Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure

Gilles Blanchard, Jean-Baptiste Fermanian

Nonasymptotic one- and two-sample tests in high dimension with unknown covariance structure

Gilles Blanchard[1] and Jean-Baptiste Fermanian[1,2]

[1] Institut de Mathématiques, CNRS, Inria, Université Paris-Saclay
[2] École Normale Supérieure de Rennes

Abstract. Let $\mathbb{X} = (X_i)_{1\leq i\leq n}$ be an i.i.d. sample of square-integrable variables in $\mathbb{R}^d$, with common expectation μ and covariance matrix Σ, both unknown. We consider the problem of testing if μ is η-close to zero, i.e. $\|\mu\| \leq \eta$ against $\|\mu\| \geq (\eta + \delta)$; we also tackle the more general two-sample mean closeness testing problem. The aim of this paper is to obtain nonasymptotic upper and lower bounds on the minimal separation distance δ such that we can control both the Type I and Type II errors at a given level. The main technical tools are concentration inequalities, first for a suitable estimator of $\|\mu\|^2$ used a test statistic, and secondly for estimating the operator and Frobenius norms of Σ coming into the quantiles of said test statistic. These properties are obtained for Gaussian and bounded distributions. A particular attention is given to the dependence in the pseudo-dimension d_* of the distribution, defined as $d_* := \|\Sigma\|_2^2/\|\Sigma\|_\infty^2$. In particular, for $\eta = 0$, the minimum separation distance is $\Theta(d_*^{1/4}\sqrt{\|\Sigma\|_\infty/n})$, in contrast with the minimax estimation distance for μ, which is $\Theta(d_e^{1/2}\sqrt{\|\Sigma\|_\infty/n})$ (where $d_e := \|\Sigma\|_1/\|\Sigma\|_\infty$). This generalizes a phenomenon spelled out in particular by Baraud (2002).

Keywords: Signal detection, Two-sample test, Minmax testing separation distance, Effective dimensionality

Contribution to a Festschrift volume in the honor of V. Spokoiny's 60th birthday

1 Introduction

We consider the following fundamental signal detection problem: given an i.i.d. sample $\mathbb{X} = (X_i)_{1\leq i\leq n}$ from a square integrable distribution $\mathbb{P}_X$ on $\mathbb{R}^d$ (or possibly a separable Hilbert space, under some conditions which will be discussed later) with $\mu = \mathbb{E}[X_1]$, test the hypothesis of "η-closeness to zero" of the mean:

$$(H_0(\eta)) : \|\mu\| \leq \eta, \text{ against } (H_1(\eta,\delta)) : \|\mu\| > \eta + \delta. \tag{1}$$

In fact, we consider the following more general two-sample mean closeness testing problem: for $\mathbb{X} = (X_i)_{1\leq i\leq n}$ and $\mathbb{Y} = (Y_i)_{1\leq i\leq m}$ two independent samples of i.i.d. variables with distributions $\mathbb{P}_X, \mathbb{P}_Y$ on $\mathbb{R}^d$ with respective means μ and ν, test the hypothesis of η-closeness (or similarity) of the two means,

$$(H_0(\eta)) : \|\mu - \nu\| \leq \eta, \text{ against } (H_1(\eta,\delta)) : \|\mu - \nu\| > \eta + \delta. \tag{2}$$

Observe that we can always formally subsume setting (1) into setting (2), by letting m go to infinity and/or assuming (if needed) that the covariance of Y_1 is zero. Therefore, in the contribution section we will concentrate mainly on setting (2).

The problem (1) (and numerous extensions thereof) has been (at least for the zero mean test problem, i.e. $\eta = 0$) a long-time subject of attention in mathematical statistics, in particular since the seminal works of Ingster (1982, 1993) in the Gaussian white noise model (see next section for a more detailed discussion of related literature). In this work, we will consider the situation where the involved distributions are either Gaussian or of bounded norm (and hence sub-Gaussian), but with unknown covariance matrix acting as a nuisance parameter.

We are interested in finding bounds on the separation distance δ, i.e. a bound on the minimum value of δ such that there exists a test with both Type I and Type II error rates bounded by a "small" prescribed quantity. Our interest here is more on the constructive side, so that we will concentrate on feasible procedures that are in particular adaptive to the covariances of the involved distributions. A matching lower bound (for any fixed covariance structure) will be provided in the Gaussian setting. We emphasize that our focus is on *finite sample* (i.e. nonasymptotic) results, as will be discussed below.

1.1 Relation to white noise model in nonparametric statistics

In the isotropic Gaussian case (white noise) with known variance , and for $\eta = 0$, the signal detection problem (1) has been studied in much generality, in particular in the infinite-dimensional setting where $\mathbb{R}^d$ is replaced by a separable Hilbert space. In this situation, due to the fact that the white noise model on an infinite-dimensional Hilbert space cannot be represented by a random variable taking values in that space, the canonical model which is considered instead is the Gaussian sequence model for the coordinates of each of the observations in an orthonormal basis (in fact the Gaussian sequence model with known variance is usually considered with a single observation of the sequence):

$$X^{(i)} = \mu^{(i)} + \sigma\varepsilon^{(i)}, \qquad i \in \mathbb{N}_{>0}, \tag{3}$$

where $(\varepsilon^{(i)})_{i\geq 1}$ is an i.i.d. standard normal sequence. This fundamental model in nonparametric statistics allows to represent in a clean way many functional spaces of interest for the signal μ through geometrical properties of its expansion coefficients $(\mu^{(i)})_{i\geq 1}$ in a suitable basis. Since in that infinite-dimensional setting the alternative $\|\mu\|^2 > \delta^2$ is "too big" and gives rise to trivial separation rates, the usual focus is on considering restricted alternatives of the form $\left\{\mu \in \mathcal{F}; \|\mu\| \geq \delta^2\right\}$, for a given nonparametric set $\mathcal{F}$. Classical alternatives of interest include in particular ℓ_2 ellipsoids (corresponding to Hilbert norms of different strengths), ℓ_p bodies, and Besov bodies. Interpreted in functional spaces, these alternatives correspond respectively to balls in Sobolev spaces (typically when considering Fourier basis coefficient expansions) or in Besov spaces (for suitable wavelet basis coefficient expansions).

The literature on these topics is profound and extensive, see e.g. Ingster and Suslina (2012) for a comprehensive overview. The case of certain classes of ℓ_2-ellipsoids appears to have been studied first by Ingster (1982) and Ermakov (1991), then a remarkable series of works of Ingster (1993) and Ingster and Suslina (1998) established minimax testing rates for general ℓ_2 ellipsoids as well as other alternatives. V. Spokoiny's contribution is prominent in this body of literature, in particular for dealing with the case of Besov bodies (Lepski and Spokoiny, 1999) as well as considering the problem of statistical adaptivity over a family of alternatives (Spokoiny, 1996).

This very limited overview of the topic of testing in the white noise model is meant to contrast with the setting considered here. On the one hand, we will not consider a particular form of alternative; on the other hand, we assume that the observations can truly be represented as elements in a possibly infinite-dimensional separable Hilbert space. Under the Gaussian assumption, this means that the covariance operator Σ of the noise process is assumed to have a finite trace, which also prevents the triviality problem mentioned above for the white Gaussian noise setting. If we represent the observation coordinates in a diagonalizing basis of Σ, our setting in the Gaussian setting amounts to the Gaussian sequence model (3) wherein the constant parameter σ is replaced by a square integrable sequence $(\sigma^{(i)})_{i\geq 1}$. Note that formally normalizing the i-th observation coordinate by $\sigma^{(i)}$ would give rise again to model (3), however the separation distance would then be measured in the weak norm $\left\|\Sigma^{1/2}\mu\right\|$.

1.2 Relation to "modern" and high-dimensional statistics

Since we only consider test separation distance without a specific alternative, the setup we consider can be considered as less elaborate, at least in the sense of asymptotic theory, than the settings with various non-parametric alternatives discussed above. On the other hand, our focus is specifically on the following points:

1. Finite-sample analysis;
2. Non-Gaussian data (we will only consider bounded data here);
3. Robustness to misspecification (here under the form of the relaxed composite null $\|\mu\|^2 \leq \eta^2$).

These features have been rightly identified by V. Spokoiny as the defining features of "modern" approach to statistics (Spokoiny, 2012; Spokoiny and Dickhaus, 2015). The problem of testing a null hypothesis defined as a neighborhood rather than an exact match has been tackled under different settings in the statistics literature, especially for the two-sample testing case. For example, motivated by bioequivalence testing between populations, Munk and Czado (1998) consider the problem of testing closeness of two real distributions as measured in Mallows distance; and Dette and Munk (1998) study the problem of testing closeness in L^2 distance of two nonparametric (Hölder regular) regression functions. In both cases, the underlying principle is to estimate the target distance

— as will be also case in the present paper — and the data is not assumed to be Gaussian, but the corresponding analysis based on Gaussian asymptotic theory.

Taking the above aspects into account in the theory, in particular non-asymptotic analysis, is motivated by a large number of high-dimensional applications, where it appears that relying on traditional asymptotic of Gaussian parametric or non-parametric theory can possibly be problematic if done without care. Finite sample theory allows to delineate more precisely in which situations traditional approximations still can be relied upon, and to study non-standard asymptotics, in particular when key parameters, such as dimensionality, can themselves depend on the sample size n. It is also of use when considering multiple testing scenarios, where multiplicity has to be taken into account precisely.

Another fruitful modern insight is that high-dimensional statistical models tend to blur the line between parametric and non-parametric point of views. Precise non-asymptotic results in a finite-dimensional setting, but where the role of key model parameters (in particular, dimensionality or effective dimensionality) is precisely analyzed, can provide key theoretical components for analyzing non-parametric settings. In the signal testing framework considered in the present paper, this way of thinking has in particular been pioneered by Baraud (2002), who obtained sharp non-asymptotic results for the problem (1) in the case $\eta = 0$, and for the finite-dimensional counterpart of the white noise model (3), i.e. the isotropic setting $\Sigma = \sigma^2 I_d$ in dimension d. Baraud further demonstrated that this result provided a valuable and versatile tool to analyze models of typical interest in high-dimensional statistics (such as sparse alternatives) as well as non-parametric alternatives (such as those mentioned in the previous section). A key insight from Baraud's work is that the minimum separation distance in that setting is $\mathcal{O}(d^{1/4}\sigma/\sqrt{n})$, in contrast with minimax estimation distance for μ, which is $\Theta(d^{1/2}\sigma/\sqrt{n})$: the testing separation distance is smaller than the minimax estimation error by a factor $d^{1/4}$.

Analyzing precisely the role of dimensionality (ambient or effective) in minimax testing separation rates and the difference with minimax estimation rates has been a subject of interest in recent literature in various settings, highlighting similar related phenomena. For instance, Lam-Weil et al. (2021) consider the problem of testing equality of two high-dimensional multinomial distributions and study the minimum ℓ_1 separation distance in a vicinity of a reference distribution π (which implicitly determines a notion of local effective dimensionality). Since this model has bounded data, our analysis could be applied in that setting, however it concerns separation in ℓ_2 distance (the separation in ℓ_1 distance exhibits considerably more involved behavior). Ostrovskii et al. (2020) consider a different type of two-sample testing problem, in a regression context, where the goal is to determine which one of the two distributions has a given (known to the user) regression vector. They give a sharp bound on the minimum separation distance between the two regression vectors including the role of the dimension, also exhibiting a difference with estimation rates.

Coming back to our model, the results of Baraud (2002) provide a sharp answer, but only in the case $\eta = 0$ and for isotropic Gaussian (white noise)

data with known variance. Still in the Gaussian isotropic case, the minimum separation rates for any value of $\eta \geq 0$ were precisely characterized by Blanchard et al. (2018). We also consider the Gaussian setting in the present work, but analyze the generalized situation where the covariance matrix Σ can be arbitrary (and unknown). In this situation, the role of the dimensionality d is played by proxy quantities depending on Σ, sometimes called effective dimensionality or effective rank. For the signal testing problem however, it turns out that the proxy dimensionalities for testing and estimation differ. Namely, for $\eta = 0$, we find that the minimax separation distance is $\mathcal{O}(d_*^{1/4}\sqrt{\|\Sigma\|_\infty/n})$, where $d_* := \|\Sigma\|_2^2/\|\Sigma\|_\infty^2$, while the minimax estimation distance for μ is $\Theta(d_e^{1/2}\sqrt{\|\Sigma\|_\infty/n})$, where $d_e := \|\Sigma\|_1/\|\Sigma\|_\infty$. (Notice that $d_* \leq d_e \leq d$ in general, while these quantities are all equal in the isotropic setting.) Furthermore, we also study the estimation of key quantities $\|\Sigma\|_\infty^{1/2}$ and $\|\Sigma\|_2$ determining the proxy dimensionality and the testing threshold[3].

A crucial mathematical tool in high-dimensional statistics is to obtain sharp concentration inequalities for quadratic forms of random vectors. These are closely related to technical tools used in the present work. An important point in such inequalities is to quantify as precisely as possible up to which point quadratic forms of non-Gaussian vectors can mimic the Gaussian behavior (i.e. that of central and non-central weighted chi-squared statistics). This topic has received a good deal of attention in the recent years and V. Spokoiny also made substantial contributions to that area (Spokoiny and Dickhaus, 2015; Spokoiny and Zhilova, 2013). In the present work, we derive from scratch the needed concentration inequalities; we discuss in more detail the relation to V. Spokoiny's own work and to related literature in Section 2.4.

1.3 Relation to machine learning and kernel mean embeddings of distributions

An application setting which motivated us to consider in detail the case of bounded data is that of testing of the data distribution via kernel mean embedding (KME) methods, a principle which has garnered a lot of attention in the machine literature since the seminal paper of Smola et al. (2007). It has been advocated in particular for two-sample (Gretton et al., 2012) and goodness-of-fit (Chwialkowski et al., 2016) testing; see Muandet et al. (2017) for a recent overview.

We describe the KME principle briefly. Assume Z is a random variable with distribution $\mathbb{P}_Z$ taking values in the measurable space $\mathcal{Z}$, and that one has at hand a fixed mapping $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space. To this mapping is associated a reproducing kernel Hilbert space (rkHs) $\mathcal{H}'$ with kernel $k(z, z') := \langle \Phi(z), \Phi(z') \rangle$.

[3] With the notation $\|\Sigma\|_p$ we mean p-Schatten norm. We will freely use in the paper the equivalent notation $\|\Sigma\|_\infty = \|\Sigma\|_{\mathrm{op}}$, $\|\Sigma\|_1 = \mathrm{Tr}(\Sigma)$, $\|\Sigma\|_2^2 = \mathrm{Tr}(\Sigma^2)$.

Assuming the variable $X = \Phi(Z)$ is Bochner integrable[4] (which is the case in particular when the mapping Φ is bounded), the kernel mean embedding of $\mathbb{P}_Z$ is defined as $\Phi(\mathbb{P}_Z) := \mathbb{E}[\Phi(Z)] \in \mathcal{H}$ (using a rather natural overload of notation for Φ). The *maximum mean discrepancy* (MMD) between distributions $\mathbb{P}, \mathbb{Q}$ in the domain of definition of Φ is defined as the semimetric

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|.$$

Since $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) > 0$ implies $\mathbb{P} \neq \mathbb{Q}$, this principle can be used for simple goodness-of-fit testing (testing for $\mathbb{P}_Z = \mathbb{P}_0$ for some known distribution $\mathbb{P}_0$, given an i.i.d. sample from $\mathbb{P}_Z$) and two-sample testing (testing for $\mathbb{P}_Z = \mathbb{P}_{Z'}$, given two independent i.i.d. samples from $\mathbb{P}_Z$ and $\mathbb{P}_{Z'}$); in each case, the test statistic is a suitable estimator of $\mathrm{MMD}_k(\mathbb{P}_Z, \mathbb{P}_0)$, resp. $\mathrm{MMD}_k(\mathbb{P}_Z, \mathbb{P}_{Z'})$ from the observed data. More generally one may want to test the relaxed null hypothesis $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) \leq \eta$ and analyze the power of the test in terms of the MMD separation itself. This is indeed a particular case of (1)-(2), when considering the Hilbert-valued variable $X = \Phi(Z)$ and, for two-sample testing, $Y = \Phi(Z')$.

A common situation is when Φ is bounded in norm by some constant L, or equivalently in terms of the kernel, $\sup_{z \in \mathcal{Z}} k(z, z) \leq L^2$. This ensures in particular that Φ is defined on all distributions. Analyzing our original setting with norm-bounded but potentially infinite-dimensional data is therefore suited to this case.

Gretton et al. (2012) derive the asymptotic distribution of the (suitably renormalized) MMD test statistic, which is identical to the one we use below (once interpreted in the KME setting). Unsurprisingly, a Gaussian limiting behavior is identified. Our study analyzes this behavior from a non-asymptotic point of view; this can be particularly of interest for situation where the mapping Φ (or equivalently the associated kernel) is to depend on the sample size, or when performing a large number of such tests in parallel: in this case uniformly valid nonasymptotic bounds are a a valuable tool for further analysis. See Marienwald et al. (2020) for such a multiple test scenario in the context of so-called multiple task averaging. Multiple tests can also be aggregated to test a global hypothesis, see Fromont et al. (2012) in the context of two-sample testing based on the KME approach.

In our study, the power of the test is investigated for alternatives of the form (1)-(2), which, interpreted in the KME setting, correspond to $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) \geq \eta + \delta$. The power of KME-based tests (in the goodness-of-fit case) was also investigated by Balasubramanian et al. (2021), but for alternatives measured in a χ^2 distance separation, more precisely, of the form $\left\{\mathbb{Q} \in \mathcal{F}; \chi^2(\mathbb{P}_0, \mathbb{Q}) \geq \delta\right\}$, where $\mathcal{F}$ is a nonparametric set of distributions whose density with respect to $\mathbb{P}_0$ is approximated at a given rate by functions in the rkHs $\mathcal{H}'$ associated to k, in the sense of interpolation with $L^2(\mathbb{P}_0)$. This is close in spirit to nonparametric points of view discussed in Section 1.1, in the sense that χ^2-separation alone is

[4] that is, the real random variable $\|\Phi(Z)\|$ is integrable, which guarantees that the integral of $\Phi(Z)$ is well-defined in a strong sense as an element of the Hilbert space; see e.g. Cohn (1980).

too weak to get nontrivial separation rates and one has to additionally consider intersection with nonparametric sets of interest. Again, because we choose to analyze alternatives measured in MMD_k-separation itself, the results we obtain in this setting have a different nature.

1.4 Overview of contributions

The main contribution of this paper is to give upper bounds on the optimal (minmax) testing separation distance for problems (1) and (2) over classes of probability distributions with fixed covariance matrix Σ for sample $\mathbb{X}$, as well as S for sample $\mathbb{Y}$ in the two-sample case. The covariance structures are considered as nuisance parameters and we investigate precisely how they influence the testing separation distance. Let $\mathcal{P}$ be a family of distributions for the two samples (we consider the Gaussian setting and the bounded setting), and $\mathcal{P}_{\Sigma,S}$ the subsets of distributions of $\mathcal{P}$ with $\mathrm{Cov}[X_1] = \Sigma$, and $\mathrm{Cov}[Y_1] = S$ (in the two-sample case). Consider the sets of distributions

$$\begin{aligned}\mathcal{H}_0(\eta, \Sigma, S) &:= \{\mathbb{P} \in \mathcal{P}_{\Sigma,S} | \mathbb{P} \text{ satisfies } H_0(\eta)\}, \\ \mathcal{A}_\delta(\eta, \Sigma, S) &:= \{\mathbb{P} \in \mathcal{P}_{\Sigma,S} | \mathbb{P} \text{ satisfies } H_1(\eta, \delta)\},\end{aligned}$$

then the optimal separation distance is, for $\alpha \in (0,1)$:

$$\delta^*(\alpha, \Sigma, S, \eta) = \inf\left\{\delta \geq 0 \middle| \exists \text{ test } T \ : \sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}(T=1) + \sup_{\mathbb{P}\in\mathcal{A}_\delta} \mathbb{P}(T=0) \leq \alpha\right\}. \tag{4}$$

In the Gaussian setting, we establish that δ^* is upper bounded up to a constant factor via

$$\delta^*(\alpha, \Sigma, S, \eta) \lesssim \sigma\kappa_\alpha \max\left(1, \min\left(d_*^{\frac{1}{4}}, d_*^{\frac{1}{2}}\frac{\sigma\kappa_\alpha}{\eta}\right)\right), \tag{5}$$

(Theorem 7) where $\kappa_\alpha := \sqrt{-\log(\alpha)}$, and, in the one-sample case, $\sigma^2 := \|\Sigma\|_{\mathrm{op}}/n$ is a scalar variance factor and $d_* := \mathrm{Tr}\,\Sigma^2/\|\Sigma\|_{\mathrm{op}}^2$ a notion of effective dimension. In the two-sample case, we obtain also (5), with $\sigma^2 := \|M_{m,n}\|_{\mathrm{op}}$, and $d_* := \mathrm{Tr}\,M_{m,n}^2/\sigma^4$, where $M_{m,n} := (\Sigma/n + S/m)$ (Theorem 8). In the one-sample case, this result can be formulated equivalently in terms of *sample complexity* n^* needed to detect at given error level α and separation distance δ for problem (1):

$$n^*(\alpha, \Sigma, S, \eta) \lesssim \|\Sigma\|_{\mathrm{op}}\kappa_\alpha\delta^{-1}\max\left(\delta^{-1}, d_*^{\frac{1}{2}}(\max(\delta,\eta))^{-1}\right). \tag{6}$$

This result is established first when assuming that Σ, S are known, then we show that it holds as well when they are unknown (under some mild assumptions on the sample size, see Corollary 14 for an explicit statement in the one-sample case and condition (27) there). Matching minimax lower bounds are given for one and two-sample problems in the Gaussian setting. In the bounded setting, we derive upper bounds only, which take the same flavor as (5) under some mild assumptions on the sample sizes.

1.5 Organization of the paper

We present in Section 2 our main results. In order to cover both the Gaussian and bounded settings under the same umbrella, we start in Section 2.1 by a generic result: assuming some suitable concentration for an estimate U of the squared signal norm $\|\mu\|^2$ holds (Assumption 1), as well as for estimators of its quantiles (Assumption 2), for the problems (1) and (2) we propose in Theorem 3 sufficient conditions on δ such that we can control the Type I and Type II errors of a test T based on U. In the following sections, the Gaussian setting and the bounded setting are considered separately. In Section 2.2, we give concentration results for U to fulfill Assumption 1. In Section 2.3 we give results to fulfill Assumption 2, which are related to the estimation of $\|\Sigma\|_\infty^{1/2}$ and $\|\Sigma\|_2$. The proofs of the corresponding results are found in Sections 3.1 to 3.6, respectively.

2 Main results

We will build a test for the model (2) based on an estimator U of the distance $\|\mu-\nu\|^2$, typically a modified U-statistic as defined below. We will first consider a general point of view to deduce bounds on the separation rate when U satisfies certain concentration properties; this will then apply both to the Gaussian and bounded settings.

2.1 A general result to upper bound separation rates

As mentioned earlier, from now on we concentrate primarily on the two-sample setting, being understood that upper bounds for the one-sample setting can be deduced readily. In order to define a general framework encompassing as particular cases the more specific settings considered below, in this section we will assume a generic statistical model $\mathcal{P}$ for the distribution of the samples $\mathbb{X}$ and $\mathbb{Y}$, which we recall we always assume to be independent and i.i.d. with respective squared integrable marginal distributions $\mathbb{P}_X, \mathbb{P}_Y$. We will thus use without comment the fact that a distribution $\mathbb{P} \in \mathcal{P}$ equivalently specifies the marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ of the samples. We will consider the covariance matrices Σ, S of $\mathbb{P}_X, \mathbb{P}_Y$ as nuisance parameters influencing the optimal separation distance, and define the sub-models

$$\mathcal{P}_{\Sigma,S} = \{\mathbb{P} \in \mathcal{P} : \mathrm{Cov}[\mathbb{P}_X] = \Sigma, \mathrm{Cov}[\mathbb{P}_Y] = S\};$$

$\mathcal{P}_\Sigma$ is defined in an analogous way for the one-sample setting.

The first property we require is a form of 2-sided concentration of U around the target quantity:

Assumption 1. For any (Σ, S) and distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$; for any given $\alpha \in (0,1)$ there exist $q_1 = q_1(\Sigma, S, \alpha), q_2 = q_2(\Sigma, S, \alpha)$ in $\mathbb{R}_+$ such that:

$$\mathbb{P}\big[\big|U - \|\mu-\nu\|^2\big| \geq \|\mu-\nu\| q_1 + q_2\big] \leq \alpha\,. \tag{7}$$

Additionally, we will consider the situation where the quantities q_1, q_2 (which are necessary to find a suitable testing threshold) are not known but must also be estimated from the data; this is the case if the covariance matrices (Σ, S) are unknown. This leads us to our second assumption:

Assumption 2. Suppose Assumption 1 holds, with the notation introduced therein. For any $\alpha \in (0,1)$ there exist two estimators $\widehat{Q}_1 = \widehat{Q}_1(\alpha)$ and $\widehat{Q}_2 = \widehat{Q}_2(\alpha)$ in $\mathbb{R}_+$ such that, for any (Σ, S) and distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$:

$$\mathbb{P}\left[\left|q_1(\Sigma, S, \alpha) - \widehat{Q}_1(\alpha)\right| \geq \frac{1}{2} q_1(\Sigma, S, \alpha)\right] \leq \alpha\,, \tag{8}$$

$$\mathbb{P}\left[\left|q_2(\Sigma, S, \alpha) - \widehat{Q}_2(\alpha)\right| \geq \frac{1}{2} q_2(\Sigma, S, \alpha)\right] \leq \alpha\,. \tag{9}$$

(In the "oracle" case where the covariances Σ, S are assumed to be known, of course Assumption 2 is trivially satisfied taking $\widehat{Q}_1 = q_1$, $\widehat{Q}_2 = q_2$.) The following generic result transforms the above assumptions into an estimate of the separation distance for setting (2).

Theorem 3. *Let $\mathcal{P}$ be a statistical model for setting (2), and U be a statistic. Let Assumptions 1 and 2 be granted. Given $\eta \geq 0$ and $\alpha \in (0,1)$, let T be the test defined by*

$$T = 1\Big\{U - \eta^2 > 2\eta \widehat{Q}_1(\alpha) + 2\widehat{Q}_2(\alpha)\Big\}\,. \tag{10}$$

Then for any (Σ, S), provided

$$\delta > 2q_1 + \min\big(2\sqrt{q_2}, 2\eta^{-1} q_2\big)\,, \tag{11}$$

it holds, for any distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$:

$$\mathbb{P}[T = 1] \leq 3\alpha\,, \ \textit{if } \mathbb{P} \textit{ satisfies } (H_0(\eta));$$
$$\mathbb{P}[T = 0] \leq 3\alpha\,, \ \textit{if } \mathbb{P} \textit{ satisfies } (H_1(\eta, \delta)).$$

2.2 Concentration properties of the test statistic

The rest of the paper is dedicated to establishing the validity of Assumptions 1 and 2 for the following statistic $U(\mathbb{X}, \mathbb{Y})$:

$$U(\mathbb{X}, \mathbb{Y}) := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \langle X_i, X_j \rangle + \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \langle Y_i, Y_j \rangle - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle X_i, Y_j \rangle\,. \tag{12}$$

Observe that provided expectations μ, ν exist, $U(\mathbb{X}, \mathbb{Y})$ is an unbiased estimator of $\|\mu - \nu\|^2$. In the KME setting as described in Section 1.3, inner products are replaced by kernel evaluations and the above statistic is the standard unbiased estimate of the squared MMD between $\mathbb{P}_X$ and $\mathbb{P}_Y$. As announced previously, we will concentrate on the following two settings:

Definition 4 (Gaussian setting). *The samples $\mathbb{X}$ and $\mathbb{Y}$ are i.i.d. Gaussian in $\mathbb{R}^d$ of marginal distributions $\mathbb{P}_X = \mathcal{N}(\mu, \Sigma)$ and $\mathbb{P}_Y = \mathcal{N}(\nu, S)$, respectively.*

In the Gaussian setting, we will assume a finite ambient dimension d for technical reasons: our proofs rely on the Gauss-Lipschitz concentration inequality, which applies in finite dimension. As will appear clearly however, all our results to come are dimension-free in the sense that d never enters the picture, instead only norms of Σ, S come into play. We surmise that our results would apply as well in the same form in the Hilbert-valued setting provided $\mathrm{Tr}(\Sigma)$ and $\mathrm{Tr}(S)$ are finite, but did not try to write down a precise approximation argument to this end.

Definition 5 (Bounded setting). *The samples $\mathbb{X}$ and $\mathbb{Y}$ are i.i.d. in a separable Hilbert space $\mathcal{H}$ with norm bounded by $L > 0$. The covariance operators for the marginal sample distributions are denoted Σ and S, respectively; observe that they have finite trace by the boundedness assumption.*

Propositions 6 and 9 give concentration bounds for the statistic U, ensuring Assumption 1 in the two above settings.

Proposition 6. *Assume the Gaussian setting holds and $n, m \geq 2$. Then with probability at least $1 - \alpha$,*

$$\left|U - \|\mu - \nu\|^2\right| \leq \|\mu - \nu\| q_1 + q_2 \,, \tag{13}$$

where U is defined in (12) *and*

$$q_1(\Sigma, S, \alpha) = \sqrt{2\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right) u(\alpha)} \,, \tag{14}$$

$$q_2(\Sigma, S, \alpha) = 32\left(\frac{\sqrt{\mathrm{Tr}\,\Sigma^2}}{n} + \frac{\sqrt{\mathrm{Tr}\,S^2}}{m}\right) u(\alpha) \,. \tag{15}$$

where $u(\alpha) := -\log\alpha + \log 8$.

Let us simplify somewhat the above expression when plugged into Theorem 3 in the case of signal detection (1). We also give a matching lower bound (up to constant factor) for the optimal separation distance.

Theorem 7. *Consider the signal detection problem* (1) *and assume the Gaussian setting with covariance matrix Σ. Then the minimum separation distance δ^* given by* (4) *so that the type I and II errors for problem* (1) *are less that $\alpha \in (0,1)$ for all $\mathbb{P} \in \mathcal{P}_\Sigma$ is upper bounded by*

$$\delta^*(\alpha, \Sigma, \eta) \lesssim \sigma_n \sqrt{u} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_* u} \cdot \frac{\sigma_n}{\eta}\right)\right), \tag{16}$$

where $u(\alpha) := -\log\alpha + \log 60$. If $d_* \geq 3$, then it is lower bounded by

$$\delta^*(\alpha, \Sigma, \eta) \geq \sigma_n \sqrt{\frac{1-\alpha}{12}} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_*(1-\alpha)} \cdot \frac{\sigma_n}{\eta}\right)\right), \tag{17}$$

where $\sigma_n^2 := \|\Sigma\|_{\mathrm{op}}/n$, *and* $d_* := \operatorname{Tr}\Sigma^2/\|\Sigma\|_{\mathrm{op}}^2$. *(The symbol* $\lesssim$ *indicates inequality up to a numerical factor).*

Observe that it holds $d_* \leq d_e$, where $d_e = \operatorname{Tr}\Sigma/\sigma^2$ is the "effective dimensionality" coming into play for signal estimation rates (namely $\mathbb{E}\big[\|\overline{X} - \mu\|^2\big]^{1/2} = \sigma\sqrt{d_e/n}$, where $\overline{X}$ is the empirical mean). In the finite d-dimensional case with $\Sigma = I_d$, it holds $d = d_e = d_*$, and the separation (16) has been shown to be optimal in the Gaussian setting for $\eta = 0$ by Baraud (2002) and for any $\eta \geq 0$ by Blanchard et al. (2018). It exhibits a continuous transition between the signal detection setting ($\eta = 0$, $\delta^* \simeq d^{1/4}\sigma/\sqrt{n}$) and the hyperplane testing setting (which is equivalent to the 1-dimensional setting by rotational invariance; $\eta \to \infty$, $\delta^* \simeq \sigma/\sqrt{n}$). In that particular situation, we observe that the signal separation distance is smaller by a factor $d^{1/4}$ than the signal estimation error, a phenomenon typical of high-dimensional statistics. In the more general setting studied here where Σ can be arbitrary, this difference between rates can be all the more marked, since in addition d_* can be much smaller than d_e.

We obtain a similar result for the two-sample problem:

Theorem 8. *Consider the two-sample mean problem* (2) *and assume the Gaussian setting with covariance matrices* Σ, S. *Then the minimum separation distance* δ^* *so that the type I and II errors for problem* (2) *is less than* $\alpha \in (0,1)$ *for all* $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$ *is upper bounded by*

$$\delta^*(\alpha, \Sigma, S, \eta) \lesssim \sigma_{n,m}\sqrt{u}\max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_* u}\cdot\frac{\sigma_{n,m}}{\eta}\right)\right), \tag{18}$$

where $u := -\log\alpha + \log 60$. *If* $d_* \geq 3$, *then it is lower bounded by*

$$\delta^*(\alpha, \Sigma, S, \eta) \geq \sigma_{n,m}\sqrt{\frac{1-\alpha}{48}}\max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_*(1-\alpha)}\cdot\frac{\sigma_{n,m}}{\eta}\right)\right), \tag{19}$$

where $\sigma_{n,m}^2 := \|M_{n,m}\|_{\mathrm{op}}$, *and* $d_* := \operatorname{Tr} M_{n,m}^2/\sigma_{n,m}^4$, *for* $M_{n,m} := \Sigma/n + S/m$. *(The symbol* $\lesssim$ *indicates inequality up to a numerical factor).*

Here the effective dimension d^* depends on the two covariance matrices Σ and S, weighted by the size of the samples.

Remark. As mentioned in the introduction, by letting m go to infinity in the two-sample case, we recover the bounds of the one sample case (up to a constant factor). It is worth examining if the converse holds, i.e. if there is an argument to reduce the two-sample problem to the simpler one-sample case (this would simplify some technical aspects of the proofs, somewhat). For the *upper* bounds on the minimum separation distance, this is the case in some specific situations: for equal sample sizes $n = m$, the two-sample case can be reduced to the one-sample problem setting by pairing the samples and considering the single sample $(X_i - Y_i)_{1\leq i\leq n}$, and one can recover this way in essence the two-sample result. If $\Sigma = S$, and for general sample sizes, we can also reduce to the single sample with size $\min(m,n)$, $(X_i - Y_i)_{1\leq i\leq \min(m,n)}$, and recover again

the two-sample result up to a numerical factor. However a reduction argument in the general case has eluded us. Concerning the *lower* bound, the argument for the two-sample case indeed hinges on a reduction the one-sample case, by considering the sub-models where one of the two sample means is known, see Section 3.4.

We now turn to the bounded setting.

Proposition 9. *Assume the bounded setting holds and $n, m \geq 2$. Then with probability at least $1-\alpha$,*

$$\left|U - \|\mu-\nu\|^2\right| \leq \|\mu-\nu\| q_1 + q_2 \,, \tag{20}$$

where U is defined in (12) *and*

$$q_1(\Sigma, S, \alpha) = 2\sqrt{2\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right)u} + \frac{4Lu}{3(n \wedge m)} \,,$$
$$q_2(\Sigma, S, \alpha) = 614\left(\frac{\sqrt{\operatorname{Tr}\Sigma^2}}{n} + \frac{\sqrt{\operatorname{Tr} S^2}}{m}\right)u + 3708\frac{L^2u^2}{(n\wedge m)^2} \,,$$

with $u(\alpha) = -\log\alpha + \log 2$.

Thus, in the bounded setting we can guarantee that the behavior of the test is qualitatively the same as in the Gaussian setting (see e.g. Theorem 7) — and this from a non-asymptotic point view, provided $n \wedge m \geq uL^2/\sigma^2$, where $\sigma^2 = \|\Sigma\|_{\mathrm{op}}$.

A special case of interest is when the data lies on the sphere of radius L, i.e. $\|X_i\| = \|Y_j\| = L$ a.s. In this case $L^2 = \operatorname{Tr}\Sigma$ and the above condition can be rewritten $n \wedge m \geq u d_e$. This situation is met in particular in the KME setting, see Section 1.3, when using a translation-invariant kernel $k(z, z') = k_\circ(z - z')$, in which case $L^2 = k_\circ(0)$.

2.3 Quantile estimation

Since we are considering the case where Σ, S can be arbitrary in this work, it is natural to assume that these are not known in advance. We study next the estimation of the quantities q_1 and q_2, in both settings (bounded and Gaussian), in order to check Assumption 2 for our generic theorem. If we can grant that assumption, Theorem 3 guarantees that the separation distance remains qualitatively the same as in the "oracle" situation where they are known. To simplify the exposition, in this section we will present results for the one-sample problem only; similar results, although slightly more technical, can be obtained for the two-sample problem. Thus, we need to have estimators of $\|\Sigma\|_{\mathrm{op}}$ and $\operatorname{Tr}\Sigma^2$ — more precisely, of their square root.

For q_1, we will use the empirical covariance operator $\widehat{\Sigma} := \widehat{\Sigma}(\mathbb{X})$:

$$\widehat{\Sigma}(\mathbb{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T \,, \tag{21}$$

where $\widehat{\mu} := \widehat{\mu}(\mathbb{X})$ is the empirical mean of the sample $\mathbb{X}$.

Proposition 10 (Gaussian setting). *Assume $\mathbb{X} = (X_i)_{1 \leq i \leq n}$ are i.i.d. Gaussian vectors of covariance Σ. For $u \geq 0$, with probability at least $1 - 3e^{-u}$:*

$$\left| \left\| \widehat{\Sigma} \right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \right| \leq 3\sqrt{2} \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \left(\sqrt{\frac{d_e}{n}} + \sqrt{\frac{u}{n}} \right), \tag{22}$$

where $\widehat{\Sigma}$ is defined in (21) *and $d_e = \operatorname{Tr} \Sigma / \|\Sigma\|_{\mathrm{op}}$.*

Proposition 11 (Bounded setting). *Assume that $\mathbb{X} = (X_i)_{1 \leq i \leq n}$ are i.i.d. bounded in norm by L and with covariance Σ. For $u \geq 0$, with probability at least $1 - 2e^{-u}$:*

$$\left| \left\| \widehat{\Sigma} \right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \right| \leq 4L \left(2\sqrt{\frac{d_e}{n}} + \sqrt{\frac{2u}{n}} + \frac{u}{3n} \right) \tag{23}$$

where $\widehat{\Sigma}$ is defined in (21) *and $d_e = \operatorname{Tr} \Sigma / \|\Sigma\|_{\mathrm{op}}$.*

These concentration bounds are not sharp in an asymptotic sense, where the main term for the scaling of the deviations is expected to follow that of asymptotic normality for eigenvalues of the empirical covariance operators, as in the classical results of Anderson (2003), but they are largely sufficient for our purposes (see Corollary 14 below). Some refined related nonasymptotic bounds can be found in the recent literature. In particular, Koltchinskii and Lounici (2017) derive nonasymptotic results for controlling $\left\| \widehat{\Sigma} - \Sigma \right\|$ in the Gaussian setting, and in the centered case where $\mu = 0$ is known. In fact, in essence the result of our technical Proposition 23 in the proof section (which is like Proposition 11 but in the centered case) can be deduced from the results of Koltchinskii and Lounici (2017) by elementary arguments. We decided to include a standalone proof here; while we do rely on the estimates of Koltchinskii and Lounici (2017) (or rather on the improved version of van Handel, 2017) for the expectation of the difference, we derive an upper bound on the deviation by a rather direct application of the Gauss-Lipschitz concentration. While Koltchinskii and Lounici (2017) also rely on such arguments, their proofs are much more involved, for the reason that they study the norm or the difference while we only are interested in the difference of the (root) norms here. Finally, we also mention very recent results of Jirak and Wahl (2018) for sharp nonasymptotic control of spectral quantities related to Σ, which could also potentially be applied here, though it seems at first glance that a logarithmic dependence in the dimension could enter into play.

For the bounded setting (Proposition 11), the bound (23) could presumably be improved to have $\sqrt{\|\Sigma\|_{\mathrm{op}}}$ instead of L for the main terms. The results of Theorem 9 of Koltchinskii and Lounici (2017) under a sub-Gaussian assumption do not seem to be able to imply Proposition 11, see the more detailed discussion below in Section 2.4.

Turning now to q_2, we will estimate $\sqrt{\operatorname{Tr}\Sigma^2}$ using the following statistic $\widehat{T} := \widehat{T}(\mathbb{X})$, which is an unbiased estimator of $\operatorname{Tr}\Sigma^2$:

$$\widehat{T}(\mathbb{X}) := \frac{1}{4n(n-1)(n-2)(n-3)} \sum_{i\neq j\neq k\neq l} \langle X_i - X_k, X_j - X_l\rangle^2 . \tag{24}$$

Proposition 12 (Gaussian setting). *Assume $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. Gaussian vectors of covariance Σ and $n \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 30\sqrt{\frac{\operatorname{Tr}\Sigma^2}{n}}u^2\right] \leq e^4 e^{-u}, \tag{25}$$

where $\widehat{T}$ is defined in (24).

Proposition 13 (Bounded setting). *Assume that $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. bounded in norm by L and with covariance Σ and $n \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 12L^2\sqrt{\frac{u}{n}}\right] \leq 2e^{-u}. \tag{26}$$

where $\widehat{T}$ is defined in (24).

Thanks to these concentration results, we can construct estimators of $q_1(\Sigma,\alpha)$ and $q_2(\Sigma,\alpha)$ satisfying Assumption 2. In the Gaussian setting, we give the following explicit corollary of Propositions 10 and 12; the proof is straightforward and omitted.

Corollary 14 (Gaussian setting). *Consider the signal detection problem* (1) *and assume the Gaussian setting holds. Let $\alpha \in (0,1)$, $u = u(\alpha) = -\log\alpha + \log 8$, and $\widehat{Q}_1(\alpha)$ and $\widehat{Q}_2(\alpha)$ be the statistics defined by*

$$\widehat{Q}_1(\alpha) = \sqrt{\frac{2\left\|\widehat{\Sigma}(\mathbb{X})\right\|_{\mathrm{op}}}{n}u}, \quad \widehat{Q}_2(u) = 32\frac{\sqrt{\widehat{T}(\mathbb{X})}}{n}u,$$

where $\widehat{\Sigma}$ is defined in (21) *and $\widehat{T}$ in* (24). *Then for any Σ, provided*

$$n \gtrsim \max\left(d_e(\Sigma), u, u^4\right), \tag{27}$$

(we recall $d_e(\Sigma) = \operatorname{Tr}\Sigma/\|\Sigma\|_{\mathrm{op}}$), then it holds, for any distribution $\mathbb{P} \in \mathcal{P}_\Sigma$:

$$\mathbb{P}\left[\left|\widehat{Q}_1(\alpha) - q_1(\Sigma,\alpha)\right| \leq q_1(\Sigma,\alpha)/2\right] \leq \alpha,$$
$$\mathbb{P}\left[\left|\widehat{Q}_2(\alpha) - q_2(\Sigma,\alpha)\right| \leq q_2(\Sigma,\alpha)/2\right] \leq \alpha,$$

where q_1, q_2 are as defined in (14),(15) *(with $m = \infty$).*

The condition (27) for n is needed to grant Assumption 2: it ensures that the deviations of the estimators $\|\widehat{\Sigma}\|_{\mathrm{op}}^{1/2}$ and $\widehat{T}$ coming from Proposition 10 and 12 are smaller than their target quantities $\|\Sigma\|_{\mathrm{op}}^{1/2}/2$ and $(\mathrm{Tr}\,\Sigma^2)^{1/2}/2$, respectively. The requirement that the size of the sample is larger than the effective dimension d_e appears mild.

For the bounded setting and the signal detection problem (1), estimators $\widehat{Q}_1$ and $\widehat{Q}_2$ satisfying Assumption 2 can also be constructed in a similar way from Propositions 11 and 13 (details omitted). In the bounded setting, the quantiles q_1 and q_2 of U are composed of two terms, the first (and larger) one gives the dependence in the covariance of the distribution, the second depends on the bound L. This additional term will have to be taken into account, and the condition on n analogous to (27) will involve L. In general this will not be a problem since L or an upper bound on L is supposed to be known, as is the case for instance in the kernel setting (see the concluding discusssion in the previous section). Finally, for the two-sample test problem (2), comparable results can be obtained using the estimators $\widehat{\Sigma}(\mathbb{Y})$ and $\widehat{T}(\mathbb{Y})$; we omit the details.

2.4 Concluding remarks

A technical discussion point: Gaussian, sub-Gaussian, and bounded vectors. The utility of our systematic distinction between the Gaussian and bounded case can be disputed in the light of recent concentration literature (see e.g. Hsu et al., 2012; Koltchinskii and Lounici, 2017 and further references therein) deriving results holding for sub-Gaussian random vectors, a seemingly more general setting emcompassing both the Gaussian and bounded settings as particular cases (since bounded variables are sub-Gaussian by Hoeffding's inequality).

This point deserves a specific discussion. The sub-Gaussianity assumption for a vector variable X (assumed centered for simplicity here) often takes the following form: for any unit vector u, denoting $X_u = \langle X, u\rangle$, it is assumed that $\|X_u\|_{\psi_2} \leq C\sqrt{\mathrm{Var}[X_u]}$ (where $\|.\|_{\psi_2}$ is the Orlicz ψ_2-norm); or equivalently in terms of Laplace transform,

$$\log(\mathbb{E}[\exp\lambda(X_u)]) \leq (C')^2\lambda^2\mathrm{Var}[X_u]/2 \text{ for all } \lambda \geq 0. \tag{28}$$

A key point is that the factors C or C' in those definitions should be independent of u, and they generally appear as global factors in the derived deviation inequalities. If the only information we have is that $\|X\|$ is bounded a.s. by L, we see that the factors C or C' should be taken of the order of $\sup_{\|u\|=1}(L/\sqrt{\mathrm{Var}[X_u]}) = L\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2}$, which is not acceptable in a high-dimensional setting, and in particular for the application to KME described in Section 1.3, where one might expect that $\|\Sigma^{-1}\|_{\mathrm{op}}$ can get arbitrarily large or even infinite.

Some works (such as Spokoiny and Zhilova, 2013 and the appendix of Spokoiny and Dickhaus, 2015) consider settings going beyond sub-Gaussianity, i.e. when (28) is only required to hold for $\lambda \leq M^{-1}$. This allows in principle for more general variables, e.g. chi-squared type statistics or variables admitting Berstein-

or Bennett-type control of their Laplace transform, while making the constant C' in (28) controlled by a fixed numerical constant. Under this assumption the "first-order" terms are of the correct order, i.e. typically only depend on the variance Σ. Unfortunately, the value of M comes up into additional terms, and since its value has to be independent of u, in the bounded setting the uniformity with respect to u means that M should be again taken of the order of $\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2}$.

To summarize, despite our best efforts we were not able to derive from existing general results, working under the (possibly extended) sub-Gaussian assumption, a concentration in the bounded setting that would not involve $\|\Sigma^{-1}\|_{\mathrm{op}}$, and this is the reason why we treated it separately with tools specific to bounded variables such as the Bousquet-Talagrand inequality. It would be of course of notable interest to obtain results under a general sub-Gaussian assumption $\sup_{\|u\|=1}\|X_u\|_{\psi_2} \leq L$, and control deviations only involving various norms of Σ for the main terms, possibly L for smaller-order terms, but not depending on $\|\Sigma^{-1}\|_{\mathrm{op}}$.

Perspectives. We finally list a few items for future developments.

- It would be interesting to obtain a version of Proposition (11) where the main term does not involve the bound L.
- A recent trend of research developed "robust" exponential concentration bounds for estimators of scalars and vectors with minimal moments assumptions (see e.g. Lugosi and Mendelson, 2019 for a survey of recent advances). It seems a very interesting question to study if such robust procedures can be pushed to the testing setting and enjoy similar nonasymptotic controls to the Gaussian and bounded settings under much relaxed distributional assumptions. Preliminary calculations seem to indicate that the "median-of-means" (MoM) approach can be applied to U-statistics without particular problems and that Assumption 1 can be granted for MoM versions of U-statistics under the assumption of existing moments of order 4, and presumably Assumption 2 under moments of order 8.
- We have analyzed here quantile estimation by direct estimation of unknown quantities coming into the quantile bounds. In practice, quantile estimation by some form of resampling procedure would be often sharper and preferred. V. Spokoiny also made notable recent contributions to this topic (Naumov et al., 2019; Spokoiny and Zhilova, 2015). In the setting of two-sample testing where the null hypothesis is strict equality, it is possible to obtain tests with exact nonasymptotic level based on permutation tests and variations thereof; see Fromont et al. (2012) for such approaches for testing equality of distributions based on the KME methodology, and Kim et al. (2020) for recent broad results on minimax optimality for the power of permutation-based tests. Estimating quantiles via bootstrap procedures is also an interesting direction to pursue in setting, in the case where the null hypothesis is based on closeness rather than equality of signals (so that exact permutation tests do not apply).

– Lower bounds establishing the optimality of the separation rates appearing have been established in the Gaussian case in Theorem 7. It would be nice find such a lower bound in the bounded case.

3 Proofs

The proofs of some of the technical results, first stated without justification along the text, can be found in Section 3.7. We first state a standard technical lemma which we will use several times in the following proofs.

Lemma 15. *Let $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$, then*

$$-\min\left(\sqrt{b}, \frac{|b|}{a}\right) \leq \sqrt{(a^2+b)_+} - a \leq \min\left(\sqrt{|b|}, \frac{|b|}{2a}\right). \tag{29}$$

3.1 Proof of Theorem 3

Let us denote $D := \|\mu - \nu\|$. Under (H_0) we have $D \leq \eta$ and thus:

$$\begin{aligned}
\mathbb{E}_{H_0}[T] &= \mathbb{P}_{H_0}\left[U - \eta^2 > 2\eta\widehat{Q}_1 + 2\widehat{Q}_2\right] \\
&\leq \mathbb{P}_{H_0}\left[U > D^2 + Dq_1 + q_2\right] \\
&\qquad + \mathbb{P}_{H_0}\left[\left|q_1 - \widehat{Q}_1\right| > q_1/2\right] + \mathbb{P}_{H_0}\left[\left|q_2 - \widehat{Q}_2\right| > q_2/2\right] \\
&\leq 3\alpha\,,
\end{aligned}$$

where we have used Assumptions 1 and 2.

We will prove below that under (H_1), we have

$$\mathbb{P}_{H_1}\left[D^2 - Dq_1(u) - q_2(u) \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] \leq 2\alpha\,, \tag{30}$$

which entails:

$$\begin{aligned}
\mathbb{P}_{H_1}[T=0] &= \mathbb{P}_{H_1}\left[U - \eta^2 \leq 2\eta\widehat{Q}_1 + 2\widehat{Q}_2\right] \\
&\leq \mathbb{P}_{H_1}\left[U \leq D^2 - Dq_1 - q_2\right] \\
&\qquad + \mathbb{P}_{H_1}\left[D^2 - Dq_1 - q_2 \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] \\
&\leq 3\alpha\,,
\end{aligned}$$

and the proof is complete. We now prove inequality (30). Let us first solve the following quadratic inequality in $Z \geq 0$:

$$Z^2 - Zq_1 - q_2 \geq \eta^2 + 3\eta q_1 + 3q_2\,. \tag{31}$$

The equation is satisfied when

$$Z \geq \frac{q_1 + \sqrt{(2\eta + 3q_1)^2 + 16q_2}}{2}\,;$$

furthermore, by Lemma 15 and the assumed inequality (11), we have that

$$\frac{q_1 + \sqrt{(2\eta + 3q_1)^2 + 16q_2}}{2} \leq \eta + 2q_1 + \min\left(2\sqrt{q_2}, \frac{2q_2}{\eta}\right) \leq \eta + \delta\,.$$

Under (H_1), $D \geq \eta + \delta$, so D satisfies equation (31). We conclude by remarking that, using Assumption 2:

$$\begin{aligned}\mathbb{P}_{H_1}\left[D^2 - Dq_1(u) - q_2(u) \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] & \\ &\leq \mathbb{P}\left[\eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2 \geq \eta^2 + 3\eta q_1 + 3q_2\right] \\ &\leq 2\alpha\,.\end{aligned}$$

□

3.2 Proof of Propositions 6 and 9

As much for the Gaussian case as for the bounded case, we will give concentration bounds for the statistic U defined in (12), by decomposing the statistic in four parts. Let us define:

$$\begin{aligned}U_{\mathbb{X}} &:= \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \langle X_i - \mu, X_j - \mu \rangle\,, \\ U_{\mathbb{Y}} &:= \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \langle Y_i - \nu, Y_j - \nu \rangle\,, \\ U_{\mathbb{X},\mathbb{Y}} &:= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle X_i - \mu, Y_j - \nu \rangle, \\ U_* &:= \left\langle \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) - \frac{1}{m} \sum_{j=1}^{m} (Y_j - \nu), \mu - \nu \right\rangle.\end{aligned}$$

We have that

$$U = \|\mu - \nu\|^2 - 2U_* + U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}\,. \tag{32}$$

Gaussian setting. We first need some results on Gaussian variables. The first result is a decoupling theorem of Vershynin (2018).

Proposition 16 (Vershynin, 2018, Theorem 6.1.1). *Let $X_1, \ldots, X_n$ be independent centered and weakly (i.e. Pettis) integrable vectors in a Hilbert space, $(a_{ij})_{1 \leq i,j \leq n}$ a family of real numbers and $F : \mathbb{R} \mapsto \mathbb{R}$ a convex function. Then*

$$\mathbb{E}\left[F\left(\sum_{i \neq j} a_{ij} \langle X_i, X_j \rangle\right)\right] \leq \mathbb{E}\left[F\left(4 \sum_{i,j} a_{ij} \langle X_i, X'_j \rangle\right)\right],$$

where (X'_i) is an independent copy of (X_i).

The following lemma is standard; see e.g. Birgé (2001), Lemma 8.2.

Lemma 17. *Let X a real random variable such that for all $0 < t < b^{-1}$:*

$$\log\bigl(\mathbb{E}\bigl[e^{tX}\bigr]\bigr) \leq \frac{(at)^2}{1-bt}\,,$$

where a and b are two positive constants. Then, for all $t \geq 0$:

$$\mathbb{P}\Bigl[X \geq 2a\sqrt{t} + bt\Bigr] \leq e^{-t}\,.$$

Proposition 18. *Let X and Y be two independent Gaussian vectors following the distributions $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(0,S)$ respectively. Then for $t < (\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}})^{-1/2}$:*

$$\log \mathbb{E}[\exp(t\langle X, Y\rangle)] \leq \frac{t^2 \operatorname{Tr}(S\Sigma)}{2(1 - t\sqrt{\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}})}\,.$$

Using Lemma 17, for all $u \geq 0$:

$$\mathbb{P}\left[\langle X, Y\rangle \geq \sqrt{2\operatorname{Tr}(S\Sigma)u} + \sqrt{\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}}u\right] \leq e^{-u}.$$

We can now prove Proposition 6. The samples $\mathbb{X}$ and $\mathbb{Y}$ have respective distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\nu,S)$. We will obtain a concentration inequality for U using its decomposition (32).

Let us first find concentration inequalities for $U_{\mathbb{X}}$ and $U_{\mathbb{Y}}$. Using decoupling (see Proposition 16) we have for all $t < (4\|\Sigma\|_{\mathrm{op}})^{-1}$:

$$\mathbb{E}[\exp(tn(n-1)U_{\mathbb{X}})] \leq \mathbb{E}\left[\exp\left(4t\left\langle \sum_{i=1}^n X_i - \mu, \sum_{i=1}^n X_i' - \mu\right\rangle\right)\right],$$

where X_i' are independent copies of the X_is. Then using Proposition 18, it holds with probability at least $1 - 2e^{-u}$:

$$n(n-1)|U_{\mathbb{X}}| \leq 4n\Bigl(\sqrt{2\operatorname{Tr}\Sigma^2 u} + \|\Sigma\|_{\mathrm{op}}u\Bigr)\,. \tag{33}$$

The same method works for $U_{\mathbb{Y}}$. The concentration of $U_{\mathbb{X},\mathbb{Y}}$ is directly obtained using Proposition 18. Finally U_* is a centered 1-dimensional Gaussian with variance $(\mu-\nu)^T\bigl(\frac{\Sigma}{n} + \frac{S}{m}\bigr)(\mu-\nu)$ and we use the classical bound $\mathbb{P}\bigl[|N| \geq \sigma\sqrt{2t}\bigr] \leq 2e^{-t}$ for $N \sim \mathcal{N}(0,\sigma^2)$. Thus we obtain that with probability at least $1 - 8e^{-u}$:

$$\begin{aligned}\bigl|U - \|\mu-\nu\|^2\bigr| \leq{}& \frac{4}{n-1}\Bigl(\sqrt{2\operatorname{Tr}\Sigma^2 u} + \|\Sigma\|_{\mathrm{op}}u\Bigr) + \frac{4}{m-1}\Bigl(\sqrt{2\operatorname{Tr}S^2 u} + \|S\|_{\mathrm{op}}u\Bigr)\\ &+ \frac{4}{\sqrt{nm}}\Bigl(\sqrt{2\operatorname{Tr}\Sigma S u} + (\|\Sigma\|_{\mathrm{op}}\|S\|_{\mathrm{op}})^{\frac{1}{2}}u\Bigr)\\ &+ \sqrt{2(\mu-\nu)^T\left(\frac{\Sigma}{n} + \frac{S}{m}\right)(\mu-\nu)u}\,.\end{aligned}$$

We conclude by upper bounding the operator norms $\|\Sigma\|_{\mathrm{op}}$ and $\|S\|_{\mathrm{op}}$ by $\sqrt{\operatorname{Tr}\Sigma^2}$ and $\sqrt{\operatorname{Tr} S^2}$ and for the third term we use that

$$(2\operatorname{Tr}(\Sigma S))^{\frac{1}{2}} \leq (4\operatorname{Tr}\Sigma^2 \operatorname{Tr} S^2)^{\frac{1}{4}} \leq (\operatorname{Tr}\Sigma^2)^{\frac{1}{2}} + (\operatorname{Tr} S^2)^{\frac{1}{2}}.$$

We finally use $(n-1)^{-1} \leq 2n^{-1}$ for $n \geq 2$ and similarly for m. It is easy to check that the fourth term is upper bounded by q_1 defined in (14). It just remains to use that $u \geq 1$ to get $u \geq \sqrt{u}$ and (13).

Bounded setting. The concentration of U is obtained in the bounded setting using a concentration inequality for degenerate U-statistics of Houdré and Reynaud-Bouret (2003). We present here a somewhat simplified version suited for our purpose[5].

Theorem 19 (Houdré and Reynaud-Bouret, 2003, Theorem 3.4). *Let $T_1, \ldots, T_N$ be independent random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Borel space $(\mathcal{T}, \mathcal{G})$. Let*

$$U_N = \sum_{i=2}^{N} \sum_{j=1}^{i-1} g_{i,j}(T_i, T_j)\,,$$

where $g_{i,j} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ are measurable Borelian functions satisfying

$$\mathbb{E}[g_{i,j}(T_i, T_j)|T_i] = \mathbb{E}[g_{i,j}(T_i, T_j)|T_j] = 0\,.$$

Let us suppose that the following quantities are finite

$$A := \sup_{t,t',i,j} |g_{i,j}(t,t')|\,,$$

$$B^2 := \max\left\{ \sup_{t,i}\left(\sum_{j=1}^{i-1} \mathbb{E}[g_{i,j}(t,T_j)^2]\right), \sup_{t,j}\left(\sum_{i=j+1}^{n} \mathbb{E}[g_{i,j}(T_i,t)^2]\right)\right\},$$

$$C^2 := \sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}[g_{i,j}(T_i,T_j)^2]\,.$$

Then for all $u > 0$:

$$\mathbb{P}\left[U_N \geq 4C(\sqrt{2u} + 2\sqrt{2}u) + 202Bu^{3/2} + 196Au^2\right] \leq 2.77e^{-u}\,. \qquad (34)$$

Let us prove Proposition 9. We recall that we suppose here that the samples $\mathbb{X}$ and $\mathbb{Y}$ are both bounded by L. To obtain a deviation inequality for the statistic U, we consider separately the statistics $U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}$ and then U_*.

[5] In the original result the u deviation term involves an additional constant D and we simply use $D \leq C$ here.

Using Theorem 19 with $N = n+m$, $T_i := X_i - \mu$ for $1 \le i \le n$ and $T_i = Y_i - \nu$ for $n+1 \le i \le n+m$, $\mathcal{T} = \{u : \|u\| \le 4L^2\}$ and

$$g_{ij}(\cdot,\cdot) = \begin{cases} \frac{1}{n(n-1)}\langle \cdot, \cdot \rangle, & \text{if } 1 \le i,j \le n, \\ \frac{1}{m(m-1)}\langle \cdot, \cdot \rangle, & \text{if } n+1 \le i,j \le n+m, \\ -\frac{1}{nm}\langle \cdot, \cdot \rangle, & \text{otherwise,} \end{cases}$$

we get that with probability greater than $1 - 5.54e^{-u}$:

$$|U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}|/2 \le 307\left(\frac{\sqrt{\operatorname{Tr}\Sigma^2}}{n} + \frac{\sqrt{\operatorname{Tr} S^2}}{m}\right)u + 1854L^2u^2\,. \tag{35}$$

To obtain the above, we have upper bounded A, B, C by:

$$A \le \frac{8L^2}{(n \wedge m)^2}\,, \quad B^2 \le \frac{8L^2}{(n \wedge m)^2}\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right),$$
$$C^2 = \frac{3}{2}\left(\frac{\operatorname{Tr}\Sigma^2}{n} + \frac{\operatorname{Tr} S^2}{m}\right);$$

then, using that $2\sqrt{ab} \le a + b$ and that $\|\Sigma\|_{\mathrm{op}} \le \sqrt{\operatorname{Tr}\Sigma^2}$, we get (35).

For U_*, we use Bernstein's inequality (i.e. combining Lemmas 32 and 17) to get that with probability at least $1 - 2e^{-u}$, it holds:

$$|U_*| \le \|\mu - \nu\|\left(\sqrt{2\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right)u} + \frac{2Lu}{3n \wedge m}\right). \tag{36}$$

Combining (35) and (36), we obtain the claim of Proposition 9. □

3.3 Proof of Theorem 7

The upper bound is directly obtained using Theorem 3. Assumption 1 is satisfied as a consequence of Proposition 6. We do not consider estimation of nuisance parameters related to the covariance matrix Σ which is assumed to be fixed and known for this result; thus Assumption 2 is trivially satisfied by taking $\widehat{Q}_1 = q_1(\Sigma, \alpha)$, $\widehat{Q}_2 = q_2(\Sigma, \alpha)$.

Let us now prove the lower bound (17). The following proof is an adaptation to the non-isotropic Gaussian setting of the proof of Theorem 5.1 in Blanchard et al. (2018). Let $\alpha \in (0,1)$, ans Σ be a positive semidefinite matrix. Without loss of generality, we can assume that Σ is diagonal: $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \ge \ldots \ge \lambda_d > 0$. Let us denote $\mathbb{P}_{\mu,\Sigma}$ the distribution of $\mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and introduce the Gaussian mixture distribution:

$$\mathbb{Q}^n_\Sigma := \frac{1}{2^{d-1}} \sum_{m \in \mathcal{M}} \mathbb{P}^{\otimes n}_{m,\Sigma}, \tag{37}$$

where

$$\mathcal{M} = \left\{(\lambda_1 v_1 h, \ldots, \lambda_{d-1} v_{d-1} h, \eta) \middle| \, v \in \{-1,1\}^{d-1}\right\}.$$

We take $h^2 := \frac{(\eta+\delta)^2-\eta^2}{\operatorname{Tr}\Sigma^2-\lambda_d^2}$. Then, for all $m \in \mathcal{M}$,

$$\|m\|_d = \sqrt{\eta^2 + (\operatorname{Tr}\Sigma^2 - \lambda_d^2)h^2} = \eta + \delta\,.$$

Let $\nu = (0, \ldots, \eta)$, it holds

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}^{\otimes n}(\phi = 1) + \sup_{\mathbb{P}\in\mathcal{A}_\delta} \mathbb{P}^{\otimes n}(\phi = 0) &\geq \mathbb{P}_{\nu,\Sigma}^{\otimes n}(\phi = 1) + \mathbb{Q}_\Sigma^n(\phi = 0) \\ &\geq 1 - \frac{1}{2}\left\|\mathbb{P}_{\nu,\Sigma}^{\otimes n} - \mathbb{Q}_\Sigma^n\right\|_{\mathrm{TV}} \\ &\geq 1 - \frac{1}{2}\left(\int_{\mathbb{R}^{d\times n}} \left(\frac{d\mathbb{Q}_\Sigma^n}{d\mathbb{P}_{\nu,\Sigma}^{\otimes n}}\right)^2 d\mathbb{P}_{\nu,\Sigma}^{\otimes n} - 1\right)^{\frac{1}{2}}.\end{aligned} \tag{38}$$

see for instance Baraud (2002). For a tensor product of Gaussian distributions with fixed, equal covariance, the empirical mean is a sufficient statistic because the Radon-Nikodym derivative of a tensor product of Gaussian measures w.r.t. the Lebesgue measure can be written for $x_1, \ldots, x_n \in \mathbb{R}^d$ as

$$\frac{d\mathbb{P}_{m,\Sigma}^{\otimes n}}{d\lambda^{\otimes n}}(x_1, \ldots, x_n) = \phi_{m,\Sigma/n}(\bar{x})F_\Sigma(x_1, \ldots, x_n)\,,$$

where $\bar{x}$ is the mean of the x_is, $\phi_{m,\Sigma/n}$ is the p.d.f. of a normal $\mathcal{N}(m, \Sigma/n)$ variable, and F_Σ is a function of $(x_1, \ldots, x_n)$ which only depends on Σ. Therefore

$$\frac{d\mathbb{Q}_\Sigma^n}{d\mathbb{P}_{\nu,\Sigma}^{\otimes n}}(x_1, \ldots, x_n) = \frac{d\mathbb{Q}_{\Sigma/n}^1}{d\mathbb{P}_{\nu,\Sigma/n}}(\bar{x}),$$

and thus

$$\int_{\mathbb{R}^{d\times n}} \left(\frac{d\mathbb{Q}_\Sigma^n}{d\mathbb{P}_{\nu,\Sigma}^{\otimes n}}\right)^2 d\mathbb{P}_{\nu,\Sigma}^{\otimes n} = \int_{\mathbb{R}^d} \left(\frac{d\mathbb{Q}_{\Sigma/n}^1}{d\mathbb{P}_{\nu,\Sigma/n}}\right)^2 d\mathbb{P}_{\nu,\Sigma/n}\,.$$

Thus the problem boils down to studying a single Gaussian vector of covariance Σ/n; for the following we will assume $n = 1$ and replace at the end Σ by Σ/n. Let us compute the densities F_ν and Q of these two distributions. For $x \in \mathbb{R}^d$:

$$F_\nu(x) = \left(\det \Sigma(2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d - \eta)^2\right) \prod_{i=1}^{d-1} \exp\left(-\frac{x_i^2}{2\lambda_i}\right),$$

and

$$\begin{aligned} Q(x) &= \left(\det \Sigma (2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d-\eta)^2\right) \\ &\qquad \times \frac{1}{2^{d-1}} \sum_{\substack{v_i\in\{-1,1\}\\ 1\le i\le d-1}} \prod_{i=1}^{d-1} \exp\left(-\frac{1}{2\lambda_i}(x_i - h\lambda_i v_i)^2\right) \\ &= \left(\det \Sigma (2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d-\eta)^2 - \frac{h^2}{2}\sum_{i=1}^{d-1}\lambda_i\right) \\ &\qquad \times \prod_{i=1}^{d-1} \exp\left(-\frac{x_i^2}{2\lambda_i}\right)\cosh(hx_i)\,. \end{aligned}$$

Using that $\mathbb{E}\big[\cosh^2(aZ)\big] = \exp(a^2\sigma^2)\cosh(a^2\sigma^2)$ when $Z\sim\mathcal{N}(0,\sigma^2)$, we have that

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{Q(x)^2}{F_\nu(x)}dx &= \left(\det \Sigma (2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-h^2\sum_{i=1}^{d-1}\lambda_i\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\lambda_d}(x_d-\eta)^2\right)dx_d \\ &\qquad \times \prod_{i=1}^{d-1}\int_{\mathbb{R}} \cosh^2(hx_i)\exp\left(-\frac{x_i^2}{2\lambda_i}\right)dx_i \\ &= \exp\left(-h^2\sum_{i=1}^{d-1}\lambda_i\right)\prod_{i=1}^{d-1}\exp(h^2\lambda_i)\cosh(h^2\lambda_i) \\ &= \prod_{i=1}^{d-1}\cosh(h^2\lambda_i). \end{aligned}$$

By Taylor expansion, we obtain the bound

$$h^2\lambda_i \le 1 \Rightarrow \cosh(h^2\lambda_i) \le 1 + \frac{e}{2}\lambda_i^2 h^4\,.$$

From this and the definition of h we deduce:

$$\log\prod_{i=1}^{d-1}\cosh(h^2\lambda_i) \le \frac{e}{2}(\operatorname{Tr}\Sigma^2 - \lambda_d^2)h^4 = \frac{e}{2(\operatorname{Tr}\Sigma^2-\lambda_d^2)}\left((\eta+\delta)^2-\eta^2\right)^2.$$

The end of the proof follows the same steps as the proof of Theorem 5.1 of Blanchard et al., 2018. That leads us to the final result: if

$$\delta \le \sqrt{\|\Sigma\|_{\mathrm{op}}\sqrt{d_*-1}s + \eta^2} - \eta \quad \text{where} \quad s := \sqrt{\frac{2}{e}\log(1+4(1-\alpha)^2)}\,,$$

and

$$d_* \ge 1 + \frac{2}{e}\ln(5), \quad \text{i.e.} \quad d_* \ge 3\,,$$

then using (38)

$$\sup_{\mathbb{P}\in\mathcal{H}_0}\mathbb{P}[\phi=1]+\sup_{\mathbb{P}\in\mathcal{A}_\delta}\mathbb{P}[\phi=0]>\alpha\,.$$

It follows

$$\begin{aligned}\delta^* &\geq \sqrt{\|\Sigma\|_{\mathrm{op}}\sqrt{d_*-1}s+\eta^2}-\eta\\ &\geq 2^{-\frac{3}{2}}\min\left(\sqrt{s\|\Sigma\|_{\mathrm{op}}}(d_*-1)^{\frac{1}{4}}, s\|\Sigma\|_{\mathrm{op}}\frac{(d_*-1)^{\frac{1}{2}}}{\eta}\right),\end{aligned}$$

and we obtain the inequality corresponding to the second part of the maximum in the right-hand side of (17) by using that $s\geq(1-\alpha)$ and that $d_*-1\geq 2d_*/3$ because $d_*\geq 3$.

Let us prove now that $\delta^*\gtrsim\sqrt{\|\Sigma\|_{\mathrm{op}}}$. Let us consider the eigenvector e_1 associated to the maximum eigenvalue $\|\Sigma\|_{\mathrm{op}}$. Then $\mathbb{P}_{(\eta+\delta)e_1,\Sigma}\in\mathcal{H}_0$ and $\mathbb{P}_{(\eta+\delta)e_1,\Sigma}\in\mathcal{A}_\delta$. Let us denote $\lambda_1=\|\Sigma\|_{\mathrm{op}}/n$, we have:

$$\begin{aligned}\int_{\mathbb{R}^d}\left(\frac{d\mathbb{P}^{\otimes n}_{(\eta+\delta)e_1,\Sigma}}{d\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma}}\right)^2 d\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma} &= \int_{\mathbb{R}^d}\left(\frac{d\mathbb{P}_{(\eta+\delta)e_1,\Sigma/n}}{d\mathbb{P}_{\eta e_1,\Sigma/n}}\right)^2 d\mathbb{P}_{\eta e_1,\Sigma/n}\\ &= \frac{e^{-\delta^2/\lambda_1}}{\sqrt{\lambda_1 2\pi}}\int_{\mathbb{R}}\exp\left(-\frac{(x-\eta)^2}{2\lambda_1}\right)\exp\left(\frac{2\delta(x-\eta)}{\lambda_1}\right)dx\\ &= \exp\left(\frac{3\delta^2}{\lambda_1}\right).\end{aligned}$$

If $\delta\leq\sqrt{\frac{\lambda_1}{3}\log(1+4(1-\alpha)^2)}$, then using (38)

$$\sup_{\mathbb{P}\in\mathcal{H}_0}\mathbb{P}[\phi=1]+\sup_{\mathbb{P}\in\mathcal{A}_\delta}\mathbb{P}[\phi=0]>\alpha\,.$$

It follows that:

$$\delta^*\geq\sqrt{\|\Sigma/n\|_{\mathrm{op}}(1-\alpha)}.$$

3.4 Proof of Theorem 8

This proof is similar to the proof of Theorem 7, so some details will be skipped. As in the one-sample case the upper bound is directly obtained using Theorem 3 and Proposition 6. We just additionally use the following upper bounds:

$$\begin{aligned}\frac{\sqrt{\mathrm{Tr}\,\Sigma^2}}{n}+\frac{\sqrt{\mathrm{Tr}\,S^2}}{m} &\leq \sqrt{2}\sqrt{\frac{\mathrm{Tr}\,\Sigma^2}{n^2}+\frac{\mathrm{Tr}\,S^2}{m^2}} \leq \sqrt{2}\sqrt{\mathrm{Tr}\left(\frac{\Sigma}{n}+\frac{S}{m}\right)^2};\\ \frac{\|\Sigma\|_{\mathrm{op}}}{n}+\frac{\|S\|_{\mathrm{op}}}{m} &\leq 2\max\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n},\frac{\|S\|_{\mathrm{op}}}{m}\right)\leq 2\left\|\frac{\Sigma}{n}+\frac{S}{m}\right\|_{\mathrm{op}},\end{aligned}$$

where the last inequality holds because Σ, S are both positive semidefinite.

The lower bound in the two-sample case is a direct consequence of the one-sample case, by reduction to the case where one of the two sample means is known, say equal to zero. More specifically, let Σ and S be two symmetric positive semidefinite matrices, we consider again the distribution $\mathbb{Q}^n_\Sigma$ defined in (37). Then

$$\int_{\mathbb{R}^{d\times(n+m)}} \left(\frac{d\mathbb{Q}^n_\Sigma \otimes \mathbb{P}^{\otimes m}_{0,S}}{d\mathbb{P}^{\otimes n}_{\nu,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S}}\right)^2 d\mathbb{P}^{\otimes n}_{\nu,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S} = \int_{\mathbb{R}^{d\times n}} \left(\frac{d\mathbb{Q}^n_\Sigma}{d\mathbb{P}^{\otimes n}_{\nu,\Sigma}}\right)^2 d\mathbb{P}^{\otimes n}_{\nu,\Sigma}\,.$$

Then using the previous results of the proof of Theorem 7 we obtain that

$$\delta^*(\alpha) \geq \left(n^{-1}\sqrt{\operatorname{Tr}\Sigma^2 - \lambda_d^2}s + \eta^2\right)^{\frac{1}{2}} - \eta, \tag{39}$$

with $s = \sqrt{\frac{2}{e}\log(1+4(1-\alpha)^2)}$. By the same token we obtain that

$$\delta^*(\alpha) \geq \left(m^{-1}\sqrt{\operatorname{Tr}S^2 - \ell_d^2}s + \eta^2\right)^{\frac{1}{2}} - \eta, \tag{40}$$

where ℓ_d is the smallest eigenvalue of the matrix S. Because $d_* \geq 3$, it holds

$$\begin{aligned}\max\big(n^{-2}\big(\operatorname{Tr}\Sigma^2 - \lambda_d^2\big), m^{-2}\big(\operatorname{Tr}S^2 - \ell_d^2\big)\big) &\geq \frac{2}{3}\max\big(n^{-2}\operatorname{Tr}\Sigma^2, m^{-2}\operatorname{Tr}S^2\big)^{\frac{1}{2}} \\ &\geq \frac{1}{6}\operatorname{Tr}\left(\frac{\Sigma}{n} + \frac{S}{m}\right)^2,\end{aligned}$$

and by combining (39) and (40), we obtain that

$$\delta^*(\alpha) \geq (2\sqrt{12})^{-1}\sigma\min\left(\sqrt{s}d_*^{\frac{1}{4}}, s\frac{\sigma d_*^{\frac{1}{2}}}{\eta}\right),$$

where $\sigma = \|\Sigma/n + S/m\|_{\mathrm{op}}$. We obtain (19) using again that $s \geq 1-\alpha$.

The last part of the lower bound is obtained as in the one-sample case using first the distributions $\mathbb{P}^{\otimes n}_{(\eta+\delta)e_1,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S}$ and $\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S}$ where e_1 is still the eigenvector associated to the biggest eigenvalue of Σ. We obtain that $\delta^*(\alpha) \gtrsim \|\Sigma/n\|^{1/2}_{\mathrm{op}}$. By the same token, we obtain that $\delta^*(\alpha) \gtrsim \|S/m\|^{1/2}_{\mathrm{op}}$ and conclude the proof using that $2\max(\|\Sigma/n\|_{\mathrm{op}}, \|S/m\|_{\mathrm{op}}) \geq \|\Sigma/n + S/m\|_{\mathrm{op}}$.

3.5 Proof of Propositions 10 and 11

We want to obtain a concentration inequality for the estimator $\sqrt{\|\widehat{\Sigma}\|_{\mathrm{op}}}$. To this end, we will first study the following:

$$\widetilde{\Sigma} := \frac{1}{n}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T\,, \tag{41}$$

where μ is the true mean of the sample $\mathbb{X}$. Then we have:

$$\big\|\widehat{\Sigma} - \widetilde{\Sigma}\big\|_{\mathrm{op}} = \big\|-(\mu - \widehat{\mu})(\mu - \widehat{\mu})^T\big\|_{\mathrm{op}} = \|\mu - \widehat{\mu}\|^2\,. \tag{42}$$

Gaussian setting. The concentration of $\|\widetilde{\Sigma}\|_{\mathrm{op}}^{1/2}$ is a consequence of the classical Lipschitz Gaussian concentration property (see e.g. Theorem 3.4 in Massart, 2003).

Theorem 20 (Gaussian Lipschitz concentration). *Let $X = (x_1, \ldots, x_d)$ be a vector of i.i.d. standard Gaussian variables, and $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a L-Lipschitz function with respect to the Euclidean norm. Then for all $t \geq 0$:*

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq e^{-\frac{t^2}{2L^2}}. \tag{43}$$

The following corollary is a direct consequence of that theorem (we provide a proof in Section 3.7), which will be used to control the term in (42).

Corollary 21. *Let X a random Gaussian vector of distribution $\mathcal{N}(\mu, \Sigma)$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\|X\| \geq \sqrt{\|\mu\|^2 + \operatorname{Tr}\Sigma} + \sqrt{2\|\Sigma\|_{\mathrm{op}} u}\right] \leq e^{-u}\,. \tag{44}$$

We will use the results of Koltchinskii and Lounici (2017) giving an upper bound of the expectation of the operator norm of the deviations of $\widetilde{\Sigma}$ from its expectation. The constants come from the improved version given by van Handel (2017).

Theorem 22 (van Handel, 2017). *Let $\mathbb{X} = (X_i)_{1 \leq i \leq n}$ a sample of independent Gaussian vectors of distribution $\mathcal{N}(0, \Sigma)$, then*

$$\mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}}\right] \leq \|\Sigma\|_{\mathrm{op}}\left((2+\sqrt{2})\sqrt{\frac{d_e}{n}} + 2\frac{d_e}{n}\right), \tag{45}$$

where $d_e = \operatorname{Tr}\Sigma / \|\Sigma\|_{\mathrm{op}}$ and $\widetilde{\Sigma}$ is defined in equation (41).

We can now prove a concentration inequality for $\|\widetilde{\Sigma}\|_{\mathrm{op}}^{1/2}$.

Proposition 23. *Let $\mathbb{X} = (X_i)_{1 \leq i \leq n}$ a sample of independent $\mathcal{N}(\mu, \Sigma)$ Gaussian vectors, then for $u \geq 0$, with probability at least $1 - 2e^{-u}$:*

$$\left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 2\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}}\,, \tag{46}$$

where $\widetilde{\Sigma}$ is defined in (41).

Remark 24. In (46), the lower and upper bounds have been brought together, but the lower bound is in fact slightly better than the upper bound. This is due to the lower bound of the expectation where $\operatorname{Tr}\Sigma$ can be replaced by $\|\Sigma\|_{\mathrm{op}}$, see (49) below.

Proof. We remark that

$$\begin{aligned}\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} &= \sup_{\|u\|_d=1} \sqrt{u^t \widetilde{\Sigma} u} \\ &= \sup_{\|u\|_d=1} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \langle u, X_i - \mu \rangle^2 \right)^{\frac{1}{2}} \\ &= \sup_{\|u\|_d=1} \sup_{\|v\|_n=1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle u, X_i - \mu \rangle v_i \\ &\overset{\text{dist}}{\sim} \sup_{\|u\|_d=1} \sup_{\|v\|_n=1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle u, \Sigma^{\frac{1}{2}} g_i \right\rangle v_i \,,\end{aligned}$$

where $(g_i)_{i=1\dots n}$ are i.i.d. standard Gaussian vectors and $\|\cdot\|_p$ for $p \in \mathbb{N}$ is defined as the Euclidean norm in $\mathbb{R}^p$. Let u and v be unit vectors in $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively and $f_{u,v} : \mathbb{R}^{d\times n} \to \mathbb{R}$:

$$f_{u,v}(y) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle u, \Sigma^{\frac{1}{2}} y_i \right\rangle v_i \,, \quad y \in \mathbb{R}^{d\times n} \,.$$

These functions are Lipschitz: indeed for all $z, y \in \mathbb{R}^{d\times n}$ we have:

$$\begin{aligned} f_{u,v}(y) - f_{u,v}(z) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\langle u, \Sigma^{\frac{1}{2}} (y_i - z_i) \right\rangle v_i \le \frac{1}{\sqrt{n}} \sum_{i=1}^n \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \|y_i - z_i\|_d |v_i| \\ &\le \frac{\|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}}{\sqrt{n}} \sqrt{\sum_{i=1}^n \|y_i - z_i\|_d^2} = \frac{\|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}}{\sqrt{n}} \|y - z\|_{d\times n} \,. \end{aligned} \tag{47}$$

A supremum of Lipschitz functions is Lipschitz, thus we can use the Gaussian Lipschitz concentration (Theorem 20), and get for all $x \ge 0$:

$$\mathbb{P}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} \right] \ge \sqrt{\frac{2x\|\Sigma\|_{\mathrm{op}}}{n}} \right] \le e^{-x} \,, \tag{48}$$

with the same control for lower deviations.

It remains to upper bound $\left| \mathbb{E}\left[\|\widetilde{\Sigma}\|_{\mathrm{op}}^{1/2} \right] - \|\Sigma\|_{\mathrm{op}}^{1/2} \right|$. For one direction, using Jensen's and triangle inequalities and inequality (29), we get:

$$\begin{aligned} \mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} \right] - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} &\le \sqrt{\|\Sigma\|_{\mathrm{op}} + \mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}} \right]} - \sqrt{\|\Sigma\|_{\mathrm{op}}} \\ &\le \min\left(\sqrt{\mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}} \right]}, \frac{\mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}} \right]}{2\sqrt{\|\Sigma\|_{\mathrm{op}}}} \right) \\ &\le 2\sqrt{\frac{2\,\mathrm{Tr}\,\Sigma}{n}} \,. \end{aligned}$$

For the last inequality, we have used Theorem 22 for the expectation, then the fact that $\min\Big((a\sqrt{x}+bx)^{1/2}, (a\sqrt{x}+bx)/2\Big) \leq \max(\sqrt{a+b}, (a+b)/2)\sqrt{x}$ where $a = 2+\sqrt{2}$, $b = 2$ and $x = d_e/n$. This is achieved by treating cases $x \leq 1$ and $x \geq 1$ separately.

For the other direction, a reformulation of (48) is that there exists a random variable $g \sim \mathrm{Exp}(1)$ such that:

$$\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}} \leq \mathbb{E}\Big[\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}}\Big] + \sqrt{\frac{2g\|\Sigma\|_{\mathrm{op}}}{n}}\,.$$

Taking the square then the expectation and then applying Jensen's inequality to the concave function $x \mapsto (a+b\sqrt{x})^2$ $(a, b \geq 0)$, we obtain:

$$\begin{aligned}\|\Sigma\|_{\mathrm{op}} \leq \mathbb{E}\Big[\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}\Big] &\leq \mathbb{E}_g\left[\left(\mathbb{E}\Big[\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}}\Big] + \sqrt{\frac{2g\|\Sigma\|_{\mathrm{op}}}{n}}\right)^2\right] \\ &\leq \left(\mathbb{E}\Big[\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}}\Big] + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}}{n}}\right)^2,\end{aligned}$$

and thus

$$\mathbb{E}\Big[\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}}\Big] - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \geq -\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}}{n}} \geq -2\sqrt{\frac{2\,\mathrm{Tr}\,\Sigma}{n}}\,. \tag{49}$$

Proof of Proposition 10. It holds

$$\Big|\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\Big| \leq \Big|\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\Big| + \big\|\widehat{\Sigma} - \widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}}\,.$$

Then, from (42):

$$\|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} \leq \|\mu - \widehat{\mu}\|\,.$$

According to Proposition 23 and Corollary 21, we obtain that for $u \geq 0$, with probability at least $1 - 3e^{-u}$:

$$\Big|\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\Big| \leq 2\sqrt{\frac{2\,\mathrm{Tr}\,\Sigma}{n}} + \sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}} + \sqrt{\frac{\mathrm{Tr}\,\Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}u}{n}}\,.$$

So, for $u \geq 0$, with probability at least $1 - 3e^{-u}$:

$$\Big|\big\|\widehat{\Sigma}\big\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\Big| \leq 3\sqrt{\frac{2\,\mathrm{Tr}\,\Sigma}{n}} + 2\sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}}\,.$$

Bounded setting. We first recall the following concentration result for bounded random vectors in the formulation of Bousquet (2002).

Theorem 25 (Talagrand-Bousquet inequality). *Assume $(X_i)_{1\leq i\leq n}$ are i.i.d. with marginal distribution $\mathbb{P}$. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to*

$\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are $\mathbb{P}$-measurable, square-integrable, bounded by M and satisfy $\mathbb{E}[f] = 0$. Then we denote

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Let σ be a positive real number such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} \operatorname{Var}[f(X_1)]$. Then for all $u \geq 0$, $\varepsilon > 0$ we have:

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z](1+\varepsilon) + \sqrt{2un\sigma^2} + \frac{Mu}{3}(1+\varepsilon^{-1}) \right] \leq e^{-u}.$$

The following corollary is a direct consequence of Theorem 25. Some refinement of this result in the same vein (including two-sided deviation control in the uncentered case) can be found in Marienwald et al. (2020) (Proposition 6.2 and Corollary 6.3).

Corollary 26. *Let X_i for $i = 1, \dots, n$ i.i.d. random vectors bounded by L with expectation μ, covariance Σ in a separable Hilbert space $\mathcal{H}$. Then for $u \geq 0$, with probability at least $1 - e^{-u}$:*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right\| \leq 2\sqrt{\frac{\operatorname{Tr} \Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} u}{n}} + \frac{4Lu}{3n}\,.$$

Lemma 27. *Let X_i for $i = 1, \dots, n$ i.i.d. random vectors bounded by L with expectation μ, covariance Σ in a separable Hilbert space $\mathcal{H}$. Then*

$$\mathbb{E}\left[\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} \right] \leq \sqrt{\frac{\operatorname{Var}[\|X_1 - \mu\|^2]}{n}}\,, \tag{50}$$

where $\widetilde{\Sigma}$ is defined in (41).

Remark 28. Using the boundedness of the variables we can upper bound this variance: $\operatorname{Var}\left[\|X_1 - \mu\|^2\right] \leq 4L^2 \operatorname{Tr} \Sigma$.

Proposition 29. *Let $(X_i)_{1 \leq i \leq n}$ be i.i.d. random vectors in a separable Hilbert space $\mathcal{H}$, with norm bounded by L and covariance Σ, then for any for $u \geq 1$, with probability at least $1 - e^{-u}$:*

$$\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} \leq 2\sqrt{\frac{\operatorname{Var}[\|X_1 - \mu\|^2]}{n}} + L\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} u}{n}} + \frac{8L^2 u}{3n}\,, \tag{51}$$

where $\widetilde{\Sigma}$ is defined in (41).

Proof. We denote in this proof $Z_i := X_i - \mu$ for $1 \leq i \leq n$. Let us first remark that if B_1 is the unit ball of $\mathcal{H}$, then:

$$\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} = \sup_{u,v \in B_1} \frac{1}{n} \sum_{i=1}^{n} \left\langle v, \left(Z_i Z_i^T - \Sigma \right) u \right\rangle =: \sup_{u,v \in B_1} \frac{1}{n} \sum_{i=1}^{n} f_{u,v}(X_i)\,.$$

Since the variables X_i have norm bounded by L, it can be assumed equivalently that they take their values in $B_L = LB_1$, and it holds $\sup_{x\in B_L}\sup_{u,v\in B_1} f_{u,v}(x) \leq 8L^2$. Furthermore, since $(u,v) \mapsto f_{u,v}(x)$ is continuous, and the Hilbert space $\mathcal{H}$ is separable, the uncountable set B_1 can be replaced by a countable dense subset. Thus we can apply Theorem 25, and obtain that with probability at least $1-e^{-x}$:

$$\left\|\widetilde{\Sigma}-\Sigma\right\|_{\mathrm{op}} \leq 2\mathbb{E}\Big[\|\widetilde{\Sigma}-\Sigma\|_{\mathrm{op}}\Big] + L\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}x}{n}} + \frac{16L^2x}{3n}\,,$$

where we have used for the variance term:

$$\begin{aligned}\sup_{u,v\in B_1}\mathbb{E}\Big[\big\langle v,\big(Z_iZ_i^T-\Sigma\big)u\big\rangle^2\Big] &\leq \sup_{u,v\in B_1}\mathbb{E}\Big[\langle v,Z_i\rangle^2\langle Z_i,u\rangle^2\Big]\\ &\leq 4nL^2\|\Sigma\|_{\mathrm{op}}\,.\end{aligned}$$

We conclude using the upper bound of the expectation from Lemma 27.

Proof of Proposition 11. As in the Gaussian case, we have:

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq \left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| + \left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

From Lemma 15 and Proposition 29, we have with probability at least $1-e^{-u}$:

$$\left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 4L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + \sqrt{\frac{16}{3}\frac{L^2u}{n}}\,,$$

where we have used that:

$$\sqrt{\frac{\operatorname{Var}[\|Z_1\|^2]}{n}} \leq \frac{2L\sqrt{\operatorname{Tr}\Sigma}}{\sqrt{n}}\,.$$

Using

$$\left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} \leq \|\mu-\widehat{\mu}\|\,,$$

and according to Corollary 26, we obtain that for $u \geq 0$, with probability at least $1-2e^{-u}$:

$$\begin{aligned}\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| &\leq \left(4L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + \sqrt{\frac{16}{3}\frac{L^2u}{n}}\right)\\ &\qquad + \left(2\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}u}{n}} + \frac{4Lu}{3n}\right)\\ &\leq 8L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + 4L\left(\sqrt{\frac{2u}{n}} + \frac{u}{3n}\right),\end{aligned}$$

where we have used for the last inequality that $\|\Sigma\|_{\mathrm{op}} \leq 4L^2$. □

3.6 Proof of Propositions 12 and 13

From a sample $\mathbb{X} = (X_i)_{1\le i\le n}$ of i.i.d. random vectors, we want to estimate $\operatorname{Tr}\Sigma^2$ where Σ is their common covariance matrix. The statistic $\widehat{T}$ defined in (24) is an unbiased estimator of $\operatorname{Tr}\Sigma^2$. This statistic is also invariant by translation.

constant ($\nabla_\mu \tau = 0$).

If we denote $\mathfrak{S}_n$ the set of permutations of $\{1,\dots,n\}$, $\widehat{T}$ can be rewritten as:

$$\widehat{T} = \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n}\frac{1}{\lfloor n/4\rfloor}\sum_{i=1}^{\lfloor n/4\rfloor}\frac{1}{4}\big\langle X_{\sigma(4i)} - X_{\sigma(4i-2)}, X_{\sigma(4i-1)} - X_{\sigma(4i-3)}\big\rangle^2; \qquad (52)$$

namely by symmetry, all the 4-tuples appear the same number of times in the right-hand side, so we just need to divide by the number of terms to obtain the identity (52). We will use this decomposition to obtain a concentration of the statistic $\widehat{T}$ for the Gaussian case and the bounded case, since the inner sum for each fixed permutation is a sum of $\lfloor n/4\rfloor$ i.i.d. terms.

Gaussian setting. Because the statistic is invariant by translation we can assume without loss of generality that $\mu = 0$. To obtain a deviation inequality for $\widehat{T}^{1/2}$, we will first find a concentration inequality for $\widehat{T}$ and then use Lemma 15. We obtain concentration via control of moments of $\widehat{T}$, so we first need some upper bounds on Gaussian moments. The following lemma is proved in Section 3.7.

Lemma 30. *Let $Z_i := \big\langle X_i^1 - X_i^3, X_i^2 - X_i^4\big\rangle^2/4$, where X_i^j for $i = 1,\dots,m$ and $1\le j\le 4$ are i.i.d. Gaussian random vectors $\mathcal{N}(0,\Sigma)$. Then for all $q\in\mathbb{N}$:*

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m Z_i - \operatorname{Tr}\Sigma^2\right)^{2q}\right] \le \left(4\sqrt{2}\phi q^2\frac{\operatorname{Tr}\Sigma^2}{\sqrt{m}}\right)^{2q}, \qquad (53)$$

where $\phi = (1+\sqrt{5})/2$ is the golden ratio.

We deduce from this lemma a concentration inequality for $\widehat{T}$.

Proposition 31. *Let $(X_i)_{1\le i\le n}$, $n\ge 4$ be i.i.d. random vectors with distribution $\mathcal{N}(\mu,\Sigma)$. Then for all $u\ge 0$:*

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge 30\frac{u^2\operatorname{Tr}\Sigma^2}{\sqrt{n}}\right] \le e^4 e^{-u}, \qquad (54)$$

where $\widehat{T}$ is defined in (24).

Proof. Using Lemma 30, (52) and the convexity of the function $x\mapsto x^{2q}$, we can upper bound the moments of $\widehat{T}$:

$$\mathbb{E}\Big[\big(\widehat{T} - \operatorname{Tr}\Sigma^2\big)^{2q}\Big] \le \left(4\sqrt{2}\phi q^2\frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4\rfloor}}\right)^{2q}. \qquad (55)$$

Let $t \geq 0$ and $q \in \mathbb{N}$, then by Markov's inequality

$$\mathbb{P}\Big[\big|\widehat{T} - \operatorname{Tr}\Sigma^2\big| \geq t\Big] \leq t^{-2q}\mathbb{E}\Big[\big(\widehat{T} - \operatorname{Tr}\Sigma^2\big)^{2q}\Big]\,. \tag{56}$$

Let us choose q as:

$$q = \left\lfloor \frac{e^{-1}}{2\sqrt{\phi}2^{\frac{1}{4}}} t^{\frac{1}{2}} \left(\frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4\rfloor}}\right)^{-\frac{1}{2}} \right\rfloor,$$

so that (55), (56) entail

$$\mathbb{P}\Big[\big|\widehat{T} - \operatorname{Tr}\Sigma^2\big| \geq t\Big] \leq e^{-4q}\,.$$

Let us now take

$$t = \frac{e^2\sqrt{2}\phi}{4} u^2 \frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4\rfloor}} \leq 30\frac{u^2 \operatorname{Tr}\Sigma^2}{\sqrt{n}}\,,$$

where we have used $\lfloor n/4\rfloor \geq n/7$ for $n \geq 4$; we obtain that for all $u \geq 0$:

$$\mathbb{P}\left[\big|\widehat{T} - \operatorname{Tr}\Sigma^2\big| \geq 30\frac{u^2 \operatorname{Tr}\Sigma^2}{\sqrt{n}}\right] \leq e^4 e^{-u}\,.$$

□

Proposition 12 directly follows from Proposition 31 and Lemma 15.

Bounded setting. As in the Gaussian case, we first obtain a concentration inequality for $\widehat{T}$ and then using Lemma 15, we obtain one for $\widehat{T}^{1/2}$. We will need the following classical Bernstein's inequality (see for instance Vershynin, 2018, Exercise 2.8.5 for the version below) which gives an upper bound on the Laplace transform of the sum of bounded random variables.

Lemma 32 (Bernstein's inequality). *Let $(X_i)_{1\leq i\leq m}$ be i.i.d. real centered random variables bounded by B such that*

$$\mathbb{E}\big[X_1^2\big] \leq \sigma^2\,.$$

Then for all $t < 3/B$:

$$\log\Big(\mathbb{E}\big[e^{t\sum X_i}\big]\Big) \leq \frac{1}{2}\frac{m\sigma^2 t^2}{1 - Bt/3}\,.$$

Via Bernstein's inequality we obtain the following result.

Proposition 33. *Let $(X_i)_{1\ \leq i\leq n}$, $n \geq 4$ be i.i.d. Hilbert-valued random variables with norm bounded by L and covariance Σ, and $\widehat{T}$ defined by* (24)*. Then for all $t \geq 0$:*

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \geq 8L^2\sqrt{\frac{\operatorname{Tr}\Sigma^2 t}{n}} + \frac{10L^4 t}{n}\right] \leq 2e^{-t}\,. \tag{57}$$

where $\widehat{T}$ is defined in (24)*.*

Proof. Let X, X', Y, Y' be i.i.d. Hilbert-valued random vectors of expectation μ, covariance Σ and with norm bounded by L, and $Z := \langle X - Y, X' - Y' \rangle^2/4$. Then it holds $0 \leq Z \leq 4L^4$, $\mathbb{E}[Z] = \operatorname{Tr} \Sigma^2$ and

$$\begin{aligned} |Z - \mathbb{E}[Z]| &\leq 4L^4; \\ \operatorname{Var}[Z] &\leq 4L^4 \mathbb{E}[Z] = 4L^4 \operatorname{Tr} \Sigma^2. \end{aligned}$$

Now using the convexity of the exponential function, (52) and then Lemma 32, we can upper bound the Laplace transform of $\widehat{T}$ as follows:

$$\log\Big(\mathbb{E}\big[e^{t\widehat{T}}\big]\Big) \leq \frac{1}{2\lfloor n/4 \rfloor} \frac{4L^4 \operatorname{Tr} \Sigma^2 t^2}{1 - 4L^4 t/(3\lfloor n/4 \rfloor)},$$

for all t such that the right-hand sise is well defined, i.e. the denominator is strictly positive. Now using Lemma 17, and $\lfloor n/4 \rfloor \geq n/7$ for $n \geq 4$, for all $t \geq 0$ it holds

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr} \Sigma^2\right| \geq 8L^2 \sqrt{\frac{\operatorname{Tr} \Sigma^2 t}{n}} + \frac{10L^4 t}{n}\right] \leq 2e^{-t}. \tag{58}$$

Proof of Proposition 13. Assuming the event entering into (58) holds, we will use the inequalities of Lemma 15:

$$\begin{aligned} \sqrt{\widehat{T}} - \sqrt{\operatorname{Tr} \Sigma^2} &\leq \sqrt{\operatorname{Tr} \Sigma^2 + 8L^2 \sqrt{\frac{\operatorname{Tr} \Sigma^2 t}{n}}} - \sqrt{\operatorname{Tr} \Sigma^2} + \sqrt{\frac{10L^4 t}{n}} \\ &\leq 4L^2 \sqrt{\frac{t}{n}} + L^2 \sqrt{\frac{10t}{n}} \leq 8L^2 \sqrt{\frac{t}{n}}. \end{aligned}$$

For the other side, we proceed analogously:

$$\begin{aligned} \sqrt{\widehat{T}} - \sqrt{\operatorname{Tr} \Sigma^2} &\geq \sqrt{\left(\operatorname{Tr} \Sigma^2 - 8L^2 \sqrt{\frac{\operatorname{Tr} \Sigma^2 t}{n}}\right)_+} - \sqrt{\operatorname{Tr} \Sigma^2} - \sqrt{\frac{10L^4 t}{n}} \\ &\geq -8L^2 \sqrt{\frac{t}{n}} - L^2 \sqrt{\frac{10t}{n}} \geq -12L^2 \sqrt{\frac{t}{n}}. \end{aligned}$$

□

3.7 Additional proofs

Proof of Lemma 15. This Lemma completes the Lemma 6.1.3 of Blanchard et al. (2018). This is its complete proof.

Let a in $\mathbb{R}_+$, it is well known that for $b \geq -a^2$:

$$a - \sqrt{|b|} \leq \sqrt{a^2 + b} \leq a + \sqrt{|b|}.$$

On the other hand, suppose that $b \geq 0$, the Taylor expansion of the function $b \mapsto \sqrt{a^2+b} - a$ gives that there exists $c \in (0,b)$ such that:

$$\sqrt{a^2+b} - a = \frac{b}{2\sqrt{a^2+c}} \leq \frac{b}{2a}\,.$$

Suppose now that $0 \geq b \geq -a^2$, then

$$\sqrt{a^2+b} \geq a + \frac{b}{a} \Leftrightarrow b \geq 2b + \frac{b^2}{a^2} \Leftrightarrow b \geq -a^2\,.$$

The equation (29) is still true when $b < -a^2$ because then:

$$-a \geq -\sqrt{|b|} \geq -\frac{|b|}{a}\,.$$

□

Proof of Proposition 18. Let g be a standard Gaussian random vector in $\mathbb{R}^d$, and $U^T D U$ be the singular value decomposition of the matrix $S^{1/2}\Sigma S^{1/2}$ where $D = \mathrm{diag}(\lambda_, \ldots, \lambda_d)$. Then we have the following equalities in distribution

$$Y^T \Sigma Y \overset{\text{dist}}{\sim} g^T S^{\frac{1}{2}} \Sigma S^{\frac{1}{2}} g \overset{\text{dist}}{\sim} g^T U^T D U g \overset{\text{dist}}{\sim} g^T D g\,.$$

The last equality is a consequence of the invariance by rotation of Gaussian vectors. Then for $t < 1/\sqrt{\|\Sigma\|_{\mathrm{op}}\|S\|_{\mathrm{op}}}$:

$$\mathbb{E}\Big[e^{t\langle X,Y\rangle}\Big] = \mathbb{E}\left[e^{\frac{t^2\|\Sigma^{\frac{1}{2}}Y\|^2}{2}}\right] = \mathbb{E}\left[e^{\frac{t^2 g^T D g}{2}}\right] = \mathbb{E}\left[\exp\left(\frac{t^2}{2}\sum_{i=1}^d \lambda_i g_i^2\right)\right].$$

Using the independence of the coordinates and that $-\log(1-x) \leq \frac{x}{1-x} \leq \frac{x}{1-\sqrt{x}}$ for $x < 1$ (the first inequality can easily be checked by termwise power series comparison), we obtain:

$$\begin{aligned}\log\Big(\mathbb{E}\Big[e^{t\langle X,Y\rangle}\Big]\Big) &= \sum_{i=1}^n -\frac{1}{2}\log\Big(1 - t\sqrt{\lambda_i}\Big) \\ &\leq \sum_{i=1}^n \frac{1}{2}\frac{t^2\lambda_i}{1-t^2\lambda_i} \leq \frac{1}{2}\frac{t^2\,\mathrm{Tr}(S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}})}{1 - t\|S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}\|_{\mathrm{op}}^{\frac{1}{2}}}\,.\end{aligned}$$

We conclude using that $\mathrm{Tr}(S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}) = \mathrm{Tr}(\Sigma S)$ and that $\|S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}\|_{\mathrm{op}} \leq \|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}$.

□

Proof of Corollary 21. We use the representation $X \overset{\text{dist}}{\sim} (\Sigma^{\frac{1}{2}} g + \mu)$, where g is a standard Gaussian random variable. We then have

$$\|X\|_d \overset{\text{dist}}{\sim} \|\Sigma^{\frac{1}{2}} g + \mu\|_d = f(g)\,,$$

where for $y \in \mathbb{R}^d$:

$$f(y) = \left\| \Sigma^{\frac{1}{2}} y + \mu \right\|_d .$$

This function f is Lipschitz with constant $\|\Sigma^{\frac{1}{2}}\|_{\mathrm{op}}$. We conclude using Theorem 20 and Jensen's inequality:

$$\mathbb{E}\big[\|X\|_d\big] \leq \sqrt{\|\mu\|_d^2 + \operatorname{Tr} \Sigma} .$$

□

Proof of Corollary 26. We apply Theorem 25, with $\varepsilon = 1$ and the set of functions $\mathcal{F} = \{f_u\}_{\|u\|_{\mathcal{H}}=1}$ where $f_u : x \in \mathcal{H} \mapsto \langle x, u \rangle_{\mathcal{H}}$ for $u \in \mathcal{H}$. We can find a countable subset of the unit sphere because $\mathcal{H}$ is separable. Then

$$Z = \sup_{\|u\|_{\mathcal{H}}=1} \sum_{i=1}^{n} \langle X_i - \mu, u \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} X_i - \mu \right\|_{\mathcal{H}} .$$

We conclude using that for all u in the unit sphere of $\mathcal{H}$, $\operatorname{Var}[\langle X_i - \mu, u \rangle_{\mathcal{H}}] \leq \|\Sigma\|_{\mathrm{op}}$ and $|\langle X_i - \mu, u \rangle_{\mathcal{H}}| \leq 2L$ a.s. We use Jensen's inequality to upper bound the expectation: $\mathbb{E}[Z] \leq (n \operatorname{Tr} \Sigma)^{\frac{1}{2}}$. □

Proof of Lemma 27. We upper bound the operator norm with the Frobenius norm. We denote in this proof $Z_i := X_i - \mu$. It holds:

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\Sigma - \widetilde{\Sigma}\big\|_{\mathrm{op}}\Big] \\
&\quad \leq \mathbb{E}\left[\sqrt{\operatorname{Tr}\big(\Sigma - \widetilde{\Sigma}\big)^2}\right] \\
&\quad \leq \left(\mathbb{E}\left[\operatorname{Tr}\left(\frac{1}{n^2}\Big(\sum_i (Z_i Z_i^T)^2 + \sum_{i \neq j} Z_i Z_i^T Z_j Z_j^T\Big) - \widetilde{\Sigma}\Sigma - \Sigma\widetilde{\Sigma} + \Sigma^2\right)\right]\right)^{\frac{1}{2}} \\
&\quad = \left(\frac{\mathbb{E}\big[\|Z\|^4\big]}{n} - \frac{\operatorname{Tr}\Sigma^2}{n}\right)^{\frac{1}{2}} \\
&\quad = \sqrt{\frac{\operatorname{Var}[\|Z\|^2]}{n}} \leq \frac{2L\sqrt{\operatorname{Tr}\Sigma}}{\sqrt{n}} .
\end{aligned}$$

□

Proof Lemma 30. First let us remark that if X and X' are independent $\mathcal{N}(0, \Sigma)$ Gaussian vectors, then

$$\langle X, X' \rangle \overset{\mathrm{dist}}{\sim} \sum_{i=1}^{d} \lambda_i g_i g_i' ,$$

where g_i and g_i' are independent standard Gaussian random variables and the λ_is are the eigenvalues of Σ. Then for $q \in \mathbb{N}$, recalling $\mathbb{E}\big[g_i^{2q}\big] = (2q!)/(2^q q!)$,

$$\begin{aligned}\mathbb{E}\Big[\langle X, X'\rangle^{2q}\Big] &= \sum_{p_1+\ldots+p_d=q} \binom{2q}{2p_1,\ldots,2p_d} \prod_{i=1}^d (\lambda_i)^{2p_i} \left(\frac{(2p_i)!}{2^{p_i} p_i!}\right)^2 \\ &\leq (2q)! \sum_{p_1+\ldots+p_d=q} \prod_{i=1}^d (\lambda_i^2)^{p_i} \\ &\leq (2q)! (\operatorname{Tr}\Sigma^2)^q,\end{aligned}$$

where we have used $(2p)! \leq 2^{2p} p!^2$. Using this bound, we upper bound the moments of the Z_is:

$$\big|\mathbb{E}[Z_i^q]\big| = 2^{-2q}\mathbb{E}\Big[\big\langle X_i^1 - X_i^3, X_i^2 - X_i^4\big\rangle^{2q}\Big] \leq (2q)!(\operatorname{Tr}\Sigma^2)^q\,.$$

We now upper bound the moments of $Z_i - \operatorname{Tr}\Sigma$. Let Z_i' be an independent copy of Z_i, then since $\mathbb{E}[Z_i'] = \operatorname{Tr}\Sigma^2$, by Jensen's inequality

$$\mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{2q}\Big] \leq \mathbb{E}\Big[\big(Z_i - Z_i'\big)^{2q}\Big] \leq 2^{2q}\mathbb{E}\Big[Z_i^{2q}\Big] \leq (4q)!(2\operatorname{Tr}\Sigma^2)^{2q}\,.$$

For the odd moments we use that the function $(\cdot)^{2q+1}$ is increasing:

$$-(\operatorname{Tr}\Sigma^2)^{2q+1} \leq \mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{2q+1}\Big] \leq \mathbb{E}\Big[Z_i^{2q+1}\Big] \leq (4q+2)!(\operatorname{Tr}\Sigma^2)^{2q+1}\,,$$

so for all $q \geq 0$:

$$\Big|\mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^q\Big]\Big| \leq (2q)!(2\operatorname{Tr}\Sigma^2)^q\,. \tag{59}$$

It remains to upper bound the moments of the sum:

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m Z_i^2 - \operatorname{Tr}\Sigma^2\right)^{2q}\right] \\ &\qquad= \frac{1}{m^{2q}} \sum_{\substack{p_1+\ldots+p_m=2q \\ p_i\neq 1}} \binom{2q}{p_1,\ldots,p_m} \prod_{i=1}^m \mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{p_i}\Big] \\ &\qquad\leq \frac{1}{m^{2q}} \sum_{\substack{p_1+\ldots+p_m=2q \\ p_i\neq 1}} \frac{(2q)!}{p_1!\ldots p_m!} \prod_{i=1}^m (2p_i)!(2\operatorname{Tr}\Sigma^2)^{p_i} \\ &\qquad\leq (2q)!\left(\frac{2\operatorname{Tr}\Sigma^2}{m}\right)^{2q} (2q)^{2q} \sum_{\substack{p_1+\ldots+p_m=2q \\ p_i\neq 1}} 1.\end{aligned}$$

Let us count the number of terms in this last sum. Consider first that we have k non-null terms $(p_{i_1},\ldots,p_{i_k})$. Their sum is equal to $2q$ but because these terms

are strictly greater than 1, we also have that $(p_{i_1}-2)+\ldots+(p_{i_k}-2)=2q-2k$, where all terms of this sum are nonnegative. The number of k-partitions of $2q-2k$ is $\binom{(2q-2k)+(k-1)}{k-1}=\binom{2q-k-1}{k-1}$ and then the number of terms in the sum is equal to:

$$\sum_{k=0}^{m}\binom{m}{k}\binom{2q-k-1}{k-1}=\sum_{k=0}^{m\wedge q}\binom{m}{k}\binom{2q-k-1}{k-1}$$
$$\leq m^q\sum_{k=0}^{q}\binom{2q-k-1}{k-1}=m^qF(2q-1)\leq m^q\phi^{2q}\,,$$

where $F(\cdot)$ is the Fibonacci sequence and $\phi=(1+\sqrt{5})/2$ is the golden ratio. So using that $(2q)!\leq(2q)^qq^q$ we obtain that

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}Z_i-\operatorname{Tr}\Sigma^2\right)^{2q}\right]\leq(2\phi^2)^q\left(\frac{\operatorname{Tr}\Sigma^2}{\sqrt{m}}\right)^{2q}(2q)^{4q}. \tag{60}$$

□

Acknowledgements. GB acknowledges support from: Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901; Agence Nationale de la Recherche (ANR), ANR-19-CHIA-0021-01 "BiSCottE"; the Franco-German University (UFA) through the binational Doktorandenkolleg CDFA 01-18. Both authors are extremely grateful to the two reviewers and to the editor, who by their very careful read of the initial manuscript and their various suggestions allowed us to improve its quality significantly.

References

Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Third edition. Wiley series in probability and mathematical statistics. Wiley.

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research* 22 (1), 1–45.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* 8 (5), 577–606.

Birgé, L. (2001). An alternative point of view on Lepski's method. In: *State of the art in probability and statistics (Leiden, 1999)*. Vol. 36. IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., 113–133.

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electronic Journal of Statistics* 12 (2), 3713–3735.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematiques de l'Académie des Sciences* 334 (6), 495–500.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. *Proc. of the 33rd International Conference on Machine Learning (ICML 2016)*. Vol. 48, 2606–2615.

Cohn, D. L. (1980). *Measure theory / Donald L. Cohn*. English. Birkhauser Boston, ix, 373 p. :

Dette, H. and Munk, A. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *The Annals of Statistics* 26 (6), 2339–2368.

Ermakov, M. S. (1991). Minimax detection of a signal in a Gaussian white noise. *Theory of Probability & Its Applications* 35 (4), 667–679.

Fromont, M., Laurent, B., Lerasle, M., and Reynaud-Bouret, P. (2012). Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by S. Mannor, N. Srebro, and R. C. Williamson. Vol. 23. Proceedings of Machine Learning Research, 1–23.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* 13 (25), 723–773.

Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In: *Stochastic Inequalities and Applications*. Progress in Probability 56, 55–69.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* 17, 6 pp.

Ingster, Y. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems of Information Transmission* 18 (2), 130–140.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I-II-III. *Mathematical Methods of Statistics* 2 (2–4), 85–114, 171–189, 249–268.

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics 169. Springer.

Ingster, Y. I. and Suslina, I. A. (1998). Minimax detection of a signal for Besov bodies and balls. *Problems of Information Transmission* 34 (1), 48–59.

Jirak, M. and Wahl, M. (2018). *Perturbation bounds for eigenspaces under a relative gap condition*. arXiv: 1803.03868 [math.PR].

Kim, I., Balakrishnan, S., and Wasserman, L. (2020). *Minimax optimality of permutation tests*. arXiv: 2003.13208 [math.ST].

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* 23 (1), 110–133.

Lam-Weil, J., Carpentier, A., and Sriperumbudur, B. K. (2021). *Local minimax rates for closeness testing of discrete distributions*. arXiv: 1902.01219 [math.ST].

Lepski, O. V. and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli* 5 (2), 333–358.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* 19 (5), 1145–1190.

Marienwald, H., Fermanian, J.-B., and Blanchard, G. (2020). High-dimensional multi-task averaging and application to kernel mean embedding. *AISTATS 2021*. arXiv: 2011.06794 [stat.ML].

Massart, P. (2003). *Concentration Inequalities and Model Selection*. Springer.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning* 10 (1-2), 1–141.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60 (1), 223–241.

Naumov, A., Spokoiny, V. G., and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields* 174 (3), 1091–1132.

Ostrovskii, D. M., Ndaoud, M., Javanmard, A., and Razaviyayn, M. (2020). *Near-Optimal Model Discrimination with Non-Disclosure*. arXiv: 2012.02901 [math.ST].

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. *Proc. International Conference on Algorithmic Learning Theory (ALT 2007)*, 13–31.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics* 24 (6), 2477–2498.

Spokoiny, V. G. (2012). Parametric estimation. Finite sample theory. *The Annals of Statistics* 40 (6), 2877–2909.

Spokoiny, V. G. and Dickhaus, T. (2015). *Basics of modern mathematical statistics*. Springer Texts in Statistics. Springer.

Spokoiny, V. G. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics* 22 (2), 100–113.

Spokoiny, V. G. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Annals of Statistics* 43 (6), 2653–2675.

van Handel, R. (2017). Structured random matrices. In: *Convexity and concentration*. Springer, 107–156.

Vershynin, R. (2018). *High-Dimensional Probability: an introduction with applications to data science*. Cambridge series in statistical and probabilistic mathematics 47. Cambridge University Press.