) + \mathbb{E}(\langle W, e_i\rangle)^2\Big] \\ &= \frac{1}{\beta}\,\mathbf{Tr}(A) + \|A^{1/2}\theta\|^2. \end{aligned}$$

□

According to Lemma 1.34 the empirical criterion r_λ rewrites as

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi \left[\int \left(\|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda \right) d\pi_\theta(\theta') \right].$$

With the help of Lemma 1.3 on page 20, we pull the expectation outside the influence function and we get

$$\begin{aligned} \psi \left(\|A^{1/2}\theta\|^2 - \lambda \right) \le \int \chi \left(\|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda \right) \mathrm{d}\pi_\theta(\theta') \\ + \min\Big\{ \log(4), \frac{1}{8}\,\mathbf{Var}\big[\|A^{1/2}\theta'\|^2 \, \mathrm{d}\pi_\theta(\theta')\big] \Big\} \end{aligned}$$

where the function χ is defined in equation (1.4) on page 18 and is such that

$$\chi(z) \le \log(1 + z + z^2/2), \qquad z \in \mathbb{R}.$$

The following lemma provides a bound on the variance.

Lemma 1.35. *We have*

$$\mathbf{Var}\Big[\|A^{1/2}\theta'\|^2 \mathrm{d}\pi_\theta(\theta')\Big] = \frac{4}{\beta}\|A\theta\|^2 + \frac{2}{\beta^2}\,\mathbf{Tr}(A^2).$$

Proof. Let us decompose A into $A = UDU^\top$, where $UU^\top = \mathrm{I}$ and $D = \mathbf{diag}(\lambda_1, \dots, \lambda_d)$. Remark that $U^\top\theta'$ has the same distribution as $U^\top\theta + W$, where $W \sim \mathcal{N}(0, \beta^{-1}\mathrm{I})$ so that

$$\begin{aligned}\mathbf{Var}\Big[\|A^{1/2}\theta'\|^2 \mathrm{d}\pi_\theta(\theta')\Big] &= \mathbf{Var}\Big(\sum_{i=1}^d ((U^\top\theta)_i + W_i)^2\lambda_i\Big)\\ &= \sum_{i=1}^d \lambda_i^2\,\mathbf{Var}\Big[((U^\top\theta)_i + W_i)^2\Big] = \sum_{i=1}^d\Big(\frac{2}{\beta^2} + \frac{4}{\beta}(U^\top\theta)_i^2\Big)\lambda_i^2\\ &= \frac{2}{\beta^2}\,\mathbf{Tr}(A^2) + \frac{4}{\beta}\|A\theta\|^2.\end{aligned}$$

□

As a consequence, we obtain that

$$\begin{aligned}\psi\Big(\|A^{1/2}\theta\|^2 - \lambda\Big) \le \int \chi\Big(\|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta')\\ + \min\Big\{\log(4), \frac{1}{2\beta}\|A\theta\|^2 + \frac{\mathbf{Tr}(A^2)}{4\beta^2}\Big\}.\end{aligned}$$

We now use Lemma 1.4 on page 20 with $m = \|A\theta\|$, $a = \log(4)$, $b = 1/(2\beta)$ and $c = \mathbf{Tr}(A^2)/(4\beta^2)$ to obtain

$$\begin{aligned}\psi\Big(\|A^{1/2}\theta\|^2 - \lambda\Big) \le \int \chi\Big(\|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta')\\ + \int \min\Big\{4\log(2), \frac{1}{\beta}\|A\theta'\|^2 + \frac{\mathbf{Tr}(A^2)}{2\beta^2}\Big\}\,\mathrm{d}\pi_\theta(\theta')\end{aligned}$$

and we conclude, by Lemma 1.5, that

$$\begin{aligned}\psi\Big(\|A^{1/2}\theta\|^2 - \lambda\Big) \le \int \log\Big[1 + \|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda + \frac{1}{2}\Big(\|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda\Big)^2\\ + \frac{c}{\beta}\Big(\|A\theta'\|^2 + \frac{\mathbf{Tr}(A^2)}{2\beta}\Big)\Big]\,\mathrm{d}\pi_\theta(\theta'),\end{aligned}$$

where $c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\Big(\frac{1+2\sqrt{2}}{2}\Big)$.

In terms of the empirical criterion r_λ we have proved that

$$\begin{aligned}r_\lambda(\theta) \le \frac{1}{n}\sum_{i=1}^n \int \log\Big[1 + \|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda + \frac{1}{2}\Big(\|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda\Big)^2\\ + \frac{c}{\beta}\Big(\|A_i\theta'\|^2 + \frac{\mathbf{Tr}(A_i^2)}{2\beta}\Big)\Big]\,\mathrm{d}\pi_\theta(\theta').\end{aligned}$$

We use the PAC-Bayesian inequality presented in Proposition 1.7 on page 22 where the family of posterior distribution is

$$\Big\{\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathbf{I}) \mid \theta \in \mathbb{R}^d,\ \beta > 0\Big\}$$

and the prior distribution is $\pi_0 \sim \mathcal{N}(0, \beta^{-1}\mathbf{I})$, to provide an upper bound for r_λ which is uniform in θ.

Proposition 1.36. *Define* $t(A,\theta') = \|A^{1/2}\theta'\|^2 - \dfrac{\mathbf{Tr}(A)}{\beta} - \lambda$. *With probability at least* $1-\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\begin{aligned} r_\lambda(\theta) &\le \int \mathbb{E}\left[t(A,\theta') + \frac{1}{2}t(A,\theta')^2 + \frac{c}{\beta}\left(\|A\theta'\|^2 + \frac{\mathbf{Tr}(A^2)}{2\beta}\right)\right] \mathrm{d}\pi_\theta(\theta') + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n} \\ &= \mathbb{E}\left[\|A^{1/2}\theta\|^2 - \lambda + \frac{1}{2}\Big(\|A^{1/2}\theta\|^2 - \lambda\Big)^2 + \frac{(c+2)\|A\theta\|^2}{\beta} + \frac{(2+3c)\,\mathbf{Tr}(A^2)}{2\beta^2}\right] \\ &\qquad + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}. \end{aligned}$$

Proof. The proof is essentially the same of the one of Proposition 1.8 on page 25. □

In analogy to the definitions given in equation (1.10) on page 26, we introduce

$$\kappa = \sup_{\substack{\theta\in\mathbb{R}^d \\ \mathbb{E}(\|A^{1/2}\theta\|^2)>0}} \frac{\mathbb{E}(\|A^{1/2}\theta\|^4)}{\mathbb{E}(\|A^{1/2}\theta\|^2)^2}.$$

Using the Cauchy-Schwarz inequality, remark that

$$\begin{aligned} \mathbb{E}(\|A\theta\|^2) \le \mathbb{E}(\|A\|_\infty \|A^{1/2}\theta\|^2) &\le \mathbb{E}(\|A\|_\infty^2)^{1/2}\mathbb{E}(\|A^{1/2}\theta\|^4)^{1/2} \\ &\le \mathbb{E}(\|A\|_\infty^2)^{1/2}\kappa^{1/2}\mathbb{E}(\|A^{1/2}\theta\|^2). \end{aligned}$$

that rewrites as

$$\mathbb{E}\big[\|A\theta\|^2\big] \le \mathbb{E}\big[\|A\|_\infty^2\big]^{1/2}\kappa^{1/2}N(\theta). \tag{1.46}$$

We then deduce from Proposition 1.36 the following result.

Proposition 1.37. *With probability at least* $1-\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\begin{aligned} r_\lambda(\theta) &\le \frac{\kappa}{2}\big[N(\theta)-\lambda\big]^2 + \left[1 + (\kappa-1)\lambda + \frac{(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}}{\beta}\right]\big[N(\theta)-\lambda\big] \\ &\quad + \frac{(\kappa-1)\lambda^2}{2} + \frac{(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}\lambda}{\beta} + \frac{(2+3c)\mathbb{E}\big[\mathbf{Tr}(A^2)\big]}{2\beta^2} + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}. \end{aligned}$$

Remark that if we consider instead of equation (1.46) any upper bound of the form

$$\mathbb{E}(\|A\theta\|^2) \le f\big[\mathbb{E}(A)\big]N(\theta),$$

Proposition 1.37 holds replacing $\mathbb{E}(\|A\|_\infty^2)^{1/2}\kappa^{1/2}$ with $f\big[\mathbb{E}(A)\big]$. Similarly we can replace $\mathbb{E}\big[\mathbf{Tr}(A^2)\big]$ with an upper bound.

We also observe that Proposition 1.37 is the analogous in the case of a symmetric positive semi-definite random matrix of Proposition 1.10 on page 27. Hence according to section 1.2.2 we define a robust estimator as follows.
Let $\Lambda \subset \left(\mathbb{R}_+ \setminus \{0\}\right)^2$ be a finite set of possible values of the couple of parameters (λ, β). We consider the family of estimators

$$\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2}, \quad (\lambda, \beta) \in \Lambda,$$

where $\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \leq 0\}$ and we define

$$(\widehat{\lambda}, \widehat{\beta}) = \arg \min_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\Big[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\Big],$$

where $B_{\lambda,\beta}$ is defined in equation (1.20) on page 31. We consider as an estimator

$$\widehat{N}(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta), \quad \theta \in \mathbb{R}^d. \tag{1.47}$$

We now directly state the analogous of Proposition 1.19 on page 38, choosing the grid

$$\Lambda = \{(\lambda_j, \beta_j) \mid 0 \leq j < K\},$$

where $a > 0$ and

$$\begin{aligned}
K &= 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil, \\
\lambda_j &= \sqrt{\frac{2}{(\kappa-1)n}\left(\frac{(2+3c)\mathbb{E}(\mathbf{Tr}(A^2))}{4(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)}\exp(ja) + \log(K/\epsilon)\right)}, \\
\beta_j &= \sqrt{2(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)\exp\left[-(j-1/2))a\right]}.
\end{aligned}$$

Proposition 1.38. *Let $\sigma \in \mathbb{R}_+$ be some energy level such that $\sigma \leq \mathbb{E}(\|A\|_\infty^2)^{1/2}$. With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \leq B_*\Big[\|\theta\|^{-2}N(\theta)\Big],$$

where B_* is defined as

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & [6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \leq \sqrt{n} \\ +\infty & \textit{otherwise} \end{cases}$$

and

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\mathbb{E}[\mathbf{Tr}(A^2)]}{4(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}}{t}}\cosh(a/2).$$

We conclude this section by making some comments on the above result.

As already observed in the case of the Gram matrix, the quantity $\kappa - 1$ corresponds to a variance term, and more precisely to

$$\mathbf{Var}\Big(\|A^{1/2}\theta\|^2\Big)\ N(\theta)^{-2}.$$

Moreover the bound does not depend explicitly on the dimension d and in particular the dimension is replaced by the entropy terms

$$\frac{\mathbb{E}(\mathbf{Tr}(A^2))}{\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}\max\{\|\theta\|^{-2}N(\theta),\sigma\}} \quad \text{and} \quad \frac{\mathbb{E}(\|A\|_\infty^2)^{1/2}}{\kappa^{1/2}\max\{\|\theta\|^{-2}N(\theta),\sigma\}}.$$

Let us remark that they satisfy

$$\frac{\mathbb{E}(\|A\|_\infty^2)^{1/2}}{\kappa^{1/2}} \leq \frac{\mathbb{E}(\mathbf{Tr}(A^2))}{\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)^{1/2}} \leq \mathbb{E}(\mathbf{Tr}(A)). \tag{1.48}$$

Indeed, $\|A\|_\infty^2 \leq \mathbf{Tr}(A^2)$,

$$\mathbb{E}(\mathbf{Tr}(A^2)) \leq \mathbb{E}(\|A\|_\infty \mathbf{Tr}(A)) \leq \mathbb{E}(\|A\|_\infty^2)^{1/2}\mathbb{E}(\mathbf{Tr}(A)^2)^{1/2},$$

and, denoting by $\{e_i\}_{i=1}^d$ an orthonormal basis of $\mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}(\mathbf{Tr}(A)^2) &= \sum_{\substack{1\leq i\leq d,\\ 1\leq j\leq d}} \mathbb{E}\Big(\|A^{1/2}e_i\|^2\|A^{1/2}e_j\|^2\Big)\\ &\leq \sum_{\substack{1\leq i\leq d,\\ 1\leq j\leq d}} \mathbb{E}\Big(\|A^{1/2}e_i\|^4\Big)^{1/2}\mathbb{E}\Big(\|A^{1/2}e_j\|^4\Big)^{1/2}\\ &\leq \kappa \sum_{\substack{1\leq i\leq d,\\ 1\leq j\leq d}} \mathbb{E}\Big(\|A^{1/2}e_i\|^2\Big)\mathbb{E}\Big(\|A^{1/2}e_j\|^2\Big) = \kappa\mathbb{E}(\mathbf{Tr}(A))^2.\end{aligned}$$

As a consequence, ζ can be bounded as shown

$$\begin{aligned}\zeta(t) \leq \sqrt{2(\kappa-1)\left(\frac{(2+3c)\mathbb{E}[\mathbf{Tr}(A)]}{4(2+c)t} + \log(K/\epsilon)\right)}\cosh(a/4)\\ + \sqrt{\frac{2(2+c)\kappa\ \mathbb{E}[\mathbf{Tr}(A)]}{t}}\cosh(a/2).\end{aligned}$$

Let us remark also that in the case when $\mathbf{rank}(A) = 1$ (as it is for the Gram matrix) then $\|A\|_\infty^2 = \mathbf{Tr}(A^2)$, so that the two entropy terms are the same. As the entropy terms are dominated by $\mathbb{E}(\mathbf{Tr}(A))$, the result can be generalized to the case where A is a random symmetric positive semi-definite operator in an infinite-dimensional Hilbert space with the only additional assumption that $\mathbb{E}[\mathbf{Tr}(A)] < +\infty$. For more detail we refer to section 1.3.

We also recall that if we choose the threshold σ such that $8\zeta(\sigma) = \sqrt{n}$, so that $B_*(t) < +\infty$ for any $t \in \mathbb{R}_+$, then σ goes to zero as the sample size grows to infinity as shown in equation (1.35).

We conclude by recalling that the estimator $\widehat{N}$ is not a quadratic form but that it is possible to deduce a positive semi-definite quadratic estimator of N as explained in section 1.2.3 on page 39.

1.5 Covariance matrix

Let $X \in \mathbb{R}^d$ be a random vector distributed according to the unknown probability measure $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$. The covariance matrix of X

$$\Sigma = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^\top\Big]$$

is a symmetric positive semi-definite random matrix. Our goal is to estimate, uniformly in θ, the quadratic form associated with Σ

$$N(\theta) = \mathbb{E}\Big(\langle \theta, X - \mathbb{E}(X)\rangle^2\Big), \quad \theta \in \mathbb{R}^d,$$

from an i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ drawn according to P. We cannot use the results we have proved for the Gram matrix, since the quadratic form N depends on the unknown quantity $\mathbb{E}(X)$. However we can find a workaround, using the results of the previous section about random symmetric matrices. Indeed, the quadratic form N can be written as

$$N(\theta) = \frac{1}{2}\,\mathbb{E}\Big(\langle \theta, X - X'\rangle^2\Big)$$

where X' is an independent copy of X. More generally, given $q \in \mathbb{N}$, we may consider q independent copies $X^{(1)}, \dots, X^{(q)}$ of X and the random matrix

$$A = \frac{1}{q(q-1)} \sum_{1 \le j < k \le q} \big(X^{(j)} - X^{(k)}\big)\big(X^{(j)} - X^{(k)}\big)^\top.$$

Remark that

$$N(\theta) = \frac{1}{q(q-1)}\,\mathbb{E}\left(\sum_{1 \le j < k \le q} \langle \theta, X^{(j)} - X^{(k)}\rangle^2\right) = \mathbb{E}\big(\,\theta^\top A\,\theta\,\big).$$

Consider the sample $A_1, \dots, A_{\lfloor n/q \rfloor}$ of independent copies of A defined as

$$A_i = \frac{1}{q(q-1)} \sum_{(i-1)q < j < k \le iq} (X_j - X_k)(X_j - X_k)^\top.$$

We can use the results of the previous section to define a robust estimator of $N(\theta)$. We will discuss later how to choose q.

Let us introduce

$$\kappa' = \sup_{\substack{\theta \in \mathbb{R}^d, \\ \mathbb{E}(\|A^{1/2}\theta\|^2) > 0}} \frac{\mathbb{E}(\|A^{1/2}\|^4)}{\mathbb{E}\big(\|A^{1/2}\theta\|^2\big)^2},$$

$$\text{and } \kappa = \sup_{\substack{\theta \in \mathbb{R}^d, \\ \mathbb{E}(\langle \theta, X - \mathbb{E}(X)\rangle^2) > 0}} \frac{\mathbb{E}\big(\langle \theta, X - \mathbb{E}(X)\rangle^4\big)}{\mathbb{E}\big(\langle \theta, X - \mathbb{E}(X)\rangle^2\big)^2}.$$

Lemma 1.39. *The two kurtosis coefficients introduced above are related by the relation*

$$\kappa' \le 1 + \tau_q(\kappa)/q,$$

where $\tau_q(\kappa) = \kappa - 1 + \dfrac{2}{q-1}$.

Proof. Replacing X with $X - \mathbb{E}(X)$ we may assume during the proof that $\mathbb{E}(X) = 0$. Recalling the definition of covariance, we have

$$\begin{aligned}
\mathbb{E}(\|A^{1/2}\theta\|^4) &= \mathbb{E}\left[\left(\frac{1}{q(q-1)}\sum_{1\le j<k\le q}\langle\theta, X^{(j)} - X^{(k)}\rangle^2\right)^2\right] \\
&= \frac{1}{q^2(q-1)^2}\sum_{\substack{1\le j<k\le q\\ 1\le s<t\le q}}\mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^2\langle\theta, X^{(s)} - X^{(t)}\rangle^2\Big) \\
&= \frac{1}{q^2(q-1)^2}\Bigg\{\sum_{\substack{1\le j<k\le q\\ 1\le s<t\le q}}\mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^2\Big)\mathbb{E}\Big(\langle\theta, X^{(s)} - X^{(t)}\rangle^2\Big) \\
&\quad + \sum_{1\le j<k\le q}\mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^4\Big) - \mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^2\Big)^2 \\
&\quad + \sum_{\substack{1\le j<k\le q\\ 1\le s<t\le q\\ |\{j,k\}\cap\{s,t\}|=1}}\Big[\mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^2\langle\theta, X^{(s)} - X^{(t)}\rangle^2\Big) \\
&\quad - \mathbb{E}\Big(\langle\theta, X^{(j)} - X^{(k)}\rangle^2\Big)\mathbb{E}\Big(\langle\theta, X^{(s)} - X^{(t)}\rangle^2\Big)\Big]\Bigg\} \\
&= \frac{1}{4}\mathbb{E}\Big(\langle\theta, X^{(2)} - X^{(1)}\rangle^2\Big)^2 + \frac{1}{2q(q-1)}\mathbb{E}\Big(\langle\theta, X^{(2)} - X^{(1)}\rangle^4\Big) - \mathbb{E}\Big(\langle\theta, X^{(2)} - X^{(1)}\rangle^2\Big)^2 \\
&\quad + \frac{q-2}{q(q-1)}\Big[\mathbb{E}\Big(\langle\theta, X^{(1)} - X^{(2)}\rangle^2\langle\theta, X^{(1)} - X^{(3)}\rangle^2\Big) - \mathbb{E}\Big(\langle\theta, X^{(1)} - X^{(2)}\rangle^2\Big)^2\Big]
\end{aligned}$$

Define $W_j = \langle\theta, X^{(j)}\rangle$.

$$\begin{aligned}
\mathbb{E}\big((W_1 - W_2)^2\big)^2 &= 4N(\theta)^2, \\
\mathbb{E}\big((W_1 - W_2)^4\big) &= \mathbb{E}(W_1^4) + 6\mathbb{E}(W_1^2)\mathbb{E}(W_2^2) + \mathbb{E}(W_2^4) \\
&= 2\mathbb{E}(W_1^4) + 6\mathbb{E}(W_1^2)^2 \le (2\kappa + 6)N(\theta)^2, \\
\mathbb{E}\Big((W_1 - W_2)^2(W_1 - W_3)^2\Big) &= \mathbb{E}(W_1^4) + 3\mathbb{E}(W_2^2)^2 \le (\kappa + 3)N(\theta)^2.
\end{aligned}$$

Therefore

$$\mathbb{E}\Big(\|A^{1/2}\theta\|^4\Big) \le \left(1 + \frac{(q-2)(\kappa-1)}{q(q-1)} + \frac{(\kappa+1)}{q(q-1)}\right)N(\theta)^2 = \left(1 + \frac{\tau_q(\kappa)}{q}\right)N(\theta)^2,$$

so that $\kappa' \le 1 + \tau_q(\kappa)/q$, since $\mathbb{E}(\|A^{1/2}\theta\|^2) = N(\theta)$. □

Remark that

$$\mathbb{E}(\mathbf{Tr}(A)) = \mathbf{Tr}(\mathbb{E}(A)) = \mathbf{Tr}(\Sigma) = \mathbb{E}\Big(\|X - \mathbb{E}(X)\|^2\Big)$$

Proposition 1.40. *Let $\widehat{N}(\theta)$ be the estimator defined in Proposition 1.38, using the entropy bound defined in terms of $\mathbb{E}(\mathbf{Tr}(A)) = \mathbf{Tr}(\Sigma)$ by equation (1.48) in the definition of the grid of parameters Λ. For any energy level $\sigma \in \mathbb{R}_+$ such that $\sigma \le \mathbf{Tr}(\Sigma)$, with probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \le B_*\Big(\|\theta\|^{-2}N(\theta)\Big),$$

where

$$B_*(t) = \begin{cases} \dfrac{(q\lfloor n/q\rfloor)^{1/2}\zeta(\max\{t,\sigma\})}{1-4(q\lfloor n/q\rfloor)^{1/2}\zeta(\max\{t,\sigma\})}, & \text{if } \Big(6+q/\tau_q(\kappa)\Big)\zeta(\max\{t,\sigma\}) \le (q\lfloor n/q\rfloor)^{1/2}, \\ +\infty, & \text{otherwise,} \end{cases}$$

$$\zeta(t) = \sqrt{2\tau_q(\kappa)\Big(\frac{(2+3c)\,\mathbf{Tr}(\Sigma)}{4(2+c)t} + \log(K/\epsilon)\Big)}\cosh(a/4) + \sqrt{\frac{2(2+c)(q+\tau_q(\kappa))\,\mathbf{Tr}(\Sigma)}{t}}\cosh(a/2),$$

$$K = 1 + \left\lceil a^{-1}\log\Big(\frac{\lfloor n/q\rfloor}{72(2+c)(1+\tau_q(\kappa)/q)^{1/2}}\Big)\right\rceil,$$

$\tau_q(\kappa) = \kappa - 1 + \dfrac{2}{q-1}$ *and* $c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\Big(\frac{1+2\sqrt{2}}{2}\Big)$.

Proof. The proof follows from Proposition 1.38 replacing κ by κ' and n by $\lfloor n/q\rfloor$. Indeed, using equation (1.48), we get

$$\zeta(t) \le \sqrt{2(\kappa'-1)\Big(\frac{(2+3c)\mathbb{E}[\mathbf{Tr}(A)]}{4(2+c)t} + \log(K/\epsilon)\Big)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa'\mathbb{E}[\mathbf{Tr}(A)]}{t}}\cosh(a/2).$$

Since by Lemma 1.39 we see that $\kappa' - 1 \le \tau_q(\kappa)/q$, we conclude the proof. □

As stated at the beginning of this proposition, we have used here rather crude entropy bounds. We can improve the constants by evaluating more carefully $\mathbb{E}(\|A\theta\|^2)$ and $\mathbb{E}[\mathbf{Tr}(A^2)]$.

Lemma 1.41. *We have*

$$\mathbb{E}\Big(\|A\theta\|^2\Big) \le \Big(1 - \frac{q-2}{q(q-1)}\Big)\|\Sigma\|_\infty N(\theta) + \frac{1}{q}\Big(\kappa + \frac{1}{q-1}\Big)\,\mathbf{Tr}(\Sigma)N(\theta) \tag{1.49}$$

$$\mathbb{E}[\mathbf{Tr}(A^2)] \le \Big(1 - \frac{q-2}{q(q-1)}\Big)\,\mathbf{Tr}(\Sigma^2) + \frac{1}{q}\Big(\kappa + \frac{1}{q-1}\Big)\,\mathbf{Tr}(\Sigma)^2. \tag{1.50}$$

Proof. Replacing X with $X - \mathbb{E}(X)$ we may assume that $\mathbb{E}(X) = 0$. Recall that

$$\mathbb{E}(\|X\|^4) \le \kappa\mathbb{E}(\|X\|^2)^2 = \kappa\,\mathbf{Tr}(\Sigma)^2$$

and $\mathbb{E}(\langle X^{(1)}, X^{(2)}\rangle^2) = \mathbf{Tr}(\Sigma^2)$. We observe that

$$\mathbb{E}(\|A\theta\|^2) = \mathbb{E}\left(\frac{1}{q^2(q-1)^2}\sum_{\substack{1\le j<k\le q\\ 1\le s<t\le q}}\langle\theta, X^{(j)}-X^{(k)}\rangle\langle X^{(j)}-X^{(k)}, X^{(s)}-X^{(t)}\rangle\langle X^{(s)}-X^{(t)},\theta\rangle\right)$$

and

$$\begin{aligned}
&\mathbb{E}\Big(\langle\theta, X^{(1)}-X^{(2)}\rangle\langle X^{(1)}-X^{(2)}, X^{(3)}-X^{(4)}\rangle\langle X^{(3)}-X^{(4)},\theta\rangle\Big)\\
&\qquad = 4\mathbb{E}\Big(\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)},\theta\rangle\Big) = 4\|\Sigma\theta\|^2 \le 4\|\Sigma\|_\infty N(\theta),\\
&\mathbb{E}\Big(\langle\theta, X^{(1)}-X^{(2)}\rangle\langle X^{(1)}-X^{(2)}, X^{(1)}-X^{(3)}\rangle\langle X^{(1)}-X^{(3)},\theta\rangle\Big)\\
&\qquad = \mathbb{E}\Big(\langle\theta, X^{(1)}\rangle^2\|X^{(1)}\|^2\Big) + 3\mathbb{E}\Big(\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)},\theta\rangle\Big)\\
&\qquad \le \kappa\,\mathbf{Tr}(\Sigma)N(\theta) + 3\|\Sigma\|_\infty N(\theta),\\
&\mathbb{E}\Big(\langle\theta, X^{(1)}-X^{(2)}\rangle\langle X^{(1)}-X^{(2)}, X^{(1)}-X^{(2)}\rangle\langle X^{(1)}-X^{(2)},\theta\rangle\Big)\\
&\qquad = 2\mathbb{E}\Big(\langle\theta, X^{(1)}\rangle^2\|X^{(1)}\|^2\Big) + 2\mathbb{E}\Big(\langle\theta, X^{(1)}\rangle^2\Big)\mathbb{E}\Big(\|X^{(1)}\|^2\Big)\\
&\qquad\quad + 4\mathbb{E}\Big(\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)},\theta\rangle\Big)\\
&\qquad \le 2(\kappa+1)\,\mathbf{Tr}(\Sigma)N(\theta) + 4\|\Sigma\|_\infty N(\theta)
\end{aligned}$$

Therefore,

$$\mathbb{E}\Big(\|A\theta\|^2\Big) \le \Big(1-\frac{q-2}{q(q-1)}\Big)\|\Sigma\|_\infty N(\theta) + \frac{1}{q}\Big(\kappa+\frac{1}{q-1}\Big)\,\mathbf{Tr}(\Sigma)N(\theta).$$

which proves the first inequality. In the same way,

$$\mathbb{E}\Big(\mathbf{Tr}(A^2)\Big) = \mathbb{E}\Bigg(\frac{1}{q^2(q-1)^2}\sum_{\substack{1\le j<k\le q\\ 1\le s<t\le q}}\langle X^{(j)}-X^{(k)}, X^{(s)}-X^{(t)}\rangle^2\Bigg),$$

and

$$\begin{aligned}
\mathbb{E}\Big(\langle X^{(1)}-X^{(2)}, X^{(3)}-X^{(4)}\rangle^2\Big) &= 4\mathbb{E}\Big(\langle X^{(1)}, X^{(2)}\rangle^2\Big)\\
&= 4\,\mathbf{Tr}(\Sigma^2),\\
\mathbb{E}\Big(\langle X^{(1)}-X^{(2)}, X^{(1)}-X^{(3)}\rangle^2\Big) &= \mathbb{E}\Big(\|X^{(1)}\|^4\Big) + 3\mathbb{E}\Big(\langle X^{(1)}, X^{(2)}\rangle^2\Big)\\
&\le \kappa\,\mathbf{Tr}(\Sigma)^2 + 3\,\mathbf{Tr}(\Sigma^2),\\
\mathbb{E}\Big(\langle X^{(1)}-X^{(2)}, X^{(1)}-X^{(2)}\rangle^2\Big) &= 2\mathbb{E}\Big(\|X^{(1)}\|^4\Big) + 2\mathbb{E}\Big(\|X^{(1)}\|^2\Big)^2 + 4\mathbb{E}\Big(\langle X^{(1)}, X^{(2)}\rangle^2\Big)\\
&\le 2(\kappa+1)\,\mathbf{Tr}(\Sigma)^2 + 4\,\mathbf{Tr}(\Sigma^2),
\end{aligned}$$

which concludes the proof. □

Using these tighter bounds, we can improve ζ to

$$\begin{aligned}
\zeta(t) = \sqrt{2\tau_q(\kappa)\Bigg(\frac{(2+3c)\big[(1-\frac{q-2}{q(q-1)})\,\mathbf{Tr}(\Sigma^2) + \frac{1}{q}(\kappa+\frac{1}{q-1})\,\mathbf{Tr}(\Sigma)^2\big]}{4(2+c)\big[(1-\frac{q-2}{q(q-1)})\|\Sigma\|_\infty + \frac{1}{q}(\kappa+\frac{1}{q-1})\,\mathbf{Tr}(\Sigma)\big]t} + \log(K/\epsilon)\Bigg)}\\
\times\cosh(a/4)\\
+\sqrt{\frac{2(2+c)\big[(1-\frac{q-2}{q(q-1)})q\|\Sigma\|_\infty + (\kappa+\frac{1}{q-1})\,\mathbf{Tr}(\Sigma)\big]}{t}}\cosh(a/2).
\end{aligned}$$

Therefore, in the case when,

$$q\|\Sigma\|_\infty \leq \mathbf{Tr}(\Sigma),$$

we can take

$$\zeta(t) = \sqrt{2\tau_q(\kappa)\left(\frac{(2+3c)\,\mathbf{Tr}(\Sigma)}{4(2+c)t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)(\kappa+1+\frac{2}{q(q-1)})\,\mathbf{Tr}(\Sigma)}{t}}\cosh(a/2).$$

If we compare this results with the bound obtained in Proposition 1.19 on page 38 for the Gram matrix estimator, we see that the first appearance of κ in the definition of ζ has been replaced with

$$\tau_q(\kappa) + 1 = \kappa + \frac{2}{q-1},$$

and that the second appearance of κ has been replaced with

$$\kappa + 1 + \frac{2}{q(q-1)}.$$

Thus, when $\|\Sigma\|_\infty \leq \mathbf{Tr}(\Sigma)/2$, that is not a very strong hypothesis, we can take at least $q = 2$, and obtain an improved bound for the estimation of Σ that is not much larger than the bound for the estimation of the centered Gram matrix, that requires the knowledge of $\mathbb{E}(X)$, since the difference between the two bounds is just a matter of replacing κ with $\kappa+2$.

1.6 Empirical results

In this section we present some empirical results that show the performance of our robust estimator Q.

We use in these experiments a simplified construction that does not lead exactly to the estimator Q, for which we have proved theoretical results in the previous sections, but should nonetheless exhibit the same kind of behaviour.

We use an iterative scheme based on the polarization formula to estimate the coefficients of the Gram matrix in an orthonormal basis of eigenvectors and then update this basis iteratively to a basis of eigenvectors of the current estimate.

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be a sample drawn according to the probability distribution P and let $\lambda > 0$. Let $p \in \mathbb{R}^n$ and define $S(p, \lambda)$ as the solution of

$$\sum_{i=1}^{n} \psi\left[\lambda\left(S(p,\lambda)^{-1} p_i^2 - 1\right)\right] = 0.$$

In practice we compute $S(p, \lambda)$ using the Newton algorithm. We observe that, when $p_i = \langle \theta, X_i \rangle$ and λ is suitably chosen, $S(p, \lambda)$ is an approximation of the estimator $\widehat{N}(\theta)$ of the quadratic form $N(\theta)$.

Define $S(p)$ as the solution obtained when the parameter λ is set to

$$\lambda = m \sqrt{\frac{1}{v}\left[\frac{2}{n}\log(\epsilon^{-1})\left(1 - \frac{2}{n}\log(\epsilon^{-1})\right)\right]}$$

where $m = \dfrac{1}{n}\sum_{i=1}^{n} p_i^2$, $v = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(p_i^2 - m\right)^2$ and $\epsilon = 0.1$.

According to [7], this value of the scale parameter λ should be close to optimal for the estimation of a single expectation from an empirical sample distribution.

Let $\bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_d \geq 0$ be the eigenvalues of the empirical Gram matrix $\bar{G}$, that will be our starting point, and let $u_1, \dots, u_d$ be a corresponding orthonormal basis of eigenvectors. We decompose the empirical Gram matrix as

$$\bar{G} = UDU^\top$$

where U is the orthogonal matrix whose columns are the eigenvectors of $\bar{G}$ and D is the diagonal matrix $D = \operatorname{diag}(\bar{\lambda}_1, \dots, \bar{\lambda}_d)$. We observe that, by the polarization formula,

$$u_i^\top G u_j = \frac{1}{4}\left[N(u_i + u_j) - N(u_i - u_j)\right], \quad i, j = 1, \dots, d,$$

where $N(u_i + \sigma u_j)$, with $\sigma \in \{+1, -1\}$, is approximated by

$$S\left(\langle u_i + \sigma u_j, X_\ell\rangle^2 \ , \ 1 \leq \ell \leq n\right).$$

Taking notation, for any $n \times d$ matrix W, we define $C(W)$ as the $d \times d$ matrix of entries

$$C(W)_{i,j} = \frac{1}{4}\left[S\Big(\left(W_{\ell,i} + W_{\ell,j}\right)^2 \mid 1 \leq \ell \leq n\Big) - S\Big(\left(W_{\ell,i} - W_{\ell,j}\right)^2 \mid 1 \leq \ell \leq n\Big)\right].$$

Let Y be the matrix whose i-th row is the vector X_i, so that

$$(YU)_{\ell,k} = \langle u_k, X_\ell \rangle, \quad 1 \leq \ell \leq n,\ 1 \leq k \leq d.$$

We update the Gram matrix estimate to

$$Q_0 = UC(YU)U^\top.$$

Then we iterate the update scheme, decomposing Q_0 as

$$Q_0 = O_0 D_0 O_0^\top,$$

where $O_0 O_0^\top = O_0^\top O_0 = \mathrm{I}$ and D_0 is a diagonal matrix and computing

$$Q_1 = O_0 C(YO_0)O_0^\top.$$

The inductive update step is more generally the following. Assuming we have constructed Q_k, we decompose it as

$$Q_k = O_k D_k O_k^\top$$

where $O_k O_k^\top = O_k^\top O_k = \mathrm{I}$ and D_k is a diagonal matrix and we define the new updated estimator of G as

$$Q_{k+1} = O_k C(YO_k)O_k^\top.$$

We stop this iterative estimation scheme when $\|Q_k - Q_{k-1}\|_F$ falls under a given threshold. In the following numerical experiment we more simply performed four updates. We take as our robust estimator of G the last update Q_k.

We now present an example of the performance of this estimator, for some i.i.d. sample of size $n = 100$ in $\mathbb{R}^{10}$ drawn according to the Gaussian mixture distribution

$$\mathrm{P} = (1-\alpha)\mathcal{N}(0, M_1) + \alpha\mathcal{N}(0, 16\,\mathbf{I}),$$

where $\alpha = 0.05$ and

$$M_1 = \left[\begin{array}{cc:ccc} 2 & 1 & & 0 & \\ 1 & 1 & & & \\ \hdashline & & 0.01 & & \\ 0 & & & \ddots & 0 \\ & & 0 & & \\ & & & & 0.01 \end{array}\right].$$

The Gram matrix of P is equal to

$$G = (1-\alpha)M_1 + 16\,\alpha\,\mathbf{I} = \left[\begin{array}{cc:ccc} 2,7 & 0.95 & & 0 & \\ 0.95 & 1.75 & & & \\ \hdashline & & 0.8095 & & \\ 0 & & & \ddots & 0 \\ & & 0 & & \\ & & & & 0.8095 \end{array}\right].$$

This example illustrates a favorable situation where the performance of the robust estimator is particularly striking when compared to the empirical Gram matrix. As it can be seen on the expression of the sample distribution as well as on the configuration plots below, this is a situation of intermittent high variance : the sample is a mixture of a rare high variance signal and a frequent low variance more structured signal.

We tested the algorithm on 500 different samples drawn according to the Gaussian mixture distribution defined above. Random sample configurations are presented in fig. 1.1.

Figure 1.2 shows that the robust estimator Q significantly improves the error in terms of square of the Frobenius norm when compared to the empirical estimator $\bar{G}$. The red solid line represents the empirical quantile function of the errors of the robust estimator, whereas the blue dotted line represents the quantiles of $\|\bar{G} - G\|_F^2$.
This quantile function is obtained by sorting the 500 empirical errors in increasing order.

The mean of the square distances $\|Q - G\|_F^2$ on 500 trials is 5.6 ± 0.4, where the indicated mean estimator and confidence interval is the non-asymptotic confidence interval given by Proposition 2.4 of [7] at confidence level 0.99. In the case of the empirical estimator, the mean is 15.5 ± 2. The empirical standard deviations are respectively 2 and 10. So we see that in this case the robust estimator reliably decreases the error by a factor larger than 2 and also produces errors with a much smaller standard deviation from sample to sample.

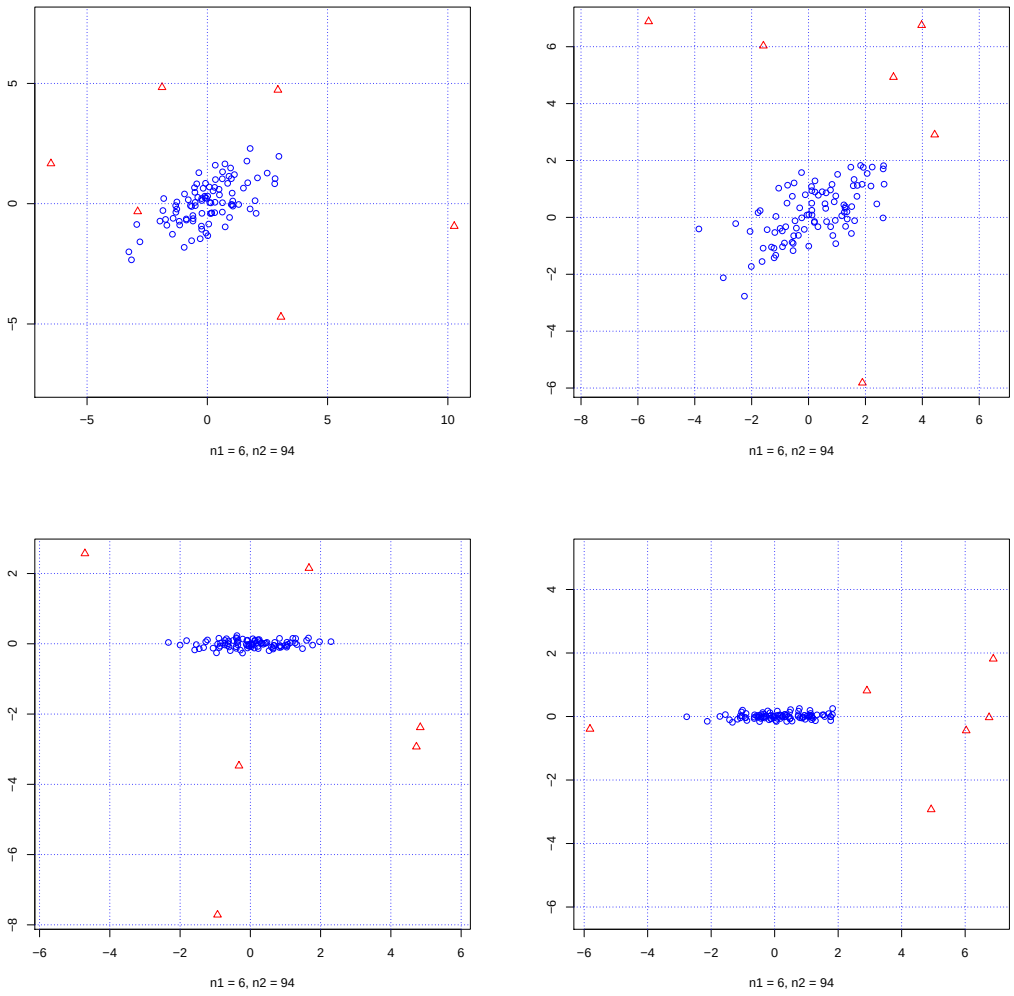

Figure 1.1: Two data samples projected onto the two first coordinates (above) and the second and third coordinates (below). Blue circles are drawn from the most frequent distribution and red triangles from the less frequent one.

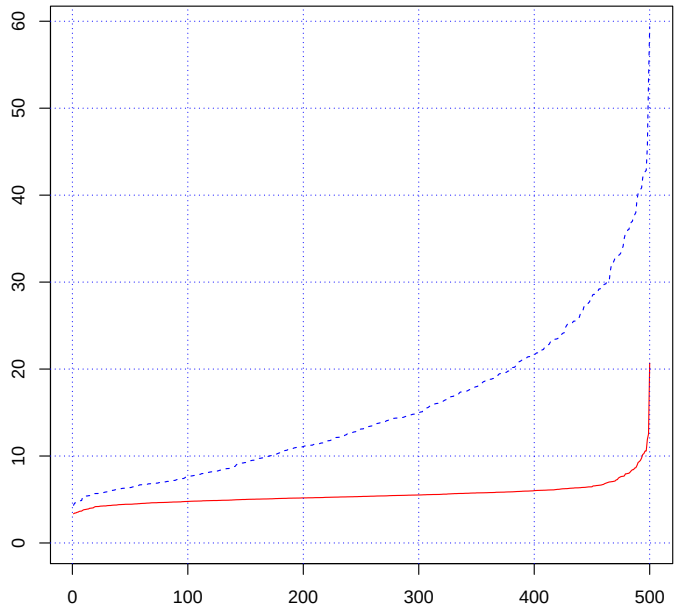

Figure 1.2: The red solid line represents the distances $\|Q - G\|_F^2$, the blue dotted line represents the distances $\|\bar{G} - G\|_F^2$.

Chapter 2

The Empirical Gram Matrix

We consider the problem of estimating the Gram matrix via the usual empirical estimator and we provide dimension-free bounds on the approximation error, using the robust estimator introduced in chapter 1 as a tool.

2.1 Introduction

Let $X \in \mathbb{R}^d$ be a random vector distributed according to the unknown probability measure $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$. We consider the problem of estimating the Gram matrix

$$G = \mathbb{E}(XX^\top) = \int xx^\top \mathrm{dP}(x)$$

from an i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ distributed according to P, by using the usual empirical estimator

$$\bar{G} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top.$$

It is known that, by the law of large numbers, the empirical Gram matrix $\bar{G}$ converges to G as n goes to infinity.

Many results concerning the Gram matrix estimate, or more generally the covariance matrix estimate, can be found in the literature, e.g. [34], [35], [25]. We briefly describe the approach presented in [34] and [35].

In Vershynin [35] the problem of estimating the Gram matrix follows from the study of spectral properties of random matrices with independent rows. The author observes that, denoting by A the matrix whose i-th row is the vector X_i, the empirical Gram matrix can be expressed as

$$\bar{G} = \frac{1}{n}A^\top A$$

where A has, by construction, independent rows. We denote by $\|\cdot\|_\infty$ the operator norm. The following result holds.

Proposition 2.1. (Corollary 5.52, [35]) *Assume that the probability distribution* P *is supported in some Euclidean ball of radius* $\sqrt{m}$. *Let* $\delta \in]0,1[$. *With probability at least* $1-\epsilon$, *if*

$$n \geq Cm\delta^{-2}\|G\|_\infty^{-1}\log(d/\epsilon),$$

where C is an absolute constant, then

$$\|\bar{G} - G\|_\infty \leq \delta \|G\|_\infty.$$

Typically the result is used with $m = \mathcal{O}(d\|G\|_\infty)$, so that the conclusion of Proposition 2.1 holds with sample size

$$n \geq C\delta^{-2} d \log(d/\epsilon).$$

This means that we need to choose the sample size $n = \mathcal{O}(d\log(d))$ to approximate the Gram matrix by the empirical Gram matrix.

A different approach is used in Tropp [34], where a bound on the operator norm of the approximation error is deduced from the matrix Bernstein inequality. The author observes that the empirical Gram matrix is an unbiased estimator of the Gram matrix, i.e. $\mathbb{E}[\bar{G}] = G$, which can be expressed as a sum of independent random matrices. Thus, a strategy to bound the approximation error is to look whether the random matrix $\bar{G}$ deviates from its mean. It is known that in the case of a random variable Z expressed as a sum of independent random variables the Bernstein inequality provides a bound on the probability that Z takes values far from its mean. We recall the Bernstein inequality.

Proposition 2.2. (Bernstein inequality) *Let $S_1, \ldots, S_n \in \mathbb{R}$ be independent random variables such that, for any $k = 1, \ldots, n$,*

$$|S_k - \mathbb{E}(S_k)| \leq R.$$

Let $Z = \sum_{k=1}^n S_k$ and $\sigma^2 = \mathbb{E}\Big[\big(Z - \mathbb{E}(Z)\big)^2\Big]$. Then, for $t \geq 0$,

$$\mathbb{P}\Big(|Z - \mathbb{E}(Z)| \geq t\Big) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

What is proved in [34] is that the scalar Bernstein inequality extends directly to matrices, that is, it is possible to obtain an exponential concentration inequality for the operator norm of a sum of independent random matrices.

Proposition 2.3. (Matrix Bernstein inequality, [34]). *Let $S_1, \ldots, S_n \in M_{d_1,d_2}(\mathbb{R})$ be independent random matrices of size $d_1 \times d_2$, such that, for any $k = 1, \ldots, n$,*

$$\big\|S_k - \mathbb{E}(S_k)\big\|_\infty \leq R.$$

Let $Z = \sum_{k=1}^n S_k$ and

$$\sigma^2 = \max\left\{\Big\|\mathbb{E}\Big[(Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^\top\Big]\Big\|_\infty, \Big\|\mathbb{E}\Big[(Z - \mathbb{E}(Z))^\top (Z - \mathbb{E}(Z))\Big]\Big\|_\infty\right\}.$$

Then, for $t \geq 0$,

$$\mathbb{P}\left[\|Z - \mathbb{E}(Z)\|_\infty \geq t\right] \leq (d_1 + d_2)\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

We now describe how to apply the matrix Bernstein inequality to study of the Gram matrix.

Let $X \in \mathbb{R}^d$ be a random vector such that $\|X\|^2 \leq B$, and let $X_1, \ldots, X_n \in \mathbb{R}^d$ be an i.i.d. sample drawn according to P. We write the approximation error as

$$\bar{G} - G = \sum_{k=1}^{n} S_k$$

where $S_k = \frac{1}{n}\left(X_k X_k^\top - G\right)$. We have $\mathbb{E}[S_k] = 0$ and

$$\|S_k\|_\infty \leq \frac{1}{n}\left(\|X_k\|^2 + \mathbb{E}[\|X\|^2]\right) \leq \frac{2B}{n}.$$

It remains to bound the variance σ^2. We observe that

$$\sigma^2 = \left\| \mathbb{E}\left[\left(\sum_{k=1}^{n} S_k \right)^2 \right] \right\|_\infty = \left\| \sum_{k=1}^{n} \mathbb{E}(S_k^2) \right\|_\infty$$

where the last identity depends on the fact that S_k are independent zero-mean matrices. Since

$$\begin{aligned}
\mathbb{E}(S_k^2) &= \frac{1}{n^2}\mathbb{E}\left[\left(X_k X_k^\top - G\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\Big(\|X_k\|^2 X_k X_k^\top - X_k X_k^\top G - G X_k X_k^\top + G^2\Big) \\
&= \frac{1}{n^2}\left(\mathbb{E}\Big(\|X_k\|^2 X_k X_k^\top\Big) - G^2\right) \preceq \frac{1}{n^2} BG,
\end{aligned}$$

where $H \preceq M$ means that $M - H$ is positive semi-definite, we conclude that

$$\sigma^2 \leq \frac{B}{n}\|G\|_\infty.$$

Applying the matrix Bernstein inequality with $d_1 = d_2 = d$ and $R = \frac{2B}{n}$, we obtain, with probability at least $1 - 2\epsilon$,

$$\|\bar{G} - G\|_\infty \leq \frac{1}{3}\left[\frac{2B}{n}\log(d/\epsilon) + \sqrt{\frac{4B^2}{n^2}\log(d/\epsilon)^2 + 18\frac{B\|G\|_\infty}{n}\log(d/\epsilon)}\right]. \tag{2.1}$$

We observe that the operator norm

$$\|\bar{G} - G\|_\infty = \sup_{\theta \in \mathbb{S}_d} |\theta^\top \bar{G}\theta - \theta^\top G\theta|$$

rewrites, according to the notation of chapter 1, as

$$\|\bar{G} - G\|_\infty = \sup_{\theta \in \mathbb{S}_d} |\bar{N}(\theta) - N(\theta)|,$$

where $\mathbb{S}_d$ denotes the unit sphere of $\mathbb{R}^d$.

In this chapter we present, with the help of the robust estimator $\widehat{N}$ introduced in chapter 1, a uniform bound on the approximation error $|N(\theta) - \bar{N}(\theta)|$ which does not depend explicitly on the dimension d (Proposition 2.7).

2.2 Estimate of the Gram matrix via the empirical Gram matrix

In this section we aim at estimating the quadratic form $N(\theta) = \theta^\top G\theta$ by the empirical estimator

$$\bar{N}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\langle\theta, X_i\rangle^2, \quad \theta \in \mathbb{R}^d,$$

from an i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ drawn according to P. We will use the robust estimator constructed in chapter 1 to provide a uniform bound on the approximation error. We briefly recall its definition.

For $\theta \in \mathbb{R}^d$ and $\lambda > 0$, we consider the empirical criterion

$$r_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\langle\theta, X_i\rangle^2 - \lambda\right),$$

where the influence function ψ is introduced in equation (1.2) on page 16, and we define $\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \le 0\}$. We consider the family of estimators

$$\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2}, \quad \theta \in \mathbb{R}^d.$$

Let $\Lambda \subset (\mathbb{R}_+ \setminus \{0\})^2$ be the finite set defined in equation (1.27) on page 34. We observe that, according to the definition of $\widehat{\alpha}(\theta)$, for any threshold $\sigma \in \mathbb{R}_+$,

$$\frac{1}{n}\sum_{i=1}^{n}\psi\Big[\lambda\Big(\max\{\widetilde{N}_\lambda(\theta), \sigma\}^{-1}\langle\theta, X_i\rangle^2 - 1\Big)\Big] \le r\Big(\lambda^{1/2}\widetilde{N}_\lambda(\theta)^{-1/2}\theta\Big) = r_\lambda\Big(\widehat{\alpha}(\theta)\,\theta\Big) \le 0,$$

where we have used the fact that the function ψ is non-decreasing. Moreover

$$r_\lambda\Big(\widehat{\alpha}(\theta)\,\theta\Big) = 0$$

as soon as $\widehat{\alpha}(\theta) < +\infty$ and this holds true, according to Proposition 1.12 on page 28, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$ and any $(\lambda, \beta) \in \Lambda$ such that $\Phi_+\big(N(\theta)\big) > 0$. Indeed, by Proposition 1.12, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$,

$$\widetilde{N}_\lambda(\theta) \ge \Phi_+\big(N(\theta)\big).$$

Let us put

$$R = \max_{i=1,\dots,n}\|X_i\| \tag{2.2}$$

and

$$\tau_\lambda(t) = \frac{\lambda^2 R^4}{3\max\{t, \sigma\}^2}, \quad t \in \mathbb{R}_+.$$

Proposition 2.4. *Let $\sigma \in \mathbb{R}_+$ be some threshold. With probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, such that $\Phi_+\big(N(\theta)\big) > 0$,*

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} \le \Big[1 - \tau_\lambda\Big(\widetilde{N}_\lambda(\theta)\Big)\Big]_+^{-1},$$

with the convention that $\frac{1}{0} = +\infty$. Moreover, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, such that $\Phi_+\big(N(\theta)\big) > 0$,

$$\frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \ge 1 - \frac{\lambda^2}{3}.$$

Proof. According to Proposition 1.12, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda,\beta) \in \Lambda$, such that $\Phi_+(N(\theta)) > 0$

$$\frac{1}{n}\sum_{i=1}^{n} \psi\Big[\lambda\Big(\widetilde{N}_\lambda(\theta)^{-1}\langle\theta, X_i\rangle^2 - 1\Big)\Big] = r_\lambda\Big(\widehat{\alpha}(\theta)\,\theta\Big) = 0.$$

Defining $g(z) = z - \psi(z)$, we get

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}} - 1 \le \frac{1}{n\lambda}\sum_{i=1}^{n} g\left[\lambda\left(\langle\theta, X_i\rangle^2 \max\{\widetilde{N}_\lambda(\theta),\sigma\}^{-1} - 1\right)\right]. \tag{2.3}$$

In the same way, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda,\beta) \in \Lambda$ such that $\Phi_+(N(\theta)) > 0$, we obtain

$$1 - \frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \le \frac{1}{n\lambda}\sum_{i=1}^{n} g\Big[\lambda\Big(1 - \langle\theta, X_i\rangle^2 \widetilde{N}_\lambda(\theta)^{-1}\Big)\Big]. \tag{2.4}$$

We remark that the derivative of g is

$$g'(z) = 1 - \psi'(z) = \begin{cases} 1 & \text{if } z \notin [-1,1] \\ \dfrac{\frac{z^2}{2}}{1 + z + \frac{z^2}{2}} & \text{if } z \in [-1,0] \\ \dfrac{\frac{z^2}{2}}{1 - z + \frac{z^2}{2}} & \text{if } z \in [0,1], \end{cases}$$

showing that $0 \le g'(z) \le z^2$, and therefore that g is a non-decreasing function satisfying

$$g(z) \le \frac{1}{3} z_+^3, \tag{2.5}$$

where $z_+ = \max\{z, 0\}$. Indeed, it is sufficient to observe that $g(z) \le 0$ if $z \le 0$, while, for $z \ge 0$,

$$g(z) = \int_0^z g'(s)\mathrm{d}s \le \int_0^z s^2 \mathrm{d}s = \frac{1}{3} z^3.$$

Applying equation (2.5) to equation (2.3) we obtain

$$\begin{aligned} \frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}} - 1 &\le \frac{\lambda^2}{3n}\sum_{i=1}^{n}\Big(\langle\theta, X_i\rangle^2 \max\{\widetilde{N}_\lambda(\theta),\sigma\}^{-1} - 1\Big)_+^3 \\ &\le \frac{\lambda^2}{3n\max\{\widetilde{N}_\lambda(\theta),\sigma\}^3}\sum_{i=1}^{n}\langle\theta, X_i\rangle^6, \end{aligned}$$

where we have used the fact $(z^2-1)_+ \le z^2$.
Since, by the Cauchy-Schwarz inequality, $\langle\theta, X_i\rangle^2 \le \|\theta\|^2 R^2$, we get

$$\begin{aligned} \frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}} - 1 &\le \frac{\lambda^2}{3n\max\{\widetilde{N}_\lambda(\theta),\sigma\}^3}\|\theta\|^4 R^4 \sum_{i=1}^{n}\langle\theta, X_i\rangle^2 \\ &= \frac{\lambda^2}{3} \times \frac{\|\theta\|^4 R^4}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}^2} \times \frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}}. \end{aligned}$$

Similarly, since g in non-decreasing, we obtain that, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$ such that $\Phi_+(N(\theta)) > 0$,

$$1 - \frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \leq \frac{1}{n\lambda} \sum_{i=1}^{n} g(\lambda) \leq \frac{\lambda^2}{3},$$

where the last inequality follows from equation (2.5). □

The above result relates the behavior of empirical estimator $\bar{N}$ to that of $\widetilde{N}_\lambda$ and the accuracy of the approximation depends on R, defined in equation (2.2). At the end of the section we mention some assumptions under which it is possible to give a non-random bound on R.

Before stating the bound on the approximation error we introduce the following result.

Proposition 2.5. *With probability at least $1-2\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, any $\sigma > 0$,*

$$\max\{\bar{N}(\theta), \sigma\} \leq \Phi_-^{-1}\Big(\max\{N(\theta), \sigma\}\Big)\Big[1 - \tau_\lambda(N(\theta))\Big]_+^{-1}$$

$$\max\{\bar{N}(\theta), \sigma\} \leq \Phi_-^{-1}\Big(\max\{N(\theta), \sigma\}\Big)\Big[1 - \tau_\lambda\Big(\bar{N}(\theta)[1 - \tau_\lambda(\sigma)]_+\Big)\Big]_+^{-1}$$

$$\bar{N}(\theta) \geq \left(1 - \frac{\lambda^2}{3}\right)_+ \Phi_+(N(\theta))$$

where Φ_+ and Φ_- are defined in Proposition 1.12 on page 28.

Proof. We consider the threshold

$$\sigma' = \Phi_-^{-1}(\max\{N(\theta), \sigma\}) \geq \max\{N(\theta), \sigma\},$$

where we have used the fact that, by definition, $\Phi_-(t)^{-1} \geq t$, for any $t \in \mathbb{R}_+$. We assume that we are in the intersection of the two events of Proposition 1.12, which holds true with probability at least $1-2\epsilon$, so that

$$\sigma' \geq \max\{N(\theta), \sigma, \widetilde{N}_\lambda(\theta)\}. \tag{2.6}$$

By Proposition 2.4, choosing as threshold $\max\{\sigma, \sigma'\}$, we get

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma, \sigma'\}} \leq \Big[1 - \tau_\lambda\big(\max\{\widetilde{N}_\lambda(\theta), \sigma'\}\big)\Big]_+^{-1},$$

(where τ_λ is still defined with respect to σ), so that, according to equation (2.6),

$$\bar{N}(\theta) \leq \sigma' \left[1 - \tau_\lambda(\sigma')\right]_+^{-1}. \tag{2.7}$$

As a consequence, recalling the definition of σ', we have

$$\bar{N}(\theta) \leq \Phi_-^{-1}\Big(\max\{N(\theta), \sigma\}\Big)\Big[1 - \tau_\lambda(N(\theta))\Big]_+^{-1}.$$

Thus, observing that

$$\sigma \leq \Phi_-^{-1}(\sigma) \leq \Phi_-^{-1}\Big(\max\{N(\theta), \sigma\}\Big)\Big[1 - \tau_\lambda(N(\theta))\Big]_+^{-1},$$

we obtain the first inequality of the proposition. To prove the second inequality, we use equation (2.7) once to see that

$$\sigma' \geq \bar{N}(\theta)\big[1-\tau_\lambda(\sigma')\big]_+ \geq \bar{N}(\theta)\big[1-\tau_\lambda(\sigma)\big]_+,$$

and we use it again to get

$$\begin{aligned}\bar{N}(\theta) &\leq \Phi_-^{-1}\big(\max\{N(\theta),\sigma\}\big)\Big[1-\tau_\lambda(\sigma')\Big]_+^{-1}\\ &\leq \Phi_-^{-1}\big(\max\{N(\theta),\sigma\}\big)\Big[1-\tau_\lambda\Big(\bar{N}(\theta)[1-\tau_\lambda(\sigma)]_+\Big)\Big]_+^{-1}.\end{aligned}$$

To complete the proof of the second inequality, it is enough to remark that

$$\sigma \leq \Phi_-^{-1}(\sigma) \leq \Phi_-^{-1}\Big(\max\{N(\theta),\sigma\}\Big)\Big[1-\tau_\lambda\Big(\bar{N}(\theta)\big[1-\tau_\lambda(\sigma)\big]_+\Big)\Big]_+^{-1}.$$

To prove the last inequality, it is sufficient to remark that $\widetilde{N}_\lambda(\theta) \geq \Phi_+(N(\theta))$ by Proposition 1.12 and hence, when $\Phi_+(N(\theta)) > 0$,

$$\bar{N}(\theta) \geq \left(1-\frac{\lambda^2}{3}\right)_+ \Phi_+(N(\theta)).$$

On the other hand, when $\Phi_+(N(\theta)) = 0$, this inequality is also obviously satisfied.

□

According to the previous result we present a bound on the approximation error which uses the notation introduced in chapter 1.

Corollary 2.6. *With probability at least $1-2\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda,\beta) \in \Lambda$, any $\sigma > 0$,*

$$\begin{aligned}&\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} - 1 \leq \widetilde{B}_{\lambda,\beta}(N(\theta)) + \frac{\tau_\lambda(N(\theta))}{\big[1-\tau_\lambda(N(\theta))\big]_+\big[1-B_{\lambda,\beta}(N(\theta))\big]_+},\\ &1 - \frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} \leq B_{\lambda,\beta}(N(\theta)) + \frac{\lambda^2}{3},\end{aligned}$$

where $B_{\lambda,\beta}$ is defined in equation (1.20) on page 31 and $\widetilde{B}_{\lambda,\beta}$ on page 33. In particular

$$\left|\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} - 1\right| \leq \widetilde{B}_{\lambda,\beta}(N(\theta)) + \frac{\tau_\lambda(N(\theta))}{\big[1-\tau_\lambda(N(\theta))\big]_+\big[1-B_{\lambda,\beta}(N(\theta))\big]_+}. \tag{2.8}$$

Proof. We observe that, by Proposition 2.5,

$$\begin{aligned}\max\{\bar{N}(\theta),\sigma\} &\leq \Phi_-^{-1}\big(\max\{N(\theta),\sigma\}\big)\big[1-\tau_\lambda(N(\theta))\big]_+^{-1}\\ &\leq \frac{\max\{N(\theta),\sigma\}}{\big[1-\tau_\lambda(N(\theta))\big]_+\big[1-B_{\lambda,\beta}(N(\theta))\big]_+},\end{aligned}$$

which implies

$$\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} - 1 \leq \frac{B_{\lambda,\beta}(N(\theta))}{\big[1-B_{\lambda,\beta}(N(\theta))\big]_+} + \frac{\tau_\lambda(N(\theta))}{\big[1-\tau_\lambda(N(\theta))\big]_+\big[1-B_{\lambda,\beta}(N(\theta))\big]_+}.$$

Applying Lemma 1.15 on page 33 we obtain the first inequality.

To prove the second inequality we observe that, by Proposition 2.5,

$$\begin{aligned}\max\{\bar{N}(\theta),\sigma\} &\geq \left(1-\frac{\lambda^2}{3}\right)_+ \Phi_+\big(\max\{N(\theta),\sigma\}\big)\\ &= \left(1-\frac{\lambda^2}{3}\right)_+ \max\{N(\theta),\sigma\}\big[1+B_{\lambda,\beta}(N(\theta))\big]^{-1},\end{aligned}$$

where we have used the fact that $\Phi_+\big(\max\{z,\sigma\}\big) = \max\{z,\sigma\}\Big(1+B_{\lambda,\beta}(z)\Big)^{-1}$ as shown in equation (1.23) on page 31. Thus we conclude that

$$1-\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} \leq \frac{B_{\lambda,\beta}(N(\theta))+\lambda^2/3}{1+B_{\lambda,\beta}(N(\theta))} \leq B_{\lambda,\beta}(N(\theta))+\frac{\lambda^2}{3}.$$

□

We observe that the fact of introducing the robust estimator $\widetilde{N}_\lambda$ in the computation of the approximation error has led to provide bounds which do not depend explicitly on the dimension d.

According to the results introduced in chapter 1 we conclude the section by presenting a more explicit (dimension-free) bound on the approximation error.

We recall some notation. Let $a>0$ and let

$$K = 1+\left\lceil a^{-1}\log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil$$

where $\kappa = \sup_{\substack{\theta\in\mathbb{R}^d\\ \mathbb{E}(\langle\theta,X\rangle^2)>0}} \frac{\mathbb{E}\big(\langle\theta,X\rangle^4\big)}{\mathbb{E}\big(\langle\theta,X\rangle^2\big)^2}$ and $c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right)$. We put

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1-4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & \big[6+(\kappa-1)^{-1}\big]\zeta(\max\{t,\sigma\}) \leq \sqrt{n}\\ +\infty & \text{otherwise}\end{cases}$$

where

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\,\mathbb{E}(\|X\|^4)^{1/2}}{4(2+c)\kappa^{1/2}t}+\log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}\mathbb{E}(\|X\|^4)^{1/2}}{t}}\cosh(a/2).$$

The following proposition holds true.

Proposition 2.7. *Consider any threshold $\sigma\in\mathbb{R}_+$ such that $\sigma\leq\mathbb{E}(\|X\|^4)^{1/2}$. Define*

$$\tau_*(t) = \frac{\lambda_*(t)^2\exp(a/2)R^4}{3\max\{t,\sigma\}^2},\quad t\in\mathbb{R}_+,$$

where R is defined in equation (2.2) and λ_* in equation (1.31) on page 36. With probability at least $1-2\epsilon$, for any $\theta\in\mathbb{S}_d$,

$$\left|\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}}-1\right| \leq B_*(N(\theta))+\frac{\tau_*(N(\theta))}{\big[1-\tau_*(N(\theta))\big]_+\big[1-B_*(N(\theta))\big]_+}.$$

Proof. The proof follows directly by applying Corollary 2.6 to $(\lambda_{\widehat{j}}, \beta_{\widehat{j}}) \in \Lambda$ defined in equation (1.27) on page 34. Indeed, by Proposition 1.17 on page 35, for any $t \in \mathbb{R}_+$,

$$B_{\lambda_{\widehat{j}}, \beta_{\widehat{j}}}(t) \leq B_*(t)$$

and, by equation (1.34) on page 36, we have $\lambda_{\widehat{j}} \leq \lambda_*(\theta) \exp(a/4)$. □

Notes: non-random bounds for R

We conclude this section by mentioning assumptions under which it is possible to give a non-random bound for R, defined in equation (2.2).

Let us assume that, for some exponent $p \geq 1$ and some positive constants α and η,

$$\mathbb{E}\Big[\exp\Big(\frac{\alpha}{2}\Big(\frac{\|X\|^{2/p}}{\mathbf{Tr}(G)^{1/p}} - 1 - \eta^{2/p}\Big)\Big)\Big] \leq 1.$$

In this case, with probability at least $1 - \epsilon$,

$$R \leq \mathbf{Tr}(G)^{1/2}\Big(1 + \eta^{2/p} + 2\alpha^{-1}\log(n/\epsilon)\Big)^{p/2}, \tag{2.9}$$

where we recall that $\mathbf{Tr}(G) = \mathbb{E}[\|X\|^2]$.
To give a point of comparison, in the centered Gaussian case where $X \sim \mathcal{N}(0, G)$ is a Gaussian vector, we have, for any $\alpha \in [0, \lambda_1^{-1}\,\mathbf{Tr}(G)[$,

$$\mathbb{E}\left[\exp\left[\frac{\alpha}{2}\left(\frac{\|X\|^2}{\mathbf{Tr}(G)} + \frac{1}{\alpha}\sum_{i=1}^{d}\log\Big(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\Big)\right)\right]\right] = 1,$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of G. Therefore, with probability at least $1 - \epsilon$,

$$R \leq \mathbf{Tr}(G)^{1/2}\left(-\frac{1}{\alpha}\sum_{i=1}^{d}\log\Big(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\Big) + \frac{2\log(n/\epsilon)}{\alpha}\right)^{1/2}.$$

Moreover we observe that

$$\lim_{\alpha \to 0_+} -\frac{1}{\alpha}\sum_{i=1}^{d}\log\Big(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\Big) = 1.$$

In order to replace equation (2.9) with some polynomial assumptions we need to replace R by

$$\widetilde{R} = \left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^6\right)^{1/6}.$$

Indeed, by the Bienaymé Chebyshev inequality, we get that, with probability at least $1-\epsilon$,

$$\widetilde{R} \leq \left(\mathbb{E}[\|X\|^6] + \left(\frac{\mathbb{E}[\|X\|^{12}]}{n\epsilon}\right)^{1/2}\right)^{1/6} \leq \Big(1 + (n\epsilon)^{-1/2}\Big)^{1/6}\,\mathbb{E}\Big[\|X\|^{12}\Big]^{1/12}$$

and hence, with probability at least $1 - n^{-1}$,

$$\widetilde{R} \leq 2^{1/6}\,\mathbb{E}\Big[\|X\|^{12}\Big]^{1/12}.$$

Also in the case where we consider this new quantity $\widetilde{R}$ we obtain a bound of the form of Proposition 2.7. Indeed we observe that another way to take advantage of equation (2.3) is to write

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}} - 1 \leq \frac{\lambda^2\|\theta\|^6}{3\max\{\widetilde{N}_\lambda(\theta),\sigma\}^3}\frac{1}{n}\sum_{i=1}^n\|X_i\|^6.$$

Thus, putting

$$\zeta_\lambda(t) = \frac{\lambda^2\widetilde{R}^6}{3\max\{t,\sigma\}^3}, \quad t \in \mathbb{R}_+,$$

we get that, for any $\theta \in \mathbb{S}_d$,

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta),\sigma\}} \leq 1 + \zeta_\lambda\big(\widetilde{N}_\lambda(\theta)\big).$$

The same reasoning used to prove Proposition 2.5 shows that, with probability at least $1-\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda,\beta) \in \Lambda$, any $\sigma > 0$,

$$\max\{\bar{N}(\theta),\sigma\} \leq \Phi_-^{-1}\big(\max\{N(\theta),\sigma\}\big)\big[1 + \zeta_\lambda\big(N(\theta)\big)\big].$$

As a consequence, with probability at least $1-2\epsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda,\beta) \in \Lambda$,

$$\left|\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} - 1\right| \leq \widetilde{B}_{\lambda,\beta}\big(N(\theta)\big) + \frac{\zeta_\lambda\big(N(\theta)\big)}{\big[1 - B_{\lambda,\beta}\big(N(\theta)\big)\big]_+}.$$

We conclude by stating the analogous of Proposition 2.7.

Proposition 2.8. *Let* $0 < \sigma \leq \mathbb{E}\big(\|X\|^4\big)^{1/2}$ *and let us put*

$$\zeta_*(t) = \frac{\lambda_*(t)^2\ \exp(a/2)\ \widetilde{R}^6}{3\max\{t,\sigma\}^3}, \quad t \in \mathbb{R}_+.$$

With probability at least $1-2\epsilon$*, for any* $\theta \in \mathbb{S}_d$*,*

$$\left|\frac{\max\{\bar{N}(\theta),\sigma\}}{\max\{N(\theta),\sigma\}} - 1\right| \leq B_*\big(N(\theta)\big) + \frac{\zeta_*\big(N(\theta)\big)}{\big[1 - B_*\big(N(\theta)\big)\big]_+}.$$

Chapter 3

Principal Component Analysis

We use the quadratic estimator introduced in chapter 1 to construct robust estimators of the eigenvalues of the Gram matrix. Based on this result we propose a new approach that qualifies the stability of principal component analysis independently of the dimension of the ambient space.

3.1 Introduction

Principal Component Analysis (PCA) is a classical tool for dimensionality reduction. The basic idea of PCA is to reduce the dimensionality of a dataset by projecting it into the space spanned by the directions of maximal variance, that are called its principal components. Since this set of directions lies in the space generated by the eigenvectors associated with the largest eigenvalues of the covariance matrix of the sample, the dimensionality reduction is achieved by projecting the dataset into the space spanned by these eigenvectors, which in the following we call *largest eigenvectors*.

Given $X \in \mathbb{R}^d$ a random vector distributed according to the unknown probability distribution $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$, the goal is to estimate the eigenvalues and eigenvectors of the covariance matrix of X

$$\Sigma = \mathbb{E}\big[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top\big]$$

from an i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ drawn according to P. We have already observed that in the case where the random vector X is centered the covariance matrix Σ is the Gram matrix

$$G = \mathbb{E}(XX^\top).$$

In this chapter we will consider the case of the Gram matrix but similar results can be deduced for the covariance matrix.

As in chapter 1 we introduce the quadratic form

$$N(\theta) = \theta^\top G\theta, \quad \theta \in \mathbb{R}^d,$$

and we observe that, denoting by $p_1, \dots, p_d$ an orthonormal basis of eigenvectors of G, the i-th eigenvalue of the Gram matrix is $\lambda_i = N(p_i)$.

From now we denote by $\lambda_1 \geq \cdots \geq \lambda_d$ the eigenvalues of G and we assume that the eigenvectors $p_1, \dots, p_d$ are ranked according to the decreasing order of their eigenvalues.

Let $\widehat{G} = Q_+$ denote the positive part of the symmetric matrix Q defined in equation (1.37) on page 40. Let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d$ be its eigenvalues and $q_1, \ldots, q_d$ the corresponding orthonormal basis of eigenvectors.

In this chapter we propose some robust versions of PCA. We first show in section 3.2 that each eigenvalue $\widehat{\lambda}_i$ of $\widehat{G}$ is a robust estimator of the corresponding eigenvalue of the Gram matrix. As a consequence, the orthogonal projector on the largest eigenvectors of G can be estimated by the projector on the largest eigenvectors of $\widehat{G}$. The behavior of this estimator is related to the size of the gap in the spectrum of the Gram matrix and in order to have a good approximation we need to have a large eigengap as shown in section 3.3. To avoid the assumption of a large gap in the spectrum of G we propose in section 3.4 a robust version of PCA which consists in performing a smooth cut-off of the spectrum of the Gram matrix via a Lipschitz function. We provide bounds on the approximation error, in terms of the operator norm (Proposition 3.8) and of the Frobenius norm (Proposition 3.9), that replace the size of the eigengap by the inverse of the Lipschitz constant.

3.2 Estimate of the eigenvalues

In this section we prove that each eigenvalue of $\widehat{G}$ is a robust estimator of the corresponding eigenvalue of the Gram matrix.

We first recall some notation. Let $a > 0$ and let

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil$$

where $c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2}\right)$ and

$$\kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}(\langle \theta, X\rangle^2) > 0}} \frac{\mathbb{E}\big(\langle \theta, X\rangle^4\big)}{\mathbb{E}\big(\langle \theta, X\rangle^2\big)^2}.$$

Let $s_4^2 = \mathbb{E}\big(\|X\|^4\big)^{1/2}$ and let $\sigma \in]0, s_4^2]$ be a threshold. We put

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\, n^{-1/2}\zeta(\max\{t,\sigma\})} & [6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \leq \sqrt{n} \\ +\infty & \text{otherwise} \end{cases}$$

where

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\ s_4^2}{4(2+c)\kappa^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}\ s_4^2}{t}}\cosh(a/2). \tag{3.1}$$

The following result holds.

Proposition 3.1. *Let us assume that $8\zeta(\sigma) \leq \sqrt{n}$, $\sigma \leq s_4^2$ and that $\kappa \geq 3/2$. With probability at least $1 - 2\epsilon$, for any $i = 1, \ldots, d$, the two following inequalities hold together*

$$\begin{aligned} &\big|\max\{\lambda_i, \sigma\} - \max\{\widehat{\lambda}_i, \sigma\}\big| \leq 2\max\{\lambda_i, \sigma\} B_*(\lambda_i) + 5\delta\|G\|_F, \\ &\big|\max\{\lambda_i, \sigma\} - \max\{\widehat{\lambda}_i, \sigma\}\big| \leq 2\max\{\widehat{\lambda}_i, \sigma\} B_*\big(\min\{\widehat{\lambda}_i, s_4^2\}\big) + 5\delta\|G\|_F. \end{aligned}$$

Proof. We observe that, for any $i \in \{1, \dots, d\}$, the vector space

$$\mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp \cap \mathbf{span}\{p_1, \dots, p_i\} \subset \mathbb{R}^d$$

is of dimension at least 1, so that the set

$$V_i = \{\theta \in \mathbb{S}_d \mid \theta \in \mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp \cap \mathbf{span}\{p_1, \dots, p_i\}\} \subset \mathbb{R}^d$$

is non-empty. Indeed, putting $A = \mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp$ and $B = \mathbf{span}\{p_1, \dots, p_i\}$, we see that $\dim(A \cap B) = \dim(A) + \dim(B) - \dim(A + B) \geq 1$, since $\dim(A + B) \leq \dim(\mathbb{R}^d) = d$ and $\dim(A) + \dim(B) = d + 1$. Hence, there exists $\theta_i \in V_i$ and for such a θ_i, we have $N(\theta_i) \geq \lambda_i$. It follows that

$$\begin{aligned}\max\{\lambda_i, \sigma\} &\leq \sup\{\max\{N(\theta), \sigma\} \mid \theta \in V_i\} \\ &\leq \sup\Big\{\max\{N(\theta), \sigma\} \mid \theta \in \mathbb{S}_d,\ \theta \in \mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp\Big\}.\end{aligned}$$

Therefore, according to Lemma 1.25 on page 43,

$$\begin{aligned}\max\{\lambda_i, \sigma\}(1 - 2B_*(\lambda_i)) &\leq \sup\Big\{\max\{\theta^\top Q\theta, \sigma\} \mid \theta \in \mathbb{S}_d \cap \mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp\Big\} + 5\delta\|G\|_F \\ &\leq \max\{\widehat{\lambda}_i, \sigma\} + 5\delta\|G\|_F.\end{aligned}$$

In the same way,

$$\begin{aligned}\max\{\widehat{\lambda}_i, \sigma\} &\leq \sup\Big\{\max\{N(\theta), \sigma\}(1 + 2B_*(N(\theta)) \mid \theta \in \mathbb{S}_d \cup \mathbf{span}\{p_1, \dots, p_{i-1}\}^\perp\Big\} \\ &\qquad + 5\delta\|G\|_F \\ &\leq \max\{\lambda_i, \sigma\}(1 + 2B_*(\lambda_i)) + 5\delta\|G\|_F, \\ \max\{\widehat{\lambda}_i, \sigma\}\Big(1 - 2B_*\big(\min\{\widehat{\lambda}_i, s_4^2\}\big)\Big) &\leq \sup\Big\{\max\{N(\theta), \sigma\} \mid \theta \in \mathbb{S}_d \cup \mathbf{span}\{p_1, \dots, p_{i-1}\}^\perp\Big\} \\ &\qquad + 5\delta\|G\|_F \\ &\leq \max\{\lambda_i, \sigma\} + 5\delta\|G\|_F, \\ \max\{\lambda_i, \sigma\} &\leq \sup\Big\{\max\{\theta^\top Q\theta, \sigma\}\Big(1 + 2B_*\big(\min\{\theta^\top Q\theta, s_4^2\}\big)\Big) \\ &\qquad \mid \theta \in \mathbb{S}_d \cup \mathbf{span}\{q_1, \dots, q_{i-1}\}^\perp\Big\} + 5\delta\|G\|_F, \\ &\leq \max\{\widehat{\lambda}_i, \sigma\}\Big(1 + 2B_*\big(\min\{\widehat{\lambda}_i, s_4^2\}\big)\Big) + 5\delta\|G\|_F.\end{aligned}$$

In all these inequalities we have used the fact that

$$\begin{aligned}t &\mapsto \max\{t, \sigma\}\Big(1 - 2B_*(\min\{t, s_4^2\})\Big) \\ t &\mapsto \max\{t, \sigma\}\Big(1 + 2B_*(\min\{t, s_4^2\})\Big)\end{aligned}$$

are non-decreasing (the latter according to Lemma 1.22 on page 41) and that $\lambda_i \leq s_4^2$. This proves the proposition for the eigenvalues of Q, and therefore also for their positive parts, that are the eigenvalues of Q_+. □

As a consequence, using the fact that

$$|\lambda_i - \widehat{\lambda}_i| \leq |\max\{\lambda_i, \sigma\} - \max\{\widehat{\lambda}_i, \sigma\}| + \sigma,$$

we obtain the following result.

Corollary 3.2. *Under the same assumptions as in Proposition 3.1, with probability at least* $1-2\epsilon$, *for any* $i=1,\dots,d$,

$$|\lambda_i - \widehat{\lambda}_i| \leq 2\max\{\lambda_i,\sigma\}B_*(\lambda_i) + 5\delta\|G\|_F + \sigma,$$
$$|\lambda_i - \widehat{\lambda}_i| \leq 2\max\{\widehat{\lambda}_i,\sigma\}B_*(\min\{\widehat{\lambda}_i, s_4^2\}) + 5\delta\|G\|_F + \sigma.$$

We present two more bounds on the estimation error that depend on the largest eigenvalues λ_1 and $\widehat{\lambda}_1$ respectively.

Corollary 3.3. *Under the same assumptions as in Proposition 3.1, with probability at least* $1-2\epsilon$, *for any* $i=1,\dots,d$,

$$|\lambda_i - \widehat{\lambda}_i| \leq 2\max\{\lambda_1,\sigma\}B_*(\lambda_1) + 5\delta\|G\|_F + \sigma,$$
$$|\lambda_i - \widehat{\lambda}_i| \leq 2\max\{\widehat{\lambda}_1,\sigma\}B_*(\min\{\widehat{\lambda}_1, s_4^2\}) + 5\delta\|G\|_F + \sigma.$$

Proof. The proof follows from Lemma 1.22 on page 41. □

In order to simplify notation, we define

$$B(t) = 2\max\{t,\sigma\}B_*\left(\min\{t, s_4^2\}\right) + 7\delta\|G\|_F + \sigma. \tag{3.2}$$

Remark that, since B_* is non-increasing, $t \mapsto \max\{t,\sigma\}B_*(\min\{t,s_4^2\})$ is non-decreasing by Lemma 1.22, and since $B_*(t) \leq 1/4$, $B(t+a) \leq B(t) + a/2$, for any $a \in \mathbb{R}_+$.

3.3 Standard Principal Component Analysis

A method to determine the number of relevant components is based on the difference in magnitude between successive eigenvalues. In this section we study the projection on the r largest eigenvectors $p_1,\dots,p_r$ of the Gram matrix, assuming that there is a gap in the spectrum of the Gram matrix, meaning that for some positive constant δ,

$$\lambda_r - \lambda_{r+1} \geq \delta. \tag{3.3}$$

We denote by Π_r the orthogonal projector on the r largest eigenvectors $p_1,\dots,p_r$ of G and similarly by $\widehat{\Pi}_r$ the orthogonal projector on the r largest eigenvectors $q_1,\dots,q_r$ of its estimate $\widehat{G}$.

Our goal is to provide a bound on the approximation $\|\Pi_r - \widehat{\Pi}_r\|_\infty$ that does not depend explicitly on the dimension d of the ambient space.

Since Π_r and $\widehat{\Pi}_r$ have the same rank, we can write

$$\|\Pi_r - \widehat{\Pi}_r\|_\infty = \sup_{\substack{\theta\in\mathbb{S}_d \\ \theta\in\mathbf{Im}(\widehat{\Pi}_r)}} \|\Pi_r\theta - \widehat{\Pi}_r\theta\|$$

as shown in Lemma B.5 in appendix B. Moreover, for any $\theta \in \mathbf{Im}(\widehat{\Pi}_r) \cap \mathbb{S}_d$, we observe that

$$\begin{aligned}\|\Pi_r\theta - \widehat{\Pi}_r\theta\|^2 &= \|\Pi_r\theta - \theta\|^2 \\ &= \|\sum_{i=1}^r \langle\theta, p_i\rangle p_i - \sum_{i=1}^d \langle\theta,p_i\rangle p_i\|^2 \\ &= \sum_{i=r+1}^d \langle\theta,p_i\rangle^2.\end{aligned}$$

Since $\theta \in \mathbf{Im}(\widehat{\Pi}_r)$, it can be written as $\theta = \sum_{k=1}^r \langle \theta, q_k \rangle q_k$ with $\sum_{k=1}^r \langle \theta, q_k \rangle^2 = 1$, so that

$$\|\Pi_r \theta - \widehat{\Pi}_r \theta\|^2 = \sum_{i=r+1}^{d} \left(\sum_{k=1}^{r} \langle \theta, q_k \rangle \langle q_k, p_i \rangle \right)^2 .$$

Hence, by the Cauchy-Schwarz inequality, we get

$$\|\Pi_r \theta - \widehat{\Pi}_r \theta\|^2 \leq \sum_{i=r+1}^{d} \left(\sum_{k=1}^{r} \langle \theta, q_k \rangle^2 \right) \left(\sum_{k=1}^{r} \langle q_k, p_i \rangle^2 \right) \tag{3.4}$$

$$= \sum_{k=1}^{r} \sum_{i=r+1}^{d} \langle q_k, p_i \rangle^2 . \tag{3.5}$$

Before stating the main result, we introduce a technical lemma.

Lemma 3.4. *With probability at least $1 - 2\epsilon$, for any $k \in \{1, \ldots, d\}$, the two inequalities hold together*

$$\sum_{i=1}^{d} (\lambda_i - \lambda_k)^2 \langle q_k, p_i \rangle^2 \leq 2B (\lambda_1)^2 \tag{3.6}$$

$$\sum_{i=1}^{d} \left(\lambda_i - \widehat{\lambda}_k \right)^2 \langle q_k, p_i \rangle^2 \leq B (\lambda_1)^2 , \tag{3.7}$$

where B is defined in equation (3.2)*.*

Proof. We observe that,

$$\begin{aligned} \|G - \widehat{G}\|_\infty &= \max\Big\{ \sup_{\theta \in \mathbb{S}_d} \theta^\top (G - \widehat{G}) \theta, \ \sup_{\theta \in \mathbb{S}_d} \theta^\top (\widehat{G} - G) \theta \Big\} \\ &= \sup_{\theta \in \mathbb{S}_d} |N(\theta) - \theta^\top \widehat{G} \theta| . \end{aligned}$$

Thus, by Corollary 1.28 on page 44, with probability at least $1 - 2\epsilon$,

$$\sup_{\theta \in \mathbb{S}_d} \|G\theta - \widehat{G}\theta\| \leq B (\lambda_1) .$$

To prove equation (3.7) it is sufficient to observe that, since

$$\|G\theta - \widehat{G}\theta\| = \| \sum_{i,j=1}^{d} (\lambda_i - \widehat{\lambda}_j) \langle \theta, q_j \rangle \langle p_i, q_j \rangle p_i \|$$

choosing $\theta = q_k$, with $k \in \{1, \ldots, d\}$,

$$\|G q_k - \widehat{G} q_k\|^2 = \sum_{i=1}^{d} (\lambda_i - \widehat{\lambda}_k)^2 \langle q_k, p_i \rangle^2 .$$

On the other hand, to prove equation (3.6), we observe that

$$\begin{aligned} \|G\theta - \widehat{G}\theta\| &= \| \sum_{i=1}^{d} \lambda_i \langle \theta, p_i \rangle p_i - \sum_{i=1}^{d} \widehat{\lambda}_i \langle \theta, q_i \rangle q_i \| \\ &= \| \sum_{i=1}^{d} \lambda_i \left(\langle \theta, p_i \rangle p_i - \langle \theta, q_i \rangle q_i \right) - \sum_{i=1}^{d} \left(\widehat{\lambda}_i - \lambda_i \right) \langle \theta, q_i \rangle q_i \| \\ &\geq \| \sum_{i=1}^{d} \lambda_i \left(\langle \theta, p_i \rangle p_i - \langle \theta, q_i \rangle q_i \right) \| - \| \sum_{i=1}^{d} \left(\widehat{\lambda}_i - \lambda_i \right) \langle \theta, q_i \rangle q_i \| \end{aligned}$$

where, by Corollary 3.3,

$$\begin{aligned}\|\sum_{i=1}^{d}\left(\widehat{\lambda}_i-\lambda_i\right)\langle\theta,q_i\rangle q_i\|^2 &= \sum_{i=1}^{d}\left(\widehat{\lambda}_i-\lambda_i\right)^2\langle\theta,q_i\rangle^2\\ &\leq B\left(\lambda_1\right)^2.\end{aligned}$$

Choosing again $\theta = q_k$, for $k \in \{1,\dots,d\}$, we get

$$\begin{aligned}\|\sum_{i=1}^{d}\lambda_i\left(\langle q_k,p_i\rangle p_i-\langle q_k,q_i\rangle q_i\right)\|^2 &= \|\sum_{i,j=1}^{d}\left(\lambda_i-\lambda_j\right)\langle q_k,q_j\rangle\langle q_j,p_i\rangle p_i\|^2\\ &= \|\sum_{i=1}^{d}\left(\lambda_i-\lambda_k\right)\langle q_k,p_i\rangle p_i\|^2\\ &= \sum_{i=1}^{d}\left(\lambda_i-\lambda_k\right)^2\langle q_k,p_i\rangle^2,\end{aligned}$$

which concludes the proof. □

We now apply Lemma 3.4 to our problem. The following proposition holds.

Proposition 3.5. *With probability at least* $1-2\epsilon$,

$$\|\Pi_r-\widehat{\Pi}_r\|_\infty \leq \frac{\sqrt{2r}}{\lambda_r-\lambda_{r+1}}B\left(\lambda_1\right)$$

where B *is defined in equation* (3.2) *and* λ_1 *is the largest eigenvalue of the Gram matrix.*

Proof. We observe that, for any $k \in \{1,\dots,r\}$,

$$\begin{aligned}\sum_{i=r+1}^{d}\left(\lambda_k-\lambda_i\right)^2\langle q_k,p_i\rangle^2 &\geq \sum_{i=r+1}^{d}\left(\lambda_r-\lambda_i\right)^2\langle q_k,p_i\rangle^2\\ &\geq \left(\lambda_r-\lambda_{r+1}\right)^2\sum_{i=r+1}^{d}\langle q_k,p_i\rangle^2.\end{aligned}$$

Then, by Lemma 3.4, with probability at least $1-2\epsilon$,

$$\left(\lambda_r-\lambda_{r+1}\right)^2\sum_{i=r+1}^{d}\langle q_k,p_i\rangle^2 \leq 2B\left(\lambda_1\right)^2.$$

Applying the above inequality to equation (3.5) we conclude the proof. □

We observe that the above proposition provides a bound on the approximation error $\|\Pi_r-\widehat{\Pi}_r\|_\infty$ that does not depend on the dimension d of the ambient space, since B is dimension-free. However the result relates the quality of the approximation of the orthogonal projector Π_r by the robust estimator $\widehat{\Pi}_r$ to the size of the spectral gap. In particular the larger the eigengap, the better the approximation is.

3.4 Robust Principal Component Analysis

In order to avoid the requirement of a large spectral gap, we replace the mere projection on the largest eigenvectors of the Gram matrix by a smooth cut-off of the spectrum of G via a Lipschitz function. In particular we consider a function which is one on the largest eigenvalues and is zero on the smallest ones. In the case there is a (sufficiently large) gap in the spectrum of G, the best Lipschitz constant is exactly the inverse size of the eigengap. In such a way we replace, in some sense, the size of the eigengap with the inverse of the Lipschitz constant.

Let f be a Lipschitz function with Lipschitz constant $1/L$. We decompose the Gram matrix as

$$G = UDU^\top$$

where $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d) \in M_d(\mathbb{R})$ is the diagonal matrix whose entries are the eigenvalues of G and $U \in M_d(\mathbb{R})$ is the orthogonal matrix of eigenvectors of G. We define $f(G)$ as

$$f(G) = Uf(D)U^\top$$

where, for any $i, j \in \{1, \dots, d\}$,

$$|f(\lambda_i) - f(\lambda_j)| \leq \frac{1}{L}|\lambda_i - \lambda_j|.$$

We provide some results on the estimate of $f(G)$, the image of the Gram matrix by the smooth cut-off f, in terms of the operator norm $\|\cdot\|_\infty$ and of the Frobenius norm $\|\cdot\|_F$. We recall that, given $M \in M_d(\mathbb{R})$ a symmetric matrix, the Frobenius norm of M is defined as

$$\|M\|_F^2 = \mathbf{Tr}(M^\top M)$$

and that $\|M\|_\infty \leq \|M\|_F$.

Proposition 3.6. *Let M, $M' \in M_d(\mathbb{R})$ be two symmetric matrices. We denote by $\mu_1, \dots, \mu_d$ the eigenvalues of M related to the orthonormal basis of eigenvectors $p_1, \dots, p_d$ and by $\mu'_1, \dots, \mu'_d$ the eigenvalues of M' related to the orthonormal basis of eigenvectors $q_1, \dots, q_d$. We have*

$$\|M - M'\|_F^2 = \sum_{i,k=1}^{d} (\mu_i - \mu'_k)^2 \langle p_i, q_k \rangle^2.$$

Proof. Since $\{p_i\}_{i=1}^d$ is an orthonormal basis of eigenvectors of M and $\{\mu_i\}_{i=1}^d$ the corresponding eigenvalues, we can write M as

$$M = \sum_{i=1}^{d} \mu_i p_i p_i^\top.$$

Similarly,

$$M' = \sum_{i=1}^{d} \mu'_i q_i q_i^\top.$$

Our goal is to evaluate $\|M - M'\|_F$ where, by definition,

$$\begin{aligned} M - M' &= \sum_{i=1}^{d} \mu_i p_i p_i^\top - \sum_{k=1}^{d} \mu'_k q_k q_k^\top \\ &= \sum_{i,k=1}^{d} (\mu_i - \mu'_k) \langle p_i, q_k \rangle q_k p_i^\top. \end{aligned}$$

We observe that $M - M'$ is a symmetric matrix and its Frobenius norm is

$$\|M - M'\|_F^2 = \mathbf{Tr}((M - M')^\top (M - M')),$$

where

$$(M - M')^\top (M - M') = \sum_{i,j,k=1}^{d} (\mu_i - \mu'_k)(\mu_i - \mu'_j)\langle p_i, q_k\rangle\langle p_i, q_j\rangle q_k q_j^\top.$$

Considering that $\mathbf{Tr}(q_k q_j^\top) = \delta_{jk}$, we conclude that

$$\begin{aligned}\|M - M'\|_F^2 &= \sum_{i,j,k=1}^{d} (\mu_i - \mu'_k)(\mu_i - \mu'_j)\langle p_i, q_k\rangle\langle p_i, q_j\rangle\delta_{jk} \\ &= \sum_{i,k=1}^{d} (\mu_i - \mu'_k)^2\langle p_i, q_k\rangle^2.\end{aligned}$$

□

Corollary 3.7. *Let f be a $1/L$-Lipschitz function. We have*

$$\|f(M) - f(M')\|_F \leq \frac{1}{L}\|M - M'\|_F.$$

Proof. It is sufficient to observe that using twice Proposition 3.6 we obtain

$$\begin{aligned}\|f(M) - f(M')\|_F^2 &= \sum_{i,k=1}^{d} (f(\mu_i) - f(\mu'_k))^2\langle p_i, q_k\rangle^2 \\ &\leq \frac{1}{L^2}\sum_{i,k=1}^{d} (\mu_i - \mu'_k)^2\langle p_i, q_k\rangle^2 = \frac{1}{L^2}\|M - M'\|_F^2.\end{aligned}$$

□

We recall that we are more particularly interested in the Lipschitz function f that is one on the largest eigenvalues and zero on the smallest ones.

Proposition 3.8. (Operator norm) *With probability at least $1 - 2\epsilon$, for any $1/L$-Lipschitz function f,*

$$\|f(G) - f(\widehat{G})\|_\infty \leq \min_{r\in\{1,\dots,d\}} L^{-1}\left(B(\lambda_1) + \sqrt{4rB(\lambda_1)^2 + 2\sum_{i=r+1}^{d}\lambda_i^2}\right),$$

where B is defined in equation (3.2) and $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of G.

Proof. We observe that all the proof holds in the event of probability at least $1 - 2\epsilon$ described in Proposition 3.1. Let $H \in M_d(\mathbb{R})$ be the matrix defined as

$$H = \sum_{k=1}^{d} \lambda_k q_k q_k^\top.$$

We observe that

$$\|f(G) - f(\widehat{G})\|_\infty \leq \|f(G) - f(H)\|_\infty + \|f(H) - f(\widehat{G})\|_\infty$$

and we look separately at the two terms. By definition of operator norm, we have

$$\begin{aligned}\|f(H) - f(\widehat{G})\|_\infty^2 &= \sup_{\theta\in\mathbb{S}_d} \|f(H)\theta - f(\widehat{G})\theta\|^2 \\ &= \sup_{\theta\in\mathbb{S}_d} \|\sum_{k=1}^{d}(f(\lambda_k) - f(\widehat{\lambda}_k))\langle\theta, q_k\rangle q_k\|^2 \\ &= \sup_{\theta\in\mathbb{S}_d} \sum_{k=1}^{d}(f(\lambda_k) - f(\widehat{\lambda}_k))^2\langle\theta, q_k\rangle^2.\end{aligned}$$

Since the function f is $1/L$-Lipschitz, we get

$$\|f(H) - f(\widehat{G})\|_\infty^2 \leq L^{-2} \sup_{\theta\in\mathbb{S}_d} \sum_{k=1}^{d}(\lambda_k - \widehat{\lambda}_k)^2\langle\theta, q_k\rangle^2$$

and then, applying Corollary 3.3, with probability at least $1 - 2\epsilon$, we obtain

$$\|f(H) - f(\widehat{G})\|_\infty^2 \leq L^{-2}B\left(\lambda_1\right)^2.$$

On the other hand, we have

$$\begin{aligned}\|f(G) - f(H)\|_\infty &\leq \|f(G) - f(H)\|_F \\ &\leq \frac{1}{L}\|G - H\|_F,\end{aligned}$$

as shown in Corollary 3.7. Hence, according to Proposition 3.6, we get

$$\|f(G) - f(H)\|_\infty^2 \leq \frac{1}{L^2}\sum_{i,k=1}^{d}(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2$$

where

$$\sum_{i,k=1}^{d}(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2 \leq \left(\sum_{i=1}^{r}\sum_{k=1}^{d} + \sum_{i=1}^{d}\sum_{k=1}^{r} + \sum_{i,k=r+1}^{d}\right)(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2.$$

Since $\lambda_i \geq 0$, for any $i \in \{1, \ldots, d\}$, we get

$$\sum_{i,k=r+1}^{d}(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2 \leq 2\sum_{i=r+1}^{d}\lambda_i^2.$$

Moreover, by Lemma 3.4, we have

$$\sum_{k=1}^{r}\sum_{i=1}^{d}(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2 \leq 2rB\left(\lambda_1\right)^2$$

and since the same bound also holds for

$$\sum_{i=1}^{r}\sum_{k=1}^{d}(\lambda_i - \lambda_k)^2\langle p_i, q_k\rangle^2,$$

we conclude the proof. □

Slightly changing the definition of the estimator we present a bound for the approximation error in terms of the Frobenius norm. Instead of considering $\widehat{G} = \sum_{i=1}^{d} \widehat{\lambda}_i q_i q_i^\top$ we consider the matrix

$$\widetilde{G} = \sum_{i=1}^{d} \widetilde{\lambda}_i q_i q_i^\top$$

with eigenvectors $q_1, \ldots, q_d$ and eigenvalues

$$\widetilde{\lambda}_i = \left[\widehat{\lambda}_i - B(\widehat{\lambda}_i)\right]_+$$

where we recall that B is defined in equation (3.2). We observe that, in the event of probability at least $1 - 2\epsilon$ described in Corollary 3.3, for any $i = 1, \ldots, d$,

$$\widetilde{\lambda}_i \leq \lambda_i.$$

We first present a result on the approximation error $\|G - \widetilde{G}\|_F$.

Proposition 3.9. (Frobenius norm) *With probability at least* $1 - 2\epsilon$,

$$\|G - \widetilde{G}\|_F \leq \min_{r \in \{1 \ldots, d\}} \sqrt{13 r B(\lambda_1)^2 + 2 \sum_{i=r+1}^{d} \lambda_i^2},$$

where B is defined in equation (3.2) *and $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of G.*

Proof. During the whole proof, we will assume that the event of probability at least $1-2\epsilon$ described in Proposition 3.1 is realized. According to Proposition 3.6, we have

$$\|G - \widetilde{G}\|_F^2 = \sum_{i,k=1}^{d} (\lambda_i - \widetilde{\lambda}_k)^2 \langle p_i, q_k \rangle^2,$$

where

$$\sum_{i,k=1}^{d} (\lambda_i - \widetilde{\lambda}_k)^2 \langle p_i, q_k \rangle^2 \leq \left(\sum_{i=1}^{r} \sum_{k=1}^{d} + \sum_{k=1}^{r} \sum_{i=1}^{d} + \sum_{i,k=r+1}^{d} \right) (\lambda_i - \widetilde{\lambda}_k)^2 \langle p_i, q_k \rangle^2.$$

Since, by definition, $\widetilde{\lambda}_i \leq \lambda_i$, it follows that

$$\begin{aligned} \sum_{i,k=r+1}^{d} (\lambda_i - \widetilde{\lambda}_k)^2 \langle p_i, q_k \rangle^2 &\leq \sum_{i,k=r+1}^{d} (\lambda_i^2 + \widetilde{\lambda}_k^2) \langle p_i, q_k \rangle^2 \\ &\leq 2 \sum_{i=r+1}^{d} \lambda_i^2. \end{aligned}$$

Furthermore, we observe that

$$\sum_{k=1}^{r} \sum_{i=1}^{d} (\lambda_i - \widetilde{\lambda}_k)^2 \langle p_i, q_k \rangle^2 \leq 2 \sum_{k=1}^{r} \sum_{i=1}^{d} (\lambda_i - \widehat{\lambda}_k)^2 \langle p_i, q_k \rangle^2 + 2 \sum_{k=1}^{r} \sum_{i=1}^{d} B(\widehat{\lambda}_k)^2 \langle p_i, q_k \rangle^2,$$

where, by Lemma 3.4,

$$\sum_{k=1}^{r} \sum_{i=1}^{d} (\lambda_i - \widehat{\lambda}_k)^2 \langle p_i, q_k \rangle^2 \leq r B(\lambda_1)^2.$$

and $B(\widehat{\lambda}_k) \leq B(\widehat{\lambda}_1)$. We have then proved that

$$\sum_{k=1}^{r}\sum_{i=1}^{d}(\lambda_i - \widetilde{\lambda}_k)^2\langle p_i, q_k\rangle^2 \leq 2rB(\lambda_1)^2 + 2rB(\widehat{\lambda}_1)^2.$$

Applying Corollary 3.3 and using the fact that $B(t+a) \leq B(t) + a/2$, as explained after equation (3.2) on page 80, we deduce that

$$B(\widehat{\lambda}_1) \leq B\big[\lambda_1 + B(\lambda_1)\big] \leq 3B(\lambda_1)/2.$$

This proves that

$$\sum_{k=1}^{r}\sum_{i=1}^{d}(\lambda_i - \widetilde{\lambda}_k)^2\langle p_i, q_k\rangle^2 \leq 2r\left[B(\lambda_1)^2 + 9B(\lambda_1)^2/4\right] = 13rB(\lambda_1)^2/2.$$

Considering that the same bound holds for

$$\sum_{i=1}^{r}\sum_{k=1}^{d}(\lambda_i - \widetilde{\lambda}_k)^2\langle p_i, q_k\rangle^2,$$

we conclude the proof.

□

To obtain a bound on $\|f(G) - f(\widetilde{G})\|_F$ it is sufficient to combine the above proposition with Corollary 3.7.

Corollary 3.10. *With the same notation as in Proposition 3.9, with probability at least* $1 - 2\epsilon$*, for any* $1/L$*-Lipschitz function* f*,*

$$\|f(G) - f(\widetilde{G})\|_F \leq \min_{r\in\{1,\ldots,d\}} L^{-1}\sqrt{13\,rB(\lambda_1)^2 + 2\sum_{i=r+1}^{d}\lambda_i^2}.$$

Notes: How to choose r in the previous bounds

In the previous bounds, the optimal choice of the dimension parameter r depends on the distribution of the eigenvalues of the Gram matrix G. Nevertheless, it is possible to upper bound what happens when this distribution of eigenvalues is the worst possible.

Observe that

$$\sum_{i=r+1}^{d}\lambda_i^2 \leq \lambda_{r+1}\,\mathbf{Tr}(G)$$

and also $r\lambda_{r+1} \leq \mathbf{Tr}(G)$, so that

$$\sum_{i=r+1}^{d}\lambda_i^2 \leq r^{-1}\,\mathbf{Tr}(G)^2.$$

Hence, if we consider for example the case where the approximation error is evaluated in terms of the Frobenius norm, the worst case formulation of Corollary 3.10 is obtained choosing

$$r = \left\lceil\sqrt{2/13}\,\mathbf{Tr}(G)B(\lambda_1)^{-1}\right\rceil$$

and, in this case, it can be restate as follows.

Corollary 3.11. *With probability at least* $1 - 2\epsilon$,

$$\|f(G) - f(\widetilde{G})\|_F \leq L^{-1}\sqrt{11\,\mathbf{Tr}(G)B(\lambda_1) + 13B(\lambda_1)^2}.$$

This proposition shows that the worst case speed is not slower than $n^{-1/4}$. We do not know whether this rate is optimal in the worst case. We could in the same way obtain a worst case corollary for Proposition 3.8.

Chapter 4

Spectral Clustering

Based on the results proved in chapter 1, we propose a new algorithm for spectral clustering. This new algorithm can be viewed as a change of representation in a reproducing kernel Hilbert space, followed by a (greedy) classification.

4.1 Introduction

Clustering is the task of grouping a set of objects into classes, called *clusters*, in such a way that objects in the same group are more similar to each other than to those in other groups. Spectral clustering algorithms use the spectrum of some data-dependent matrices to perform clustering. These matrices can be either the affinity (or similarity) matrix [10] or the Laplacian matrix [11].

Given $X_1, \ldots, X_n$, a set of points to cluster, the affinity matrix $A \in M_n(\mathbb{R})$ measures the similarity, in terms of distance, between each pair of points and a common definition is

$$A_{i,j} = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \tag{4.1}$$

where σ is a free scale parameter. The Laplacian matrix is obtained by rescaling the affinity matrix by its row sums. Among all the definitions of the Laplacian matrix, we choose the normalized Laplacian matrix

$$L = I - D^{-1/2} A D^{-1/2},$$

where D is the diagonal matrix whose i-th entry is $D_{ii} = \sum_j A_{ij}$. The Laplacian matrix L is positive semi-definite, its smallest eigenvalue is zero and the corresponding eigenvector is the constant vector $\mathbb{1} = (1, \ldots, 1)^\top$.

We observe that the Laplacian matrix has the same eigenvectors as $D^{-1/2} A D^{-1/2}$ (and related eigenvalues) and in the following we consider for simplicity

$$L = D^{-1/2} A D^{-1/2}$$

and we refer to it as the Laplacian matrix.

Many spectral clustering algorithms have been proposed, e.g. [32], [20], [22]. We briefly describe one of the most successful ones, introduced by Ng, Jordan and Weiss [22].

The goal is to cluster a set of points into c classes according to their similarity. Assuming that the number c of clusters is known, the authors use simultaneously the c largest eigenvectors of the Laplacian matrix to group the data points into classes. More precisely, denoting by $X_1, \ldots, X_n$ the dataset, the algorithm goes as follows

1. Form the affinity matrix A with entries $A_{i,j} = \exp\left(-\|X_i - X_j\|^2/2\sigma^2\right)$ if $i \neq j$ and $A_{i,i} = 0$.
2. Construct the Laplacian matrix L.
3. Compute the c largest eigenvectors of L and form the $n \times c$ matrix X whose columns are these largest eigenvectors.
4. Renormalize each row of X to have unit length and treat each row as a point in $\mathbb{R}^c$.
5. Cluster points according to the new representation.

Spectral clustering can also be related to Markov chains as shown in [20] and in particular the affinity matrix A can be converted into a transition probability matrix

$$R = D^{-1}A.$$

The Laplacian matrix $L = D^{1/2}RD^{-1/2}$ has the same eigenvalues as R and related eigenvectors.

Our approach relies on viewing those matrices as empirical versions of underlying integral operators.

Let $\mathcal{X} \subset \mathbb{R}^d$ be a set endowed with the probability distribution $\mathrm{P} \in \mathcal{M}_+^1(\mathcal{X})$ and let L^2_{P} be the space of square integrable functions on $\mathcal{X}$ with respect to P. The operator related to the matrix L can be interpreted as the discrete version of the integral operator $L_K : L^2_{\mathrm{P}} \to L^2_{\mathrm{P}}$ defined by

$$L_K f(x) = \int f(y) \frac{K(x,y)}{\left(\int K(x,z)\mathrm{dP}(z)\right)^{\frac{1}{2}} \left(\int K(y,z)\mathrm{dP}(z)\right)^{\frac{1}{2}}} \,\mathrm{dP}(y), \quad f \in L^2_{\mathrm{P}},$$

where $K \in L^2_{\mathrm{P}} \times L^2_{\mathrm{P}}$ is a symmetric positive semi-definite kernel on $\mathcal{X}$ and in particular

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

in the case where the affinity matrix A has the form described in equation (4.1).

Connections between empirical operators and their continuous counterpart have been studied in many works, e.g. [24], [36].

Starting from the kernel

$$\bar{K}(x,y) = \frac{K(x,y)}{\left(\int K(x,z)\mathrm{dP}(z)\right)^{\frac{1}{2}} \left(\int K(y,z)\mathrm{dP}(z)\right)^{\frac{1}{2}}}$$

we introduce, in section 4.2, an ideal spectral clustering algorithm that would require the exact knowledge of the law P. It consists in replacing the projection on the c largest eigenvectors of L done in step 3 of the algorithm introduced in [22] by computing some suitable power of the operator L_K. Indeed, $\bar{K}$ is related to a Markov chain, and iterating $\bar{K}$ (or rather the operator L_K) performs some kind of soft truncation of its eigenvalues and is related to the computation of the marginal distribution at time t of the Markov chain. This leads to a natural dimensionality reduction and allows us to propose an algorithm that automatically estimates the number c of clusters when it is not known in advance.

In order to obtain a practical algorithm, all we have to do is to compute some estimate

of the operator L_K, using the estimators and the generalization bounds introduced in the previous chapters, as shown in section 4.3.

We conclude providing in section 4.4 some experiments in the setting of image analysis. We will show that a very simple greedy classification algorithm can be used to perform the final classification, once the change of representation has been operated.

For more details on reproducing kernel Hilbert spaces we refer to appendix C.

4.2 Description of an ideal algorithm

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, or more generally of some separable Hilbert space of possibly infinite dimension, endowed with the (unknown) probability distribution $\mathrm{P} \in \mathcal{M}_+^1(\mathcal{X})$ and let $\widetilde{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive semi-definite kernel. We define

$$K(x, y) = \widetilde{K}(x, y)^2, \quad x, y \in \mathcal{X},$$

and we observe that K is itself a symmetric positive semi-definite kernel.

Our approach relies on interpreting clustering as a change of representation in a reproducing kernel Hilbert space, described by a change of kernel.

We start with an ideal version of this algorithm which uses the unknown distribution P. For any $x, y \in \mathcal{X}$,

1. Form the kernel (compute the Laplace operator)

$$\bar{K}(x, y) = \frac{K(x, y)}{(\int K(x, z)\ \mathrm{dP}(z))^{\frac{1}{2}}\ (\int K(y, z)\ \mathrm{dP}(z))^{\frac{1}{2}}}$$

2. Construct the new kernel (iterate the Markov chain)

$$\bar{K}_m(x, y) = \int \bar{K}(y, z_1)\bar{K}(z_1, z_2) \dots \bar{K}(z_{m-1}, x)\ \mathrm{dP}^{\otimes(m-1)}(z_1, \dots, z_{m-1}),$$

   where $m > 0$ is a free parameter

3. Make a last change of kernel (normalize the kernel, or equivalently project the representation on the unit sphere)

$$K_m(x, y) = \frac{\bar{K}_{2m}(x, y)}{\bar{K}_{2m}(x, x)^{\frac{1}{2}} \bar{K}_{2m}(y, y)^{\frac{1}{2}}}$$

4. Cluster points according to the new representation defined by the kernel K_m.

We will see that this new representation sends clusters to the neighborhood of an orthonormal basis, and therefore to a neighborhood of the vertices of a simplex, making subsequent classification an easy task.

4.2.1 Analysis of the algorithm

We first observe that the kernel $\bar{K}$ is symmetric and positive semi-definite. According to the Moore-Aronszajn theorem (Proposition C.14), it defines a reproducing kernel Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$, such that

$$\bar{K}(x, y) = \langle \phi(y), \phi(x) \rangle_{\mathcal{H}}. \tag{4.2}$$

We introduce the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{G}v = \int \langle v, \phi(z)\rangle_{\mathcal{H}} \ \phi(z) \ \mathrm{dP}(z) \tag{4.3}$$

and we observe that the kernel $\bar{K}_m$ can be written in terms of some power of $\mathcal{G}$.

Proposition 4.1. *We have*

$$\bar{K}_m(x, y) = \langle \mathcal{G}^{\frac{m-1}{2}}\phi(x), \mathcal{G}^{\frac{m-1}{2}}\phi(y)\rangle_{\mathcal{H}}.$$

Proof. Since $\mathcal{G}$ is a positive operator, it is sufficient to prove, by induction, that

$$\bar{K}_m(x, y) = \langle \mathcal{G}^{m-1}\phi(x), \phi(y)\rangle_{\mathcal{H}}. \tag{4.4}$$

Case $m = 1$. We observe that by definition

$$\bar{K}_1(x, y) = \bar{K}(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}.$$

We now assume that equation (4.4) holds for any $\ell < m$ and we prove the identity for $\bar{K}_m$. We observe that, by definition,

$$\begin{aligned}\bar{K}_m(x, y) &= \int \bar{K}(y, z_1) \left(\int \bar{K}(z_1, z_2) \dots \bar{K}(x, z_{m-1}) \ \mathrm{dP}^{\otimes(m-2)}(z_2, \dots, z_{m-1}) \right) \ \mathrm{dP}(z_1) \\ &= \int \bar{K}(y, z_1) \ \bar{K}_{m-1}(x, z_1) \ \mathrm{dP}(z_1).\end{aligned}$$

Applying the induction hypothesis we obtain that

$$\begin{aligned}\bar{K}_m(x, y) &= \int \langle \phi(y), \phi(z_1)\rangle_{\mathcal{H}} \langle \mathcal{G}^{m-2}\phi(x), \phi(z_1)\rangle_{\mathcal{H}} \ \mathrm{dP}(z_1) \\ &= \langle \mathcal{G}^{m-2}\phi(x), \int \langle \phi(y), \phi(z_1)\rangle_{\mathcal{H}} \phi(z_1) \ \mathrm{dP}(z_1)\rangle_{\mathcal{H}} \\ &= \langle \mathcal{G}^{m-2}\phi(x), \mathcal{G}\phi(y)\rangle_{\mathcal{H}},\end{aligned}$$

which concludes the proof, since $\mathcal{G}$ is self-adjoint. □

As a consequence, the kernel K_m rewrites as

$$K_m(x, y) = \Big\langle \frac{\mathcal{G}^{m-1/2}\phi(x)}{\|\mathcal{G}^{m-1/2}\phi(x)\|_{\mathcal{H}}}, \frac{\mathcal{G}^{m-1/2}\phi(y)}{\|\mathcal{G}^{m-1/2}\phi(y)\|_{\mathcal{H}}} \Big\rangle_{\mathcal{H}}.$$

Accordingly, the representation of $x \in \mathcal{X}$ defined by the kernel K_m is isometric to the representation $\|\mathcal{R}(x)\|_{\mathcal{H}}^{-1}\mathcal{R}(x) \in \mathcal{H}$ where

$$\mathcal{R}(x) = \mathcal{G}^{m-1/2}\phi(x) \in \mathbf{Im}(\mathcal{G}) \subset \mathcal{H}. \tag{4.5}$$

In the sequel of this section we describe other isometric representations of x that can be seen as different realizations of the kernel space defined by K_m.

The representation of x in the kernel space defined by K_m is isometric to the representation $\|\mathbf{R}(x)\|^{-1}\mathbf{R}(x) \in L^2_{\mathrm{P}}$, where $\mathbf{R}(x) \in L^2_{\mathrm{P}}$ is defined as

$$\mathbf{R}(x)(y) = \bar{K}_m(x, y) = \langle \mathcal{G}^{m-1}\phi(x), \phi(y)\rangle_{\mathcal{H}}. \tag{4.6}$$

This is a consequence of the fact that

$$\langle \mathbf{R}(x), \mathbf{R}(z) \rangle_{L^2_{\mathrm{P}}} = \int \bar{K}_m(x,y)\bar{K}_m(y,z)\,\mathrm{dP}(y) = \bar{K}_{2m}(x,z).$$

We now describe a third isometric representation, related to a Markov chain, that is helpful to understand the clustering effect observed in favorable situations.

Let us assume that the support of P is made of several compact connected components and let

$$K(x,y) = \exp(-\beta\|x-y\|^2),$$

where $\beta > 0$ is large enough. The Markov chain related to the operator

$$M(f) : x \mapsto \frac{\displaystyle\int K(x,y)f(y)\,\mathrm{dP}(y)}{\displaystyle\int K(x,z)\,\mathrm{dP}(z)}$$

has a small probability to jump from one component to another one. Thus, for reasons that we will not try to prove here, for suitable values of m, the measure $M^m_x : f \mapsto M^m(f)(x)$ is close to be supported by the connected component which x belongs to and more precisely it is close to the restriction of the invariant measure of the operator M restricted to this component. As a result, the function $x \mapsto M^m(f)(x)$ is approximately constant on each connected component.

More precisely, let us put

$$\mu(x) = \int K(x,z)\,\mathrm{dP}(z). \tag{4.7}$$

Lemma 4.2. *We have*

$$M^m(f)(x) = \mu(x)^{-1/2} \int \bar{K}_m(x,y)\mu(y)^{1/2} f(y)\,\mathrm{dP}(y).$$

Proof. We prove the result by induction.
Case $m = 1$. Recalling the definition of the kernel $\bar{K}$ we get

$$\begin{aligned} Mf(x) &= \int \mu(x)^{-1/2}\bar{K}(x,y)\mu(y)^{1/2}f(y)\ \mathrm{dP}(y) \\ &= \int \mu(x)^{-1/2}\bar{K}_1(x,y)\mu(y)^{1/2}f(y)\ \mathrm{dP}(y). \end{aligned}$$

We now assume that the identity holds for any $\ell < m$. It follows that

$$\begin{aligned} M^m f(x) &= \mu(x)^{-1} \int K(x,z)\mu(z)^{-1/2}\bar{K}_{m-1}(z,y)\mu(y)^{1/2}f(y)\,\mathrm{dP}(y)\mathrm{dP}(z) \\ &= \mu(x)^{-1/2} \int \bar{K}(x,z)\bar{K}_{m-1}(z,y)\mu(y)^{1/2}f(y)\,\mathrm{dP}(y)\mathrm{dP}(z) \\ &= \mu(x)^{-1/2} \int \bar{K}_m(x,y)\mu(y)^{1/2}f(y)\,\mathrm{dP}(y) \end{aligned}$$

which concludes the proof. □

As a consequence, the probability measure $M^m_x : f \mapsto M^m(f)(x)$ has density

$$\frac{\mathrm{d}M^m_x}{\mathrm{dP}}(y) = \mu(x)^{-1/2}\bar{K}_m(x,y)\mu(y)^{1/2}.$$

Moreover, if we denote by $M(x, y)$ the kernel corresponding to the operator M, we observe that

$$\int \mu(x)M(x,y)f(y)\,\mathrm{dP}(x)\mathrm{dP}(y) = \int K(x,y)f(y)\,\mathrm{dP}(x)\mathrm{dP}(y) = \int \mu(y)f(y)\,\mathrm{dP}(y).$$

This shows that the invariant measure Q of the operator M has a density with respect to P equal to

$$\frac{\mathrm{dQ}}{\mathrm{dP}} = \mu.$$

Lemma 4.3. *We have*

$$\Big\langle \frac{\mathrm{d}M_x^m}{\mathrm{dQ}}, \frac{\mathrm{d}M_z^m}{\mathrm{dQ}} \Big\rangle_{L^2_{\mathrm{Q}}} = \mu(x)^{-1/2}\bar{K}_{2m}(x,z)\mu(z)^{-1/2}.$$

Proof. The proof follows from the identity

$$\frac{\mathrm{d}M_x^m}{\mathrm{dQ}}(y) = \mu(x)^{-1/2}\bar{K}_m(x,y)\mu(y)^{-1/2}.$$

□

Consider the representation of $x \in \mathcal{X}$ given by $\|\mathfrak{R}(x)\|^{-1}\mathfrak{R}(x) \in L^2_Q$, where

$$\mathfrak{R}(x) = \mu(x)^{1/2}\frac{\mathrm{d}M_x^m}{\mathrm{dQ}} \in L^2_{\mathrm{Q}}.$$

Lemma 4.3 shows that the representation of $x \in \mathcal{X}$ in the kernel space defined by K_m is also isometric to $\mathfrak{R}(x)$.

We conclude the section observing that

$$K_m(x,z) = \bar{K}_{2m}(x,x)^{-1/2}\bar{K}_{2m}(x,z)\bar{K}_{2m}(z,z)^{-1/2}$$

is the cosine of the angle formed by the two vectors representing x and z. Moreover, as explained above, when β and m are chosen in a suitable way, the supports of the probability measures M_x^m and M_z^m are almost disjoint if x and z belong to two different clusters, whereas M_x^m and M_z^m are almost the same when x and z belong to the same cluster. Therefore, in this setting, the cosine K_m is either close to 0 or close to 1.

This means that in the Hilbert space defined by the normalized kernel K_m, the c clusters are concentrated around c orthogonal unit vectors, forming the vertices of a regular simplex. Equation (4.5) shows that this simplex is necessarily (in first approximation) contained in the linear span of the c largest eigenvectors of the Gram operator.

4.2.2 Choice of the scale parameter

In the case when the influence kernel K is of the form

$$K_\beta(x,y) = \exp(-\beta\|x-y\|^2),$$

we propose to choose β as the solution of the equation

$$F(\beta) := \int K_\beta(x,y)^2\,\mathrm{dP}(x)\mathrm{dP}(y) = h,$$

where h is a suitable parameter which measures the probability that two independent points drawn according to the probability P are close to each other.
Introducing the Gram operator $\mathbf{G}_\beta : L^2_{\mathrm{P}} \to L^2_{\mathrm{P}}$ defined by K_β such that

$$\mathbf{G}_\beta(f)(x) = \int K_\beta(x,y) f(y)\, \mathrm{dP}(y), \quad x \in \mathcal{X},$$

the parameter h is equal to the square of the Hilbert Schmidt norm of $\mathbf{G}_\beta$. As reminded in appendix C (Proposition C.2), $\mathbf{G}_\beta$ has a discrete spectrum $\lambda_1 \geq \lambda_2 \geq \cdots$. Since $K_\beta(x,x) = 1$, it satisfies

$$\sum_{i=1}^{+\infty} \lambda_i(\beta) = \int K_\beta(x,x)\, \mathrm{dP}(x) = 1,$$
$$\sum_{i=1}^{+\infty} \lambda_i^2(\beta) = F(\beta) \leq 1.$$

Therefore, $F(\beta)$ governs the spread of the eigenvalues of $\mathbf{G}_\beta$. We observe that

$$\lim_{\beta \to 0} F(\beta) = 1$$

which implies that $\lambda_1(\beta) \underset{\beta \to 0}{\longrightarrow} 1$ whereas $\lambda_i(\beta) \underset{\beta \to 0}{\longrightarrow} 0$ for $i \geq 2$. On the other hand we have

$$\lim_{\beta \to +\infty} F(\beta) = 0,$$

so that when β grows, the eigenvalues are spread more and more widely. For these reasons, the value of the parameter $h = F(\beta)$ controls the effective dimension of the representation of the distribution of points P in the reproducing kernel Hilbert space defined by K_β. Experiments show that we want this effective dimension to be pretty large, meaning that we will impose a small value of the parameter h.

4.3 Estimation of the ideal algorithm by an empirical one

By empirical algorithm we mean an algorithm based on the empirical distribution

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where $X_1, \ldots, X_n$ is an i.i.d. sample drawn from the probability measure P. The algorithm described in section 4.2 is *ideal* since it depends on the unknown probability distribution P. In this section we use the results obtained in the previous chapters to present an estimated version of this ideal algorithm and to prove its convergence by non-asymptotic bounds.

Let $\mathcal{X}$ be a compact subset of a separable Hilbert space. In this section we consider the Gaussian kernel

$$K(x,y) = \exp\left(-\beta \|x-y\|^2\right), \quad x, y \in \mathcal{X},$$

where $\beta > 0$ is a free parameter. However the construction we are going to present holds for any kernel $K = \widetilde{K}^2$ where $\widetilde{K}$ is a symmetric positive semi-definite kernel, normalized in such a way that $\widetilde{K}(x,x) = 1$, for any x in the state space $\mathcal{X}$.

In this setting the kernel $\widetilde{K}$ is defined by

$$\widetilde{K}(x,y) = \exp\left(-\frac{\beta}{2}\|x-y\|^2\right), \quad x,y \in \mathcal{X}.$$

To fix β we observe that, denoting by A the matrix of entries

$$A_{i,j} = \exp\left(-\beta\|X_i - X_j\|^2\right), \quad i,j = 1,\dots,n,$$

we have, with the notation of section 4.2.2,

$$F(\beta) = \frac{1}{n^2}\,\mathbf{Tr}(A^2) = \frac{1}{n} + f(\beta)$$

where $f(\beta) = \frac{1}{n^2}\sum_{\substack{i,j \\ i\neq j}} \exp\left(-2\beta\|X_i - X_j\|^2\right)$. We choose β as the solution of $f(\beta) = 0.005$.

Let $\widetilde{\mathcal{H}}$ be the reproducing kernel Hilbert space defined by $\widetilde{K}$ and let $\widetilde{\phi} : \mathcal{X} \to \widetilde{\mathcal{H}}$ be the corresponding feature map, so that

$$K(x,z) = \langle \widetilde{\phi}(x), \widetilde{\phi}(z)\rangle^2_{\widetilde{\mathcal{H}}}.$$

Lemma 4.4. *The density $\mu(x)$ of the invariant measure, defined in equation (4.7), can be written as*

$$\mu(x) = \int \langle \widetilde{\phi}(x), v\rangle^2_{\widetilde{\mathcal{H}}}\, \mathrm{d}\widetilde{\mathrm{P}}(v) =: N_{\widetilde{\mathcal{H}}}(x), \quad x \in \mathcal{X},$$

where $\widetilde{\mathrm{P}} = \mathrm{P} \circ \widetilde{\phi}^{-1} \in \mathcal{M}^1_+(\widetilde{\mathcal{H}})$.

Proof. It is sufficient to observe that

$$\begin{aligned}
\mu(x) &= \int K(x,z)\, \mathrm{dP}(z) \\
&= \int \langle \widetilde{\phi}(x), \widetilde{\phi}(z)\rangle^2_{\widetilde{\mathcal{H}}}\, \mathrm{dP}(z) \\
&= \int \langle \widetilde{\phi}(x), v\rangle^2_{\widetilde{\mathcal{H}}}\, \mathrm{d}\widetilde{\mathrm{P}}(v) = N_{\widetilde{\mathcal{H}}}(x).
\end{aligned}$$

□

Therefore, according to its definition on page 91, the kernel $\bar{K}$ can be expressed as

$$\bar{K}(x,y) = N_{\widetilde{\mathcal{H}}}(x)^{-1/2} K(x,y) N_{\widetilde{\mathcal{H}}}(y)^{-1/2}.$$

Rather than an estimate of $\bar{K}$ itself, we provide an estimate of its thresholded approximation

$$\bar{K}_\sigma(x,y) = \max\{N_{\widetilde{\mathcal{H}}}(x), \sigma\}^{-1/2} K(x,y) \max\{N_{\widetilde{\mathcal{H}}}(y), \sigma\}^{-1/2}, \quad \sigma > 0.$$

When the threshold σ is suitably small this approximation retains the clustering properties of $\bar{K}$. To support this claim, at least on a heuristic basis, we remark that $\bar{K}_\sigma$ is related to the sub-Markov operator

$$M_\sigma(f) : x \mapsto \frac{\displaystyle\int K(x,y) f(y)\, \mathrm{dP}(y)}{\max\left\{\displaystyle\int K(x,z)\, \mathrm{dP}(z), \sigma\right\}}$$

and when σ is small, M_σ modifies M only in the regions where the density of P is the smallest. The fact that M_σ kills the diffusion of M in regions of low probability density may be expected to preserve an interesting clustering effect.

It should also be remarked that since $x \mapsto K(x,y)$ is continuous, also $x \mapsto \int K(x,y)\,\mathrm{dP}(z)$ is continuous, by the Lebesgue dominated convergence theorem, and hence, as we assume that $\mathcal{X}$ is compact,

$$\sigma_* = \inf_{x \in \mathrm{supp(P)}} \int K(x,z)\,\mathrm{dP}(z) > 0.$$

This means that $M_\sigma(f) = M(f)$ in L^2_{P} when $0 < \sigma \le \sigma_*$. As a consequence the Markov chains defined by the two operators (M and M_σ) have the same distribution for any starting point in the support of P and have the same eigenvalues and eigenvectors (since L^2_{P} identifies functions that coincide on the support of P). In other words, $\bar{K}_\sigma(x,y) = \bar{K}(x,y)$ for any $x, y \in \mathrm{supp(P)}$.

From now on we assume that $0 < \sigma \le \sigma_*$.

As a consequence of Lemma 4.4 we can estimate $\mu(x) = N_{\widetilde{\mathcal{H}}}(x)$ by the empirical estimator

$$\widehat{\mu}(x) = \bar{N}_{\widetilde{\mathcal{H}}}(x) = \frac{1}{n}\sum_{i=1}^{n} \langle \widetilde{\phi}(x), \widetilde{\phi}(X_i)\rangle^2 = \frac{1}{n}\sum_{i=1}^{n} K(x, X_i). \tag{4.8}$$

We introduce the estimated kernel

$$\widehat{K}(x,y) = \max\{\bar{N}_{\widetilde{\mathcal{H}}}(x), \sigma\}^{-1/2} K(x,y) \max\{\bar{N}_{\widetilde{\mathcal{H}}}(y), \sigma\}^{-1/2}$$

which provides an estimate of the thresholded kernel $\bar{K}_\sigma$. We remark that an explicit change of feature maps allows us to identify the Hilbert space defined by the kernel $\bar{K}$ (which we have denoted by $\mathcal{H}$) and the one defined by $\widehat{K}$. Indeed, we have

$$\widehat{K}(x,y) = \langle \widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}}, \qquad x, y \in \mathcal{X},$$

where the estimated feature map $\widehat{\phi}$ is defined as

$$\widehat{\phi}(x) = \chi(x)\phi(x), \tag{4.9}$$

with $\chi(x) = \left(\dfrac{N_{\widetilde{H}}(x)}{\max\{\overline{N}_{\widetilde{H}}(x), \sigma\}}\right)^{1/2}$. Further, when $x \in \mathrm{supp(P)}$ and $\sigma \le \sigma_*$, we have

$$\chi(x) = \left(\frac{\max\{N_{\widetilde{H}}(x), \sigma\}}{\max\{\bar{N}_{\widetilde{H}}(x), \sigma\}}\right)^{1/2}. \tag{4.10}$$

Next proposition provides a bound on the accuracy of the estimation of $\bar{K}_\sigma$ by $\widehat{K}$.

Let us first recall some notation. Let $a > 0$ and let

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil$$

where $c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2}\right)$ and

$$\kappa = \sup_{\substack{\theta \in \widetilde{\mathcal{H}} \\ \mathbb{E}[\langle \theta, \widetilde{\phi}(X)\rangle^2_{\widetilde{\mathcal{H}}}] > 0}} \frac{\mathbb{E}\big[\langle \theta, \widetilde{\phi}(X)\rangle^4_{\widetilde{\mathcal{H}}}\big]}{\mathbb{E}\big[\langle \theta, \widetilde{\phi}(X)\rangle^2_{\widetilde{\mathcal{H}}}\big]^2}$$

with $X \in \mathcal{X}$ a random vector of law P. We put

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & [6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \le \sqrt{n} \\ +\infty & \text{otherwise} \end{cases} \tag{4.11}$$

with

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\ \mathbf{Tr}(\widetilde{\mathcal{G}})}{4(2+c)t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa\ \mathbf{Tr}(\widetilde{\mathcal{G}})}{t}}\cosh(a/2)$$

where $\widetilde{\mathcal{G}}v = \int \langle v, \widetilde{\phi}(z)\rangle_{\widetilde{\mathcal{H}}}\ \widetilde{\phi}(z)\ \mathrm{dP}(z)$ is the Gram operator on $\widetilde{\mathcal{H}}$. According to Proposition 2.7 on page 74 we define

$$\tau_*(t) = \frac{\lambda_*(t)^2 \exp(a/2) R^4}{3\max\{t,\sigma\}^2}, \quad t \in \mathbb{R}_+,$$

where λ_* in equation (1.31) on page 36 and $R = \max_{i=1,\dots,n} \|\widetilde{\phi}(X_i)\|_{\widetilde{\mathcal{H}}}$.
In this setting we take $R = 1$ since we are working with a normalized kernel $\widetilde{K}$ which sends the observations on the unit sphere of the corresponding Hilbert space $\widetilde{\mathcal{H}}$.

Proposition 4.5. *Assume $0 < \sigma \le \sigma_*$ and let*

$$\xi(t) = B_*(t) + \frac{\tau_*(t)}{[1-\tau_*(t)]_+[1-B_*(t)]_+}.$$

With probability at least $1 - 2\epsilon$, for any $x \in \mathrm{supp}(\mathrm{P})$,

$$\left|\frac{\bar{K}_\sigma(x,y)}{\widehat{K}(x,y)} - 1\right| \le \frac{\xi(\mu(x)) + \xi(\mu(y))}{2} + \frac{\xi(\mu(x))^2 + \xi(\mu(y))^2}{2}$$

$$|\chi(x) - 1| \le \frac{\xi(\mu(x)) + \xi(\mu(x))^2}{2}\left[1 - \frac{\xi(\mu(x)) + \xi(\mu(x))^2}{2}\right]_+^{-1}.$$

Proof. By Proposition 2.7, with probability at least $1 - 2\epsilon$, for any $x \in \mathcal{X}$, we have

$$1 - \frac{\xi(\mu(x)) + \xi(\mu(x))^2}{2} \le [1 - \xi(\mu(x))]^{1/2} \le \frac{\phi(x)}{\widehat{\phi}(x)} \le [1 + \xi(\mu(x))]^{1/2} \le 1 + \frac{1}{2}\xi(\mu(x)),$$

which proves the second inequality. To obtain the first bound we observe that, for any $x, y \in \mathcal{X}$,

$$\begin{aligned} \frac{\bar{K}_\sigma(x,y)}{\widehat{K}(x,y)} &\le [1 + \xi(\mu(x))]^{1/2}[1 + \xi(\mu(y))]^{1/2} \\ &\le [1 + 2^{-1}\xi(\mu(x))]\,[1 + 2^{-1}\xi(\mu(y))] \\ &\le 1 + 2^{-1}[\xi(\mu(x)) + \xi(\mu(x))] + 4^{-1}\xi(\mu(x))\xi(\mu(y)) \\ &\le 1 + 2^{-1}[\xi(\mu(x)) + \xi(\mu(x))] + 8^{-1}[\xi(\mu(x))^2 + \xi(\mu(y))^2] \end{aligned}$$

and moreover

$$\frac{\bar{K}_\sigma(x,y)}{\widehat{K}(x,y)} \geq \big[1-\xi(\mu(x))\big]^{1/2}\big[1-\xi(\mu(y))\big]^{1/2}$$
$$\geq \left(1-\frac{\xi(\mu(x))+\xi(\mu(x))^2}{2}\right)\left(1-\frac{\xi(\mu(y))+\xi(\mu(y))^2}{2}\right)$$
$$\geq 1-\frac{\xi(\mu(x))+\xi(\mu(y))}{2}-\frac{\xi(\mu(x))^2+\xi(\mu(y))^2}{2}.$$

□

Replacing $\bar{K}$ by the estimated kernel $\widehat{K}$ in the definition of $\bar{K}_m$ on page 91, we obtain the new kernel

$$\bar{H}_m(x,y)=\int \widehat{K}(y,z_1)\widehat{K}(z_1,z_2)\dots\widehat{K}(x,z_{m-1})\,\mathrm{dP}^{\otimes(m-1)}(z_1,\dots,z_{m-1}).$$

This new kernel is still not observable since it makes use again of the sample distribution P. To make the mathematical discussion simpler, we study the estimation of $\bar{H}_m$ from a second i.i.d. sample $X_{n+1},\dots,X_{2n}\in\mathbb{R}^d$ drawn according to P and independent from the first sample $X_1,\dots,X_n$.

In this simplified split sample setting, $\widehat{K}$ can be treated as a non-random kernel since the distribution of the second sample, conditioned on the value of the first sample, $\mathrm{P}_{X_{n+1:2n}|X_{1:n}}$, is the same product distribution $\mathrm{P}^{\otimes n}$ as the distribution of the second sample $\mathrm{P}_{X_{n+1:2n}}$.

Similarly to Proposition 4.1, we observe that the kernel $\bar{H}_m$ can be written as

$$\bar{H}_m(x,y)=\langle\widehat{\mathcal{G}}^{m-1}\widehat{\phi}(x),\widehat{\phi}(y)\rangle_{\mathcal{H}}, \tag{4.12}$$

where $\widehat{\mathcal{G}}:\mathcal{H}\to\mathcal{H}$ is the Gram operator defined by

$$\widehat{\mathcal{G}}v=\int\langle v,\widehat{\phi}(z)\rangle_{\mathcal{H}}\;\widehat{\phi}(z)\;\mathrm{dP}(z). \tag{4.13}$$

We remark that, since by definition $\chi(x)>0$,

$$\mathbf{ker}(\widehat{\mathcal{G}})=\left\{u\in\mathcal{H},\int\langle u,\widehat{\phi}(x)\rangle^2\,\mathrm{dP}(x)=0\right\}$$
$$=\left\{u\in\mathcal{H},\int\langle u,\phi(x)\rangle^2\,\mathrm{dP}(x)=0\right\}=\mathbf{ker}(\mathcal{G}).$$

Consequently, since $\mathbf{Im}(\mathcal{G})=\mathbf{ker}(\mathcal{G})^\perp$, we get

$$\overline{\mathbf{span}}\Big(\phi(\mathrm{supp}(P))\Big)=\mathbf{ker}(\mathcal{G})^\perp=\mathbf{Im}(\mathcal{G})=\mathbf{Im}(\widehat{\mathcal{G}})=\overline{\mathbf{span}}\Big(\widehat{\phi}(\mathrm{supp}(P))\Big). \tag{4.14}$$

To provide an estimator of the Gram operator $\widehat{\mathcal{G}}$ we use the construction done in section 1.3. We consider a confidence region for $\langle\widehat{\mathcal{G}}\theta,\theta\rangle_{\mathcal{H}}$ that we denote by

$$B_-(\theta)\leq\langle\widehat{\mathcal{G}}(\theta),\theta\rangle_{\mathcal{H}}\leq B_+(\theta),\quad\theta\in\mathcal{H},$$

and we define the estimator $\mathcal{Q}$ of $\widehat{\mathcal{G}}$ as in equation (1.45) on page 48. Let $\widehat{\mathcal{Q}}=\mathcal{Q}_+$ be the positive part of $\mathcal{Q}$ so that, according to Proposition 1.33 on page 49, with probability at least $1-2\epsilon$, for any $\theta\in\mathbb{S}_{\mathcal{H}}$ (the unit sphere of $\mathcal{H}$),

$$\big|\max\{\langle\theta,\widehat{\mathcal{Q}}\theta\rangle,\sigma\}-\max\{\langle\theta,\widehat{G}\theta\rangle,\sigma\}\big|\leq 2\max\{\langle\theta,\widehat{\mathcal{G}}\theta\rangle,\sigma\}B_*(\langle\theta,\widehat{\mathcal{G}}\theta\rangle)+\eta, \tag{4.15}$$

for a small parameter η. Hence we get the estimated kernel

$$\widehat{H}_m(x,y) = \langle \widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}},$$

and we consider as an estimator of $K_m(x,y) = \bar{K}_{2m}(x,x)^{-1/2}\bar{K}_{2m}(x,y)\bar{K}_{2m}(y,y)^{-1/2}$ the kernel

$$H_m(x,y) = \widehat{H}_{2m}(x,x)^{-1/2}\widehat{H}_{2m}(x,y)\widehat{H}_{2m}(y,y)^{-1/2}.$$

The accuracy of this estimation is proved in the following proposition. We introduce

$$\mathcal{E}_k(x,y) = \big|\widehat{H}_{2m}(x,y) - \bar{K}_{2m}(x,y)\big|, \quad x,y \in \mathrm{supp}(\mathrm{P}),$$

and the non-normalized quadratic error on the non-normalized kernel produced by the clustering algorithm

$$\mathcal{E}_k(x) = \left(\int \mathcal{E}_k(x,y)^2 \,\mathrm{dP}(y)\right)^{1/2}, \quad x \in \mathrm{supp}(\mathrm{P}).$$

We recall that

$$\bar{K}_{2m}(x,y) = \langle \mathcal{G}^{2m-1}\phi(x), \phi(y)\rangle_{\mathcal{H}}$$

where $\mathcal{G}$ is the Gram operator defined in equation (4.3).

Proposition 4.6. *For any $x, y \in \mathrm{supp}(\mathrm{P})$,*

$$\mathcal{E}_k(x,y) \le \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}\mu(y)^{1/2}}\Big(\|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty + 2\|\chi - 1\|_\infty\Big)$$
$$\mathcal{E}_k(x) \le \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}}\Big(\|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty + 2\|\chi - 1\|_\infty\Big),$$

where χ and μ are defined in equation (4.10) *and in equation* (4.7) *respectively.*

Proof. We observe that, by definition,

$$\begin{aligned}\mathcal{E}_k(x,y) &= |\langle \widehat{\mathcal{Q}}^{2m-1}\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}} - \langle \mathcal{G}^{2m-1}\phi(x), \phi(y)\rangle_{\mathcal{H}}| \\ &\le |\langle (\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1})\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}}| + |\langle \mathcal{G}^{2m-1}(\widehat{\phi}(x) - \phi(x)), \widehat{\phi}(y)\rangle_{\mathcal{H}}| \\ &\qquad + \langle \mathcal{G}^{2m-1}\phi(x), (\widehat{\phi}(y) - \phi(y))\rangle_{\mathcal{H}}|.\end{aligned}$$

Recalling the definition of $\widehat{\phi}$ we get

$$\begin{aligned}\mathcal{E}_k(x,y) \le \|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty \|\chi\|_\infty^2 \|\phi(x)\|_{\mathcal{H}}\|\phi(y)\|_{\mathcal{H}} \\ + \|\chi - 1\|_\infty(\|\chi\|_\infty + 1)\|\phi(x)\|_{\mathcal{H}}\|\phi(y)\|_{\mathcal{H}}.\end{aligned}$$

Since, by definition, $K(x,x) = 1$, we obtain

$$\|\phi(x)\|_{\mathcal{H}}^2 = \bar{K}(x,x) = \frac{K(x,x)}{\mu(x)^{1/2}\mu(x)^{1/2}} = \frac{1}{\mu(x)}$$

which proves the first bound. We now consider $\mathcal{E}_k$. We have

$$\begin{aligned}\mathcal{E}_k(x) &= \left(\int \left(\langle \widehat{\mathcal{Q}}^{2m-1}\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}} - \langle \mathcal{G}^{2m-1}\phi(x), \phi(y)\rangle_{\mathcal{H}}\right)^2 \mathrm{dP}(y)\right)^{1/2} \\ &\leq \left(\int \langle (\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1})\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}}^2 \,\mathrm{dP}(y)\right)^{1/2} \\ &\quad + \left(\int \langle \mathcal{G}^{2m-1}(\widehat{\phi}(x) - \phi(x)), \widehat{\phi}(y)\rangle_{\mathcal{H}}^2 \,\mathrm{dP}(y)\right)^{1/2} \\ &\quad + \left(\int \langle \mathcal{G}^{2m-1}\phi(x), \widehat{\phi}(y) - \phi(y)\rangle_{\mathcal{H}}^2 \,\mathrm{dP}(y)\right)^{1/2}.\end{aligned}$$

Recalling that

$$\|\widehat{\mathcal{G}}^{1/2}u\|_{\mathcal{H}}^2 = \langle \widehat{\mathcal{G}}u, u\rangle_{\mathcal{H}} = \int \langle u, \widehat{\phi}(y)\rangle_{\mathcal{H}}^2 \,\mathrm{dP}(y) \leq \|\chi\|_\infty^2 \langle \mathcal{G}u, u\rangle, \quad u \in \mathcal{H},$$

we conclude

$$\begin{aligned}\mathcal{E}_k(x) \leq \|\widehat{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1})\widehat{\phi}(x)\|_{\mathcal{H}} + \|\widehat{\mathcal{G}}^{1/2}\mathcal{G}^{2m-1}(\widehat{\phi}(x) - \phi(x))\|_{\mathcal{H}} \\ + \|\chi - 1\|_\infty \|\mathcal{G}^{2m-1/2}\phi(x)\|_{\mathcal{H}}.\end{aligned}$$

□

Before providing a bound on the quantities that appear at the right-hand side of the above inequalities, we present another way to construct the estimator H_m that will lead to obtain tighter bounds on the approximation error.

We recall that the kernel K_m can be written as

$$K_m(x, y) = \left\langle \|\mathbf{R}(x)\|^{-1}\mathbf{R}(x), \|\mathbf{R}(y)\|^{-1}\mathbf{R}(y)\right\rangle_{L^2_{\mathrm{P}}},$$

where the representation $\mathbf{R}(x)$ is defined in equation (4.6), and we separately estimate the representation $\|\mathbf{R}(x)\|^{-1}\mathbf{R}(x)$ and the scalar product in L^2_{P}.

Our estimated representation of x in L^2_{P} is $\widehat{\mathbf{N}}(x)^{-1}\widehat{\mathbf{R}}(x)$, where

$$\widehat{\mathbf{R}}(x)(y) = \langle \widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x), \phi(y)\rangle_{\mathcal{H}} \quad \text{and} \quad \widehat{\mathbf{N}}(x) = \langle \widehat{\mathcal{Q}}^{2m-1}\widehat{\phi}(x), \widehat{\phi}(x)\rangle_{\mathcal{H}}^{1/2}.$$

This representation is not fully observable because of the presence of the ideal feature map ϕ in the definition of $\widehat{\mathbf{R}}$. Nevertheless, it can be used in practice since, as we will explain in the following, it is possible to derive an observable estimate of the norm $\|\widehat{\mathbf{R}}(x) - \mathbf{R}(x)\|_{L^2_{\mathrm{P}}}$.

More precisely we show that we can estimate the norm in $\mathbf{Im}\,(\mathbf{G}) \subset L^2_{\mathrm{P}}$, where $\mathbf{G} : L^2_{\mathrm{P}} \to L^2_{\mathrm{P}}$ is the Gram operator defined by the formula

$$\mathbf{G}(f)(x) = \int \bar{K}(x, y) f(y) \,\mathrm{dP}(y)$$

and such that

$$\|\mathbf{G}\|_\infty = \|\mathcal{G}\|_\infty = 1.$$

A tool for representing functions in $\mathbf{Im}(\mathbf{G})$ is the operator $\mathcal{S} : \mathcal{H} \to L^2_{\mathrm{P}}$ defined by

$$\mathcal{S}(u)(x) = \langle u, \phi(x) \rangle_{\mathcal{H}}, \quad x \in \mathcal{X}.$$

We observe that with this notation

$$\mathbf{G} = \mathcal{S}\mathcal{S}^* \quad \text{and} \quad \mathcal{G} = \mathcal{S}^*\mathcal{S} \tag{4.16}$$

where we recall that $\mathcal{H}$ is the reproducing kernel Hilbert space defined by $\bar{K}$ and ψ the corresponding feature map.

Lemma 4.7. *We have*

$$\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G})) = \mathcal{S}(\mathbf{Im}(\widehat{\mathcal{G}})).$$

Proof. We observe that, since $\mathbf{G}$ is symmetric, $\mathbf{Im}(\mathbf{G}) = \mathbf{Im}(\mathbf{G}^2)$. By equation (4.16), we deduce that $\mathbf{G}^2 = \mathcal{S}\mathcal{G}\mathcal{S}^*$, showing that $\mathbf{Im}(\mathbf{G}) = \mathbf{Im}(\mathbf{G}^2) \subset \mathcal{S}(\mathbf{Im}(\mathcal{G}))$. Moreover, as $\mathcal{S}\mathcal{G} = \mathbf{G}\mathcal{S}$, we conclude that $\mathcal{S}(\mathbf{Im}(\mathcal{G})) \subset \mathbf{Im}\,\mathbf{G}$. □

Therefore any $f \in \mathbf{Im}(\mathbf{G})$ is of the form $f = \mathcal{S}u$, with $u \in \mathbf{Im}(\mathcal{G})$, so that we can estimate

$$\|f\|^2_{L^2_{\mathrm{P}}} = \langle \mathcal{G}u, u \rangle_{\mathcal{H}}$$

by $\langle \widehat{\mathcal{Q}}u, u \rangle_{\mathcal{H}}$ and the estimation error is bounded as described in the following lemma.

Lemma 4.8. *For any $u \in \mathcal{H}$, we have*

$$\left| \|\mathcal{S}u\|^2_{L^2_{\mathrm{P}}} - \langle \widehat{\mathcal{Q}}u, u \rangle_{\mathcal{H}} \right| \leq \|\mathcal{G} - \widehat{\mathcal{Q}}\|_\infty \|u\|^2_{\mathcal{H}}.$$

More generally, for any $u, v \in \mathcal{H}$,

$$\left| \langle \mathcal{S}u, \mathcal{S}v \rangle_{L^2_{\mathrm{P}}} - \langle \widehat{\mathcal{Q}}u, v \rangle_{\mathcal{H}} \right| \leq \|\mathcal{G} - \widehat{\mathcal{Q}}\|_\infty \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}.$$

Proof. It is sufficient to observe that

$$\begin{aligned}
\left| \langle \mathcal{S}u, \mathcal{S}v \rangle_{L^2_{\mathrm{P}}} - \langle \widehat{\mathcal{Q}}u, v \rangle_{\mathcal{H}} \right| &= \left| \int \langle u, \phi(x) \rangle_{\mathcal{H}} \langle \phi(x), v \rangle_{\mathcal{H}} \, \mathrm{dP}(x) - \langle \widehat{\mathcal{Q}}u, v \rangle_{\mathcal{H}} \right| \\
&= \left| \langle \mathcal{G}u, v \rangle_{\mathcal{H}} - \langle \widehat{\mathcal{Q}}u, v \rangle_{\mathcal{H}} \right| \\
&\leq \left\| (\mathcal{G} - \widehat{\mathcal{Q}})u \right\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \\
&\leq \|\mathcal{G} - \widehat{\mathcal{Q}}\|_\infty \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}.
\end{aligned}$$

□

Let us now come back to $\mathbf{R}(x)$ and $\widehat{\mathbf{R}}(x)$.

Lemma 4.9. *We have*

$$\begin{aligned}
&\mathbf{R}(x) = \mathcal{S}\mathcal{G}^{m-1}\phi(x) \in \mathbf{Im}(\mathbf{G}) \\
&\widehat{\mathbf{R}}(x) = \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x) \in \mathbf{Im}(\mathbf{G}) \quad \textit{almost surely.}
\end{aligned}$$

Proof. We observe that, by definition,

$$\widehat{\mathbf{R}}(x) = \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x) \quad \text{and} \quad \mathbf{R}(x) = \mathcal{S}\mathcal{G}^{m-1}\phi(x).$$

Hence, since

$$\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G})) = \mathcal{S}\Big(\overline{\mathbf{span}}\,\phi(\mathrm{supp}(\mathrm{P}))\Big),$$

we conclude that $\mathbf{R}(x) \in \mathbf{Im}(\mathbf{G})$. Moreover, since

$$\begin{aligned}\mathbf{span}\{\widehat{\phi}(X_n), \dots, \widehat{\phi}(X_{2n})\} &= \mathbf{span}\{\phi(X_n), \dots, \phi(X_{2n})\} \\ &\subset \overline{\mathbf{span}}\Big(\phi(\mathrm{supp}(\mathrm{P}))\Big) \quad \text{almost surely,}\end{aligned}$$

also $\widehat{\mathbf{R}}(x) \in \mathbf{Im}(\mathbf{G})$. □

As a consequence, combining the two previous results, the kernel

$$\langle \widehat{\mathbf{R}}(x), \widehat{\mathbf{R}}(y)\rangle_{L^2_{\mathrm{P}}} = \langle \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x), \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(y)\rangle_{L^2_{\mathrm{P}}}$$

is estimated by

$$\widehat{H}_{2m}(x,y) = \langle \widehat{\mathcal{Q}}^{2m-1}\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}}$$

and therefore we obtain, as before, as an estimator of K_m the kernel

$$H_m(x,y) = \widehat{H}_{2m}(x,x)^{-1/2}\widehat{H}_{2m}(x,y)\widehat{H}_{2m}(y,y)^{-1/2}.$$

Let us mention that the representation $\widehat{\mathbf{R}}(x) \in L^2_{\mathrm{P}}$ is isometric to the fully observed representation

$$\widehat{\mathcal{R}}(x) = \widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x), \qquad x \in \mathcal{X},$$

in the Hilbert space $(\mathbf{Im}(\mathcal{G}), \|\cdot\|_{\mathcal{G}}) = (\mathbf{Im}(\widehat{\mathcal{G}}), \|\cdot\|_{\mathcal{G}})$ with the non-observed Hilbert norm $\|u\|_{\mathcal{G}} = \langle \mathcal{G}u, u\rangle^{1/2}$ (that could be estimated by $\langle \widehat{\mathcal{G}}u, u\rangle^{1/2}$ according to Lemma 4.8). However we do not consider the representation

$$\widehat{\mathcal{Q}}^{m-1/2}\widehat{\phi}(x) \in (\mathcal{H}, \|\cdot\|_{\mathcal{H}})$$

because its study would require a bound for the operator norm $\|\widehat{\mathcal{Q}}^{3/2} - \widehat{\mathcal{G}}^{3/2}\|_\infty$ and the one we obtained in previous chapters is less precise than what we got for $\|\widehat{\mathcal{Q}} - \widehat{\mathcal{G}}\|_\infty$ itself (that is instead what we get choosing the representation $\widehat{\mathcal{R}}(x)$).

In the following we provide non-asymptotic bounds showing that the estimated non-normalized representation $\widehat{\mathbf{R}}$ converges towards the ideal non-normalized representation $\mathbf{R}$, defined in equation (4.5), when the sample size goes to infinity.

More specifically we give non-asymptotic bounds for the non-normalized representation error in L^2_{P} norm

$$\mathcal{E}_r(x) = \big\|\mathbf{R}(x) - \widehat{\mathbf{R}}(x)\big\|_{L^2_{\mathrm{P}}}, \quad x \in \mathrm{supp}(\mathrm{P}),$$

and

$$\mathcal{E}_c(f) = \left(\int \Big\langle \mathbf{R}(x) - \widehat{\mathbf{R}}(x), f\Big\rangle^2_{L^2_{\mathrm{P}}} \mathrm{dP}(x)\right)^{1/2}, \quad f \in L^2_{\mathrm{P}},$$

that qualifies the convergence of the coordinates of the representation. We introduce these two errors because we will be able to produce a tighter bound for $\mathcal{E}_c$. In favorable situations the ideal representation $\mathbf{R}$ lives in a neighborhood of a low-dimensional space so that the speed of convergence is well reflected by the speed of convergence along those few dimensions.

Proposition 4.10. *The two errors defined above are such that, for any $x \in \mathrm{supp}(\mathrm{P})$, any $f \in L^2_{\mathrm{P}}$,*

$$\mathcal{E}_r(x) \leq \mu(x)^{-1/2}\Big(\|\chi\|_\infty \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_\infty + \|\chi - 1\|_\infty\Big)$$
$$\mathcal{E}_c(f) \leq \|f\|_{L^2_{\mathrm{P}}}\Big(\|\chi\|_\infty \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_\infty + \|\chi - 1\|_\infty\Big),$$

where χ and μ are defined in equation (4.10) *and in equation* (4.7) *respectively.*

Proof. In order to prove the first inequality, we observe that, by definition,

$$\begin{aligned}\mathcal{E}_r(x) &= \big\|\mathcal{S}\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x) - \mathcal{S}\mathcal{G}^{m-1}\phi(x)\big\|_{L^2_{\mathrm{P}}} \\ &\leq \big\|\mathcal{S}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\widehat{\phi}(x)\big\|_{L^2_{\mathrm{P}}} + \big\|\mathcal{S}\mathcal{G}^{m-1}(\widehat{\phi}(x) - \phi(x))\big\|_{L^2_{\mathrm{P}}}.\end{aligned}$$

Moreover, since $\|\mathcal{S}u\|^2_{L^2_{\mathrm{P}}} = \langle \mathcal{S}^*\mathcal{S}u, u\rangle_{\mathcal{H}} = \langle \mathcal{G}u, u\rangle_{\mathcal{H}}$, we get $\|\mathcal{S}\|_\infty = \|\mathcal{G}\|_\infty^{1/2} = 1$. As a consequence, recalling the definition of $\widehat{\phi}$, we get

$$\mathcal{E}_r(x) \leq \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_\infty \|\chi\|_\infty \|\phi(x)\|_{\mathcal{H}} + \|\chi - 1\|_\infty \|\phi(x)\|_{\mathcal{H}}.$$

We now prove the second bound. Let $\mathbf{\Phi} : L^2_{\mathrm{P}} \to L^2_{\mathrm{P}}$ be the orthogonal projector on $\mathbf{Im}(\mathbf{G})$. Since, according to Lemma 4.9, almost surely, $\widehat{\mathbf{R}}(x) - \mathbf{R}(x) \in \mathbf{Im}(\mathbf{G})$, for any $x \in \mathcal{X}$, then

$$\Big\langle \widehat{\mathbf{R}}(x) - \mathbf{R}(x), f\Big\rangle_{L^2_{\mathrm{P}}} = \Big\langle \widehat{\mathbf{R}}(x) - \mathbf{R}(x), \mathbf{\Phi}(f)\Big\rangle_{L^2_{\mathrm{P}}} \quad \text{almost surely.}$$

Moreover, since $\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G}))$, there is $u \in \mathbf{Im}(\mathcal{G})$ such that $\mathbf{\Phi}(f) = \mathcal{S}u$. We can then write

$$\begin{aligned}\Big\langle \widehat{\mathbf{R}}(x) - \mathbf{R}(x), f\Big\rangle_{L^2_{\mathrm{P}}} &= \Big\langle \widehat{\mathbf{R}}(x) - \mathbf{R}(x), \mathcal{S}u\Big\rangle_{L^2_{\mathrm{P}}} \\ &= \Big\langle \mathcal{S}(\widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x) - \mathcal{G}^{m-1}\phi(x)), \mathcal{S}u\Big\rangle_{L^2_{\mathrm{P}}} \\ &= \langle \widehat{\mathcal{Q}}^{m-1}\widehat{\phi}(x) - \mathcal{G}^{m-1}\phi(x), \mathcal{G}u\rangle_{\mathcal{H}} \\ &= \langle (\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\widehat{\phi}(x), \mathcal{G}u\rangle_{\mathcal{H}} + \langle \mathcal{G}^{m-1}(\widehat{\phi}(x) - \phi(x)), \mathcal{G}u\rangle_{\mathcal{H}}.\end{aligned}$$

Therefore, similarly as before, we get

$$\begin{aligned}\mathcal{E}_c(f) &\leq \|\widehat{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\mathcal{G}u\|_{\mathcal{H}} + \|\chi - 1\|_\infty \left(\int \langle \mathcal{G}^{m-1}\phi(x), \mathcal{G}u\rangle^2_{\mathcal{H}} \,\mathrm{dP}(x)\right)^{1/2} \\ &= \|\widehat{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\mathcal{G}u\|_{\mathcal{H}} + \|\chi - 1\|_\infty \|\mathcal{G}^{m+1/2}u\|_{\mathcal{H}} \\ &\leq \|\chi\|_\infty \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_\infty \|\mathcal{G}^{1/2}u\|_{\mathcal{H}} + \|\chi - 1\|_\infty \|\mathcal{G}^{1/2}u\|_{\mathcal{H}}.\end{aligned}$$

We conclude the proof observing that

$$\|\mathcal{G}^{1/2}u\|_{\mathcal{H}} = \|\mathbf{\Phi}(f)\|_{L^2_{\mathrm{P}}} \leq \|f\|_{L^2_{\mathrm{P}}},$$

where we recall that $\mathcal{G} = \mathcal{S}^*\mathcal{S}$ and $\mathbf{\Phi}(f) = \mathcal{S}u$. □

In order to provide observable bounds for the four errors defined previously on page 100 and on page 104 we need to introduce some technical results.

Lemma 4.11. *Let $m > 0$. We have*

$$\|\widehat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty \le \Big(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\Big)^m - 1 \le m\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\Big(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\Big)^{m-1}.$$

Proof. We recall that $\|\mathcal{G}\|_\infty = 1$. Thus, since

$$\widehat{\mathcal{Q}}^m - \mathcal{G}^m = \sum_{k=0}^{m-1} \widehat{\mathcal{Q}}^k(\widehat{\mathcal{Q}} - \mathcal{G})\mathcal{G}^{m-k-1},$$

we get

$$\|\widehat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty \le \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \sum_{k=0}^{m-1} \|\widehat{\mathcal{Q}}\|_\infty^k.$$

Hence, since $\|\widehat{\mathcal{Q}}\|_\infty \le 1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty$, we have

$$\|\widehat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty \le \Big(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\Big)^m - 1 = \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \sum_{k=0}^{m-1} \big(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\big)^k \le m\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \big(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\big)^{m-1}.$$

□

We introduce

$$\delta_1 = \xi\big(\operatorname*{essinf}_{\mathrm{P}} \mu\big) + \xi\big(\operatorname*{essinf}_{\mathrm{P}} \mu\big)^2,$$

where we recall that ξ is defined in Proposition 4.5 and $\mu(x) = \int K(x,z)\,\mathrm{dP}(z)$. We also put

$$\delta_2 = 2\max\{\|\widehat{\mathcal{G}}\|_\infty, \sigma\} B_*(\|\widehat{\mathcal{G}}\|_\infty) + \sigma + \eta,$$

where B_* is defined in equation (4.11), $\widehat{\mathcal{G}}$ is the Gram operator defined in equation (4.13) and $\eta > 0$ is introduced in equation (4.15) on page 99.

Proposition 4.12. *Assume $0 < \sigma \le \sigma_*$. With probability at least $1 - 2\epsilon$, the following inequalities hold together*

$$\|\chi - 1\|_\infty \le \frac{\delta_1}{(1-\delta_1)_+},$$
$$\|\widehat{\mathcal{G}} - \mathcal{G}\|_\infty \le \frac{\delta_1(2-\delta_1)}{(1-\delta_1)_+^2},$$

so that

$$\|\chi\|_\infty \le (1-\delta_1)_+^{-1} \quad \textit{and} \quad \|\widehat{\mathcal{G}}\|_\infty \le (1-\delta_1)_+^{-2}$$

where χ is defined in equation (4.10). Moreover, with probability at least $1 - 2\epsilon$,

$$\|\widehat{\mathcal{Q}} - \widehat{\mathcal{G}}\|_\infty \le \delta_2.$$

Proof. According to Proposition 4.5, we have

$$\|\chi - 1\|_\infty \leq \frac{\delta_1}{(1-\delta_1)_+} \tag{4.17}$$

so that $\|\chi\|_\infty \leq (1-\delta_1)_+^{-1}$. Moreover, we observe that

$$\begin{aligned}
|\langle(\widehat{\mathcal{G}} - \mathcal{G})u, u\rangle_{\mathcal{H}}| &= \left|\int \left(\langle u, \widehat{\phi}(x)\rangle^2 - \langle u, \phi(x)\rangle^2\right) \mathrm{dP}(x)\right| \\
&= \left|\int (\chi(x)^2 - 1)\langle u, \phi(x)\rangle^2 \,\mathrm{dP}(x)\right| \\
&\leq \|\chi^2 - 1\|_\infty \int \langle u, \phi(x)\rangle^2 \,\mathrm{dP}(x) = \|\chi^2 - 1\|_\infty \langle \mathcal{G}u, u\rangle.
\end{aligned}$$

Therefore, since $\|\mathcal{G}\|_\infty = 1$,

$$\|\widehat{\mathcal{G}} - \mathcal{G}\|_\infty \leq \|\chi^2 - 1\|_\infty \|\mathcal{G}\|_\infty \leq \|1 - \chi\|_\infty \Big(2 + \|1 - \chi\|_\infty\Big).$$

The last inequality follows immediately from Proposition 1.33 on page 49. □

Corollary 4.13. *With probability at least $1 - 4\epsilon$, in addition to the previous inequalities,*

$$\delta_2 \leq 2(1-\delta_1)_+^{-2} B_*\left((1-\delta_1)_+^{-2}\right) + \sigma + \eta$$

and

$$\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \leq \frac{\delta_1(2-\delta_1)}{(1-\delta_1)_+^2} + \delta_2.$$

Therefore

$$\begin{aligned}
\mathcal{E}_k(x,y) &\leq \mu(x)^{-1/2}\mu(y)^{-1/2} 2m(2\delta_1 + \delta_2)\frac{(1 + 2\delta_1 + \delta_2)^{2m-2}}{(1-\delta_1)_+^{4m}}, && x, y \in \mathrm{supp}(\mathrm{P}), \\
\mathcal{E}_k(x) &\leq \mu(x)^{-1/2} 2m(2\delta_1 + \delta_2)\frac{(1 + 2\delta_1 + \delta_2)^{2m-2}}{(1-\delta_1)_+^{4m}}, && x \in \mathrm{supp}(\mathrm{P}), \\
\mathcal{E}_r(x) &\leq \mu(x)^{-1/2}(m - 1/2)(2\delta_1 + \delta_2)\frac{(1 + 2\delta_1 + \delta_2)^{m-2}}{(1-\delta_1)_+^{2m-1}}, && x \in \mathrm{supp}(\mathrm{P}), \\
\mathcal{E}_c(f) &\leq \|f\|_{L^2_{\mathrm{P}}}(m - 1/2)(2\delta_1 + \delta_2)\frac{(1 + 2\delta_1 + \delta_2)^{m-2}}{(1-\delta_1)_+^{2m-1}}, && f \in L^2_{\mathrm{P}},
\end{aligned}$$

where the four errors above are defined on page 100 and on page 104.

Proof. The bound for δ_2 is a consequence of Lemma 1.22 on page 41. The second bound comes from the triangle inequality applied to the operator norm

$$\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \leq \|\widehat{\mathcal{Q}} - \widehat{\mathcal{G}}\|_\infty + \|\widehat{\mathcal{G}} - \mathcal{G}\|_\infty.$$

Combining previous bounds, we get

$$\begin{aligned}
\mathcal{E}_k(x,y) \leq \mu(x)^{-1/2}\mu(y)^{-1/2}(1-\delta_1)^{-2}\Bigg[&(2m-1)\left(\frac{2\delta_1}{(1-\delta_1)^2} + \delta_2\right) \\
&\times \left(1 + \frac{2\delta_1}{(1-\delta_1)^2} + \delta_2\right)^{2m-2} + \frac{2\delta_1}{1-\delta_1}\Bigg],
\end{aligned}$$

which is lower than the simplified bound stated in the corollary. A similar computation holds for $\mathcal{E}_k(x)$, replacing $\mu(y)^{-1/2}$ by the constant one. Concerning $\mathcal{E}_r(x)$ we get

$$\mathcal{E}_r(x) \leq \mu(x)^{-1/2}\left[\frac{(m-1)}{(1-\delta_1)_+}\left(\frac{2\delta_1}{(1-\delta_1)_+^2}+\delta_2\right)\right.$$
$$\left.\times\left(1+\frac{2\delta_1}{(1-\delta_1)_+^2}+\delta_2\right)^{m-2}+\frac{\delta_1}{(1-\delta_1)_+}\right],$$

leading to the simplified bound presented in the corollary. The computation for $\mathcal{E}_c(f)$ is the same replacing $\mu(x)^{-1/2}$ with $\|f\|_{L^2_{\mathrm{P}}}$. □

To conclude we observe that the last renormalization to obtain the kernel H_m can also be seen as a projection on the sphere. Indeed the feature map of H_m can be taken to be the projection on the sphere of the feature map of $\widehat{H}_{2m}$.

From this kernel we derive the new classification by thresholding the distances between the points. We have already said that taking a suitable power of the kernel $\bar{K}$, which is here approximated by $\widehat{K}$, will threshold to zero its smallest eigenvalues. This leads to a natural dimensionality reduction which allows us to automatically estimate the number c of classes. We produce a greedy classification algorithm which consists in taking an index $i \in \{1, \ldots, n\}$ at random, constructing the corresponding class

$$C_i = \{j \mid H_m(X_i, X_j) \geq s\},$$

where s is a fixed threshold, and then starting again with the remaining indices. In such a way we construct a sequence of indices $i_1, \ldots, i_c$, where the number c of classes is then estimated automatically. The k-th class is hence defined by

$$C_k = C_{i_k} \setminus \bigcup_{\ell<k} C_{i_\ell}, \quad k = 1, \ldots, c.$$

In some configurations, this obviously yields an unstable classification, but in practice, when the classes are clearly separated from each other, this simple scheme is successful.

4.3.1 Implementation of the algorithm

Here we describe a simplified algorithm freely inspired by the algorithms for which we proved theoretical learning bounds in the previous sections.

First recall that we estimate $\mu(x)$ by $\widehat{\mu}(x)$ as described in equation (4.8) on page 97 to form the estimated kernel

$$\widehat{K}(x,y) = \max\{\widehat{\mu}(x), \sigma\}^{-1/2} K(x,y) \max\{\widehat{\mu}(y), \sigma\}^{-1/2}$$

of $\bar{K}$. Then we construct an estimator of the kernel

$$\bar{H}_m(x,y) = \int \widehat{K}(y,z_1)\widehat{K}(z_1,z_2)\ldots\widehat{K}(x,z_{m-1})\,\mathrm{dP}^{\otimes(m-1)}(z_1,\ldots,z_{m-1}),$$

based on a second independent sample $X_{n+1}, \ldots, X_{2n} \in \mathbb{R}^d$ drawn according to P.

According to equation (4.12), we have

$$\bar{H}_m(x,y) = \langle \widehat{\mathcal{G}}^{m-1}\widehat{\phi}(x), \widehat{\phi}(y)\rangle_{\mathcal{H}}$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space defined by $\bar{K}$, the Gram operator $\widehat{\mathcal{G}}$ is defined in equation (4.13) and the feature map $\widehat{\phi}$ in equation (4.9).
Instead of estimating $\widehat{\mathcal{G}}$ by $\widehat{\mathcal{Q}}$ constructed as explained on page 99 (according to equation (1.45) on page 48) we use here a simpler iterative scheme to compute $\widehat{\mathcal{Q}}$. More precisely, we use the polarization formula to estimate the coefficients of $\widehat{\mathcal{Q}}$ in an orthonormal basis of

$$\mathbf{span}\{\widehat{\phi}(X_{n+1}), \dots, \widehat{\phi}(X_{2n})\}$$

and then update this basis by replacing it iteratively by a diagonalization basis of the previous estimate.

We introduce the $n \times n$ matrix M defined as

$$M_{i,j} = \widehat{K}(X_{n+i}, X_{n+j}), \qquad i, j = 1, \dots, n.$$

We denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ its eigenvalues and we decompose M as

$$M = UDU^\top$$

with $UU^\top = U^\top U = \mathrm{I}$ and $D = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. Let r be the rank of M so that $\lambda_r > 0$ and $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_n = 0$. We define

$$V_{i,j} = \lambda_i^{-1/2} U_{j,i}, \qquad 1 \leq i \leq r, \quad 1 \leq j \leq n.$$

Lemma 4.14. *An orthonormal basis of* $\mathbf{span}\{\widehat{\phi}(X_{n+1}), \dots, \widehat{\phi}(X_{2n})\}$ *is given by*

$$q_i = \sum_{j=1}^{n} V_{i,j}\ \widehat{\phi}(X_{n+j}), \quad i = 1, \dots, r,$$

where the sum is performed in $\mathcal{H}$*. Moreover, we can reconstruct* $\widehat{\phi}(X_{n+i})$*, for* $i = 1, \dots, n$*, as*

$$\widehat{\phi}(X_{n+i}) = \sum_{k=1}^{r} U_{i,k} \lambda_k^{1/2} q_k. \tag{4.18}$$

Proof. To prove the first part of the lemma, it is sufficient to observe that the number of vectors is r and that, for any $i, j \in \{1, \dots, r\}$,

$$\langle q_i, q_j \rangle_{\mathcal{H}} = (VMV^\top)_{i,j} = \lambda_i^{-1/2} (U^\top U D U^\top U)_{i,j} \lambda_j^{-1/2} = \lambda_i^{-1/2} D_{i,j} \lambda_j^{-1/2} = \mathrm{I}_{i,j}.$$

To prove equation (4.18), we observe that

$$\begin{aligned}\widehat{\phi}(X_{n+i}) = \sum_{k=1}^{r} \langle \widehat{\phi}(X_{n+i}), q_k \rangle q_k = \sum_{k=1}^{r} \sum_{j=1}^{n} M_{i,j} U_{j,k} \lambda_k^{-1/2} q_k \\ = \sum_{k=1}^{r} (UDU^\top U)_{i,k} \lambda_k^{-1/2} q_k = \sum_{k=1}^{r} U_{i,k} \lambda_k \lambda_k^{-1/2} q_k.\end{aligned}$$

□

We start by estimating in the basis $\{q_i\}_{i=1}^r$ the Gram matrix

$$G_{i,j} = \langle \widehat{\mathcal{G}}(q_i), q_j \rangle_{\mathcal{H}}, \quad i, j = 1, \dots, r.$$

Introducing the Gram kernel

$$H_{i,j} = \langle \widehat{\mathcal{G}}\widehat{\phi}(X_{n+i}), \widehat{\phi}(X_{n+j})\rangle_{\mathcal{H}}, \quad i,j = 1,\dots,n,$$

we see, from the definition of q_i, that

$$G = VHV^\top. \tag{4.19}$$

We then use the polarization formula to get

$$G_{i,j} = \frac{1}{4}\Big[\langle \widehat{\mathcal{G}}(q_i+q_j), q_i+q_j\rangle_{\mathcal{H}} - \langle \widehat{\mathcal{G}}(q_i-q_j), q_i-q_j\rangle_{\mathcal{H}}\Big], \quad i,j = 1,\dots,r.$$

Given a vector $p = (p_1,\dots,p_n)$ of real values (that will be of the form $\langle \theta, \widehat{\phi}(X_{n+i})\rangle_{\mathcal{H}}$) and a scale parameter $\lambda > 0$ (that we choose uniform here for simplicity), we consider the solution $S(p,\lambda)$ of the equation

$$\sum_{i=1}^{n} \psi\Big[\lambda\Big(S(p,\lambda)^{-1}p_i^2 - 1\Big)\Big] = 0, \tag{4.20}$$

where the function ψ is defined in equation (1.2) on page 16. In practice we use a few iterations of the Newton algorithm to solve this equation. We estimate

$$\begin{aligned}\langle \widehat{\mathcal{G}}(q_i+\sigma q_j), q_i+\sigma q_j\rangle_{\mathcal{H}} &= \int \langle q_i+\sigma q_j, \widehat{\phi}(x)\rangle_{\mathcal{H}}^2 \,\mathrm{dP}(x)\\ &= \int \Big(\sum_{k=1}^{n}(V_{i,k}+\sigma V_{j,k})\langle \widehat{\phi}(X_{n+k}), \widehat{\phi}(x)\rangle_{\mathcal{H}}\Big)^2 \mathrm{dP}(x),\end{aligned}$$

where $\sigma \in \{-1,+1\}$, by $\widehat{N}(q_i+\sigma q_j)$, which we approximate by

$$S\left[\left(\sum_{k=1}^{n}(V_{i,k}+\sigma V_{j,k})M_{k,\ell}, 1 \le \ell \le n\right), \lambda\right]. \tag{4.21}$$

We then remark that for any couple of indices (i,ℓ), such that $1 \le i \le r$ and $1 \le \ell \le n$,

$$\sum_{k=1}^{n} V_{i,k}M_{k,\ell} = \big(D^{-1/2}U^\top M\big)_{i\ell} = \big(D^{-1/2}U^\top UDU^\top\big)_{i,\ell} = \big(D^{1/2}U^\top\big)_{i,\ell},$$

so that equation (4.21) becomes

$$S\left[\left(\big(D^{1/2}U^\top\big)_{i,\ell} + \sigma\big(D^{1/2}U^\top\big)_{j,\ell}, 1 \le \ell \le n\right), \lambda\right].$$

Taking notations, for any $r \times n$ matrix W, we denote by $C(W)$ the $r \times r$ matrix of entries

$$C(W)_{i,j} = \frac{1}{4}\left[S\left(\Big(W_{i,\ell}+W_{j,\ell},\ 1 \le \ell \le n\Big), \lambda\right) - S\left(\Big(W_{i,\ell}-W_{j,\ell},\ 1 \le \ell \le n\Big), \lambda\right)\right].$$

We define a first estimate Q_0 of the $r \times r$ matrix G of entries $G_{i,j} = \langle \widehat{\mathcal{G}}(q_i), q_j\rangle$ by

$$Q_0 = C\Big(\Pi_{r,n}D^{1/2}U^\top\Big),$$

where $\Pi_{r,n}$ is the $r \times n$ matrix of the projection on the r first coordinates, that is $(\Pi_{r,n})_{i,j} = \delta_{i,j}$, for $1 \leq i \leq r$ and $1 \leq j \leq n$. We then start the iterative scheme. We decompose

$$Q_0 = O_0 D_0 O_0^\top,$$

where $O_0 O_0^\top = O_0^\top O_0 = \mathrm{I}$ and D_0 is a diagonal matrix, and we observe that D_0 is an estimator of

$$O_0^\top G O_0 = O_0^\top V H V^\top O_0,$$

according to equation (4.19), which can be re-estimated by $C\Big(O_0^\top \Pi_{r,n} D^{1/2} U^\top\Big)$, since, by definition, $V = \Pi_{r,n} D^{1/2} U^\top$. This yields a new estimator of $G = O_0 O_0^\top G O_0 O_0^\top$ equal to

$$Q_1 = O_0 C\Big(O_0^\top \Pi_{r,n} D^{1/2} U^\top\Big) O_0^\top.$$

The inductive update step is more generally the following. We decompose

$$Q_k = O_k D_k O_k^\top,$$

where $O_k O_k^\top = O_k^\top O_k = \mathrm{I}$ and D_k is a diagonal matrix, and we define the new updated estimator of G as

$$Q_{k+1} = O_k \; C\Big(O_k^\top \Pi_{r,n} D^{1/2} U^\top\Big) \; O_k^\top.$$

We stop this iterative estimation scheme when $\|Q_k - Q_{k-1}\|_F$ falls under a suitable threshold and we take as our robust estimator of $\langle \widehat{\mathcal{G}}^m(q_i), q_j \rangle_{\mathcal{H}}$ the last update $(Q_k^m)_{ij}$.

The choice of the number of iterations is done automatically. We denote by $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_r$ the eigenvalues of Q_k. Let p be the maximum number of classes we consider. As it can be seen in the simulations, the choice of p is robust, meaning that p can be harmlessly overestimated. We choose the number m of iterations by solving

$$\left(\frac{\widehat{\lambda}_p}{\widehat{\lambda}_1}\right)^m \simeq \zeta$$

where $\zeta > 0$ is a given small parameter. The choice of ζ is also robust and $1/100$ is a reasonable value for it.

Once we have estimated $\langle \widehat{\mathcal{G}}^m(q_i), q_j \rangle_{\mathcal{H}}$ by $(Q_k^m)_{i,j}$, the matrix

$$\bar{H}_{2m}(X_{n+i}, X_{n+j}) = \langle \widehat{\mathcal{G}}^{2m-1} \widehat{\phi}(X_{n+i}), \widehat{\phi}(X_{n+j}) \rangle_{\mathcal{H}}, \quad i, j = 1, \ldots, n,$$

is approximated by

$$\begin{aligned} \widetilde{H} &= U D^{1/2} \Pi_{r,n}^\top Q_k^{2m-1} \Pi_{r,n} D^{1/2} U^\top \\ &= U D^{1/2} \Pi_{r,n}^\top O_k D_k^{2m-1} O_k^\top \Pi_{r,n} D^{1/2} U^\top, \end{aligned}$$

according to equation (4.18).

As a last step we renormalize $\widetilde{H}$ to form the new kernel

$$\widehat{C}_{i,j} = \big(\widetilde{H}_{i,i}\big)^{-1/2} \widetilde{H}_{i,j} \big(\widetilde{H}_{j,j}\big)^{-1/2}, \quad i, j = 1, \ldots, n. \tag{4.22}$$

This renormalization can also be interpreted as the projection on the sphere of the embedded sample, or equivalently the kernel $\widehat{C}$ can be seen as the correlation matrix of the embedded dataset.

A further (optional) step is to center the representation by computing

$$\overline{C} = \widehat{C} - n^{-1}\mathbb{1}_{n,n}\widehat{C} - n^{-1}\widehat{C}\mathbb{1}_{n,n} + n^{-2}\mathbb{1}_{n,n}\widehat{C}\mathbb{1}_{n,n}$$

where $\mathbb{1}_{n,n}$ is the $n \times n$ matrix whose entries are all equal to one, and by normalizing once again to form

$$\widetilde{C}_{i,j} = \left(\overline{C}_{i,i}\right)^{-1/2} \overline{C}_{i,j} \left(\overline{C}_{j,j}\right)^{-1/2}, \quad i,j = 1,\ldots,n. \tag{4.23}$$

From the kernel $\widehat{C}$ (or as an alternative from $\widetilde{C}$), we derive the new classification

$$\widehat{C}_i = \Big\{ j \mid \widehat{C}_{i,j} \geq s \Big\},$$

where s is a threshold that we took equal to 0.1 in all our experiments.

This may not be symmetric or transitive, but this is not a problem in practice. We recall that our (greedy) classification algorithm consists in taking an index $i_1 \in \{1,\ldots,n\}$ at random, form the corresponding class $\widehat{C}_{i_1}$ and then start again with the remaining indices. In such a way, denoting by c the number of classes, we construct a sequence of indices $i_1,\ldots,i_c$, and we define the k-th class as

$$\widetilde{C}_k = \widehat{C}_{i_k} \setminus \bigcup_{\ell<k} \widehat{C}_{i_\ell}.$$

It is worth underlying that in our method the number of clusters is estimated automatically, while in the spectral clustering algorithms mentioned above, [32], [20], [22], it is assumed to be known in advance.

4.4 Empirical results

In this section we give some examples. We first show how the algorithm described in the previous section groups the points to cluster at the vertices of a simplex, simplifying the geometry of the classes. Then, we test the algorithm in the setting of image analysis.

4.4.1 A first example

We briefly describe the setting in which we test the algorithm introduced in the previous section. We consider an i.i.d. set of $n = 900$ points to cluster, whose configuration is shown in figure 4.1 and we fix the maximum number of classes $p = 7$.

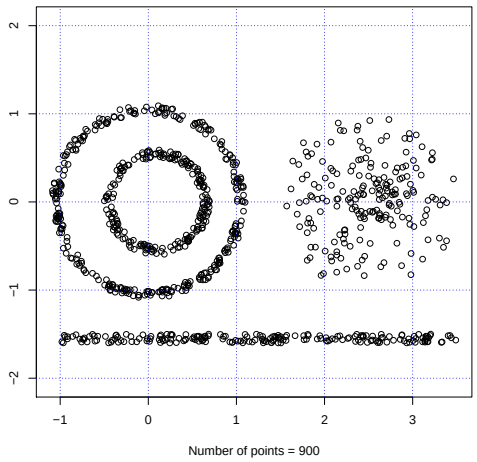

Figure 4.1: The data configuration.

In these experiments, we have not used a split sample scheme and we have recycled the same sample twice to compute the normalized kernel as well as to estimate the Gram operator. This means that in the implementation of the algorithm described in section 4.3.1, we took $X_{n+i} = X_i$, for any $i = 1, \ldots, n$. With the same notation, the matrix $M_{ij} = \widehat{K}(X_{n+i}, X_{n+j})$ is computed as

$$M = D^{-1/2} A D^{-1/2} \tag{4.24}$$

where A is the kernel matrix $A_{i,j} = n^{-1} \exp\left(-\beta \|X_i - X_j\|^2\right)$ and D is the diagonal matrix with entries

$$D_{i,i} = \max\left\{ \frac{1}{n} \sum_{j=1}^{n} A_{i,j}, \sigma \right\},$$

with $\sigma = 0.001$.

Figure 4.2 shows that the new representation induced by the change of kernel described in equation (4.23) groups the data points at the vertices of the simplex generated by the largest eigenvectors of the matrix $\widetilde{C}$. This simple configuration makes it possible to compute the classification, including the number of clusters, using a straightforward greedy algorithm.

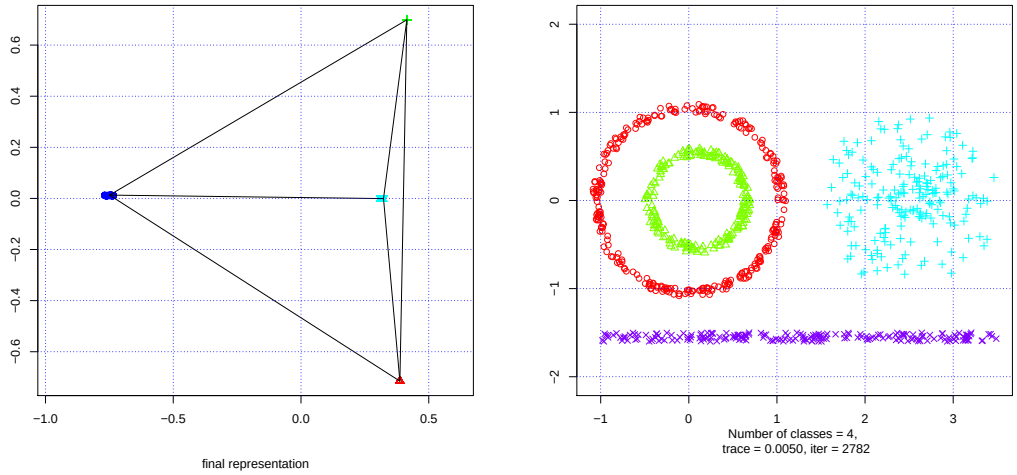

Figure 4.2: On the left the simplex generated by the eigenvectors of $\widetilde{C}$, on the right classification performed on $\widetilde{C}$.

In figure 4.3 we plotted the first eigenvalues of M in black (joined by a solid line), the eigenvalues of its iteration $\langle \widehat{\mathcal{G}}^{2m-1}\widehat{\phi}(X_{n+i}), \widehat{\phi}(X_{n+j})\rangle_{\mathcal{H}}$ estimated by $\widetilde{H}_{i,j}$ in blue (joined by a dashed line) and the eigenvalues of the covariance matrix of the final representation (defined by $\widetilde{C}$) in red (joined by a dash dotted line). We observe that the first eigenvalues of M are close to one, while there is a remarkable gap between the eigenvalues of its iteration. In particular the size of the gap is larger once we have renormalized using the matrix $\widetilde{C}$.

The number of iterations is automatically estimated and it is equal to 1441.

We obtain similar results by considering the matrix $\widehat{C}$ instead of $\widetilde{C}$.

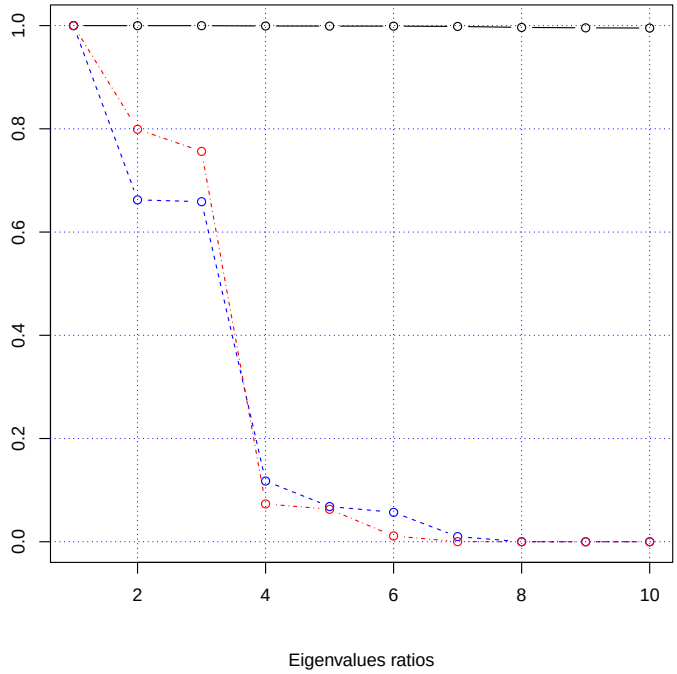

Figure 4.3: In black (solid line) the eigenvalues of M, in blue (dashed line) those of $\widetilde{H}$ and in red (dash-dot line) the eigenvalues of $\widetilde{C}$.

We can also test the simpler clustering algorithm obtained by estimating the Gram operator by its empirical counterpart. More precisely, instead of computing the robust estimator with the iterative scheme described in the previous section, we simply construct the matrix M as in equation (4.24) and we raise it to the power m to obtain the new matrix

$$\widetilde{M}_{i,j} = (M^m)_{i,i}^{-1/2}(M^m)_{i,j}(M^m)_{j,j}^{-1/2}.$$

We then apply the classification algorithm to the matrix $\widetilde{M}$.

Figure 4.4 shows that in this special case the use of a robust estimator is not needed. Indeed, in the representation defined by $\widetilde{M}$, the data points are also grouped at the vertices of the simplex generated by the first eigenvectors of $\widetilde{M}$ and the same classification is computed.

The number of iterations is automatically estimated and it is equal to 2090.

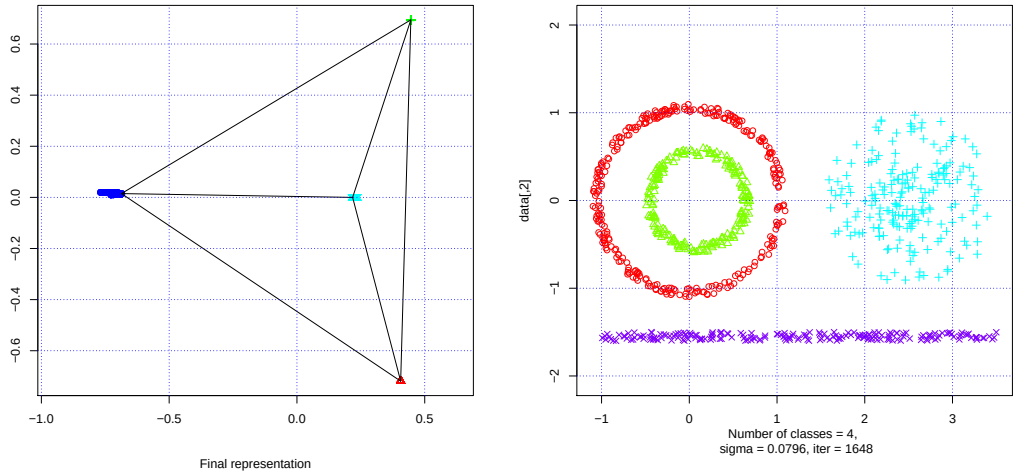

Figure 4.4: On the left the simplex generated by the eigenvectors of $\widetilde{M}$, on the right classification performed on $\widetilde{M}$.

4.4.2 Some perspectives on invariant shape analysis

In this section we present a small example about image classification. The goal of image analysis is to recognize and classify patterns. These patterns may have been transformed by some set of transformations, such as translations, rotations or scaling and are more generally affected by the conditions in which the images to be analysed have been taken. The challenge is then to find a representation that is not hampered by the variations of the same pattern from image to image. Making a representation invariant to a set of transformations is a challenging task, even in the simpler case of translations. Indeed, it is sufficient to observe that the set of functions obtained by translating a single pattern in various directions typically spans a vector space of high dimension, meaning that the shapes (here the functions) that we would like to put in the same category do not even live in a common low-dimensional subspace.

A possible approach is to study representations that leave invariant some group of transformations. For instance, the Fourier transform has translation invariant properties, since its modulus is translation invariant. However it is unstable to small deformations at high frequencies. Wavelet transforms provide a workaround. Scattering representations proposed by Mallat [18] compute translation invariant representations by cascading wavelet transforms and modulus pooling operators. They bring improvements for audio classification [1] and for image classification [3].

This kind of careful mathematical study has to be repeated for any kind of transformations and in image analysis the pattern transformations we would like to take into account are numerous and not easy to formalize, since they may come from changes in the perspective or in illumination, from partial occlusions, from object deformations, etc.

Coupling spectral clustering with a change of representation in a reproducing kernel Hilbert space may lead to a generic and rather radical alternative. Instead of deriving the representation of a pattern from a mathematical study, the idea is to learn the representation itself from examples of patterns that sample in a sufficiently dense way the orbits of the set of transformations at stake.

Developing a convincing toolbox for unsupervised invariant shape analysis is beyond the scope of this study and it will be carried on elsewhere. Our purpose in this section is just to hint that spectral clustering, coupled with some preliminary change of representation in a reproducing kernel Hilbert space, can be a winning tool to bring down the representation of classes to a low-dimensional space. We suggest with a small example how this approach may lead to learn transformation invariant representations from datasets containing small successive transformations of a same pattern.

We briefly describe this approach to get a hint of its potential. We consider two images (figure 4.5) and we create our patterns by translating a subwindow of a given size in each image repeatedly, using a translation vector smaller than the subwindow size. In such a way we create a sample consisting of two classes of connected images, as shown in figure 4.6. This notion of translation cannot be grasped easily by a mathematical definition, because we do not translate a function (here the image defined as a function sampled on a rectangular grid of pixels), but a window of observation. Hence in this case the translation depends on the image content and it may not be easy to model in any realistic situation.

We present on an example a successful scenario using first a change of representation in a reproducing kernel Hilbert space to better separate the two classes, and then spectral clustering to shrink each class to a tight blob. We show that the so called kernel trick, introduced to better separate classes in the supervised learning framework addressed by support vector machines (SVMs), also works in an unsupervised context. In this setting, we do not separate classes using hyperplanes, since we do not know class labels which would be necessary to run a SVM, but, instead, we use spectral clustering to finishing the work.

Figure 4.5: The two original images.

Figure 4.6: Our sample consisting of two classes of connected images, the first sequence is obtained with a horizontal translation, the second one with a diagonal translation.

Let $X_1, \ldots, X_n$ be the sample of images shown in figure 4.6. Each photo is represented

as a matrix whose entries are the gray values of the corresponding pixels. We apply twice the change of representation described by the change of kernel. We first consider the reproducing kernel Hilbert space $\mathcal{H}_1$ defined by

$$k_1(x,y) = \exp\left(-\beta_1\|x-y\|^2\right)$$

and then the reproducing kernel Hilbert space $\mathcal{H}_2$ defined by

$$k_2(x,y) = \exp\left(-\beta_2\|x-y\|^2_{\mathcal{H}_1}\right) = \exp\Big(-2\beta_2(1-k_1(x,y))\Big).$$

Recalling the definition of k_1, the kernel k_2 rewrites as

$$k_2(x,y) = \exp\left[-2\beta_2\left(1-\exp\left(-\beta_1\|x-y\|^2\right)\right)\right]$$

where $\beta_1, \beta_2 > 0$ are obtained as described in section 4.2.2, based on the trace estimation. We define the new kernel

$$K(x,y) = \exp\left(-\beta\|x-y\|^2_{\mathcal{H}_2}\right),$$

where the parameter $\beta > 0$ is chosen again as in section 4.2.2, and we introduce

$$\bar{K}(x,y) = \frac{K(x,y)}{\left(\int K(x,z)\ \mathrm{dP}(z)\right)^{\frac{1}{2}}\left(\int K(y,z)\ \mathrm{dP}(z)\right)^{\frac{1}{2}}}.$$

Proceeding as already done in the previous sections, we estimate $\bar{K}$ with the kernel $\widehat{K}$ and we apply the classification algorithm to $\widehat{K}^m$. As in the framework of SVMs, we use the kernel trick to embed the sample in a higher-dimensional space in which the geometry of the classes is simpler. However, as already said, we do not use hyperplanes but spectral clustering to separate the two classes.

In figure 4.7 we compare the representation of the images in the initial space and in the space $\mathcal{H}_2$.

On the left we present the projection of the sample onto the space spanned by the first two largest eigenvectors of the matrix of inner products between images $\langle X_i, X_j\rangle$. On the right we plot the projection onto the space spanned by the two largest eigenvectors of the matrix of inner products $k_2(X_i, X_j)$ in $\mathcal{H}_2$.

We observe that in the first representation the two classes intersect each other while in the second one, after the change of representation, they are already separated.

To conclude, figure 4.8 shows the final representation. Here the data points are projected onto the space spanned by the two largest eigenvectors of the matrix $\widehat{K}(X_i, X_j)^m$.

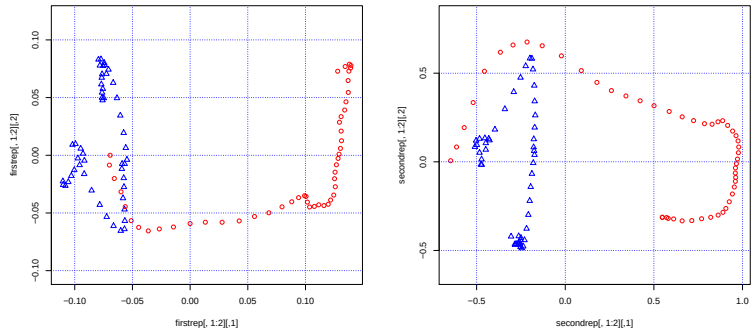

Figure 4.7: On the left the projection onto the space spanned by the two largest eigenvectors of $\langle X_i, X_j \rangle$, on the right the projection onto the space spanned by the two largest eigenvectors of $k_2(X_i, X_j)$.

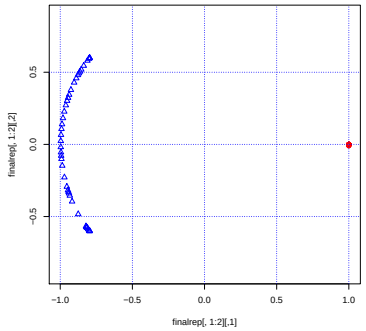

Figure 4.8: The projection onto space spanned by the two largest eigenvectors of $\widehat{K}(X_i, X_j)^m$.

Appendix A

Density Estimation

In this appendix we consider the problem of estimating the density function of an unknown probability distribution and we present an estimator, based on the results obtained in chapter 1, that is more robust than the classical kernel density estimator.

A.1 Introduction

Let $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$ be an unknown probability measure and let us assume it has a density with respect to the Lebesgue measure. The goal of density estimation is to construct an estimator of the density function f from an i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ distributed according to P. To achieve this aim, we need to introduce the concept of *approximate identities* (also called approximations to the identity). Let φ be a function defined on $\mathbb{R}^d$ and let $\epsilon > 0$. We assume that φ is a non-negative measurable function such that

$$\int \varphi(x)\mathrm{d}x = 1.$$

Let us denote by $L^p(\mathbb{R}^d)$ the space of p-integrable functions with respect to the Lebesgue measure.

Proposition A.1. (Theorem 1.18. [33]) *For any $f \in L^p(\mathbb{R}^d)$, $p \in [1, +\infty[$, we have*

$$\lim_{\epsilon \to 0} \|f * \varphi_\epsilon - f\|_{L^p(\mathbb{R}^d)} = 0,$$

where $\varphi_\epsilon(x) = \epsilon^{-d}\varphi(x/\epsilon)$ are called approximate identities. Moreover, for any $f \in L^p(\mathbb{R}^d)$, $p \in [1, +\infty]$,

$$\lim_{\epsilon \to 0} f * \varphi_\epsilon(x) = f(x) \quad \textit{almost surely.}$$

As a consequence, assuming $f \in L^p(\mathbb{R}^d)$, the convolution

$$\begin{aligned} f * \varphi_\epsilon(x) &= \int \varphi_\epsilon(z - x) f(z)\ \mathrm{d}z \\ &= \int \varphi_\epsilon(z - x)\ \mathrm{dP}(z) \end{aligned}$$

converges to f almost surely as ϵ goes to zero.

We now consider as approximate identities the family of Gaussian distributions $\pi_\alpha \sim \mathcal{N}(0, \alpha^{-1}\mathrm{I})$ of means 0 and covariance matrix $\alpha^{-1}\mathrm{I}$. By the properties of approximate

identities, the convolution

$$f * \pi_\alpha(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \int \exp\left(-\frac{\alpha}{2}\|z-x\|^2\right) \, \mathrm{dP}(z)$$

converges to f as α grows to infinity. However, since P is unknown, also $f * \pi_\alpha$ is unknown. A possible approach to estimate $f * \pi_\alpha$ is to replace the distribution P by the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. This leads to the kernel density estimator

$$\bar{f}_\alpha(x) = \frac{1}{n}\left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \sum_{i=1}^n \exp\left(-\frac{\alpha}{2}\|X_i - x\|^2\right),$$

where $1/\sqrt{\alpha}$ is usually called *bandwidth*. We discuss the use of more general kernels in appendix A.3.

In the next section we construct, with the help of the results presented in chapter 1, a robust estimator $\widehat{f}_\alpha$ of the convolution $f * \pi_\alpha$.

A.2 Estimate of the density function

In this section, our goal is construct a robust estimator of

$$f * \pi_\alpha(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \int \exp\left(-\frac{\alpha}{2}\|z-x\|^2\right) \, \mathrm{dP}(z), \quad x \in \mathbb{R}^d,$$

from the i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to P. We introduce the kernel

$$k_\alpha(x, x') = \exp\left(-\frac{\alpha}{4}\|x' - x\|^2\right), \quad x, x' \in \mathbb{R}^d,$$

and, according to the Moore-Aronszajn theorem (Proposition C.14), it defines a reproducing kernel Hilbert space $\mathcal{H}$ and a feature map $\phi_\alpha : \mathbb{R}^d \to \mathcal{H}$ such that

$$k_\alpha(x, x') = \langle \phi_\alpha(x'), \phi_\alpha(x) \rangle_{\mathcal{H}}.$$

It follows that the convolution rewrites as

$$f * \pi_\alpha(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \int \langle \phi_\alpha(z), \phi_\alpha(x) \rangle_{\mathcal{H}}^2 \, \mathrm{dP}(z).$$

In analogy to the notation of chapter 1, we introduce

$$N_\alpha(\phi_\alpha(x)) = \int \langle \phi_\alpha(z), \phi_\alpha(x) \rangle_{\mathcal{H}}^2 \, \mathrm{dP}(z),$$

so that

$$f * \pi_\alpha(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \; N_\alpha(\phi_\alpha(x)).$$

Let $\widehat{N}_\alpha$ be the estimator of N_α introduced in section 1.3 on page 46. We define

$$\widehat{f}_\alpha(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \; \widehat{N}_\alpha(\phi_\alpha(x)).$$

Next proposition provides a uniform bound on the approximation error $\left| f * \pi_\alpha(x) - \widehat{f}_\alpha(x) \right|$.

Proposition A.2. *Let $X \in \mathbb{R}^d$ be a random vector of law* P. *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$. Let $a > 0$ and let*

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}} \right) \right\rceil$$

with $c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2}\right)$ *and*

$$\kappa = \sup_{x \in \mathcal{X}} \frac{\mathbb{E}\Big[\exp\Big(-\alpha\|x - X\|^2\Big)\Big]}{\mathbb{E}\Big[\exp\Big(-\frac{\alpha}{2}\|x - X\|^2\Big)\Big]^2}.$$

Define

$$\zeta(t) \;=\; \sqrt{2(\kappa-1)\left(\frac{(2+3c)}{4(2+c)\kappa^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) \;+\; \sqrt{\frac{2(2+c)\kappa^{1/2}}{t}}\cosh(a/2)$$

and the bound

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})} & [6 + (\kappa-1)^{-1}]\zeta(\max\{t,\sigma\}) \le \sqrt{n} \\ +\infty & \textit{otherwise,} \end{cases}$$

depending on the threshold σ. For any choice of $\sigma \in]0,1]$, with probability at least $1 - 2\epsilon$, for any $x \in \mathcal{X}$,

$$\left| \frac{\max\{f * \pi_\alpha(x), \sigma\big(\alpha/(2\pi)\big)^{d/2}\}}{\max\{\widehat{f}_\alpha(x), \sigma\big(\alpha/(2\pi)\big)^{d/2}\}} - 1 \right| \le B_*\Big\{\mathbb{E}\Big[\exp\Big(-\frac{\alpha}{2}\|x - X\|^2\Big)\Big]\Big\}, \tag{A.1}$$

and consequently

$$\Big| f * \pi_\alpha(x) - \widehat{f}_\alpha(x) \Big| \le B\Big\{\mathbb{E}\Big[\exp\Big(-\frac{\alpha}{2}\|x - X\|^2\Big)\Big]\Big\},$$

where

$$B(t) = \left(\frac{\alpha}{2\pi}\right)^{d/2}\left(\max\{t,\sigma\}\frac{B_*(t)}{1 - B_*(t)} + \sigma\right).$$

Proof. Remark that $s_4^2 = \mathbb{E}(\|\phi_\alpha(X)\|^4)^{1/2} = 1$, since $\|\phi_\alpha(x)\|_{\mathcal{H}} = 1$, for any $x \in \mathbb{R}^d$. To prove the result, it is sufficient to bound, uniformly on $x \in \mathcal{X}$, the quantity

$$\left| \frac{\max\{N_\alpha(\phi_\alpha(x)), \sigma\}}{\max\{\widehat{N}_\alpha(\phi_\alpha(x)), \sigma\}} - 1 \right|.$$

We can then remark that if in Proposition 1.17 on page 35 and consequently in Proposition 1.31 on page 47 we restrict the statement to any $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is some subset of $\mathbb{R}^d$, then, in the definition of κ (equation (1.10) on page 26), we can also restrict the supremum to the same subset. Using this variant of Proposition 1.31 on page 47 where we restrict to $\theta \in \phi_\alpha(\mathcal{X})$, we obtain that with probability at least $1 - 2\epsilon$, for any $x \in \mathcal{X}$,

$$\left| \frac{\max\{N_\alpha(\phi_\alpha(x)), \sigma\}}{\max\{\widehat{N}_\alpha(\phi_\alpha(x)), \sigma\}} - 1 \right| \le B_*[N_\alpha(\phi_\alpha(x))],$$

which concludes the proof. □

Let us remark that we can see the influence of the dimension in this result by looking at equivalents when α goes to $+\infty$. In this case (under suitable regularity assumptions)

$$\mathbb{E}\Big[\exp\Big(-\frac{\alpha}{2}\|x - X\|^2\Big)\Big] \sim \left(\frac{2\pi}{\alpha}\right)^{d/2} f(x),$$

so that

$$\kappa \sim \left(\frac{\alpha}{4\pi}\right)^{d/2} \sup_{x\in\mathcal{X}} f(x)^{-1}.$$

A.3 Kernel density estimation

Kernel density estimation is a non-parametric method of estimating the probability density function of a continuous random variable which consists in associating to each point of a given dataset a kernel function centered on that point. The kernel density estimator is defined as the sum of the centered kernel functions, scaled by a parameter h, called bandwidth, to have unit area. To be more precise, given an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ and a kernel $\mathbf{K}$ such that $\mathbf{K}(x, y) = \mathbf{k}(x - y)$, the kernel density estimator has the form

$$\frac{h^{-d}}{n} \sum_{i=1}^{n} \mathbf{k}\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}^d.$$

In the previous section we have considered the case of a Gaussian kernel with bandwidth $1/\sqrt{\alpha}$. We now observe that, up to the scaling parameter, the kernel density estimator is the empirical version of

$$\int \mathbf{K}(x, z) \, \mathrm{dP}(z) = \int \mathbf{k}(x - z) \, \mathrm{dP}(z), \quad x \in \mathbb{R}^d. \tag{A.2}$$

In this section we provide a more robust estimator of the quantity in equation (A.2). Let $\mathcal{X}$ be a subset of some separable Hilbert space endowed with the unknown probability distribution $\mathrm{P} \in \mathcal{M}_+^1(\mathcal{X})$. We assume that the kernel $\mathbf{K}$ is such that

$$\mathbf{K}(x, y) = \widetilde{\mathbf{K}}(x, y)^2, \quad x, y \in \mathcal{X},$$

where $\widetilde{\mathbf{K}}$ is a symmetric positive semi-definite kernel on $\mathcal{X}$. Our goal is to estimate

$$\int \mathbf{K}(x, z) \, \mathrm{dP}(z) = \int \widetilde{\mathbf{K}}(x, z)^2 \, \mathrm{dP}(z) \tag{A.3}$$

from an i.i.d. sample $X_1, \ldots, X_n \in \mathcal{X}$ drawn according to P. By the Moore-Aronszajn theorem (Proposition C.14) $\widetilde{\mathbf{K}}$ defines a reproducing kernel Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$\widetilde{\mathbf{K}}(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{H}}.$$

Thus, the quantity in equation (A.3) rewrites as

$$\int \langle \phi(x), \phi(z) \rangle_{\mathcal{H}}^2 \, \mathrm{dP}(z),$$

which we denote by $N(\phi(x))$ in analogy to the notation used in chapter 1. We can see that N is the quadratic form associated with the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{G}\theta = \int \langle \theta, \phi(z) \rangle_{\mathcal{H}} \, \phi(z) \, \mathrm{dP}(z).$$

Let $\widehat{N}(\theta)$ be the estimator of the Gram operator defined in Proposition 1.31 on page 47, with θ restricted to be in $\phi(\mathcal{X})$. We can estimate

$$N(\phi(x)) = \int \mathbf{K}(x,z)\ \mathrm{dP}(z)$$

by $\widehat{N}\big(\phi(x)\big)$ with the following bound on the estimation error.

Proposition A.3. *With the above notation, let $X \in \mathcal{X}$ be a random vector of law* P. *We define, for any $t \in \mathbb{R}_+$,*

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)\mathbb{E}\big(\mathbf{K}(X,X)^2\big)^{1/2}}{4(2+c)\kappa^{1/2}t} + \log(K/\epsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}\mathbb{E}\big(\mathbf{K}(X,X)^2\big)^{1/2}}{t}}\cosh(a/2)$$

where K and c are defined in Proposition A.2 and

$$\kappa = \sup_{\substack{x\in\mathcal{X},\\ \mathbb{E}\left(\mathbf{K}(x,X)\right)>0}} \frac{\mathbb{E}\big(\mathbf{K}(x,X)^2\big)}{\mathbb{E}\big(\mathbf{K}(x,X)\big)^2}$$

Let $\sigma \in \mathbb{R}_+$ be such that $8\zeta(\sigma) \le \sqrt{n}$ and $\sigma \le \mathbb{E}\big[\mathbf{K}(X,X)^2\big]^{1/2}$. Consider the bound

$$B_*(t) = \frac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1 - 4\,n^{-1/2}\zeta(\max\{t,\sigma\})}, \quad t \in \mathbb{R}_+.$$

With probability at least $1-2\epsilon$, for any $x \in \mathcal{X}$,

$$\left|\frac{\max\{\mathbb{E}\big(\mathbf{K}(x,X)\big),\sigma\}}{\max\{\widehat{N}\big(\phi(x)\big),\sigma\}} - 1\right| \le B_*\big(\mathbb{E}\big(\mathbf{K}(x,X)\big)\big).$$

Appendix B

Orthogonal Projectors

In this appendix we introduce some results on orthogonal projectors.

Let $P,\ Q : \mathbb{R}^d \to \mathbb{R}^d$ be two orthogonal projectors. We denote by $\mathbb{S}_d$ the unit sphere of $\mathbb{R}^d$. By definition,

$$\|P - Q\|_\infty = \sup_{x \in \mathbb{S}_d} \|Px - Qx\|$$

where, without loss of generality, we can take the supremum over the normalized eigenvectors of $P - Q$.

A good way to describe the geometry of $P - Q$ is to consider the eigenvectors of $P + Q$.

Lemma B.1. *Let $x \in \mathbb{S}_d$ be an eigenvector of $P + Q$ with eigenvalue λ.*

1. *In the case when $\lambda = 0$, then $Px = Qx = 0$, so that $x \in \mathbf{ker}(P) \cap \mathbf{ker}(Q)$;*
2. *in the case when $\lambda = 1$, then $PQx = QPx = 0$, so that*

$$x \in \mathbf{ker}(P) \cap \mathbf{Im}(Q) \oplus \mathbf{Im}(P) \cap \mathbf{ker}(Q);$$

3. *in the case when $\lambda = 2$, then $x = Px = Qx$, so that $x \in \mathbf{Im}(P) \cap \mathbf{Im}(Q)$;*
4. *otherwise $\lambda \in]0, 1[\cup]1, 2[$,*

$$(P - Q)^2 x = (2 - \lambda)\lambda x \neq 0,$$

so that $(P - Q)x \neq 0$. Moreover

$$(P + Q)(P - Q)x = (2 - \lambda)(P - Q)x,$$

so that $(P - Q)x$ is an eigenvector of $P + Q$ with eigenvalue $2 - \lambda$. Moreover

$$0 < \|Px\| = \|Qx\| < \|x\|,$$

$x - Px \neq 0$, and $\big(Px, x - Px\big)$ is an orthogonal basis of $\mathbf{span}\{x, (P - Q)x\}$.

Proof. The operator $P + Q$ is symmetric, positive semi-definite, and $\|P + Q\| \leq 2$, so that there is a basis of eigenvectors and all eigenvalues are in the intervall $[0, 2]$.

In case 1, $0 = \langle Px + Qx, x\rangle = \|Px\|^2 + \|Qx\|^2$, so that $Px = Qx = 0$.

In case 2, $PQx = P(x - Px) = 0$ and similarly $QPx = Q(x - Qx) = 0$.

In case 3,

$$\|Px\|^2 + \|Qx\|^2 = \langle (P + Q)x, x\rangle = 2\langle x, x\rangle = \|Px\|^2 + \|x - Px\|^2 + \|Qx\|^2 + \|x - Qx\|^2,$$

so that $\|x - Px\| = \|x - Qx\| = 0$.
In case 4, remark that

$$PQx = P(\lambda x - Px) = (\lambda - 1)Px$$

and similarly $QPx = Q(\lambda x - Qx) = (\lambda - 1)Qx$. Consequently

$$(P - Q)(P - Q)x = (P - QP - PQ + Q)x = (2 - \lambda)(P + Q)x = (2 - \lambda)\lambda x \neq 0,$$

so that $(P - Q)x \neq 0$. Moreover

$$(P + Q)(P - Q)x = (P - PQ + QP - Q)x = (2 - \lambda)(P - Q)x.$$

Therefore $(P - Q)x$ is an eigenvector of $P + Q$ with eigenvalue $2 - \lambda \neq \lambda$, so that $\langle x, (P - Q)x\rangle = 0$, since $P + Q$ is symmetric. As $\langle x, (P - Q)x\rangle = \|Px\|^2 - \|Qx\|^2$, this proves that $\|Px\| = \|Qx\|$. Since $(P + Q)x = \lambda x \neq 0$, necessarily $\|Px\| = \|Qx\| > 0$. Observe now that

$$\|Px\|^2 = \frac{1}{2}(\|Px\|^2 + \|Qx\|^2) = \frac{1}{2}\langle x, (P + Q)x\rangle = \frac{\lambda}{2}\|x\|^2 < \|x\|^2.$$

Therefore $\|x - Px\|^2 = \|x\|^2 - \|Px\|^2 > 0$, proving that $x - Px \neq 0$. Similarly, since P and Q play symmetric roles, $\|Qx\| < \|x\|$ and $x - Qx \neq 0$.
As P is an orthogonal projector, $(Px, x - Px)$ is an orthogonal pair of non-zero vectors. Moreover

$$x = x - Px + Px \in \mathbf{span}\{Px, x - Px\}$$

and

$$(P - Q)x = 2Px - \lambda x = (2 - \lambda)Px - \lambda(x - Px) \in \mathbf{span}\{Px, x - Px\}$$

therefore, $(Px, x - Px)$ is an orthogonal basis of $\mathbf{span}\{x, (P - Q)x\}$. □

Lemma B.2. *There is an orthonormal basis $(x_i)_{i=1}^d$ of eigenvectors of $P + Q$ with corresponding eigenvalues $\{\lambda_i,\ i = 1, \dots d\}$ and indices $2m \leq p \leq q \leq s$, such that*

1. $\lambda_i \in]1, 2[$, *if* $1 \leq i \leq m$,
2. $\lambda_{m+i} = 2 - \lambda_i$, *if* $1 \leq i \leq m$, *and* $x_{m+i} = \|(P - Q)x_i\|^{-1}(P - Q)x_i$,
3. $x_{2m+1}, \dots, x_p \in \big(\mathbf{Im}(P) \cap \mathbf{ker}(Q)\big)$, *and* $\lambda_{2m+1} = \cdots = \lambda_p = 1$,
4. $x_{p+1}, \dots, x_q \in \big(\mathbf{Im}(Q) \cap \mathbf{ker}(P)\big)$, *and* $\lambda_{p+1} = \cdots = \lambda_q = 1$,
5. $x_{q+1}, \dots, x_s \in \mathbf{Im}(P) \cap \mathbf{Im}(Q)$, *and* $\lambda_{q+1} = \cdots = \lambda_s = 2$,
6. $x_{s+1}, \dots, x_d \in \mathbf{ker}(P) \cap \mathbf{ker}(Q)$, *and* $\lambda_{s+1} = \cdots = \lambda_d = 0$.

Proof. There exists a basis of eigenvectors of $P + Q$ (as already explained at the beginning of proof of Lemma B.1). From the previous lemma, we learn that all eigenvectors in the kernel of $P + Q$ are in $\mathbf{ker}(P) \cap \mathbf{ker}(Q)$, as on the other hand obviously $\mathbf{ker}(P) \cap \mathbf{ker}(Q) \subset \mathbf{ker}(P + Q)$ we get that

$$\mathbf{ker}(P + Q) = \mathbf{ker}(P) \cap \mathbf{ker}(Q).$$

In the same way the previous lemma proves that the eigenspace corresponding to the eigenvalue 2 is equal to $\mathbf{Im}(P) \cap \mathbf{Im}(Q)$. It also proves that the eigenspace correponding to the eigenvalue 1 is included in and consequently is equal to $\big(\mathbf{Im}(P) \cap \mathbf{ker}(Q)\big) \oplus \big(\mathbf{ker}(P) \cap \mathbf{Im}(Q)\big)$, so that we can form an orthonormal basis of this eigenspace by taking the union

of an orthonormal basis of $\mathbf{Im}(P) \cap \mathbf{ker}(Q)$ and an orthonormal basis of $\mathbf{ker}(P) \cap \mathbf{Im}(Q)$. Consider now an eigenspace corresponding to an eigenvalue $\lambda \in]0,1[\cup]1,2[$ and let x, y be two orthonormal eigenvectors in this eigenspace. Remark that (still from the previous lemma)

$$\langle (P-Q)x, (P-Q)y\rangle = \langle (P-Q)^2 x, y\rangle = (2-\lambda)\lambda\langle x, y\rangle = 0.$$

Therefore, if $x_1, \ldots, x_k$ is an orthonormal basis of the eigenspace V_λ corresponding to the eigenvalue λ, then $(P-Q)x_1, \ldots, (P-Q)x_k$ is an orthogonal system in $V_{2-\lambda}$. If this system was not spanning $V_{2-\lambda}$, we could add to it an orthogonal unit vector $y_{k+1} \in V_{2-\lambda}$ so that $x_1, \ldots, x_k, (P-Q)y_{k+1}$ would be an orthogonal set of non-zero vectors in V_λ, which would contradict the fact that $x_1, \ldots, x_k$ was supposed to be an orthonormal basis of V_λ. Therefore,

$$\Big(\|(P-Q)x_i\|^{-1}(P-Q)x_i,\ 1 \le i \le k \Big)$$

is an orthonormal basis of $V_{2-\lambda}$. Doing this construction for all the eigenspaces V_λ such that $\lambda \in]0,1[$ achieves the construction of the orthonormal basis described in the lemma. □

Lemma B.3. *Consider the orthonormal basis of the previous lemma. The set of vectors*

$$(Px_1, \ldots, Px_m, x_{2m+1}, \ldots, x_p, x_{q+1}, \ldots, x_s)$$

is an orthogonal basis of $\mathbf{Im}(P)$*. The set of vectors*

$$(Qx_1, \ldots, Qx_m, x_{p+1}, \ldots, x_q, x_{q+1}, \ldots, x_s)$$

is an orthogonal basis of $\mathbf{Im}(Q)$*.*

Proof. According to Lemma B.1 on page 125, $(Px_i, x_i - Px_i)$ is an orthogonal basis of $\mathbf{span}\{x_i, x_{m+i}\}$, so that

$$(Px_1, \ldots, Px_m, x_1 - Px_1, \ldots, x_m - Px_m, x_{2m+1}, \ldots, x_d)$$

is another orthogonal basis of $\mathbb{R}^d$. Each vector of this basis is either in $\mathbf{Im}(P)$ or in $\mathbf{ker}(P)$ and more precisely

$$\begin{aligned} Px_1, \ldots, Px_m, x_{2m+1}, \ldots, x_p, x_{q+1}, \ldots, x_s &\in \mathbf{Im}(P), \\ x_1 - Px_1, \ldots, x_m - Px_m, x_{p+1}, \ldots, x_q, x_{s+1}, \ldots, x_d &\in \mathbf{ker}(P). \end{aligned}$$

This proves the claim of the lemma concerning P. Since P and Q play symmetric roles, this proves also the claim concerning Q, *mutatis mutandis*. □

Lemma B.4. *The projectors P and Q have the same rank if and only if*

$$p - 2m = q - p.$$

Lemma B.5. *Assume that* $\mathbf{rank}(P) = \mathbf{rank}(Q)$*. Then*

$$\|P-Q\|_\infty = \sup_{\theta \in \mathbf{Im}(Q) \cap \mathbb{S}_d} \|(P-Q)\theta\|.$$

Proof. As $P-Q$ is a symmetric operator, we have

$$\begin{aligned} \sup_{\theta \in \mathbb{S}_d} \|(P-Q)\theta\|^2 &= \sup\Big\{ \langle (P-Q)^2\theta, \theta\rangle \mid \theta \in \mathbb{S}_d \Big\} \\ &= \sup\Big\{ \langle (P-Q)^2\theta, \theta\rangle \mid \theta \in \mathbb{S}_d \text{ is an eigenvector of } (P-Q)^2 \Big\}. \end{aligned}$$

Remark that the basis described in Lemma B.2 is also a basis of eigenvectors of $(P-Q)^2$. More precisely, according to Lemma B.1

$$\begin{aligned}
(P-Q)^2 x_i &= \lambda_i(2-\lambda_i)x_i, && 1 \le i \le m,\\
(P-Q)^2 x_{m+i} &= \lambda_i(2-\lambda_i)x_{m+i}, && 1 \le i \le m,\\
(P-Q)^2 x_i &= x_i, && 2m < i \le q,\\
(P-Q)^2 x_i &= 0, && q < i \le d.
\end{aligned}$$

If $q-2m>0$, then $\|P-Q\|_\infty = 1$, and $q-p>0$, according to Lemma B.4, so that $\|(P-Q)x_{p+1}\| = 1$, where $x_{p+1} \in \mathbf{Im}(Q)$. If $q=2m$ and $m>0$, there is $i \in \{1,\dots,m\}$ such that $\|P-Q\|_\infty^2 = \lambda_i(2-\lambda_i)$. Since x_i and x_{m+i} are two eigenvectors of $(P-Q)^2$ corresponding to this eigenvalue, all the non-zero vectors in $\mathbf{span}\{x_i, x_{m+i}\}$ (including Qx_i) are also eigenvectors of the same eigenspace. Consequently $(P-Q)^2 Qx_i = \lambda_i(2-\lambda_i)Qx_i$, proving that

$$\left\|(P-Q)\frac{Qx_i}{\|Qx_i\|}\right\|^2 = \lambda_i(2-\lambda_i),$$

and therefore that $\sup_{\theta\in\mathbb{S}_d}\|(P-Q)\theta\|$ is reached on $\mathbf{Im}(Q)$. Finally, if $q=0$, then $P-Q$ is the null operator, so that $\sup_{\theta\in\mathbb{S}_d}\|(P-Q)\theta\|$ is reached everywhere, including on $\mathbf{Im}(Q)\cap\mathbb{S}_d$. □

Appendix C

Reproducing Kernel Hilbert Spaces

C.1 Operators on Hilbert spaces

For more details on the results presented in this section we refer to [23].

Let $\mathcal{H}$ be a Hilbert space. We denote by $\mathcal{L}(\mathcal{H})$ the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$.

Definition C.1. Let X and Y be two Banach spaces. An operator $T \in \mathcal{L}(X, Y)$ is called *compact* if for very bounded sequence $\{x_n\} \subset X$, the sequence $\{Tx_n\}$ has a subsequence convergent in Y.

Proposition C.2. (Riesz-Schauder theorem, Theorem VI.15, [23]) *Let T a compact operator on $\mathcal{H}$. The spectrum of T is a discrete set having no limit points except perhaps zero. Moreover, any non-zero eigenvalue of T has finite multiplicity.*

An operator $T \in \mathcal{L}(\mathcal{H})$ is called *self-adjoint* if $T^* = T$, where $T^* \in \mathcal{L}(\mathcal{H})$ is defined by

$$\langle T^* x, y\rangle_{\mathcal{H}} = \langle x, Ty\rangle_{\mathcal{H}}.$$

Proposition C.3. (Hilbert-Schmidt theorem, Theorem VI.16, [23]) *Let T be a self-adjoint compact operator on $\mathcal{H}$. There exists an orthonormal basis $\{e_n\}_n$ of $\mathcal{H}$ such that $Te_n = \lambda_n e_n$ and $\lambda_n \to 0$ as $n \to +\infty$.*

The canonical form for compact operators is given by the following result.

Proposition C.4. (Theorem VI.17, [23]) *Let T be a compact operator on $\mathcal{H}$. There exist orthonormal sets (not necessarily complete) $\{\psi_n\}_n$, $\{\phi_n\}_n$ and positive real number $\{\lambda_n\}_n$ with $\lambda_n \to 0$ such that*

$$T = \sum_n \lambda_n \langle \psi_n, \cdot\rangle \phi_n.$$

The sum may be finite or infinite and converges in norm. The numbers $\{\lambda_n\}_n$ are called *singular values* of T and precisely they are the eigenvalues of $|T| = \sqrt{T^*T}$. The orthonormal set $\{\psi_n\}_n$ is such that

$$T^*T\psi_n = \lambda_n^2 \psi_n$$

and $\phi_n = T\psi_n/\lambda_n$. Hence, if the operator T is self-adjoint, $\phi_n = \psi_n$ for any n.

We are now going to introduce some criterium for determining when a given operator is compact.

Let $\mathcal{H}$ be a separable Hilbert space and let $\{e_n\}_n$ be an orthonormal basis of $\mathcal{H}$. For any positive operator $T \in \mathcal{L}(\mathcal{H})$, we define the *trace* of T as

$$\mathbf{Tr}(T) = \sum_n \langle Te_n, e_n \rangle_{\mathcal{H}}.$$

Definition C.5. An operator $T \in \mathcal{L}(\mathcal{H})$ is called *trace class* if

$$\mathbf{Tr}\,|T| = \sum_n \langle (T^*T)^{1/2} e_n, e_n \rangle_{\mathcal{H}} < +\infty.$$

Proposition C.6. (Theorem VI.21, [23]) *Every trace class operator is compact. Vice versa, a compact operator T is trace class if and only if $\sum_n \lambda_n < +\infty$ where λ_n are the singular values of T.*

Definition C.7. An operator $T \in \mathcal{L}(\mathcal{H})$ is called Hilbert-Schmidt if

$$\mathbf{Tr}(T^*T) < +\infty.$$

Proposition C.8. (Theorem VI.22, [23]) *Every Hilbert-Schmidt operator is compact. Vice versa, a compact operator T is Hilbert-Schmidt if and only if $\sum_n \lambda_n^2 < +\infty$ where λ_n are the singular values of T. Moreover,*

$$\|T\|_{HS} \leq \mathbf{Tr}\,|T|$$

where $\|T\|_{HS}^2 = \mathbf{Tr}(T^*T)$ is the Hilbert-Schmidt norm.

Proposition C.9. (Theorem VI.23, [23]) *Let (X, μ) be a measure space and $\mathcal{H} = L^2(X, \mu)$. An operator $T \in \mathcal{L}(\mathcal{H})$ is Hilbert-Schmidt if and only if there is a function $K \in L^2(X \times X, \mu \otimes \mu)$ with*

$$Tf(x) = \int K(x, z) f(z)\, \mathrm{d}\mu(z).$$

Combining Proposition C.4 and Proposition C.9 we obtain that the function K can be written as

$$K(x, y) = \sum_n \lambda_n \psi_n(x) \psi_n(y)$$

where ψ_n are the eigenfunctions of T and the sum converges in $L^2(X \times X, \mu \otimes \mu)$.

C.2 Definitions and main properties

Definition C.10. Let X be a set. A (real) *reproducing kernel Hilbert space* on X is a Hilbert space $\mathcal{K}$ such that

1. the element of $\mathcal{K}$ are functions defined on X;
2. (reproducing property) for all $x \in X$ there exists $K_x \in \mathcal{K}$ such that

$$f(x) = \langle f, K_x \rangle_{\mathcal{K}}, \quad f \in \mathcal{K}. \tag{C.1}$$

The reproducing property is equivalent to requiring that, for any $x \in X$, there exists a positive constant C_x such that

$$|f(x)| \leq C_x \|f\|_{\mathcal{K}}, \quad f \in \mathcal{K}.$$

The reproducing kernel corresponding to $\mathcal{K}$ is the function $K : X \times X \to \mathbb{R}$ defined by

$$K(x, x') = \langle K_{x'}, K_x \rangle_{\mathcal{K}}.$$

The kernel K is of *positive type*, that is, it is symmetric and for any $x_1, \ldots, x_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

We now present some example of reproducing kernels:

1) **Linear kernels**: $K(x, x') = x \cdot x'$

2) **Gaussian kernels**: $K(x, x') = e^{-\|x-x'\|^2/\sigma^2}, \quad \sigma > 0$

3) **Polynomial kernels**: $K(x, x') = (x \cdot x' + 1)^k, \quad k \in \mathbb{N}$.

Next result provides a characterization of reproducing kernel Hilbert spaces.

Lemma C.11. *Let $\mathcal{K}$ be a reproducing kernel Hilbert space. We have*

$$\mathcal{K} = \overline{\text{span}\{K_x \mid x \in X\}}.$$

Another way to characterize reproducing kernel Hilbert spaces is via the so called feature map.

Definition C.12. Let X be a set and $\mathcal{H}$ be a (real) Hilbert space, a *feature map* is a map $\phi : X \to \mathcal{H}$ such that

$$\text{if } v \in \mathcal{H} \text{ is such that } \langle v, \phi(x) \rangle_{\mathcal{H}} = 0 \ \forall x \in X, \text{ then } v = 0. \tag{C.2}$$

We observe that condition (C.2) is equivalent to

$$\overline{\text{span}\{\phi(x) \mid x \in X\}} = \mathcal{H}.$$

We now show that there is a one to one correspondence between reproducing kernel Hilbert spaces and feature maps. The first result shows that a feature map $\phi : X \to \mathcal{H}$ univocally determines a reproducing kernel Hilbert space. We define the set

$$\mathcal{K} = \{f_v \mid v \in \mathcal{H}\} \tag{C.3}$$

where $f_v : X \to \mathbb{R}$ is defined by

$$f_v(x) = \langle v, \phi(x) \rangle_{\mathcal{H}}.$$

The proposition below shows that $\mathcal{K}$ endowed with the product

$$\langle f_v, f_w \rangle_{\mathcal{K}} = \langle v, w \rangle_{\mathcal{H}}, \tag{C.4}$$

is a reproducing kernel Hilbert space.

Proposition C.13. *Given $\phi : X \to \mathcal{H}$ a feature map, the set $\mathcal{K}$ defined in equation* (C.3) *with the inner product* (C.4) *is a reproducing kernel Hilbert space with reproducing kernel*

$$K(x, x') = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}}.$$

Moreover, the map $W : \mathcal{H} \to \mathcal{K}$ defined by

$$Wv = f_v$$

is a unitary operator and in particular $K_x = f_{\phi(x)}$.

The space $\mathcal{K}$ is univocally determined by the kernel K, in the sense that, given $\phi' : X \to \mathcal{H}'$ another feature map such that

$$K(x, x') = \langle \phi'(x'), \phi'(x) \rangle_{\mathcal{H}},$$

then $\{f_{v'} \mid v' \in \mathcal{H}'\}$ and $\mathcal{K}$ coincides as Hilbert spaces.

Vice versa, any reproducing kernel Hilbert space $\mathcal{K}$ defines a feature map $\phi : X \to \mathcal{K}$ putting

$$\phi(x) = K_x, \quad x \in X,$$

whose feature operator is the identity, by the reproducing property.

We conclude the section stating the Moore-Aronszajn theorem.

Proposition C.14. (Moore-Aronszajn theorem) *Let $K : X \times X \to \mathbb{R}$ be a kernel of positive type. There exists a unique reproducing kernel Hilbert space with K as reproducing kernel.*

We then conclude that a kernel $K : X \times X \to \mathbb{R}$ is of positive type if and only if there exist a Hilbert space $\mathcal{H}$ and a function $\phi : X \to \mathcal{H}$ such that, for any $x, x' \in X$,

$$K(x, x') = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}}.$$

C.3 The Mercer theorem

In this section we introduce the Mercer theorem, which characterizes continuous kernels on compact domains.

Let $\phi : X \to \mathcal{H}$ be a feature map and $\mathcal{K}$ the corresponding reproducing kernel Hilbert space with reproducing kernel K. We are going to make the following assumptions:

1. X is a compact space which satisfies the second axiom of countability, i.e. its topology has a countable basis;
2. the reproducing kernel K is continuous;
3. $\mu \in \mathcal{M}_+^1(X)$ is a probability measure such that $\operatorname{supp}(\mu) = X$.

We consider the continuous operator $\mathcal{S} : \mathcal{K} \to L^2(X, \mu)$ defined by

$$\mathcal{S}f(x) = f(x) = \langle f, K_x \rangle_{\mathcal{K}}, \quad x \in X,$$

and we denote by $\mathcal{S}^* : L^2(X, \mu) \to \mathcal{K}$ its adjoint. In order to state some properties of these operators, we introduce the operator $K_x \otimes K_x : \mathcal{K} \to \mathcal{K}$ defined as

$$K_x \otimes K_x f = \langle f, K_x \rangle_{\mathcal{K}} K_x.$$

Proposition C.15. *The following properties hold*

1. *$\mathcal{S}$ and $\mathcal{S}^*$ are Hilbert-Schmidt operators.*
2. *$\mathcal{S}$ is injective*
3. *The operator $\mathcal{S}\mathcal{S}^* : L^2(X,\mu) \to L^2(X,\mu)$ is such that*

$$\mathcal{S}\mathcal{S}^* f(x) = \int K(x,z) f(z) \, \mathrm{d}\mu(z)$$

   and it is a positive trace class operator.

4. *$\mathcal{S}^*\mathcal{S}$ is the expectation with respect to μ of the operator $K_x \otimes K_x$, that is, it is such that*

$$\langle \mathcal{S}^*\mathcal{S} f, g\rangle_{\mathcal{K}} = \int \langle f, K_x\rangle_{\mathcal{K}} \langle g, K_x\rangle_{\mathcal{K}} \, \mathrm{d}\mu(x)$$

   and it is a positive trace class operator.

In particular, since $\mathcal{S}\mathcal{S}^*$ is a self-adjoint positive trace class operator, it is compact. Hence, by the Hilbert-Schmidt therorm (Proposition C.3), there exists an orthonormal basis $\{\varphi_i\}_i$ of $L^2(X,\mu)$ consisting of eigenvectors of $\mathcal{S}\mathcal{S}^*$ with positive eigenvalues, that is

$$\mathcal{S}\mathcal{S}^*\varphi_i = \sigma_i^2 \varphi_i, \quad \sigma_i \geq 0.$$

By the Riesz-Schauder theorem (Proposition C.2) every non-zero eigenvalue has finite multiplicity and 0 is (eventually) the only limit point. We introduce $I = \{i \mid \sigma_i > 0\}$ the set of indices corresponding to the non-zero eigenvalues and for any $i \in I$, we define

$$f_i = \frac{1}{\sigma_i} \mathcal{S}^* \varphi_i.$$

Proposition C.16. (Mercer theorem) *Let K be a continuous and of positive type kernel and let $\mathcal{K}$ be a reproducing kernel Hilbert space with reproducing kernel K. There exists a unique choice of the orthonormal basis $\{\varphi_i\}_i$ such that the eigenfunctions corresponding to non-zero eigenvalues are continuous. With this choice,*

$$f_i(x) = \sigma_i \varphi_i(x), \quad i \in I,$$

and $\{f_i\}_{i\in I}$ is an orthonormal basis of $\mathcal{K}$. Moreover, the kernel K has the representation

$$K(x,x') = \sum_i \sigma_i^2 \varphi_i(x)\varphi_i(x'), \quad x, x' \in X,$$

where the convergence is absolute and uniform.

Bibliography

[1] Joakim Andén and Stéphane Mallat. Multiscale scattering for audio classification. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 657–662, Miami (Florida), USA, October 24-28 2011.

[2] Gérard Biau and André Mas. PCA-kernel estimation. *Stat. Risk Model.*, 29(1):19–46, 2012.

[3] Joan Bruna and Stephane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, August 2013.

[4] Chris Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[6] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[7] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[8] Olivier Catoni. Estimating the gram matrix and least square regression through pac-bayes bounds. *preprint*, 2015.

[9] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, 2000.

[10] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[11] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Math. J.*, 25(100)(4):619–633, 1975.

[12] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[13] Ian Jolliffe. Principal component analysis. *Wiley StatsRef: Statistics Reference Online*, 2002.

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

[15] Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for spectral projectors of sample covariance. *preprint arXiv:1408.4643*, 2014.

[16] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *preprint arXiv:1405.2468*, 2014.

[17] Vladimir Koltchinskii and Karim Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *preprint arXiv:1504.07333*, 2015.

[18] Stéphane Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, 2012.

[19] David A. McAllester. Pac-bayesian model averaging. In *In Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170. ACM Press, 1999.

[20] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *In Advances in Neural Information Processing Systems*, pages 873–879. MIT Press, 2001.

[21] Stanislav Minsker. Geometric median and robust estimation in banach spaces. to appear.

[22] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[23] Michael Reed and Barry Simon. *Methods of modern mathematical physics. I.* Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, second edition, 1980. Functional analysis.

[24] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

[25] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

[26] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond.* Adaptive computation and machine learning. MIT Press, 2002.

[27] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, Cambridge, MA, USA, 1999.

[28] Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.*, 3:233–269, March 2003.

[29] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[30] John Shawe-taylor, Chris Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *Eds.): ALT 2002, LNAI 2533*, pages 23–40. Springer-Verlag, 2002.

[31] John Shawe-taylor, Christopher K. I. Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *IEEE Transactions on Information Theory*, 51:2005, 2005.

[32] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[33] E. Stein and G. Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, Princenton, N.J., 1971.

[34] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[35] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[36] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.

[37] Laurent Zwald, Olivier Bousquet, and Gilles Blanchard. Statistical properties of kernel principal component analysis. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 594–608. Springer, Berlin, 2004.