Proceedings of Machine Learning Research vol 75:1–36, 2018 31st Annual Conference on Learning Theory

Polynomial Time and Sample Complexity for Non-Gaussian Component Analysis: Spectral Methods

Yan Shuo Tan YANSHUO@UMICH.EDU
Department of Mathematics, University of Michigan

Roman Vershynin RVERSHYN@UCI.EDU
Department of Mathematics, University of California, Irvine

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

The problem of Non-Gaussian Component Analysis (NGCA) is about finding a maximal low-dimensional subspace E in $\mathbb{R}^n$ so that data points projected onto E follow a non-Gaussian distribution. Vempala and Xiao (2011) proposed a local search algorithm, and showed that it was able to estimate E accurately with polynomial time and sample complexity, if the dimension of E is treated as a constant and with the assumption that all one-dimensional marginals of the non-Gaussian distribution over E have non-Gaussian moments. In this paper, we propose a simple spectral algorithm called REWEIGHTED PCA, and prove that it possesses the same guarantee. The principle that underlies this approach is a new characterization of multivariate Gaussian distributions.

1. Introduction

1.1. Non-Gaussian Component Analysis

Dimension reduction is a necessary step for much of modern data analysis, the principle being that the structure or "interestingness" of a collection of data points is contained in a geometric structure which has much lower dimension than the ambient vector space. We consider the case where the geometric structure in question is a linear subspace. In other words, we are in the situation where the variation of the data points within this subspace contains some information which we would like to extract, while their variation in the complementary directions constitutes mere noise.

In many cases, it is reasonable to think of the noise as being Gaussian. Formally, we then have the following generative model. Let E be an unknown d-dimensional subspace of $\mathbb{R}^n$, and let $E^\perp$ be the orthogonal complement of E. Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$, which we can decompose into two independent components: a non-Gaussian component $\tilde{\mathbf{X}}$ that takes values in E, and a Gaussian component $\mathbf{g}$ that takes values in $E^\perp$. In other words, we let $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{g}) \in E \oplus E^\perp$.[1]

Our goal is to recover the subspace E from a sample of independent realizations of $\mathbf{X}$. This is precisely the framework of the problem of Non-Gaussian Component Analysis (NGCA). We make no assumption on the relative magnitudes of $\tilde{\mathbf{X}}$ and $\mathbf{g}$. When the noise component is much smaller, which is a reasonable assumption in some real world applications, E can be recovered using the standard Principal Component Analysis (PCA). However, PCA manifestly fails when the signal to noise ratio is small, i.e. when $\tilde{\mathbf{X}}$ has lower magnitude than $\mathbf{g}$.

1. It is not necessary to assume that the Gaussian and non-Gaussian subspaces are perpendicular. They automatically become perpendicular if we apply a whitening transformation.

With mild distributional assumptions, applying a whitening transformation to the data points can be done efficiently with sample size linear in the dimension (see Vershynin (2011)). As such, we might as well assume that the distribution is already whitened (i.e. isotropic). In other words, for the rest of this paper, we work with the model:

Definition 1 (Isotropic NGCA model)

$$\boldsymbol{X} = (\tilde{\boldsymbol{X}}, \boldsymbol{g}) \in E \oplus E^{\perp}, \quad \mathbb{E}\boldsymbol{X} = 0, \quad \mathbb{E}\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}_n. \tag{1.1}$$

The NGCA problem is closely related to the problem of Independent Component Analysis (ICA), but generalizes it in a crucial way. ICA assumes the existence of a latent variable s with independent coordinates, whereas in our case, the distribution of $\tilde{\mathbf{X}}$ is allowed to have any manner of dependencies amongst its entries.

1.2. Quantifying "non-Gaussianity"

In order to provide a guarantee for an algorithm for NGCA, one needs to quantify the deviation of $\tilde{\mathbf{X}}$ from being Gaussian. We will do so in terms of its moments.

Definition 2 *We say that $\tilde{\boldsymbol{X}}$ is (m, η)-moment-identifiable along a unit vector $\boldsymbol{v} \in E$ if there is some $1 \le r \le m$ for which*

$$\left|\mathbb{E}\{\langle \tilde{\boldsymbol{X}}, \boldsymbol{v}\rangle^r\} - \gamma_r\right| \ge \eta. \tag{1.2}$$

Here γ_r is the r-th moment[2] of a $\mathcal{N}(0,1)$ random variable. The r-th moment distance of $\tilde{\boldsymbol{X}}$ from a standard Gaussian is defined as the quantity

$$D_{\tilde{\boldsymbol{X}},r} := \sup_{\boldsymbol{v} \in S^{n-1} \cap E} \left|\mathbb{E}\{\langle \tilde{\boldsymbol{X}}, \boldsymbol{v}\rangle^r\} - \gamma_r\right|. \tag{1.3}$$

There are two reasons why we take such an approach. First, it allows us to analyze our proposed algorithm more easily, since the algorithm is a moment method, and second, by the classical moment problem, if $D_{\tilde{\mathbf{X}},r} = 0$ for all positive integers r, then $\tilde{\mathbf{X}}$ has the standard Gaussian distribution.

Nonetheless, readers may be concerned about how the moment-identifiability condition squares with other notions of distribution distance. This was investigated somewhat by Vempala and Xiao (2011), who proved the following result for log-concave distributions on $\mathbb{R}$.

Fact 3 (Lemma 1 in Vempala and Xiao (2011)) *Let G be the density of a standard Gaussian random variable, F the density of an isotropic log-concave distribution. Suppose G is not (m, η)-moment-identifiable, i.e. for $r = 1, \ldots, m$, $|\mathbb{E}_F\{X^r\} - \gamma_r| < \eta$. Then there is a universal constant C such that*

$$\|F - G\|_1 \le C\frac{\log m}{m^{1/16}} + \eta m e^m.$$

We note that the log-concave assumption is simply to obtain a tail bound for the characteristic function for F. Hence, the result also holds for any distribution with a C^1 density, albeit with possibly a different constant in the bound. Furthermore, the method for proving the result can easily be generalized to multivariate distributions.

2. One can check that $\gamma_r = (r-1)!!$ when r is even, and is zero for r odd.

1.3. Previous work on NGCA

As far as we know, the NGCA problem was first formulated and studied by Blanchard et al. (2006). They observed that whenever $\mathbf{X}$ satisfies the NGCA model (1.1), then for any smooth function h, we have

$$\boldsymbol{\beta}(h) := \mathbb{E}\{\mathbf{X}h(\mathbf{X})\} - \mathbb{E}\{\nabla h(\mathbf{X})\} \in E. \tag{1.4}$$

This suggests that if we can find a rich enough collection of functions $\mathcal{H}$, then one should be able to recover E as the span of $\{\boldsymbol{\beta}(h)\colon h \in \mathcal{H}\}$. Hence, the authors proposed first forming empirical estimates $\hat{\boldsymbol{\beta}}(h)$ using the given i.i.d. samples of $\mathbf{X}$, and then running PCA on this collection of vectors. Inspired by the FastICA algorithm of Hyvarinen (1999), they suggested picking test functions of the form $h_{a,\boldsymbol{\omega}}(\mathbf{x}) = \tilde{h}_a(\langle \mathbf{x}, \boldsymbol{\omega}\rangle)$ where $\boldsymbol{\omega} \in S^{n-1}$ and $\{\tilde{h}_a\colon a \in \mathbb{R}\}$ is a one-parameter family of smooth functions. They called this approach *Multi-index Projection Pursuit*.

Subsequent papers have built upon this in several ways. Kawanabe et al. (2006) investigated the situation when the contrast functions h_i's are chosen to be radial kernel functions, and when these are adapted to the data in an iterative fashion. Diederichs et al. (2010, 2013) replaced the PCA step with a semidefinite program, thereby yielding an approach they call *Sparse NGCA*.

All the papers in this line of research suffer from the defect that the performance of the algorithms all depend experimentally and theoretically on some "good" behavior of the $\boldsymbol{\beta}(h)$'s. Clearly, how "good" the $\boldsymbol{\beta}(h)$'s are depends intimately on how the chosen contrast functions interact with the particular way in which $\tilde{\mathbf{X}}$ deviates from being Gaussian. None of these papers are able to quantify this dependence theoretically, and instead simply assume the "good" behavior (see for instance Assumption 1 in Diederichs et al. (2013)), so their proposed algorithms cannot be said to have polynomial time and sample complexity guarantees.

Indeed, prior to our work, the only algorithm with such guarantees was proposed and studied by Vempala and Xiao (2011). Their strategy was to adapt Frieze et al. (1996)'s work on ICA to higher moments. For each positive integer r, they defined the marginal moment function $f_r(\mathbf{v}) := \mathbb{E}\{\langle \mathbf{X}, \mathbf{v}\rangle^r\}$, and noted that the strict local optima of f_r would have to lie in E. Furthermore, for each r, the r-th moment tensors of $\mathbf{X}$ defining f_r can be approximated up to ϵ accuracy in each of its entries with enough samples. These therefore yield empirical estimates $\hat{f}_r$ that have local optima that are close to those of f_r. Finally, they showed how to identify a local optima of $\hat{f}_r$ using a 2nd order local search. The samples are then projected onto the orthogonal complement of this direction, and the algorithm is applied recursively on the projection. They were able to prove that whenever $\tilde{\mathbf{X}}$ is (m, η)-moment-identifiable along all unit vectors $\mathbf{v} \in E$, then their algorithm recovers a subspace $\hat{E}$ close enough to E with time and sample complexity polynomial in n, η, $1/\epsilon$, and $\log(1/\delta)$, where δ is the failure probability. The degree of the polynomial however grows linearly in m and d. [3]

Other work on NGCA include M. Kawanabe (2005); Kawanabe et al. (2006); Kawanabe and Theis (2006); Sasaki et al. (2016). These papers have limited theoretical analysis, and we omit a discussion of these because of space constraints.

3. We are of course omitting numerous details of their work. In addition, their statement of their guarantee (see Theorem 1 in their paper) is also somewhat different from how we have stated it here: they have both a slightly weaker assumption on $\tilde{\mathbf{X}}$ ((m, η)-moment-distinguishability) and a slightly weaker conclusion on $\hat{E}$ (in terms of moment distance). Nonetheless, their intermediate results are sufficient to prove the version that we have stated in the main text.

2. Main results

The principle that underlies our approach to NGCA is a new characterization of multivariate Gaussian distributions. Throughout this section, $\mathbf{X}$ denotes a random vector in $\mathbb{R}^n$ and $\mathbf{g}$ is a standard Gaussian random vector in $\mathbb{R}^n$. By $\mathbf{X}'$ we will always denote an independent copy of $\mathbf{X}$.

Theorem 4 (First Gaussian test) *Suppose $\boldsymbol{X}$ has the same radial distribution as $\boldsymbol{g}$, i.e. $\|\boldsymbol{X}\|_2$ and $\|\boldsymbol{g}\|_2$ are identically distributed. If $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ has the same distribution as $\langle \boldsymbol{g}, \boldsymbol{g}' \rangle$, then $\boldsymbol{X}$ has the same distribution as $\boldsymbol{g}$, i.e. the standard Gaussian distribution.*

We will prove this theorem in Section 3 via a decomposition of moment tensors and a resulting energy minimization property of the Gaussian measure. The theorem guarantees that any non-Gaussianity of $\mathbf{X}$ is always captured by either the norm $\|\mathbf{X}\|_2$ or the dot product pairings $\langle \mathbf{X}, \mathbf{X}' \rangle$. It is clear that the norm condition on its own is not sufficient to guarantee that $\mathbf{X} \overset{d}{=} \mathbf{g}$. For instance, let $\boldsymbol{\theta}$ have any non-uniform distribution on the sphere. Then $\|\mathbf{g}\|_2 \boldsymbol{\theta}$ has the same radial distribution as $\mathbf{g}$, but is not itself Gaussian.

This result by itself does not address the NGCA problem, in which we are looking for non-Gaussian *directions* in the distribution of $\mathbf{X}$. To this end, we propose a matrix version of the first Gaussian test. Pick a parameter $\alpha > 0$ and consider the *test matrices*

$$\boldsymbol{\Phi}_{\mathbf{X},\alpha} := \frac{1}{Z_{\boldsymbol{\Phi}}} \mathbb{E}\{e^{-\alpha \|\mathbf{X}\|_2^2} \mathbf{X}\mathbf{X}^T\} \quad \text{and} \quad \boldsymbol{\Psi}_{\mathbf{X},\alpha} := \frac{1}{Z_{\boldsymbol{\Psi}}} \mathbb{E}\{e^{-\alpha \langle \mathbf{X}, \mathbf{X}' \rangle} \mathbf{X}(\mathbf{X}')^T\}, \tag{2.1}$$

where the normalizing quantities $Z_{\boldsymbol{\Phi}} = Z_{\boldsymbol{\Phi},\mathbf{X}}(\alpha) := \mathbb{E}\{e^{-\alpha \|\mathbf{X}\|_2^2}\}$ and $Z_{\boldsymbol{\Psi}} = Z_{\boldsymbol{\Psi},\mathbf{X}}(\alpha) := \mathbb{E}\{e^{-\alpha \langle \mathbf{X}, \mathbf{X}' \rangle}\}$ are the moment generating functions of $\|\mathbf{X}\|_2^2$ and $\langle \mathbf{X}, \mathbf{X}' \rangle$ respectively. We also remark that they resemble partition functions in statistical mechanics.

For a standard Gaussian random vector $\mathbf{g}$, a straightforward computation (see Lemma 31) shows that both test matrices are multiples of the identity, namely

$$\boldsymbol{\Phi}_{\mathbf{g},\alpha} = (2\alpha + 1)^{-1} \mathbf{I}_n \quad \text{and} \quad \boldsymbol{\Psi}_{\mathbf{g},\alpha} = \alpha(\alpha^2 - 1)^{-1} \mathbf{I}_n. \tag{2.2}$$

Our second test guarantees that the non-Gaussianity of $\mathbf{X}$ is captured by one of the test matrices, and moreover that their eigenvectors reveal the non-Gaussian directions of $\mathbf{X}$.

Theorem 5 (Second Gaussian test) *Consider a random vector $\boldsymbol{X}$ which follows the isotropic NGCA model (1.1). Then, for any $|\alpha|$ small enough, either $\boldsymbol{\Phi}_{\boldsymbol{X},\alpha}$ has an eigenvalue not equal to $(2\alpha+1)^{-1}$ or $\boldsymbol{\Psi}_{\boldsymbol{X},\alpha}$ has an eigenvalue not equal to $\alpha(\alpha^2-1)^{-1}$. Furthermore, all eigenvectors corresponding to such eigenvalues lie in E.*[4]

In Section 4, we will show how to derive the second Gaussian test from the first using a block diagonalization formula for each of the matrices $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ and $\boldsymbol{\Psi}_{\mathbf{X},\alpha}$. Again, it is easy to see that $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ is not sufficient by itself to identify non-Gaussian directions: Take $\mathbf{X} = \|\mathbf{g}\|_2 \boldsymbol{\theta}$ as before, and this

4. The matrix $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ always exists, but when $\tilde{\mathbf{X}}$ is not sub-Gaussian (i.e. can be rescaled so that marginals have tails lighter than a standard Gaussian), $\boldsymbol{\Psi}_{\mathbf{X},\alpha}$ may not be well-defined even for small α. In that case, $\|\tilde{\mathbf{X}}\|_2$ has a different distribution from $\|\mathbf{g}\|_2$, so that $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ has non-Gaussian eigenvalues. We can hence think of $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ as the primary test matrix, and $\boldsymbol{\Psi}_{\mathbf{X},\alpha}$ being an auxiliary that is only required in hard (effectively adversarial) cases.

time that assume that $\boldsymbol{\theta}$ is uniform on $\{\pm\mathbf{e}_i\}_{1=1}^N$. The symmetry implies that $\boldsymbol{\Phi}_{\mathbf{X},\alpha}$ is a scalar matrix, and computing its trace shows that it is equal to $(2\alpha+1)^{-1}\mathbf{I}_n$.

For simplicity, we stated both Gaussian tests for population rather than for finite samples; they involve taking expectations over the entire distribution of $\mathbf{X}$ which is typically unknown in practice. However, both tests are quite robust and work provably well on finite (polynomially large) samples. Robust versions of Gaussian tests can be formulated in terms of our definition of moment distance (see (1.3)).

Theorem 6 (First Gaussian test, robust) *There is a universal constant $c > 0$ such that for each positive integer r, we have either*

$$|\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}| \geq c\eta_r^2/\tilde{\gamma}_r \quad \textit{or} \quad \left|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}\right| \geq c\eta_r^2.$$

Here $\tilde{\gamma}_r = \mathbb{E}\{|g|^r\}$ is the r-th absolute moment of a standard Gaussian random variable, and $\eta_r = \min\{D_{\boldsymbol{X},r}, \tilde{\gamma}_r\}$.

As with the non-robust version, we will prove this theorem in Section 3 using a decomposition of moment tensors. There is a similar robust version of the second Gaussian test, which we will skip here but state and prove in Section 4.

Robustness allows us to use finite sample averages instead of expectations in the Gaussian tests, which is critical for practical applications. Indeed, consider a sample $\mathbf{X}_1, \dots, \mathbf{X}_N, \mathbf{X}'_1, \dots, \mathbf{X}'_N$ of $2N$ i.i.d. realizations of a random variable $\mathbf{X}$. We can then define the sample versions of the test matrices in (2.1) in an obvious way:

$$\hat{\boldsymbol{\Phi}}_{\mathbf{X},\alpha} = \frac{1}{\hat{Z}_{\boldsymbol{\Phi}}} \sum_{i=1}^N e^{-\alpha\|\mathbf{X}_i\|_2^2} \mathbf{X}_i\mathbf{X}_i^T \quad \text{and} \quad \hat{\boldsymbol{\Psi}}_{\mathbf{X},\alpha} = \frac{1}{\hat{Z}_{\boldsymbol{\Psi}}} \sum_{i=1}^N e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}'_i\rangle} (\mathbf{X}_i(\mathbf{X}'_i)^T + \mathbf{X}'_i\mathbf{X}_i^T), \quad (2.3)$$

with the normalizing quantities $\hat{Z}_{\boldsymbol{\Phi}} := \sum_{i=1}^N e^{-\alpha\|\mathbf{X}_i\|_2^2}$ and $\hat{Z}_{\boldsymbol{\Psi}} := 2\sum_{i=1}^N e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}'_i\rangle}$.

The second Gaussian test leads to the following straightforward algorithm for solving NGCA problem based on a finite sample: Use the sample to compute the test matrices $\hat{\boldsymbol{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\boldsymbol{\Psi}}_{\mathbf{X},\alpha}$; select the eigenspaces corresponding to the eigenvalues that significantly deviate from the Gaussian eigenvalues. Then all vectors in both eigenspaces will be close to the non-Gaussian subspace E which we are trying to find. Let us state this algorithm and its guarantee precisely.

Algorithm 1 REWEIGHTED PCA($\mathbf{X}$,α_1,α_2,β_1,β_2)

Input: Data points $\mathbf{X}_1, \dots, \mathbf{X}_N, \mathbf{X}'_1, \dots, \mathbf{X}'_N$, scaling parameters $\alpha_1, \alpha_2 \in \mathbb{R}$, tolerance parameters $\beta_1, \beta_2 > 0$.

Output: Two estimates $\hat{E}_{\boldsymbol{\Phi}}$ and $\hat{E}_{\boldsymbol{\Psi}}$ for E.

1: Compute test matrices $\hat{\boldsymbol{\Phi}}_{\mathbf{X},\alpha_1}$ and $\hat{\boldsymbol{\Psi}}_{\mathbf{X},\alpha_2}$.

2: Compute the eigenspace $\hat{E}_{\boldsymbol{\Phi}}$ of $\hat{\boldsymbol{\Phi}}_{\mathbf{X},\alpha_1}$ corresponding to the nonzero eigenvalues that are farther than β_1 from the value $(2\alpha_1+1)^{-1}$.

3: Compute the eigenspace $\hat{E}_{\boldsymbol{\Psi}}$ of $\hat{\boldsymbol{\Psi}}_{\mathbf{X},\alpha_2}$ corresponding to the nonzero eigenvalues that are farther than β_2 from the value $\alpha_2(\alpha_2^2-1)^{-1}$.

Theorem 7 (Finding one non-Gaussian direction) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector (see Appendix B) which follows the isotropic NGCA model (1.1), and with sub-Gaussian norm bounded*

above by $K \geq 1$. Let r be the minimum integer for which the r-th moment distance $D_{\tilde{\boldsymbol{X}},r} =: D > 0$. Then for any $\delta, \epsilon \in (0,1)$, with probability at least $1-\delta$, if we run Reweighted PCA *with a choice of parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ that is optimal up to constant multiples, at least one of $\hat{E}_{\boldsymbol{\Phi}}$ and $\hat{E}_{\boldsymbol{\Psi}}$ is non-trivial, and any unit vector in their union is ϵ-close to one in E, so long as the sample size N is greater than $\text{poly}_r(n, 1/\epsilon, \log(1/\delta), 1/D, K)$. Here, poly_r is a polynomial whose total degree depends linearly on r.*

The idea of the proof is to use eigenvector perturbation theory from Davis and Kahan (1970). The robust version of the second Gaussian test asserts the existence of a gap between Gaussian and non-Gaussian eigenvalues. By bounding the deviation of the test matrix estimators $\hat{\boldsymbol{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\boldsymbol{\Psi}}_{\mathbf{X},\alpha}$ from their expectation, we can thus show that their eigenstructures are similar. We will prove this theorem formally in Section 5.

The next step is to obtain a good estimate for the entire non-Gaussian subspace. To do so, we follow Vempala and Xiao (2011)'s strategy of projecting the sample points onto the orthogonal complement of the found directions, and recursing our algorithm on the new sample. After a set number of iterations, we collate all the found directions into a basis spanning candidate subspace $\hat{E}$. To state our guarantee for this procedure, we use the following notion of distance between subspaces.

Definition 8 (Subspace distance) *Let F and F' be subspaces of $\mathbb{R}^n$ of dimensions m. Let $\boldsymbol{U}$ and $\boldsymbol{U}'$ be matrices whose columns form an orthonormal basis for F and F' respectively. The distance between F and F' is defined to be $d(F, F') := \|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}'(\boldsymbol{U}')^T\|_F$.*

Theorem 9 (Finding all non-Gaussian directions) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector which follows the isotropic NGCA model (1.1), and with sub-Gaussian norm bounded above by $K \geq 1$. Suppose that $\tilde{\boldsymbol{X}}$ is (m, η)-moment-identifiable along all unit vectors $\boldsymbol{v} \in E$. Then running* Reweighted PCA *recursively (i.e. Algorithm 2) produces an estimate $\hat{E}$ such that $d(\hat{E}, E) < \epsilon$ so long as the sample size N is greater than $\text{poly}_{m,d}(n, 1/\epsilon, \log(1/\delta), 1/D, K)$. Here, $\text{poly}_{m,d}$ is a polynomial whose total degree depends linearly on m and d.*

We shall prove this theorem in Appendix H. The theorem gives a polynomial time and sample complexity guarantee that REWEIGHTED PCA solves the NGCA problem, so long as m and d are assumed to be constants, while making exactly the same assumptions as Vempala and Xiao (2011). This means that theoretically, both algorithms do just as well. On the other hand, REWEIGHTED PCA is a simple spectral algorithm, which is easier and faster to implement than local search.

Furthermore, while local search discovers one non-Gaussian direction at a time, REWEIGHTED PCA possibly discovers multiple directions in each iteration. Most importantly, there is hope that all non-Gaussian directions can be discovered in the very first iteration. This is probably what will happen in practice with real data, and we may moreover prove that this is the case for special distributions. For instance, we can prove the following guarantee for finding a planted sphere.

Corollary 10 (Finding a sphere) *Let $\tilde{\boldsymbol{X}}$ be uniformly distributed on the scaled unit sphere $\sqrt{d}S^{d-1}$ in E. Suppose we are given a sample of size $N \gtrsim dn^2(n + \log(1/\delta))/\epsilon^2$, then running the first two steps of* REWEIGHTED PCA *with a choice of $\alpha \in [c_1/n, c_2/n]$, and $\beta = \alpha/3$ yields a subspace $\hat{E}_{\boldsymbol{\Phi}}$ so that $d(\hat{E}_{\boldsymbol{\Phi}}, E) < \epsilon$. Here, c_1 and c_2 are absolute constants.*

2.1. Reweighted PCA in other contexts

The name of the algorithm stems from the first test matrix, which can be seen as a PCA matrix for the reweighted sample obtained when each point $\mathbf{X}_i$ is given the weight $e^{-\alpha\|\mathbf{X}_i\|_2^2}$. As mentioned in the previous section, $\mathbf{\Phi}_{\mathbf{X},\alpha}$ reveals at least one non-Gaussian direction in all but adversarial situations, and so can be considered the primary test matrix.

The idea of doing PCA with weight functions that are non-linear in the sample points can be traced back at least as far as Brubaker and Vempala (2008). In that paper, the authors similarly use Gaussian weights, but do so in order to handle clustering for Gaussian mixture models that are highly non-spherical. In a later paper, Goyal et al. (2014) used Fourier weights to handle ICA. While our analysis is radically different, the idea for the algorithm was directly inspired by these two papers.

2.2. Organization of paper and notation

In Section 3, we will prove the first Gaussian test and its robust version. In Section 4, we will prove the second Gaussian test and state a robust version needed for proving our guarantee for Reweighted PCA. The guarantee for finding one direction is proved in Section 5 and Appendix G. The guarantees for finding all directions, and the special case of finding a sphere are proved in Appendices H and I respectively. For the sake of space, many technical details are also deferred to the appendix. Throughout the paper, scalars are denoted in standard font, while vectors and matrices are denoted with bold font. C and c denote absolute constants whose value may change from line to line. We let $\mathbf{g}_n$ denote the standard Gaussian vector in $\mathbb{R}^n$. The subscript is omitted whenever the dimension is obvious. In addition, for each r, we let γ_r and $\tilde{\gamma}_r$ denote the r-th moment and r-th absolute moment of a standard Gaussian random variable.

3. Proof of the first Gaussian test

The first Gaussian test is based on the work of Tan (2017). For completeness, we will repeat the key arguments. The statements and proofs in this section are valid more generally for random variables $\mathbf{X}$ with finite moments of all orders (not necessarily sub-Gaussian).

Recall the following fact from multilinear algebra. For any positive integer r, we may identify the r-th tensor product $T^r(\mathbb{R}^n) = \mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n$ with $\mathbb{R}^{n^r}$ by picking as a basis the vectors $\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \cdots \otimes \mathbf{e}_{i_r}\}_{1\le i_1,\dots i_r\le n}$. With this choice of Euclidean structure, the Euclidean inner product between any two pure tensors $\mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_r$ and $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_r$ (treated as vectors) can be written as

$$\langle \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_r, \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_r\rangle = \prod_{i=1}^{r} \langle \mathbf{u}_i, \mathbf{v}_i\rangle.$$

In particular, for power tensors $\mathbf{u}^{\otimes r}$ and $\mathbf{v}^{\otimes r}$, we have the formula $\langle \mathbf{u}^{\otimes r}, \mathbf{v}^{\otimes r}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle^r$.

Now let $\mathbf{X}$ and $\mathbf{Y}$ be two independent random vectors. The above formula allows us to rewrite the r-th moment of their inner product as an inner product between their r-th moment tensors. Namely, we have

$$\mathbb{E}\{\langle \mathbf{X}, \mathbf{Y}\rangle^r\} = \mathbb{E}\{\langle \mathbf{X}^{\otimes r}, \mathbf{Y}^{\otimes r}\rangle\} = \langle \mathbf{M}_{\mathbf{X}}^r, \mathbf{M}_{\mathbf{Y}}^r\rangle, \tag{3.1}$$

where we define $\mathbf{M}^r_{\mathbf{X}} := \mathbb{E}\mathbf{X}^{\otimes r}$. For independent copies $\mathbf{X}$, $\mathbf{X}'$ of the same random vector having distribution μ, $\mathbf{M}^r_{\mathbf{X}} = \mathbf{M}^r_{\mathbf{X}'}$, so

$$\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}' \rangle^r\} = \|\mathbf{M}^r_{\mathbf{X}}\|^2. \tag{3.2}$$

Next, for any random vector $\mathbf{X}$, let $\mathbf{X}_{rot}$ denote a random vector that is independent of $\mathbf{X}$, has the same radial distribution as $\mathbf{X}$, and whose distribution is rotationally invariant. We call $\mathbf{X}_{rot}$ the *rotational symmetrization* of $\mathbf{X}$. Comparing the moment tensors of a random vector and those of its rotational symmetrization gives rise to what we shall call eccentricity tensors. Specifically, for any positive integer r, we define the r-th *eccentricity tensor* of $\mathbf{X}$ to be $\mathbf{E}^r_{\mathbf{X}} = \mathbf{M}^r_{\mathbf{X}} - \mathbf{M}^r_{\mathbf{X}_{rot}}$.

Since $\mathbf{X} \overset{d}{=} \mathbf{X}_{rot}$ if and only if $\mathbf{X}$ is rotationally invariant, we see that the eccentricity tensors of $\mathbf{X}$ are quantitative measures of how far its distribution is from being rotationally invariant. This interpretation is further supported by the following observation.

Lemma 11 (Orthogonality) *The eccentricity tensors of a random vector $\boldsymbol{X}$ are orthogonal to the moment tensors of its rotational symmetrization. In other words, for any positive integer r,*

$$\langle \boldsymbol{E}^r_{\boldsymbol{X}}, \boldsymbol{M}^r_{\boldsymbol{X}_{rot}} \rangle = 0 \quad \textit{and} \quad \|\boldsymbol{M}^r_{\boldsymbol{X}}\|_2^2 = \left\|\boldsymbol{M}^r_{\boldsymbol{X}_{rot}}\right\|_2^2 + \|\boldsymbol{E}^r_{\boldsymbol{X}}\|_2^2. \tag{3.3}$$

Proof Let $\mathbf{Q}$ be a random orthogonal matrix chosen according to the Haar measure on $O(n)$. For any fixed vector $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{Qv}$ is uniformly distributed on the sphere of radius $\|\mathbf{v}\|_2$, so if $\mathbf{Y}$ is any random vector independent of $\mathbf{Q}$, applying $\mathbf{Q}$ to $\mathbf{Y}$ preserves its radial distribution but makes $\mathbf{QY}$ rotationally invariant.

Now choose $\mathbf{Q}$ to be independent of $\mathbf{X}$ and $\mathbf{X}_{rot}$. Our previous discussion implies that $\mathbf{Q}^T\mathbf{X} \overset{d}{=} \mathbf{QX}_{rot} \overset{d}{=} \mathbf{X}_{rot}$. We use this to compute

$$\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}_{rot} \rangle^r\} = \mathbb{E}\{\langle \mathbf{X}, \mathbf{QX}_{rot} \rangle^r\} = \mathbb{E}\{\langle \mathbf{Q}^T\mathbf{X}, \mathbf{X}_{rot} \rangle^r\} = \mathbb{E}\{\langle \mathbf{X}'_{rot}, \mathbf{X}_{rot} \rangle^r\}, \tag{3.4}$$

where $\mathbf{X}'_{rot}$ is an independent copy of $\mathbf{X}_{rot}$. We may then apply identities (3.1) and (3.2) to rewrite the above equation as

$$\left\langle \mathbf{M}^r_{\mathbf{X}}, \mathbf{M}^r_{\mathbf{X}_{rot}} \right\rangle = \left\langle \mathbf{M}^r_{\mathbf{X}_{rot}}, \mathbf{M}^r_{\mathbf{X}_{rot}} \right\rangle. \tag{3.5}$$

Subtracting the right hand side from the left hand side gives (3.3). ■

Theorem 12 *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$ with finite moments of all orders. Then*

a) (Minimization) If $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$, and $\boldsymbol{X}_{rot}, \boldsymbol{X}'_{rot}$ are independent copies of its rotational symmetrization, we have

$$\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^r\} \geq \mathbb{E}\{\langle \boldsymbol{X}_{rot}, \boldsymbol{X}'_{rot} \rangle^r\} \tag{3.6}$$

for any positive integer r.

b) (Uniqueness) Furthermore, if equality holds in (3.6) for all r and we further assume that $\boldsymbol{X}$ has a subexponential distribution[5], then $\boldsymbol{X}$ is rotationally invariant.

5. For an introduction to the properties of subexponential distributions, we again refer the reader to Vershynin (2011).

Proof Using identity (3.2), we rewrite the first claim as

$$\|\mathbf{M}^r_{\mathbf{X}}\|_2^2 \geq \left\|\mathbf{M}^r_{\mathbf{X}_{rot}}\right\|_2^2,$$

and this follows immediately from equation (3.3).

If equality holds for all positive integers r, then by (3.3), $\mathbf{E}^r_{\mathbf{X}} = 0$ for all r, implying that $\mathbf{X}$ and $\mathbf{X}_{rot}$ have the same moment tensors of all orders. Since subexponential random variables are characterized by their moments (see Lemma 24), $\mathbf{X}$ and $\mathbf{X}_{rot}$ have the same distribution. ■

Proof [Proof of Theorem 4] If $\mathbf{X}$ has the same radial distribution as $\mathbf{g}$, then $\mathbf{g}$ is the rotational symmetrization of $\mathbf{X}$. The claim is then a direct application of the uniqueness portion of Theorem 12. ■

We now move on to proving the robust version of the test, namely Theorem 6.

Lemma 13 *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. Let $\boldsymbol{\theta}$ be uniformly distributed on the sphere S^{n-1}. Then the following hold for any positive integer r:*

a) $\boldsymbol{M}^r_{\boldsymbol{X}_{rot}} = \mathbb{E}\{\|\boldsymbol{X}\|_2^r\}\boldsymbol{M}^r_{\boldsymbol{\theta}}$.

b) $\|\boldsymbol{E}^r_{\boldsymbol{X}}\|_2^2 = (\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle\}^r) - (\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^2(\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle\}^r)$.

c) For any unit vector $v \in \mathbb{R}^n$,

$$\begin{aligned}|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \leq |\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}|\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \\ + \left(\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - (\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\right)^{1/2}.\end{aligned} \tag{3.7}$$

d) In particular, when r is odd,

$$|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \leq \left(\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\}\right)^{1/2} = \left|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \left(\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}\right)\right|^{1/2}. \tag{3.8}$$

Proof Deferred to Appendix B. ■

By balancing the two terms on the right hand side in part c), we obtain the following lemma, whose proof is again deferred to Appendix B.

Lemma 14 *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$ for $n \geq 2$. Suppose there is a unit vector $v \in S^{n-1}$, an even integer $r \geq 2$, and a positive number $0 < \delta \leq 1$ such that $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{v}\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}| \geq \delta\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\}$. Then either*

$$|\mathbb{E}\{\|\boldsymbol{X}\|_2^r\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^r\}| \geq \frac{\delta^2}{4}\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\} \quad \textit{or} \quad \left|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}\right| \geq \frac{15\delta^2}{64}(\mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{v}\rangle^r\})^2.$$

Proof [Proof of Theorem 6] If r is odd, then the statement follows from (3.8). If r is even, set $\delta = \frac{D_{\mathbf{X},r}}{\mathbb{E}\langle \mathbf{g}, \mathbf{v}\rangle^r}$ in the previous theorem. ■

4. Proof of the second Gaussian test

In this section, we return to the setting where $\mathbf{X}$ follows the NGCA model (1.1). We further assume that the non-Gaussian component $\tilde{\mathbf{X}}$ is a sub-Gaussian random vector with sub-Gaussian norm bounded by K. In order not to break the flow of the paper, most of the proofs are deferred to Appendix C.

The first step in proving the test is to notice that the independence of the Gaussian and non-Gaussian components allows us to block diagonalize the test matrices.

Lemma 15 (Block diagonalization for $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$) *Assume E is spanned by the first d basis vectors. Then the test matrices $\mathbf{\Phi}_{\boldsymbol{X},\alpha}$ and $\mathbf{\Psi}_{\boldsymbol{X},\alpha}$ decompose into blocks in the following manner:*

$$\mathbf{\Phi}_{\boldsymbol{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha} & 0 \\ \hline 0 & \mathbf{\Phi}_{\boldsymbol{g},\alpha} \end{array}\right), \qquad \mathbf{\Psi}_{\boldsymbol{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha} & 0 \\ \hline 0 & \mathbf{\Psi}_{\boldsymbol{g},\alpha} \end{array}\right). \tag{4.1}$$

We then observe that the trace of the test matrices are conveniently equal to the negated log derivatives of their respective partition functions.

Lemma 16 (Trace of $\mathbf{\Phi}_{\mathbf{Y},\alpha}$ and $\mathbf{\Psi}_{\mathbf{Y},\alpha}$) *Let $\boldsymbol{Y}$ be any random vector in $\mathbb{R}^n$. Then $\operatorname{Tr}(\mathbf{\Phi}_{\boldsymbol{Y},\alpha}) = -(\log Z_{\mathbf{\Phi},\boldsymbol{Y}})'(\alpha)$ and $\operatorname{Tr}(\mathbf{\Psi}_{\boldsymbol{Y},\alpha}) = -(\log Z_{\mathbf{\Psi},\boldsymbol{Y}})'(\alpha)$.*

Our next lemma shows that for α small enough, the partition functions themselves differentiate between Gaussian and non-Gaussian random vectors. This is obvious once we realize that they are just the moment generating functions of $\|\mathbf{X}\|_2^2$ and $\langle \mathbf{X}, \mathbf{X}' \rangle$, and that these are analytic in a small neighborhood around 0.

Lemma 17 (Partition functions characterize Gaussian distributions) *The following hold for any sub-Gaussian random vector $\boldsymbol{Y}$:*

a) If $Z_{\mathbf{\Phi},\boldsymbol{Y}}(\alpha_k) = Z_{\mathbf{\Phi},\boldsymbol{g}}(\alpha_k)$ for a sequence of values α_k converging to 0, then $\boldsymbol{Y}$ has the same radial distribution as $\boldsymbol{g}$.

b) If in addition, $Z_{\mathbf{\Psi},\boldsymbol{Y}}(\beta_k) = Z_{\mathbf{\Psi},\boldsymbol{g}}(\beta_k)$ for a sequence of values β_k converging to 0, then $\boldsymbol{X}$ has the standard Gaussian distribution.

We are now in a position to prove the second Gaussian test.

Proof [Proof of Theorem 5] Let $\mathbf{g}_d$ denote the standard Gaussian in $\mathbb{R}^d$. By Lemma 17, either $Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha) \neq Z_{\mathbf{\Phi},\mathbf{g}_d}(\alpha)$ for $|\alpha|$ small enough, or $Z_{\mathbf{\Psi},\tilde{\mathbf{X}}}(\alpha) \neq Z_{\mathbf{\Psi},\mathbf{g}_d}(\alpha)$ for $|\alpha|$ small enough. As such, either $(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Phi},\mathbf{g}_d})'(\alpha)$ or $(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Psi},\mathbf{g}_d})'(\alpha)$. Assume the former holds, and let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha}$. Since we may write $\mathbf{\Phi}_{\mathbf{X},\alpha}$ in a block form, these eigenvalues are either those of $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}$ or $\mathbf{\Phi}_{\mathbf{g},\alpha}$. Without loss of generality, we may assume that $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}$, and $\lambda_{d+1}, \ldots, \lambda_n$ are those of $\mathbf{\Phi}_{\mathbf{g},\alpha}$.

Lemma 31 tells us that $\lambda_{d+1} = \cdots = \lambda_n = (2\alpha+1)^{-1}$. On the other hand, by Lemma 16,

$$\sum_{i=1}^{d} \lambda_i = \operatorname{Tr}(\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}) = -(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'(\alpha).$$

By Lemma 30, $-(\log Z_{\mathbf{\Phi},\mathbf{g}_d})'(\alpha) = d(2\alpha+1)^{-1}$, so we have $\sum_{i=1}^{d} \lambda_i \neq d(2\alpha+1)^{-1}$. Dividing through by d, we get $\frac{1}{d}\sum_{i=1}^{d} \lambda_i \neq (2\alpha+1)^{-1}$, which implies that at least one λ_i differs from this value for $1 \leq i \leq d$.

If it were the case that $(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'(\alpha) \neq (\log Z_{\mathbf{\Psi},\mathbf{g}_d})'(\alpha)$, a similar argument involving $\mathbf{\Psi}_{\mathbf{X},\alpha}$ gives the alternate conclusion. ■

It is tedious but not too difficult to make the second Gaussian test quantitative. We do this by tracking how the non-Gaussian moments for $\|\tilde{\mathbf{X}}\|_2$ and $\langle \tilde{\mathbf{X}}, \tilde{\mathbf{X}}' \rangle$ contribute to the power series expansions for $-(\log Z_{\mathbf{\Phi},\tilde{\mathbf{X}}})'$ and $-(\log Z_{\mathbf{\Psi},\tilde{\mathbf{X}}})'$ around 0. This yields the following theorem.

Theorem 18 (Second Gaussian test, robust) *Let r be the integer such that $D_{\tilde{\boldsymbol{X}},r} > 0$ and $D_{\tilde{\boldsymbol{X}},r'} = 0$ for all $r' < r$. Then either*

a) for $|\alpha| \leq \eta_r^2 r/[(CK^2)^r(d^{r+1} + (r+1)!)]$, we have

$$\left| \frac{1}{d} \sum_{i=1}^{d} \lambda_i(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{\alpha}{\alpha^2 - 1} \right| \geq \frac{c\eta_r^2}{d(r-1)!} |\alpha|^{r-1}, \tag{4.2}$$

b) or for $|\alpha| \leq \eta_r^2 r/[(CK^2)^{r/2}\tilde{\gamma}_r(d^{r/2+1} + (r/2+1)!)]$, we have

$$\left| \frac{1}{d} \sum_{i=1}^{d} \lambda_i(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{1}{2\alpha + 1} \right| \geq \frac{c\eta_r^2}{d(r/2-1)!\tilde{\gamma}_r} |\alpha|^{r/2-1}. \tag{4.3}$$

Here $\tilde{\gamma}_r = \mathbb{E}|\langle \boldsymbol{g}, \boldsymbol{v} \rangle|^r$ for an arbitrary vector $v \in S^{n-1}$ and $\eta_r = \min\{D_{\boldsymbol{X},r}, \tilde{\gamma}_r\}$.

5. Proof of guarantee for Reweighted PCA

The second Gaussian test tells us how we can recover non-Gaussian directions from $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$. Our guarantee for Reweighted PCA algorithm shows that we can do the same with the plug-in estimators $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$. To this end, we first provide concentration bounds for these estimators, whose proofs can be found in Appendix E.

Theorem 19 (Concentration for $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$) *There is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, and any $0 \leq \alpha < 1/[CK^2 n]$, we have $\mathbb{P}\{\|\hat{\mathbf{\Phi}}_{\boldsymbol{X},\alpha} - \mathbf{\Phi}_{\boldsymbol{X},\alpha}\| > \epsilon\} \leq \delta$ so long as $N \geq CK^2(n + \log(1/\delta))\epsilon^{-2}$.*

Theorem 20 (Concentration for $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha}$) *There is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, if $N \geq CK^2(n+\log(1/\delta))\epsilon^{-2}$ and $|\alpha| \leq 1/[CK^2\tau(n+\tau)]$, we have $\mathbb{P}\{\|\hat{\mathbf{\Psi}}_{\boldsymbol{X},\alpha} - \mathbf{\Psi}_{\boldsymbol{X},\alpha}\| > \epsilon\} \leq \delta$. Here, $\tau = \log^{1/2}(N/\min\{\delta, K\epsilon\})$.*

Lemma 21 (Guarantee for $\hat{E}_{\mathbf{\Phi}}$) *Suppose the moments of $\|\tilde{\boldsymbol{X}}\|_2^2$ and $\|\boldsymbol{g}_d\|_2^2$ agree up to order $r-1$, but there is a number $\Delta > 0$ such that $\left|\mathbb{E}\{\|\tilde{\boldsymbol{X}}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}_d\|_2^{2r}\}\right| \geq \Delta$. For any $\delta, \epsilon \in (0,1)$, pick α_1 such that $0 < \alpha_1 < \min\{\Delta r/[(CK^2)^r(d^{r+1} + (r+1)!)], 1/[CK^2 n]\}$, and $\beta_1 = \Delta\alpha_1^{r-1}/[4d(r-1)!]$. Then with probability at least $1-\delta$, Reweighted PCA with $2N \geq CK^2 d^{3/2}(n+\log(1/\delta))/\beta_1^2\epsilon^2$ samples together with this choice of α_1 and β_1 produces a nontrivial estimate $\hat{E}_{\mathbf{\Phi}}$ of dimension $1 \leq \hat{d}_{\mathbf{\Phi}} \leq d$, such that there is a $\hat{d}_{\mathbf{\Phi}}$-dimensional subspace $E_{\mathbf{\Phi}} \subset E$ satisfying $d(\hat{E}_{\mathbf{\Phi}}, E_{\mathbf{\Phi}}) \leq \epsilon$.*

Proof Deferred to Appendix G ■

Lemma 22 (Guarantee for $\hat{E}_{\mathbf{\Psi}}$) *Suppose the moments of $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ and $\langle \boldsymbol{g}, \boldsymbol{g}' \rangle$ agree up to order $r-1$ but $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}' \rangle^r\}| \geq \Delta$. For any $\delta, \epsilon, \tau \in (0,1)$, pick $0 < \alpha_2 < \min\{\Delta r/[(CK^2)^r(d^{r+1} + (r+1)!)], 1/[CK^2 n^{1+\tau}]\}$, and $\beta_2 = \Delta \alpha_2^{r-1}/4d(r-1)!$. Then with probability at least $1-\delta$, Reweighted PCA with sample size $2N$ satisfying $\exp(n^{2\tau}) \min\{\delta, K\epsilon\} \geq 2N \geq CK^2 d^{3/2}(n + \log(1/\delta))/\beta_2^2 \epsilon^2$, together with this choice of α_2 and β_2 produces a nontrivial estimate $\hat{E}_{\mathbf{\Psi}}$ of dimension $1 \leq \hat{d}_{\mathbf{\Psi}} \leq d$, such that there is a $\hat{d}_{\mathbf{\Psi}}$-dimensional subspace $E_{\mathbf{\Psi}} \subset E$ satisfying $d(\hat{E}_{\mathbf{\Psi}}, E_{\mathbf{\Psi}}) \leq \epsilon$.*

Proof The proof is completely analogous to that for the previous theorem, except that we replace our estimates and identities for $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ with those for $\mathbf{\Psi}_{\mathbf{X},\alpha_2}$ wherever necessary. ■

Proof [Proof of Theorem 7] Combine the last two lemmas with Theorem 18 from the last section. ■

Remark 23 (Selecting optimal parameters) *If the problem parameters d, n, r, K and $D_{\tilde{\boldsymbol{X}},r}$ were known before hand, then in principle, one could compute the optimal tuning parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$. In practice, however, one rarely is in this situation, so one would have to estimate the problem parameters as a first step to solving the NGCA problem. Nonetheless, one can do this by the doubling/halving trick. In other words, we start with some fixed initial choice of α_1 and α_2. Using Theorems 19 and 20, we can detect whether there are any outlier eigenvalues with high probability. If there are none, we halve α_1 and α_2 and try again, repeating this process until outliers show up. The number of iterations is then the base 2 logarithm of the final α_1 and α_2, plus an additive constant. This is at most polynomial in all the problem parameters, so the algorithm remains efficient.*

6. Discussion

We have presented and analyzed an algorithm that is guaranteed to return at least one non-Gaussian direction efficiently, with sample and time complexity a polynomial in the problem parameters for a fixed r, where r is the smallest order at which $\tilde{\mathbf{X}}$ has positive r-th moment distance from a standard Gaussian. Furthermore, if $\tilde{\mathbf{X}}$ is (m, η)-moment-identifiable, then the algorithm estimates the d-dimensional non-Gaussian subspace efficiently with polynomial time and sample complexity for fixed m and d.

Since the degree of the polynomial increases linearly in r, it would seem that the algorithm is practically useless if r is larger than a small constant. However, note that having all third and fourth moments equal those of a Gaussian is a condition that is already stringent in one dimension, and which becomes even more so in higher dimensions. As such, unless $\tilde{\mathbf{X}}$ has some kind of adversarial distribution, r will be either 4 or 3, depending on whether $\tilde{\mathbf{X}}$ is centrally symmetric or not.

The algorithm also often delivers much more than is guaranteed for several reasons. First, in order to bound the subspace perturbation by ϵ, we used a very crude estimate of the eigengap, bounding it from below using the pigeonhole principle, which in the worst case assumes that the eigenvalues are spread out at regular intervals. This should not happen in practice, and we expect the non-Gaussian eigenvalues to instead cluster relatively tightly around their average. If this happens, the sample complexity requirement can be relaxed by a factor of d.

Second, just as it is extremely unlikely for r to be higher than 4, for a general non-Gaussian $\tilde{\mathbf{X}}$ and a small, random α, it is extremely unlikely for any of the non-Gaussian values of $\mathbf{\Phi}_{\mathbf{X},\alpha}$ to be equal to the Gaussian one on the dot. This means that even though the guarantee for a single run of the base algorithm is for one direction, in practice we most probably can recover the entire subspace E simultaneously with just $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$ alone (as in the case in Corollary 10), albeit with a more sophisticated truncation technique.

6.1. Conjectures and questions

We conjecture that REWEIGHTED PCA actually recovers the entire non-Gaussian subspace E with in polynomial time and sample complexity if we fix m, but now allow d to vary. This would improve upon both our result and that of Vempala and Xiao (2011). The first Gaussian test for a random vector $\mathbf{X}$ using the distribution of its norm and dot product pairing also leads to further questions. For a fixed nonzero real number t, both of these appear in the formula for $\|\mathbf{Y}_t\|_2^2$, where we set $\mathbf{Y}_t := \mathbf{X} + t\mathbf{X}'$, so it is natural to ask whether Reweighted PCA works with $\mathbf{\Phi}_{\mathbf{Y}_t,\alpha}$ alone for some t. In particular, does it work for $t = -1$? It is also an open question whether $\langle \mathbf{X}, \mathbf{X}' \rangle$ alone is sufficient to test whether $\mathbf{X}$ is standard Gaussian.

Acknowledgements

Both authors are partially supported by NSF Grant DMS 1265782 and U.S. Air Force Grant FA9550-14-1-0009.

References

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with Unknown Gaussian Noise, and Implications for Gaussian Mixtures and Autoencoders. *Algorithmica*, 72 (1):215–236, 2015.

Patrick Billingsley. *Probability and Measure - Third Edition*. 1995.

Gilles Blanchard, Motoaki Kawanabe, Masashi Sugiyama, Vladimir Spokoiny, and Klaus-Robert Müller. In Search of Non-Gaussian Components of a High-Dimensional Distribution. *Journal of Machine Learning Research*, 7(2):247–282, 2006.

S. Charles Brubaker and Santosh S. Vempala. Isotropic PCA and affine-invariant clustering. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 551–560, 2008.

E. Çınlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer New York, 2011.

Pierre Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Elmar Diederichs, Anatoli Juditsky, Vladimir Spokoiny, and Christof Schütte. Sparse non-Gaussian component analysis. *IEEE Transactions on Information Theory*, 56(6):3033–3047, 2010.

Elmar Diederichs, Anatoli Juditsky, Arkadi Nemirovski, and Vladimir Spokoiny. Sparse non Gaussian component analysis by semidefinite programming. *Machine Learning*, 91(2):211–238, 2013.

A. Frieze, M. Jerrum, and Ravi Kannan. Learning linear transformations. *Proceedings of 37th Conference on Foundations of Computer Science*, pages 359–368, 1996.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing - STOC '14*, number c, pages 584–593, New York, New York, USA, 2014. ACM Press.

Peter J Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435–475, 2007.

Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, may 1999.

Motoaki Kawanabe and Fabian J. Theis. Estimating non-Gaussian subspaces by characteristic functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3889 LNCS:157–164, 2006.

Motoaki Kawanabe and Fabian J. Theis. Joint low-rank approximation for extracting non-Gaussian subspaces. *Signal Processing*, 87(8):1890–1903, 2007.

Motoaki Kawanabe, Masashi Sugiyama, Gilles Blanchard, and Klaus-Robert Müller. A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75, 2006.

M. Kawanabe. Linear dimension reduction based on the fourth-order cumulant tensor. *Proc. of Artifical Neural Networks – ICANN 2005*, pages 151–156, 2005.

Hiroaki Sasaki, Aapo Hyvärinen, and Masashi Sugiyama. Clustering via mode seeking by direct estimation of the gradient of a log-density. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8726 LNAI (PART 3):19–34, 2014.

Hiroaki Sasaki, Gang Niu, and Masashi Sugiyama. Non-Gaussian Component Analysis with Log-Density Gradient Estimation. *International Conference on Artificial Intelligence and Statistics*, 51:1–20, 2016.

Yan Shuo Tan. Energy optimization for distributions on the sphere and improvement to the welch bounds. *Electron. Commun. Probab.*, 22:12 pp., 2017.

Santosh S. Vempala and Ying Xiao. Structure from Local Optima: Learning Subspace Juntas via Higher Order PCA. *arXiv:1108.3329*, 2011.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, 2011.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Appendix A. Equivalence of NGCA models

In this section, we note the equivalence of several formulations of the NGCA model used in the literature. First, the isotropic NGCA model (1) can be written equivalently as

$$F(\mathbf{x}) = H(\mathbf{P}_E(\mathbf{x}))G(\mathbf{P}_{E^\perp}(\mathbf{x})),$$

where F is the distribution of $\mathbf{X}$, H is the distribution of $\tilde{\mathbf{X}}$, and G is the standard normal distribution. This is the way in which Vempala and Xiao (2011) stated the NGCA model.

Next, consider the model

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{g},$$

where now $\tilde{\mathbf{X}} \in E$ as before, but $\mathbf{g}$ is a centered Gaussian in $\mathbb{R}^n$ with arbitrary covariance. As a special case of this, we have $\mathbf{X} = (\tilde{\mathbf{X}}, \mathbf{g}) \in E \oplus E'$, where E and E' are complementary but not necessarily orthogonal. Let $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X})$, and consider the whitened distribution $\mathbf{\Sigma}^{-1/2}\mathbf{X} = \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}} + \mathbf{\Sigma}^{-1/2}\mathbf{g}$. Now the non-Gaussian subspace is $\mathbf{\Sigma}^{-1/2}E$, which we assume without loss of generality to be the span of the first d coordinate vectors. This means that $\mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}})$ only has nonzero entries in its top left d by d block. Since we can decompose

$$\mathbf{I}_n = \mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{X}) = \mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}}) + \mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{g}),$$

this in turn implies that we can write

$$\mathrm{Cov}(\mathbf{\Sigma}^{-1/2}\mathbf{g}) = \left(\begin{array}{c|c} \mathbf{A} & 0 \\ \hline 0 & \mathbf{I}_{n-d} \end{array}\right),$$

where $\mathbf{A}$ is a PSD matrix such that $\mathbf{A} = \mathbf{I}_d - \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}}$. Because of this structure, we have $\mathbf{\Sigma}^{-1/2}\mathbf{g} = (\tilde{\mathbf{h}}, \mathbf{h}) \in E \oplus E^\perp$, with $\tilde{\mathbf{h}} \sim \mathcal{N}(\mathbf{0}, \mathbf{A})$ and $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-d})$. Since these two Gaussian components have zero correlation, they are independent. Since a non-Gaussian distribution remains non-Gaussian after convolution with a Gaussian, if we set $\tilde{\mathbf{Y}} := \mathbf{\Sigma}^{-1/2}\tilde{\mathbf{X}} + \tilde{\mathbf{h}}$ to be our new non-Gaussian component, we see that we have again produced an instance of (1).

This additive model seems to be the most common formulation of NGCA in the literature (see Blanchard et al. (2006); Kawanabe et al. (2006), etc.). It can also be equivalently written as

$$F(\mathbf{x}) = H(\mathbf{P}_E(\mathbf{x}))G(\mathbf{x}), \tag{A.1}$$

where G is now a centered Gaussian density with arbitrary covariance, and H is now just some function. See Lemma 1 in Blanchard et al. (2006) for more details.

Appendix B. Details for Section 3

Let $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ be a convex, increasing function with $\psi(0) = 0$. We define the *Orlicz norm* of a random variable X with respect to ψ as

$$\|X\|_\psi := \inf\{\lambda > 0 : \mathbb{E}\{\psi(|X|/\lambda)\} \leq 1\}. \tag{B.1}$$

Equipped with this norm, the space of random variables with finite norm forms a Banach space, called an *Orlicz space*. For $\alpha \geq 1$, set $\psi_\alpha(x) := \exp(x^\alpha) - 1$. Elements of the ψ_1 Orlicz space are called *subexponential* random variables. Similarly, elements of the ψ_2 Orlicz space are called

sub-Gaussian random variables. The reason for this terminology is that the finiteness of (B.1) is equivalent to tail decay conditions for $\psi = \psi_\alpha$ (see Vershynin (2011).) Finally, we say that a random vector $\mathbf{X}$ in $\mathbb{R}^n$ is sub-Gaussian (respectively subexponential) if all its 1-dimensional marginals are sub-Gaussian (respectively subexponential).

Lemma 24 *Let $\mathbf{X}$ be a subexponential random vector in $\mathbb{R}^n$. Then the distribution of $\mathbf{X}$ is determined by its moment tensors.*

Proof Let $\phi_{\mathbf{X}}(\mathbf{v}) = \mathbb{E}\{e^{i\langle \mathbf{X},\mathbf{v}\rangle}\}$ denote the characteristic function of $\mathbf{X}$, and let $K = \|\mathbf{X}\|_{\psi_1}$ denote the subexponential norm of $\mathbf{X}$. We then have the following moment growth condition (see Vershynin (2011)):

$$\sup_{\mathbf{v}\in S^{n-1}} \limsup_{r\to\infty} \frac{(\mathbb{E}\{|\langle \mathbf{X},\mathbf{v}\rangle|^r\})^{1/r}}{r} \lesssim K. \tag{B.2}$$

This condition implies that for each $\mathbf{v} \in S^{n-1}$, the function $t \mapsto \mathbb{E}\{e^{it\langle \mathbf{X},\mathbf{v}\rangle}\}$ can be written as a power series with coefficients $\frac{\mathbb{E}\{\langle \mathbf{X},\mathbf{v}\rangle^r\}}{r!}$ (see Billingsley (1995)), so $\phi_{\mathbf{X}}(\mathbf{v})$ is determined by the moments $\mathbb{E}\{\langle \mathbf{X},\mathbf{v}\rangle^r\}$. By (3.1), $\mathbb{E}\{\langle \mathbf{X},\mathbf{v}\rangle^r\} = \langle \mathbf{M}^k_{\mathbf{X}}, \mathbf{v}^{\otimes r}\rangle$, so these are functions of the moment tensors. Finally, it is a fact from elementary probability that $\mathbf{X}$ is determined by its characteristic function (see Çınlar (2011)). ■

Proof [Proof of Lemma 13] For the first statement, observe that $\mathbf{X}_{rot} = \|\mathbf{X}\|_2\boldsymbol{\theta}$, with $\|\mathbf{X}\|_2$ and $\boldsymbol{\theta}$ independent. We thus have

$$\mathbf{M}^r_{\mathbf{X}_{rot}} = \mathbb{E}\{(\|\mathbf{X}\|_2\boldsymbol{\theta})^{\otimes r}\} = \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbb{E}\{\boldsymbol{\theta}^{\otimes r}\} = \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}}.$$

Next, rewrite (3.3) as $\|\mathbf{E}^r_{\mathbf{X}}\|_2^2 = \|\mathbf{M}^r_{\mathbf{X}}\|_2^2 - \|\mathbf{M}^r_{\mathbf{X}_{rot}}\|_2^2$. By (3.2), we have $\|\mathbf{M}^r_{\mathbf{X}}\|_2^2 = \mathbb{E}\{\langle \mathbf{X},\mathbf{X}'\rangle^r\}$ and using a), we get $\|\mathbf{M}^r_{\mathbf{X}_{rot}}\|_2^2 = (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2\mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\}$.

To prove part c), fix v and write

$$\mathbb{E}\{\langle \mathbf{X},\mathbf{v}\rangle^r\} - \mathbb{E}\{\langle \mathbf{g},\mathbf{v}\rangle^r\} = \langle \mathbf{M}^r_{\mathbf{X}} - \mathbf{M}^r_g, \mathbf{v}^{\otimes r}\rangle = \langle \mathbf{M}^r_{\mathbf{X}_{rot}} - \mathbf{M}^r_g, \mathbf{v}^{\otimes r}\rangle + \langle \mathbf{E}^r_{\mathbf{X}}, \mathbf{v}^{\otimes r}\rangle.$$

We use a) to write

$$\langle \mathbf{M}^r_{\mathbf{X}_{rot}} - \mathbf{M}^r_g, \mathbf{v}^{\otimes r}\rangle = \langle \mathbb{E}\{\|\mathbf{X}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}\mathbf{M}^r_{\boldsymbol{\theta}}, \mathbf{v}^{\otimes r}\rangle = (\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\})\mathbb{E}\{\langle \boldsymbol{\theta},\mathbf{v}\rangle^r\}.$$

Notice that $\mathbb{E}\{\langle \boldsymbol{\theta},\mathbf{v}\rangle^r\} = \mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\}$. We then combine the last two equations with b) and Cauchy-Schwarz to get (3.7). Finally, to get the last claim, we use the fact that $\mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} = \mathbb{E}\{\langle \mathbf{g},\mathbf{g}'\rangle^r\} = 0$ whenever r is odd. ■

Proof [Proof of Lemma 14] Observe that (3.7) gives the bound

$$\delta\mathbb{E}\{\langle \mathbf{g},\mathbf{v}\rangle^r\} \le |\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}|\mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} + \Big(\mathbb{E}\{\langle \mathbf{X},\mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2\big(\mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\}\big)\Big)^{1/2}. \tag{B.3}$$

Suppose $|\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}| \le \frac{\delta^2}{4}\mathbb{E}\{\langle \mathbf{g},\mathbf{v}\rangle^r\}$. Then the second term on the right in equation (B.3) has to be large. Indeed, since $\delta \le 1$ and

$$\mathbb{E}\{\langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle^r\} = \mathbb{E}\{\theta_1^r\} \le \mathbb{E}\{\theta_1^2\} = \frac{1}{n},$$

we have, for $n \geq 2$, that

$$\left(\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\right)^{1/2} \geq \delta \mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} - \frac{\delta^2}{4}\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}$$
$$\geq \frac{7\delta}{8}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}.$$

Now, applying the fact that $\mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\} = (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2 \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$, we use the reverse triangle inequality and the above bound to write

$$\left|\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - \mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\}\right| \geq \left|\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}\right| \tag{B.4}$$
$$- \left|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\right| \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$$
$$\geq \left(\frac{7\delta}{8}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}\right)^2 - \left|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\right| \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}. \tag{B.5}$$

Next, notice that

$$\left|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\right| = |\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}|(\mathbb{E}\{\|\mathbf{X}\|_2^r\} + \mathbb{E}\{\|\mathbf{g}\|_2^r\})$$
$$= |\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}| \cdot 2\mathbb{E}\{\|\mathbf{g}\|_2^r\} + (\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\})^2,$$

so by the assumption on $|\mathbb{E}\{\|\mathbf{X}\|_2^r\} - \mathbb{E}\{\|\mathbf{g}\|_2^r\}|$, we have

$$\left|(\mathbb{E}\{\|\mathbf{X}\|_2^r\})^2 - (\mathbb{E}\{\|\mathbf{g}\|_2^r\})^2\right| \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \leq \frac{\delta^2}{4}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\} \cdot 2\mathbb{E}\{\|\mathbf{g}\|_2^r\} \cdot \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\} \tag{B.6}$$
$$+ \left(\frac{\delta^2}{4}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}\right)^2 \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$$
$$= \frac{\delta^2}{2}(\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\})^2 + \left(\frac{\delta^2}{4}\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\}\right)^2 \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{\theta}'\rangle^r\}$$
$$\leq \frac{17\delta^2}{32}(\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\})^2. \tag{B.7}$$

We can now substitute (B.6) into (B.4) to get

$$\left|\mathbb{E}\{\langle \mathbf{X}, \mathbf{X}'\rangle^r\} - \mathbb{E}\{\langle \mathbf{g}, \mathbf{g}'\rangle^r\}\right| \geq \frac{15\delta^2}{64}(\mathbb{E}\{\langle \mathbf{g}, \mathbf{v}\rangle^r\})^2.$$

■

Appendix C. Details for Section 4

Proof [Proof of Lemma 15] The decompositions follow easily from the independence of the two components of the mixed vector, $\tilde{\mathbf{X}}$ and $\mathbf{g}$, as well as the unconditional symmetry of the Gaussian

component. Let us illustrate this by proving the decomposition for $\mathbf{\Phi}_{\mathbf{X},\alpha}$. First, note that $e^{-\alpha\|\mathbf{X}\|_2^2} = e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}e^{-\alpha\|\mathbf{g}\|_2^2}$, so that $Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) = Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha)Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)$. The top left d by d block is hence given by

$$\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\}}{Z_{\mathbf{\Phi},\tilde{\mathbf{X}}}(\alpha)} = \mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha}.$$

The bottom right d' by d' block is also computed similarly. Finally, any entry outside these two blocks is of the form

$$\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i\mathbf{g}_j\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = \frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i(-\mathbf{g}_j)\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = -\frac{\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\tilde{\mathbf{X}}_i\mathbf{g}_j\}}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = 0.$$

■

Proof [Proof of Lemma 16] We have

$$\mathrm{Tr}(\mathbf{\Phi}_{\mathbf{X},\alpha}) = \frac{\mathbb{E}\|\mathbf{X}\|_2^2 e^{-\alpha\|\mathbf{X}\|_2^2}}{\mathbb{E}e^{-\alpha\|\mathbf{X}\|_2^2}} = \frac{-Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} = -(\log Z_{\mathbf{\Phi},\mathbf{X}})'(\alpha).$$

The calculation for $\mathbf{\Psi}_{\mathbf{X},\alpha}$ is similar. ■

In order to prove Lemma 17, we first need to establish the analyticity for the two partition functions.

Lemma 25 (Analyticity for $Z_{\mathbf{\Phi},\mathbf{X}}$ and $Z_{\mathbf{\Psi},\mathbf{X}}$) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded by $K \geq 1$. The functions $Z_{\mathbf{\Phi},\boldsymbol{X}}$ and $Z_{\mathbf{\Psi},\boldsymbol{X}}$ are both analytic on $(-1/CK^2, 1/CK^2)$. They are given by the formulae $Z_{\mathbf{\Phi},\boldsymbol{X}}(\alpha) = \sum_{r=0}^{\infty}\mathbb{E}\{\|\boldsymbol{X}\|_2^{2r}\}(-\alpha)^r/r!$ and $Z_{\mathbf{\Psi},\boldsymbol{X}}(\alpha) = \sum_{r=0}^{\infty}\mathbb{E}\{\langle\boldsymbol{X},\boldsymbol{X}'\rangle^r\}(-\alpha)^r/r!$. Furthermore, by choosing C sufficiently large, on this interval they satisfy the bounds*

$$|Z_{\mathbf{\Phi},\boldsymbol{X}}(\alpha)|, |Z_{\mathbf{\Psi},\boldsymbol{X}}(\alpha)| \leq e^{CK^2n|\alpha|} + \frac{CK^2|\alpha|}{1-CK^2|\alpha|}. \tag{C.1}$$

Proof Let us first prove the bounds in (C.1). Observe that

$$\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\} \leq \mathbb{E}\{e^{|\alpha|\|\mathbf{X}\|_2^2}\} = \sum_{n=0}^{\infty}\frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2n}\}}{n!}|\alpha|^n. \tag{C.2}$$

Here, Tonelli allows us to interchange the sum and expectation. We next use Lemma 32 to bound the terms of this series. Indeed, using the equivalent estimate (E.5), we have

$$\mathbb{E}\{\|\mathbf{X}\|_2^{2r}\} \leq C^r K^{2r}(n^r + r!)$$

for some universal constant C. Substituting this into (C.2) and using $|\alpha| \leq 1/CK^2$, we have

$$\begin{aligned}\mathbb{E}\{e^{-\alpha\|\mathbf{X}\|_2^2}\} &\leq \sum_{r=0}^{\infty}\frac{(CK^2)^r(n^r+r!)}{r!}|\alpha|^r \\ &= \sum_{r=0}^{\infty}\frac{(CK^2n|\alpha|)^r}{r!} + \sum_{r=1}^{\infty}(CK^2|\alpha|)^r \\ &= e^{CK^2n|\alpha|} + \frac{CK^2|\alpha|}{1-CK^2|\alpha|}.\end{aligned}$$

One may prove the bound for $Z_{\mathbf{\Psi},\mathbf{X}}$ by doing the same computation but using (E.2) instead of (E.1).

We next handle analyticity of $Z_{\mathbf{\Phi},\mathbf{X}}$. We shall prove by induction on r that we may differentiate under the integral sign to get the formula

$$Z^{(r)}_{\mathbf{\Phi},\mathbf{X}}(\alpha) = (-1)^r \mathbb{E}\{\|\mathbf{X}\|_2^{2r} e^{-\alpha\|\mathbf{X}\|_2^2}\}. \tag{C.3}$$

Assume the formula is true for all $r' < r$. Then

$$Z^{(r)}_{\mathbf{\Phi},\mathbf{X}}(\alpha) = (-1)^{r-1} \lim_{h\to 0} \mathbb{E}\{\frac{\|\mathbf{X}\|_2^{2r-2} e^{-(\alpha+h)\|\mathbf{X}\|_2^2} - \|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2}}{h}\} \tag{C.4}$$

$$= (-1)^r \lim_{h\to 0} \mathbb{E}\{\|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2} \frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h}\} \tag{C.5}$$

Next, note that the integrand is positive and by the mean value theorem, for a fixed value of $\|\mathbf{X}\|_2^2$, we have

$$\frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h} = \|\mathbf{X}\|_2^2 e^{-h'\|\mathbf{X}\|_2^2}$$

for some $h' \in [0, h]$ if $h > 0$ and $h' \in [h, 0]$ otherwise. As such, we have

$$\|\mathbf{X}\|_2^{2r-2} e^{-\alpha\|\mathbf{X}\|_2^2} \frac{1 - e^{-h\|\mathbf{X}\|_2^2}}{h} \leq \|\mathbf{X}\|_2^{2r} e^{(|h|-\alpha)\|\mathbf{X}\|_2^2}$$

For $|h| - \alpha \leq 1/CK^2$, one can easily show that this is integrable by expanding this as a power series in $\|\mathbf{X}\|_2^2$ and bounding the growth of the coefficients as above. As such, we may apply the Dominated Convergence Theorem to push the limit inside the expectation in (C.4), thereby yielding (C.3).

In particular, differentiating $Z_{\mathbf{\Phi},\mathbf{X}}$ at 0, we see that its Taylor series at 0 is given by

$$Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) \sim \sum_{r=0}^{\infty} \frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2r}\}}{r!}(-\alpha)^r. \tag{C.6}$$

The formula above shows that the Taylor series is absolutely convergent on our chosen interval. We next need to show that $Z_{\mathbf{\Phi},\mathbf{X}}$ agrees with its Taylor series on this interval, meaning we have to show that the remainder term for the r-th Taylor polynomial goes to zero pointwise. The Lagrange form of the remainder term is written as

$$R_{Z_{\mathbf{\Phi},\mathbf{X}},r}(\alpha) = \frac{Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}(\alpha')}{(r+1)!}\alpha^{r+1}$$

where $0 < |\alpha'| < |\alpha|$. Applying Cauchy-Schwarz to the formula (C.3), we get

$$|Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}(\alpha)| \leq \left(\mathbb{E}\{\|\mathbf{X}\|_2^{4r+2}\}\right)^{1/2} \left(\mathbb{E}\{e^{-2\alpha\|\mathbf{X}\|_2^2}\}\right)^{1/2}. \tag{C.7}$$

Lemma 32 again allows us to compute

$$\left(\mathbb{E}\{\|\mathbf{X}\|_2^{4r+2}\}\right)^{1/2} \leq (CK^2)^{r+1}(n^{r+1} + (r+1)!).$$

This implies that for any $C' > 2C$,

$$\begin{aligned}\|R_{Z_{\mathbf{\Phi},\mathbf{X},r}}\|_{L_\infty([-1/C'K^2,1/C'K^2])} &\leq \frac{\|Z^{(r+1)}_{\mathbf{\Phi},\mathbf{X}}\|_{L_\infty([-1/C'K^2,1/C'K^2])}}{(C'K^2)^{r+1}(r+1)!} \\ &\leq \frac{(CK^2)^{r+1}(n^{r+1}+(r+1)!)}{(C'K^2)^{r+1}(r+1)!}\left(\mathbb{E}\{e^{2\|\mathbf{X}\|_2^2/C'K^2}\}\right)^{1/2} \\ &\leq \left(\frac{C}{C'}\right)^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\left(e^{2nC/C'}+\frac{2C/C'}{1-2C/C'}\right).\end{aligned}$$

Using the fact that $r! \sim (\frac{r}{e})^r$, this last expression decays to zero as r tends to ∞. Finally, to prove the claim for $Z_{\mathbf{\Psi},\mathbf{X}}$, we repeat the same arguments. ■

Note that in the course of proving the last lemma, we have also proved the following result to be used elsewhere in the paper.

Lemma 26 (Taylor remainder terms for $Z_{\mathbf{\Phi},\mathbf{X}}$ and $Z_{\mathbf{\Psi},\mathbf{X}}$) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded above by $K \geq 1$. There is an absolute constant C such that for all $0 < \alpha < 1/CK^2$, on the interval $[-\alpha, \alpha]$, the remainder terms for the r-th degree Taylor polynomials for $Z_{\mathbf{\Phi},\boldsymbol{X}}$ and $Z_{\mathbf{\Psi},\boldsymbol{X}}$ at 0 satisfy the uniform bound*

$$\|R_{Z_{\mathbf{\Phi},\boldsymbol{X},r}}\|_\infty, \|R_{Z_{\mathbf{\Psi},\boldsymbol{X},r}}\|_\infty \leq (CK^2)^{r+1}\alpha^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\left(e^{CK^2\alpha n}+\frac{CK^2\alpha}{1-CK^2\alpha}\right) \tag{C.8}$$

Proof [Proof of Lemma 17] By Lemma 25, all four functions are analytic in a neighborhood of 0. Now recall that two different analytic functions cannot agree on a sequence with an accumulation point. ■

We now move on to proving Theorem 18. This requires the following technical lemma.

Lemma 27 *Let $\boldsymbol{X}$ be sub-Gaussian random vector in $\mathbb{R}^n$ with sub-Gaussian norm bounded above by $K \geq 1$. Suppose the moments of $\|\boldsymbol{X}\|_2^2$ and $\|\boldsymbol{g}\|_2^2$ agree up to order $r-1$, but there is a number $\Delta > 0$ such that $\left|\mathbb{E}\{\|\boldsymbol{X}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^{2r}\}\right| \geq \Delta$, then there is an absolute constant C such that for $|\alpha| \leq \Delta r/(CK^2)^r(n^{r+1}+(r+1)!)$, we have*

$$\left|(\log Z_{\mathbf{\Phi},\boldsymbol{X}})'(\alpha) - (\log Z_{\mathbf{\Phi},\boldsymbol{g}})'(\alpha)\right| \geq \frac{\Delta}{2(r-1)!}|\alpha|^{r-1}. \tag{C.9}$$

Similarly, suppose the moments of $\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$ and $\langle \boldsymbol{g}, \boldsymbol{g}'\rangle$ agree up to order $r-1$ but $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}'\rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}'\rangle^r\}| \geq \Delta$, then for $|\alpha| \leq \Delta r/(CK^2)^r(n^{r+1}+(r+1)!)$, we have

$$\left|(\log Z_{\mathbf{\Psi},\boldsymbol{X}})'(\alpha) - (\log Z_{\mathbf{\Psi},\boldsymbol{g}})'(\alpha)\right| \geq \frac{\Delta}{2(r-1)!}|\alpha|^{r-1}. \tag{C.10}$$

Proof Let us first prove (C.9). For every positive integer k, let $p_{\mathbf{X},k}(\alpha) = \sum_{j=0}^k \mathbb{E}\{\|\mathbf{X}\|_2^{2j}\}\alpha^j/j!$ denote the k-th Taylor polynomial of $Z_{\mathbf{\Phi},\mathbf{X}}$, and define $p_{\mathbf{g},k}$ analogously. For convenience, also denote the k-th Taylor remainder term as $R_{\mathbf{X},k} := R_{Z_{\mathbf{\Phi},\mathbf{X}},k}$. For any α, we then have

$$(\log Z_{\mathbf{\Phi},\mathbf{X}})'(\alpha) - (\log Z_{\mathbf{\Phi},\mathbf{g}})'(\alpha) = \frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}, \tag{C.11}$$

which we can then bound using

$$\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right| \geq \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}-\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}\right| - \left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| - \left|\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}-\frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right|. \tag{C.12}$$

We now bound each of these three terms individually. First, we need upper and lower bounds for $p_{\mathbf{X},k}(\alpha)$. Using the $\|\mathbf{X}\|_2^2$ moment bound (E.5), we have

$$\begin{aligned}
|p_{\mathbf{X},k}(\alpha)-1| &\leq \sum_{j=1}^{k}\frac{\mathbb{E}\{\|\mathbf{X}\|_2^{2j}\}|\alpha|^j}{j!} \\
&\leq \sum_{j=1}^{k}\frac{(CK^2)^j(n^j+j!)|\alpha|^j}{j!} \\
&= \sum_{j=1}^{k}\frac{(CK^2n|\alpha|)^j}{j!}+\sum_{j=1}^{k}(CK^2|\alpha|)^j \\
&\leq e^{CK^2n|\alpha|}-1+\frac{CK^2|\alpha|}{1-CK^2|\alpha|}.
\end{aligned}$$

By sharpening the constant C in our assumption on $|\alpha|$ if necessary, we may thus ensure that

$$|p_{\mathbf{X},k}(\alpha)-1| \leq \frac{1}{2} \tag{C.13}$$

By the same argument, we can also ensure that

$$\left|p'_{\mathbf{X},k}(\alpha)-\mathbb{E}\|\mathbf{X}\|_2^2\right| \leq \frac{1}{2}. \tag{C.14}$$

By our assumptions on the moments of $\|\mathbf{X}\|_2^2$ and $\|\mathbf{g}\|_2^2$, we have $p_{\mathbf{X},r-1} \equiv p_{\mathbf{g},r-1}$. Furthermore, only the leading terms of $p'_{\mathbf{X},r}$ and $p'_{\mathbf{g},r}$ differ. This, together with (C.13) implies that

$$\begin{aligned}
\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}-\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}\right| &\geq \frac{2}{3}\left|p'_{\mathbf{X},r}(\alpha)-p'_{\mathbf{g},r}(\alpha)\right| \\
&\geq \frac{2\Delta|\alpha|^{r-1}}{3(r-1)!}.
\end{aligned} \tag{C.15}$$

Next, we have

$$\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| \leq \left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| + \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right|. \tag{C.16}$$

Again we bound these two terms individually. Using the identity $p_{\mathbf{X},r}(\alpha) = p_{\mathbf{X},r-1}(\alpha) + \mathbb{E}\|\mathbf{X}\|_2^{2r}(-\alpha)^r/r!$, we get

$$\begin{aligned}
\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| &= \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right|\left|1-\frac{p_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| \\
&= \left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)p_{\mathbf{X},r-1}(\alpha)}\right|\frac{\mathbb{E}\|\mathbf{X}\|_2^{2r}|\alpha|^r}{r!}
\end{aligned} \tag{C.17}$$

Using the bounds on $p_{\mathbf{X},r}$ and $p'_{\mathbf{X},r}$ (C.13) and (C.14), together with the $\|\mathbf{X}\|_2^2$ moment bound (E.5), we get

$$\left|\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)p_{\mathbf{X},r-1}(\alpha)}\right|\frac{\mathbb{E}\|\mathbf{X}\|_2^{2r}|\alpha|^r}{r!} \leq \frac{3}{8}\frac{(CK^2)^r(n^r+r!)|\alpha|^r}{r!}. \tag{C.18}$$

For the first term in (C.16), we write

$$\begin{aligned}\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)} - \frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| &= \left|(\log Z_{\mathbf{\Phi},\mathbf{X}}(\alpha))' - (\log p_{\mathbf{X},r}(\alpha))'\right| \\ &= \left|\frac{d}{d\alpha}\log\left(\frac{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right| \\ &= \left|\frac{d}{d\alpha}\log\left(1+\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right|. \end{aligned} \tag{C.19}$$

Using Lemma 26 together with our assumptions on $|\alpha|$, we observe that

$$\begin{aligned}|R_{\mathbf{X},r}(\alpha)| &\leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\left(e^{CK^2|\alpha|n}+\frac{CK^2|\alpha|}{1-CK^2|\alpha|}\right) \\ &\leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right). \end{aligned} \tag{C.20}$$

In particular, by sharpening the constant C in our assumption on $|\alpha|$ if necessary, we can ensure that this quantity is less than $\frac{1}{4}$. In this case, we have

$$\left|\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| \leq \frac{1}{2},$$

so that

$$\begin{aligned}\left|\frac{d}{d\alpha}\log\left(1+\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right| &= \left|\log'\left(1+\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)\right|\left|\left(\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)'\right| \\ &\leq 2\left|\left(\frac{R_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right)'\right| \\ &\leq 2\left(\left|\frac{R'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| + \left|\frac{R_{\mathbf{X},r}(\alpha)p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)^2}\right|\right). \end{aligned} \tag{C.21}$$

By our bounds on these functions (C.13), (C.14), and (C.20), we have

$$\left|\frac{R_{\mathbf{X},r}(\alpha)p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)^2}\right| \leq (CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right). \tag{C.22}$$

Furthermore, by using the moment bounds (E.5) as before, one can show that

$$|R'_{\mathbf{X},r}(\alpha)| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right).$$

so that the first term is also bounded according to

$$\left|\frac{R'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \tag{C.23}$$

As such, combining (C.19) and (C.21) tells us that

$$\begin{aligned}\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r}(\alpha)}\right| &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right)+(CK^2)^{r+1}|\alpha|^{r+1}\left(\frac{n^{r+1}}{(r+1)!}+1\right)\\ &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right).\end{aligned} \tag{C.24}$$

We can now use this estimate together with (C.18) to continue (C.16), writing

$$\begin{aligned}\left|\frac{Z'_{\mathbf{\Phi},\mathbf{X}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)}-\frac{p'_{\mathbf{X},r}(\alpha)}{p_{\mathbf{X},r-1}(\alpha)}\right| &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right)+\frac{(CK^2)^r(n^r+r!)|\alpha|^r}{r!}\\ &\leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right).\end{aligned} \tag{C.25}$$

Notice that same methods also give us

$$\left|\frac{p'_{\mathbf{g},r}(\alpha)}{p_{\mathbf{g},r-1}(\alpha)}-\frac{Z'_{\mathbf{\Phi},\mathbf{g}}(\alpha)}{Z_{\mathbf{\Phi},\mathbf{g}}(\alpha)}\right| \leq (CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \tag{C.26}$$

We may therefore finally substitute these last two bounds, together with (C.15), into (C.12). This yields

$$\left|(\log Z_{\mathbf{\Phi},\mathbf{X}})'(\alpha)-(\log Z_{\mathbf{\Phi},\mathbf{g}})'(\alpha)\right| \geq \frac{2\Delta|\alpha|^{r-1}}{3(r-1)!}-C(CK^2)^r|\alpha|^r\left(\frac{n^{r+1}}{r!}+r+1\right). \tag{C.27}$$

We now claim that with our assumptions on $|\alpha|$, the first term dominates the second. This is a simple calculation, thereby competing the proof of (C.9). To prove (C.10), we repeat the entire argument, but using the relevant estimates for $Z_{\mathbf{\Psi},\mathbf{X}}$ instead of those for $Z_{\mathbf{\Phi},\mathbf{X}}$. ■

Applying the previous lemma in the setting of our NGCA model, we get the following result.

Theorem 28 (Robustness for non-Gaussian eigenvalues) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector satisfying the NGCA model (1.1), and with sub-Gaussian norm bounded above by $K \geq 1$. Let $\lambda_1(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}),\ldots,\lambda_d(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha})$ denote the eigenvalues of $\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha}$. Suppose the moments of $\|\tilde{\boldsymbol{X}}\|_2^2$ and $\|\boldsymbol{g}_d\|_2^2$ agree up to order $r-1$, but there is a number $\Delta > 0$ such that $\left|\mathbb{E}\{\|\tilde{\boldsymbol{X}}\|_2^{2r}\}-\mathbb{E}\{\|\boldsymbol{g}_d\|_2^{2r}\}\right| \geq \Delta$, then there is an absolute constant C such that for $|\alpha| \leq \Delta r/(CK^2)^r(d^{r+1}+(r+1)!)$, we have*

$$\left|\frac{1}{d}\sum_{i=1}^{d}\lambda_i(\mathbf{\Phi}_{\tilde{\boldsymbol{X}},\alpha})-\frac{1}{2\alpha+1}\right| \geq \frac{\Delta}{2d(r-1)!}|\alpha|^{r-1}. \tag{C.28}$$

Similarly, let $\lambda_1(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}), \ldots, \lambda_d(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha})$ *denote the eigenvalues of* $\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}$*, and suppose the moments of* $\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$ *and* $\langle \boldsymbol{g}, \boldsymbol{g}' \rangle$ *agree up to order* $r-1$ *but* $|\mathbb{E}\{\langle \boldsymbol{X}, \boldsymbol{X}' \rangle^r\} - \mathbb{E}\{\langle \boldsymbol{g}, \boldsymbol{g}' \rangle^r\}| \geq \Delta$*. Then for* $|\alpha| \leq \Delta r/(CK^2)^r(d^{r+1} + (r+1)!)$*, we have*

$$\left| \frac{1}{d} \sum_{i=1}^{d} \lambda_i(\mathbf{\Psi}_{\tilde{\boldsymbol{X}},\alpha}) - \frac{\alpha}{\alpha^2 - 1} \right| \geq \frac{\Delta}{2d(r-1)!} |\alpha|^{r-1}. \tag{C.29}$$

Proof This is simply a translation of the previous theorem with the help of Lemma 16, which tells us that the log derivatives of the partition functions are equal to the traces of $\mathbf{\Phi}_{\mathbf{X},\alpha}$ and $\mathbf{\Psi}_{\mathbf{X},\alpha}$, and that of Lemma 31, which tells us what the Gaussian eigenvalue is. ■

Proof [Proof of Theorem 18] Combine the previous Corollary with Theorem 6. ■

Appendix D. Identities for Gaussian test matrices

In this section, we let g denote a standard Gaussian random variable, and $\mathbf{g}_n$, a standard Gaussian random vector in $\mathbb{R}^n$. First, notice that independence gives $Z_{\mathbf{\Phi},\mathbf{g}_n}(\alpha) = Z_{\mathbf{\Phi},g}(\alpha)^n$ and $Z_{\mathbf{\Psi},\mathbf{g}_n}(\alpha) = Z_{\mathbf{\Psi},g}(\alpha)^n$.

Lemma 29 *We have the identities* $Z_{\mathbf{\Phi},\boldsymbol{g}_n}(\alpha) = (2\alpha+1)^{-n/2}$ *when* $\alpha > -1/2$ *and* $Z_{\mathbf{\Psi},\boldsymbol{g}_n}(\alpha) = \left(1-\alpha^2\right)^{-n/2}$ *when* $|\alpha| < 1$*.*

Proof By the remarks above, it suffices to prove the formula when $n = 1$. These are then simple exercises in calculus. Notice that

$$Z_{\mathbf{\Phi},g}(\alpha) = \mathbb{E}\{e^{-\alpha g^2}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha t^2} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(2\alpha+1)t^2}{2}} dt.$$

Now substitute $u = \sqrt{2\alpha+1} \cdot t$ to arrive at the formula for $Z_{\mathbf{\Phi},g}$. For the next formula, we use conditional expectations to write

$$Z_{\mathbf{\Psi},g}(\alpha) = \mathbb{E}\{e^{-\alpha g g'}\} = \mathbb{E}\{\mathbb{E}\{e^{-\alpha g g'} \big| g\}\}. \tag{D.1}$$

The inner expectation can be computed as

$$\mathbb{E}\{e^{-\alpha g g'} \big| g\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha g t} e^{-\frac{t^2}{2}} dt = e^{\frac{(\alpha g)^2}{2}}.$$

Substituting this back into (D.1) and using the same technique as above gives us what we want. ■

Lemma 30 *We have the identities* $-(\log Z_{\mathbf{\Phi},\boldsymbol{g}_n})'(\alpha) = n(2\alpha+1)^{-1}$ *when* $\alpha > -1/2$ *and* $-(\log Z_{\mathbf{\Psi},\boldsymbol{g}_n})'(\alpha) = n\alpha(\alpha^2-1)^{-1}$ *when* $|\alpha| < 1$*.*

Lemma 31 *We have the identities* $\mathbf{\Phi}_{\boldsymbol{g}_n,\alpha} = (2\alpha+1)^{-1}\boldsymbol{I}_n$ *when* $\alpha > -1/2$ *and* $\mathbf{\Psi}_{\boldsymbol{g}_n,\alpha} = \alpha(\alpha^2-1)^{-1}\boldsymbol{I}_n$ *when* $|\alpha| < 1$*. Here,* $\boldsymbol{I}_n$ *is the* n*-dimensional identity matrix.*

Proof By rotational symmetry, we know that both matrices are multiples of the identity. To compute these scalars, it hence suffices to find the trace of both matrices. But

$$\mathrm{Tr}(\mathbf{\Phi}_{\mathbf{g}_n,\alpha}) = \frac{\mathbb{E}\{e^{-\alpha\|\mathbf{g}_n\|_2^2}\|\mathbf{g}_n\|_2^2\}}{\mathbb{E}\{e^{-\alpha\|\mathbf{g}_n\|_2^2}\}} = -(\log Z_{\mathbf{\Phi},\mathbf{g}_n})'(\alpha).$$

Dividing by n and using the previous lemma gives us what we want. ■

Appendix E. Concentration and moment bounds

Theorem 32 (Concentration of norm for general sub-Gaussian vectors) *Let $\boldsymbol{X}$ be a sub-Gaussian random vector in $\mathbb{R}^n$, with $\|\boldsymbol{X}\|_{\psi_2} \leq K$. There is a universal constant C such that for each positive integer $r > 0$, the moments of $\|\boldsymbol{X}\|_2$ and $\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$ satisfy*

$$(\mathbb{E}\{\|\boldsymbol{X}\|_2^r\})^{1/r} \leq CK(\sqrt{n} + \sqrt{r}) \tag{E.1}$$

$$\left(\mathbb{E}\{|\langle \boldsymbol{X}, \boldsymbol{X}'\rangle|^r\}\right)^{1/2r} \leq CK(\sqrt{n} + \sqrt{r}). \tag{E.2}$$

Proof The second bound follows from the first, since by Cauchy-Schwarz,

$$\left(\mathbb{E}\{|\langle \mathbf{X}, \mathbf{X}'\rangle|^r\}\right)^{1/2r} \leq \left(\mathbb{E}\{\|\mathbf{X}\|_2^r\|\mathbf{X}'\|_2^r\}\right)^{1/2r} = (\mathbb{E}\{\|\mathbf{X}\|_2^r\})^{1/r}$$

To prove (E.1), pick a $\frac{1}{2}$-net $\mathcal{N}$ on S^{n-1}. A volumetric argument shows that one may pick $\mathcal{N}$ to have size no more than 5^n (see Vershynin (2011)). We then have

$$\|\mathbf{X}\|_2 = \sup_{\mathbf{v}\in S^{n-1}} \langle \mathbf{X}, \mathbf{v}\rangle \leq 2 \sup_{\mathbf{v}\in\mathcal{N}} \langle \mathbf{X}, \mathbf{v}\rangle.$$

By definition, there is a universal constant c such that for any fixed unit vector $v \in S^{d-1}$, $\mathbb{P}\{\langle \mathbf{X}, \mathbf{v}\rangle > t\} \leq 2\exp\left(-\frac{ct^2}{K^2}\right)$. Taking a union bound over the net thus gives

$$\mathbb{P}\{\|\mathbf{X}\|_2 > 2t\} \leq 2\exp\left(n\log 5 - \frac{ct^2}{K^2}\right). \tag{E.3}$$

Next, we integrate out the tail bound (E.3) to obtain bounds for the moments. Observe that if $\frac{ct^2}{2K^2} \geq n\log 5$, we have $n\log 5 - \frac{ct^2}{K^2} \leq -\frac{ct^2}{2K^2}$. This condition on t is equivalent to $t \geq CK\sqrt{n}$, so we have

$$\mathbb{P}\{\|\mathbf{X}\|_2 > 2t\} \leq \begin{cases} 1 & t < CK\sqrt{n} \\ 2\exp\left(-\frac{ct^2}{K^2}\right) & t \geq CK\sqrt{n} \end{cases} \tag{E.4}$$

For any positive integer r, we integrate this bound to get

$$\begin{aligned}
\mathbb{E}\{\|\mathbf{X}\|_2^r\} &= \int_0^\infty rt^{r-1}\mathbb{P}\{\|\mathbf{X}\| > t\}dt \\
&\leq \int_0^{CK\sqrt{n}} rt^{r-1}dt + \int_{CK\sqrt{n}}^\infty 2rt^{r-1}\exp\left(-\frac{ct^2}{K^2}\right)dt \\
&\leq C^rK^rn^{r/2} + C^rK^rr\int_0^\infty t^{r/2-1}e^{-t}dt.
\end{aligned}$$

The integral in the last line is the gamma function, so in short, we have shown that

$$\mathbb{E}\{\|\mathbf{X}\|_2^r\} \leq C^r K^r (n^{r/2} + \Gamma(r/2+1)). \tag{E.5}$$

Taking r-th roots of both sides and using Hölder, together with the fact that $\Gamma(x)^{1/x} \lesssim x$, gives (E.1). ■

Lemma 33 (Covariance estimation for sub-Gaussian random vectors) *Let $\boldsymbol{X}$ be a centered sub-Gaussian random vector in $\mathbb{R}^n$ with covariance matrix $\boldsymbol{\Sigma}$ and sub-Gaussian norm satisfying $\|\boldsymbol{X}\|_{\psi_2} \leq K$ for some $K \geq 1$. Let $\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i \boldsymbol{X}_i^T$ denote the sample covariance matrix from N independent samples. Then there is an absolute constant C such that for any $0 < \epsilon, \delta < 1$, we have $\mathbb{P}\left\{\left\|\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}\right\| > \epsilon\right\} \leq \delta$ so long as $N \geq CK^2(n + \log(1/\delta))\epsilon^{-2}$.*

Proof Refer to Vershynin (2011). ■

Lemma 34 (Moments of spherical marginals) *Let $\boldsymbol{\theta}$ be uniformly distributed on the sphere S^{n-1}. Then for any unit vector $v \in S^{n-1}$ and any positive integer k, we have*

$$\mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{v}\rangle^{2k}\} = \frac{1 \cdot 3 \cdots (2k-1)}{n \cdot (n+2) \cdots (n+2k-2)} \tag{E.6}$$

Proof There are several ways to prove this identity. We shall prove this by computing Gaussian integrals. Let g and g_n denote standard Gaussians in 1 dimension and n dimensions respectively. Then using the radial symmetry of $\mathbf{g}$, we have

$$\mathbb{E}\{g^{2k}\} = \mathbb{E}\{\langle \mathbf{g}_n, \mathbf{v}\rangle^{2k}\} = \mathbb{E}\{\langle \|\mathbf{g}_n\|_2 \boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\} = \mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\}\mathbb{E}\{\langle \boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\}.$$

Rearranging gives

$$\mathbb{E}\{\langle \boldsymbol{\theta}, \mathbf{v}\rangle^{2k}\} = \frac{\mathbb{E}\{g^{2k}\}}{\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\}}.$$

We then compute

$$\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\} = \frac{\omega_n}{(2\pi n)^{n/2}} \int_0^\infty r^{2k} r^{n-1} e^{-r^2/2} dr, \tag{E.7}$$

where ω_n is the volume of the sphere S^{n-1}. It is well known that

$$\omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)},$$

while we also have

$$\int_0^\infty r^{2k} r^{n-1} e^{-r^2/2} dr = 2^{n/2+k-1}\Gamma(n/2+k).$$

Substituting these back into (E.7) gives

$$\mathbb{E}\{\|\mathbf{g}_n\|_2^{2k}\} = 2^k \frac{\Gamma(n/2+k)}{\Gamma(n/2)} = n \cdot (n+2) \cdots (n+2k-2). \tag{E.8}$$

This yields the denominator in (E.6). A similar calculation for $\mathbb{E}\{g^{2k}\}$ yields the numerator. ■

Proof [Proof of Theorem 19] Let $\mathbf{Y} = e^{-\alpha\|\mathbf{X}\|_2^2}\mathbf{X}$. Then $\mathbf{Y}$ is a sub-Gaussian random vector with $\|\mathbf{Y}\|_{\psi_2} \leq K$. Let $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ denote its covariance and empirical covariance matrices respectively. Then $\|\mathbf{\Sigma}\| \leq 1$ and by Lemma 33, we have $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq \epsilon/2$ with probability at least $1 - \delta/2$. Next, observe that $\mathbf{\Phi}_{\mathbf{X},\alpha} = Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}\mathbf{\Sigma}$ and $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} = \hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}\hat{\mathbf{\Sigma}}$, where $\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha) = \sum_{j=1}^N e^{-\alpha\|\mathbf{X}_j\|_2^2}/N$. As such, we have

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq |\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\|. \tag{E.9}$$

Combining our lower bound on α with the power series formula for $Z_{\mathbf{\Phi}}$ from Lemma 25, we have $Z_{\mathbf{\Phi},\mathbf{X}}(\alpha) \geq 1/2$. Furthermore, we may apply Hoeffding's inequality to see that $|\hat{Z}_{\mathbf{\Phi},\mathbf{X}}(\alpha) - Z_{\mathbf{\Phi},\mathbf{X}}(\alpha)| \leq \epsilon/2$ with probability at least $1 - \delta/2$. We can now combine all of this together to get the probability bound. ■

Proof [Proof of Theorem 20] First, define $\mathbf{\Sigma} = \mathbb{E}\{e^{-\alpha\langle\mathbf{X},\mathbf{X}'\rangle}\mathbf{X}(\mathbf{X}')^T\}$ and $\hat{\mathbf{\Sigma}} = \sum_{i=1}^N e^{-\alpha\langle\mathbf{X}_i,\mathbf{X}_i'\rangle}(\mathbf{X}_i(\mathbf{X}_i')^T + \mathbf{X}_i'\mathbf{X}_i^T)/2N$, so that $\mathbf{\Psi}_{\mathbf{X},\alpha} = Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}\mathbf{\Sigma}$ and $\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} = \hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}\hat{\mathbf{\Sigma}}$. As in the previous theorem, we can write

$$\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| \leq |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\|. \tag{E.10}$$

This time however, we cannot immediately invoke Lemma 33 because we can no longer view $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ as the covariance and empirical covariance matrices of a random vector. Nonetheless, we can follow the same proof scheme with a few adjustments.

The basic idea is to use a net argument to transform the operator deviation bound into a scalar bound for random variables. Let $\mathcal{N}$ be a $\frac{1}{4}$-net on S^{n-1}. By a volumetric argument, we may pick $\mathcal{N}$ to have size no more than 9^n (see Vershynin (2011)). For any n by n real symmetric matrix $\mathbf{M}$, we then have

$$\|\mathbf{M}\| = \sup_{\mathbf{v}\in S^{n-1}} |\langle\mathbf{v}, \mathbf{M}\mathbf{v}\rangle| \leq 2\sup_{\mathbf{v}\in\mathcal{N}} |\langle\mathbf{v}, \mathbf{M}\mathbf{v}\rangle|. \tag{E.11}$$

As such, by taking a union bound, we can hope to bound $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$ by bounding $|\langle\mathbf{v}, (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\rangle|$ for a fixed unit vector $v \in S^{n-1}$. Let us do just this. We have

$$\langle\mathbf{v}, \hat{\mathbf{\Sigma}}\mathbf{v}\rangle = \frac{1}{N}\sum_{i=1}^N e^{-\alpha\langle\mathbf{X}_i,\mathbf{X}_i'\rangle}\langle\mathbf{X}_i, \mathbf{v}\rangle\langle\mathbf{X}_i', \mathbf{v}\rangle,$$

so that

$$\langle\mathbf{v}, (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\rangle = \frac{1}{N}\sum_{i=1}^N (Y_i - \mathbb{E}Y_i), \tag{E.12}$$

where

$$Y_i = e^{-\alpha\langle\mathbf{X}_i,\mathbf{X}_i'\rangle}\langle\mathbf{X}_i, \mathbf{v}\rangle\langle\mathbf{X}_i', \mathbf{v}\rangle. \tag{E.13}$$

Observe that the Y_i's are i.i.d. random variables. At this point in the proof of covariance estimation, one observes that the resulting random variables are subexponential, so one may apply Bernstein's inequality. Unfortunately, our Y_i's are not subexponetial because of the $e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}$ factor. The way we overcome this is to condition on the size of these factors being uniformly small. Indeed, by Lemma 35 to come, we have $e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle} \leq e$ for all samples i with probability at least $1-\delta$. We call this event A.

Next, define $\tilde{Y}_i := Y_i 1_A$. The $\tilde{Y}_i$'s are i.i.d random variables with subexponential norm bounded by eK^2. We can then apply Bernstein and our assumption on the sample size N to get

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i - \mathbb{E}\tilde{Y}_i)\right| > \epsilon\right\} \leq e^{-N\epsilon^2/CK^4} \leq \frac{\delta}{9^n}. \tag{E.14}$$

Conditioning on the set A, we have $Y_i = \tilde{Y}_i$ for each i. We can also rewrite the bound on the right hand side using our assumption on N. Doing this gives us

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}\tilde{Y}_i)\right| > \epsilon \;\middle|\; A\right\} \leq \frac{\delta}{9^n}. \tag{E.15}$$

We would like to replace $\mathbb{E}\tilde{Y}_i$ with $\mathbb{E}Y_i$, but the two quantities are not necessarily equal. Nonetheless, we can bound their difference as follows. We have

$$\mathbb{E}Y_i - \mathbb{E}\tilde{Y}_i = \mathbb{E}\{Y 1_{A^c}\} = \mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle 1_{A^c}\}. \tag{E.16}$$

We apply generalized Hölder to write

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle 1_{A^c}\}| \leq \left(\mathbb{E}\{e^{-4\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\}\right)^{1/4}\left(\mathbb{E}\{\langle \mathbf{X}_i, \mathbf{v}\rangle^4\langle \mathbf{X}_i', \mathbf{v}\rangle^4\}\right)^{1/4}\mathbb{P}\{A^c\}^{1/2}. \tag{E.17}$$

We now use the moment bounds for sub-Gaussian random variables and Lemma 36 to bound the first two multiplicands on the right. This gives us

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i, \mathbf{X}_i'\rangle}\langle \mathbf{X}_i, \mathbf{v}\rangle\langle \mathbf{X}_i', \mathbf{v}\rangle 1_{A^c}\}| \leq CK^2\mathbb{P}\{A^c\}^{1/2}. \tag{E.18}$$

Next, we use Lemma 35 together with our assumption on $|\alpha|$, tightening the constant if necessary, to see that $\mathbb{P}\{A^c\} \leq \epsilon^2/C^2K^4$. We combine this together with the last few equations to obtain $|\mathbb{E}Y_i - \mathbb{E}\tilde{Y}_i| \leq \epsilon$, and combining this with (E.15), we obtain

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mathbb{E}Y_i)\right| > 2\epsilon \;\middle|\; A\right\} \leq \frac{\delta}{9^n}. \tag{E.19}$$

Recall that Y_i's were defined for a fixed $\mathbf{v} \in \mathcal{N}$. We can take a union bound over all vectors in $\mathcal{N}$ to get

$$\mathbb{P}\left\{\sup_{\mathbf{v}\in\mathcal{N}}|\langle \mathbf{v}, (\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{v}\rangle| > 2\epsilon \;\middle|\; A\right\} \leq \delta. \tag{E.20}$$

Combining this with (E.11) then gives

$$\mathbb{P}\Big\{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| > 4\epsilon \;\Big|\; A\Big\} \leq \delta. \tag{E.21}$$

Let us continue to bound the other terms in (E.10) conditioned on the set A. Notice that on this set, $\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)$ is an average of terms that are each bounded in absolute value by e. Using Hoeffding's inequality together with a similar argument as above to bound $|\mathbb{E}\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)1_A - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)|$, one may show that

$$\mathbb{P}\Big\{|\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha) - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)| > \epsilon/2 \;\Big|\; A\Big\} \leq \delta. \tag{E.22}$$

We may also use the power series formula for $Z_{\mathbf{\Psi},\mathbf{X}}$ from Lemma 25 together with our bound on $|\alpha|$ to show that $Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) \geq \frac{1}{2}$.

It remains to bound $\|\mathbf{\Sigma}\|$. To do this, we let v again be an arbitrary unit vector, and use Cauchy-Schwarz to compute

$$|\mathbb{E}\{e^{-\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\langle \mathbf{X}_i,\mathbf{v}\rangle\langle \mathbf{X}_i',\mathbf{v}\rangle\}| \leq \Big(\mathbb{E}e^{-2\alpha\langle \mathbf{X}_i,\mathbf{X}_i'\rangle}\Big)^{1/2}\Big(\mathbb{E}\{\langle \mathbf{X}_i,\mathbf{v}\rangle^2\langle \mathbf{X}_i',\mathbf{v}\rangle^2\}\Big)^{1/2}. \tag{E.23}$$

We have already seen that moment bounds and Lemma 36 imply that this is bounded by an absolute constant C. In fact, we can take $C = 3$.

Putting everything together, we see that on the set A, we can continue writing (E.10) as

$$\begin{aligned}\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| &\leq |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + |\hat{Z}_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1} - Z_{\mathbf{\Psi},\mathbf{X}}(\alpha)^{-1}|\|\mathbf{\Sigma}\| \\ &\leq C\epsilon.\end{aligned}$$

Using our bound for $\mathbb{P}\{A\}$, we can therefore uncondition to get

$$\mathbb{P}\Big\{\|\hat{\mathbf{\Psi}}_{\mathbf{X},\alpha} - \mathbf{\Psi}_{\mathbf{X},\alpha}\| > C\epsilon\Big\} \leq \delta + \mathbb{P}\{A\} \leq 2\delta. \tag{E.24}$$

Finally, note that we can massage the constants so that the multiplying constants in front of ϵ and δ disappear. ■

Lemma 35 *For any $0 < \delta < 1$ and $N \in \mathbb{N}$, if $|\alpha| \leq \Big(CK^2\sqrt{\log(N/\delta)}(\sqrt{n} + \sqrt{\log(N/\delta)})\Big)^{-1}$, then*

$$\mathbb{P}\left\{\sup_{1\leq i\leq N} e^{-\alpha\langle \boldsymbol{X}_i,\boldsymbol{X}_i'\rangle} > e\right\} \leq \delta. \tag{E.25}$$

Proof Without loss of generality, assume that $\alpha > 0$. Using the union bound, it suffices to prove that

$$\mathbb{P}\big\{\langle \mathbf{X},\mathbf{X}'\rangle < -1/\alpha\big\} = \mathbb{P}\Big\{e^{-\alpha\langle \mathbf{X},\mathbf{X}'\rangle} > e\Big\} \leq \frac{\delta}{N}. \tag{E.26}$$

To compute this, we first condition on $\mathbf{X}'$ and use the sub-Gaussian tail of $\mathbf{X}$ to get

$$\mathbb{P}\big\{\langle \mathbf{X},\mathbf{X}'\rangle < -1/\alpha \mid \mathbf{X}'\big\} \leq \exp\left(-\frac{1}{CK^2\alpha^2\|\mathbf{X}'\|_2^2}\right),$$

and integrating out $\mathbf{X}'$, then gives

$$\mathbb{P}\{\langle \mathbf{X}, \mathbf{X}' \rangle < -1/\alpha\} \leq \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}}\}. \tag{E.27}$$

To compute this expectation, let A be the event that $\|\mathbf{X}'\|_2 \leq CK(\sqrt{n} + \sqrt{\log(N/\delta)})$. Then by equation (E.4) in Theorem 32, we have $\mathbb{P}\{A^c\} \leq \delta/N$. As such, we can break up the expectation into the portion over A and the the portion over A^c to obtain

$$\begin{aligned}
\mathbb{E}e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} &= \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A\}\mathbb{P}\{A\} + \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A^c\}\mathbb{P}\{A^c\} \\
&\leq \mathbb{E}\{e^{-(CK^2\alpha^2\|\mathbf{X}'\|_2^2)^{-1}} \mid A\} + \mathbb{P}\{A^c\} \\
&\leq \exp\left(-\frac{1}{CK^4\alpha^2(n + \log(N/\delta))}\right) + \frac{\delta}{N}.
\end{aligned} \tag{E.28}$$

As such, we just need the first term to be less than δ/N, which corresponds to the requirement that

$$\frac{1}{CK^4\alpha^2(n + \log(N/\delta))} \geq \log(N/\delta).$$

This is simply a rearrangement of our assumption on $|\alpha|$. ■

Lemma 36 (Better bound for $Z_{\mathbf{\Psi}}$) *There is an absolute constant C such that if $|\alpha| \leq 1/CK^2\sqrt{n}$, then $Z_{\mathbf{\Psi},\boldsymbol{X}}(\alpha) \leq 3$.*

Proof The idea of the proof is similar to that of the previous lemma. We first condition on $\mathbf{X}'$ and use the sub-Gaussian nature of $\mathbf{X}$ to bound its Laplace transform, thereby obtaining

$$\mathbb{E}\{e^{-\alpha\langle \mathbf{X}, \mathbf{X}' \rangle} \mid \mathbf{X}'\} \leq e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2}.$$

Integrating out $\mathbf{X}'$ gives

$$\begin{aligned}
Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) &\leq \mathbb{E}\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2}\} \\
&= \int_0^\infty \mathbb{P}\left\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\right\} dt \\
&\leq e + \int_e^\infty \mathbb{P}\left\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\right\} dt
\end{aligned} \tag{E.29}$$

Next, we use our assumption on $|\alpha|$ to write

$$\begin{aligned}
\mathbb{P}\left\{e^{CK^2\alpha^2\|\mathbf{X}'\|_2^2} > t\right\} &= \mathbb{P}\left\{\|\mathbf{X}'\|_2 > \frac{\sqrt{\log t}}{CK|\alpha|}\right\} \\
&\leq \mathbb{P}\left\{\|\mathbf{X}'\|_2 > \sqrt{\log t} CK\sqrt{n}\right\}.
\end{aligned} \tag{E.30}$$

For $t > e$, we have $\sqrt{\log t} > 1$, so we may apply (E.4) to get

$$\mathbb{P}\left\{\|\mathbf{X}'\|_2 > \sqrt{\log t} CK\sqrt{n}\right\} \leq e^{-\log t Cn} = t^{-Cn}. \tag{E.31}$$

Plugging this into (E.29) gives

$$Z_{\mathbf{\Psi},\mathbf{X}}(\alpha) \leq e + \frac{e^{-Cn}}{Cn} \leq 3 \tag{E.32}$$

if we choose C to be large enough. ■

Appendix F. Eigenvector perturbation theory

If two n by n matrices are close in spectral norm, one can use minimax identities to show that their eigenvalues are also close. It is less trivial to show that their eigenvectors are also close, which is the case in the presence of an "eigengap". This was addressed by Davis and Kahan (1970).

Definition 37 *Let E and $\hat{E}$ be two subspaces of $\mathbb{R}^n$ of dimension d. Let $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$ be n by d matrices with orthonormal columns forming a basis for E and $\hat{E}$ respectively. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ be the singular values of $\boldsymbol{V}^T\hat{\boldsymbol{V}}$. We define the* principal angles *of E and $\hat{E}$ to be $\boldsymbol{\theta}_i(E, \hat{E}) = \arccos \sigma_i$ for $1 \leq i \leq d$.*

Lemma 38 *Let E, $\hat{E}$, $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$ be as in the previous definition. We have*

$$\|\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F^2 = 2\sum_{i=1}^{d} \sin^2 \boldsymbol{\theta}_i(E, \hat{E}). \tag{F.1}$$

In particular, the quantity depends only on E and $\hat{E}$ and not the choice of bases.

Proof We expand

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 = \|\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|_F^2 + \|\mathbf{V}\mathbf{V}^T\|_F^2 - 2\langle \mathbf{V}\mathbf{V}^T, \hat{\mathbf{V}}\hat{\mathbf{V}}^T\rangle. \tag{F.2}$$

Observe that

$$\|\mathbf{V}\mathbf{V}^T\|_F^2 = \mathrm{Tr}(\mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T) = \mathrm{Tr}(\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}) = \mathrm{Tr}(\mathbf{I}_d) = d. \tag{F.3}$$

Similarly, we have

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|_F^2 = d. \tag{F.4}$$

Next, we compute

$$\langle \mathbf{V}\mathbf{V}^T, \hat{\mathbf{V}}\hat{\mathbf{V}}^T\rangle = \mathrm{Tr}(\mathbf{V}\mathbf{V}^T\hat{\mathbf{V}}\hat{\mathbf{V}}^T) = \mathrm{Tr}(\hat{\mathbf{V}}^T\mathbf{V}\mathbf{V}^T\hat{\mathbf{V}}) = \|\hat{\mathbf{V}}^T\mathbf{V}\|_F^2. \tag{F.5}$$

Next, we use the fact that the squared Frobenius norm of a matrix is the sum of squares of its singular values to write

$$\|\hat{\mathbf{V}}^T\mathbf{V}\|_F^2 = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \cos^2 \boldsymbol{\theta}_i(E, \hat{E}). \tag{F.6}$$

We may then combine these identities to write

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 = 2\sum_{i=1}^{d}(1 - \cos^2 \boldsymbol{\theta}_i(E, \hat{E})) = 2\sum_{i=1}^{d} \sin^2 \boldsymbol{\theta}_i(E, \hat{E}). \tag{F.7}$$

as was to be shown. ■

Using the previous lemma, it is easy to see that the distance between subspaces is preserved under taking orthogonal complements.

Lemma 39 *Let F and F' be subspaces of $\mathbb{R}^n$ of dimensions m, and let F' and $F'^{\perp}$ denote their orthogonal complements. We have $d(F, F') = d(F^{\perp}, F'^{\perp})$.*

We can now use these observations to state Theorem 2 from Yu et al. (2015) in a convenient form.

Theorem 40 *Let $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ be two n by n symmetric real matrices, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$. Fix $1 \leq r \leq s \leq n$, and assume that $\min\{\lambda_r - \lambda_{r+1}, \lambda_s - \lambda_{s+1}\} > 0$, where we define $\lambda_0 = \infty$ and $\lambda_{n+1} = -\infty$. Let $d = r + n - s$, and let $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r, \boldsymbol{v}_{s+1}, \ldots, \boldsymbol{v}_n)$ and $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \ldots, \hat{\boldsymbol{v}}_r, \hat{\boldsymbol{v}}_{s+1}, \ldots, \hat{\boldsymbol{v}}_n)$ be n by d matrices whose columns are orthonormal eigenvectors to $\lambda_1, \lambda_2, \ldots, \lambda_r, \lambda_{s+1}, \ldots, \lambda_n$ and $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_r, \hat{\lambda}_{s+1}, \ldots, \hat{\lambda}_n$ respectively. Then*

$$\|\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F \leq \frac{2\sqrt{2d}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{\min\{\lambda_r - \lambda_{r+1}, \lambda_s - \lambda_{s+1}\}}. \tag{F.8}$$

Appendix G. Proof of Lemma 21

Proof Combining Lemmas 15, 16, and 31 tells us that in the right coordinates, $\mathbf{\Phi}_{\mathbf{X},\alpha}$ block diagonalizes as

$$\mathbf{\Phi}_{\mathbf{X},\alpha} = \left(\begin{array}{c|c} \mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha} & 0 \\ \hline 0 & (2\alpha+1)^{-1}\mathbf{I}_{n-d,} \end{array}\right). \tag{G.1}$$

Next, label the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We can find $0 \leq p \leq q \leq n$ such that the eigenvalues corresponding to the $\mathbf{\Phi}_{\tilde{\mathbf{X}},\alpha_1}$ block are $\lambda_1, \lambda_2, \ldots, \lambda_p, \lambda_{q+1}, \ldots, \lambda_n$. Using Theorem 28, we then have

$$\left|\frac{1}{d}\left(\sum_{i=1}^{p}\lambda_i + \sum_{i=q+1}^{n}\lambda_i\right) - \frac{1}{2\alpha_1+1}\right| \geq \frac{\Delta}{2d(r-1)!}\alpha_1^{r-1} = 2\beta_1. \tag{G.2}$$

In particular, we have $\frac{1}{p}\sum_{i=1}^{p}\lambda_i - 1/(2\alpha_1+1) \geq 2\beta_1$, and $1/(2\alpha_1+1) - \frac{1}{n-q}\sum_{i=q+1}^{n}\lambda_i \geq 2\beta_1$. Since at least one of these sums of eigenvalues is non-empty, truncating the eigenvalues of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ at the β_1 level gives us a non-trivial subspace of E.

In order to show that our empirical estimate $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ also has an approximation to this property, we will need to use the eigenvector perturbation theory explained in Appendix F. First, we need to bound from below the "eigengap" in $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$. Suppose first that $p \geq 1$, i.e. that there are eigenvalues larger than $(2\alpha_1+1)^{-1}$. Then by the pigeonhole principle, one can find i such that $(2\alpha_1+1)^{-1} + \beta_1/2 \geq \lambda_{i+1} \geq (2\alpha_1+1)^{-1}$ and $\lambda_i - \lambda_{i+1} \geq \beta_1/2d$. Similarly, if $q \leq n-1$, then we can find j such that $(2\alpha_1+1)^{-1} \geq \lambda_{j-1} \geq (2\alpha_1+1)^{-1} - \beta_1/2$ and $\lambda_{j-1} - \lambda_j \geq \beta_1/2d$.

Now let F be the span of the eigenvectors of $\mathbf{\Phi}_{\mathbf{X},\alpha_1}$ corresponding to $\lambda_1, \ldots, \lambda_i, \lambda_j, \ldots, \lambda_n$, and let $\hat{F}$ be the eigenvectors of $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha_1}$ corresponding to $\hat{\lambda}_1, \ldots, \hat{\lambda}_i, \hat{\lambda}_j, \ldots, \hat{\lambda}_n$. By Theorem 19, with probability at least $1-\delta$, we have

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq \frac{\beta_1\epsilon}{4\sqrt{2}d^{3/2}}. \tag{G.3}$$

We may then use Theorem 40 to see that $d(\hat{F}, F) \leq \epsilon$.

We are not yet done, because we do not have access to $\hat{F}$. Nonetheless, we can show that $\hat{F}$ contains $\hat{E}_{\mathbf{\Phi}}$. Using eigenvalue perturbation inequalities together with equation (G.3) tells us that we have

$$\hat{\lambda}_{i+1} \leq \lambda_{i+1} + \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} + \frac{\beta_1}{2} + \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} + \frac{2\beta_1}{3}, \tag{G.4}$$

and similarly that

$$\hat{\lambda}_{j-1} \leq \lambda_{j-1} - \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} - \frac{\beta_1}{2} - \frac{\beta_1 \epsilon}{2d} \leq (2\alpha_1 + 1)^{-1} - \frac{2\beta_1}{3}. \tag{G.5}$$

Let $\hat{I}_{\mathbf{\Phi}} = \{i : |\hat{\lambda}_i - (1 - 2\alpha_1)^{-1}| > \beta_1\}$. We see that this set does not contain any index between $i+1$ and $j-1$, so $\hat{E}_{\mathbf{\Phi}}$, which comprises the span of the eigenvectors to these eigenvalues, does not contain any eigenvector that $\hat{F}$ does not contain, as was to be shown. The inclusion then implies that we may find a subspace $E_{\mathbf{\Phi}} \subset F$ such that $d(\hat{E}_{\mathbf{\Phi}}, E_{\mathbf{\Phi}}) \leq \epsilon$.

Finally, we observe that $\dim \hat{E}_{\mathbf{\Phi}} \geq 1$, since

$$\frac{1}{p}\sum_{i=1}^{p} \hat{\lambda}_i - \frac{1}{2\alpha_1 + 1} \geq \frac{1}{p}\sum_{i=1}^{p} \lambda_i - \frac{\beta_1 \epsilon}{2d} - \frac{1}{2\alpha_1 + 1} > \beta_1, \tag{G.6}$$

and

$$\frac{1}{2\alpha_1 + 1} - \frac{1}{n-q}\sum_{i=q+1}^{n} \hat{\lambda}_i \geq \frac{1}{2\alpha_1 + 1} - \frac{1}{n-q}\sum_{i=q+1}^{n} \lambda_i - \frac{\beta_1 \epsilon}{2d} > \beta_1. \tag{G.7}$$

■

Appendix H. Proof of Theorem 9

Before we prove the guarantee, we state our proposed algorithm more formally.

Algorithm 2 ITERATED REWEIGHTED PCA($\mathbf{X}$,d,$\boldsymbol{\alpha}_1$,$\boldsymbol{\alpha}_2$,$\boldsymbol{\beta}_1$,$\boldsymbol{\beta}_2$)

Input: Data points $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}'_1, \ldots, \mathbf{X}'_N]$, scaling parameters $\alpha_1, \alpha_2 \in \mathbb{R}$, tolerance parameters $\beta_1, \beta_2 > 0$.

Output: Output $\hat{E}$ for E.

1: Initialize $\check{E} := 0$.
2: **for** $k = 1, \ldots, d$ **do**:
3: $F_1, F_2 :=$ REWEIGHTED PCA($\mathbf{P}_{\check{E}^\perp}\mathbf{X}$,$\alpha_1^{(k)}$,$\alpha_2^{(k)}$,$\beta_1^{(k)}$,$\beta_2^{(k)}$).
4: **if** $F_1 \neq 0$, then $\check{E} := \check{E} \oplus F_1$.
5: **else** $\check{E} := \check{E} \oplus F_2$.
6: **if** $\dim(\check{E}) = d$ **return** $\hat{E} := \check{E}$.

Proof We provide an outline of the proof, omitting details that are similar to those in the proof of Theorem 7. Suppose we are at Step 3, having just completed k iterations, and have found $\check{E}$ so that $\dim(\check{E}) = d_k$ and $d(\check{E}, E_k) < \epsilon_k$ for some subspace $E_k \subset E$. Call $\mathbf{Y} := \mathbf{P}_{E^\perp}\mathbf{X}$, and $\check{\mathbf{Y}} := \mathbf{P}_{\check{E}^\perp}\mathbf{X}$.

By Lemma 42, the remaining non-Gaussian part of $\mathbf{Y}$ is either $(m, c\eta^2/\tilde{\gamma}_m)$-norm-moment-identifiable or it is $(m, c\eta^2)$-product-moment-identifiable (see Definition 41 below). Let us assume

that the former holds since the other case is similar. For convenience, we denote $\alpha = \alpha_1^{(k+1)}$, $\beta = \beta_1^{(k+1)}$ to be the scaling and tolerance parameters for the $k+1$-th iteration.

By Theorem 28, we observe the existence of non-Gaussian eigenvalues in the $\mathbf{\Phi}$ matrix for $\mathbf{Y}$ for α small enough (specifically, $\alpha < \min\{c\eta^2 r/(CK^2)^r \tilde{\gamma}_m (d^{r+1} + (r+1)!), 1/CK^2 n\}$):

$$\left| \frac{1}{d-d_0} \sum_{i=1}^{d} \lambda_i(\mathbf{\Phi}_{\mathbf{P}_E \mathbf{Y}, \alpha}) - \frac{1}{2\alpha + 1} \right| \geq \frac{c\eta^2}{d(m-1)!\tilde{\gamma}_m} \alpha^{m-1}. \tag{H.1}$$

It remains to see that this signal is not destroyed by the noise stemming from our estimation of $\mathbf{\Phi}_{\mathbf{Y},\alpha}$ by $\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}$. Note that we have

$$\begin{aligned}
|\mathbb{E}\{e^{-\alpha\|\mathbf{Y}\|_2^2}\} - \mathbb{E}\{e^{-\alpha\|\check{\mathbf{Y}}\|_2^2}\}| &\leq \mathbb{E}\{(e^{-\alpha\|\mathbf{Y}\|_2^2} + e^{-\alpha\|\check{\mathbf{Y}}\|_2^2})|\alpha\|\mathbf{Y}\|_2^2 - \alpha\|\check{\mathbf{Y}}\|_2^2|\} \\
&\leq \alpha\mathbb{E}\{|\|\mathbf{Y}\|_2^2 - \|\check{\mathbf{Y}}\|_2^2|\} \\
&= \alpha\mathbb{E}\{|\mathbf{X}^T(\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp})\mathbf{X}|\} \\
&\leq \alpha\|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\|\mathbb{E}\{\|\mathbf{X}\|_2^2\} \\
&\leq n\alpha\epsilon_k.
\end{aligned}$$

Here, the first inequality follows from Lemma 43, while the last one follows from the fact that

$$\|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\| \leq \|\mathbf{P}_{E_k^\perp} - \mathbf{P}_{\check{E}^\perp}\|_F = d(\check{E}, E_k).$$

By doing several computations similar to the above, we obtain

$$\|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \leq \text{poly}_m(n)\epsilon_k. \tag{H.2}$$

Meanwhile, Theorems 19 and 20 imply that with high probability,

$$\|\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \leq \epsilon_0 \tag{H.3}$$

We may combine (H.2) and (H.3) to get

$$\begin{aligned}
\|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}\| &\leq \|\mathbf{\Phi}_{\mathbf{Y},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| + \|\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha} - \mathbf{\Phi}_{\check{\mathbf{Y}},\alpha}\| \\
&\leq \text{poly}_m(n)\epsilon_k + \epsilon_0.
\end{aligned}$$

Suppose ϵ_0 and ϵ_k are small enough so that

$$\text{poly}_m(n)\epsilon_k + \epsilon_0 \lesssim \frac{\eta^2}{d(m-1)!\tilde{\gamma}_m} \alpha^{m-1}. \tag{H.4}$$

Then the non-Gaussian eigenvalues continue to be outlier eigenvalues of $\hat{\mathbf{\Phi}}_{\check{\mathbf{Y}},\alpha}$, and can be discovered via truncation. One can formalize this using same argument as in the proof of Lemma 21. Finally, we again imitate the proof of Lemma 21 and appeal to Theorem 40. This tells us that the eigenspace F_1 corresponding to the found eigenvalues is ϵ' close to that of the "true" eigenspace F' in E if

$$\text{poly}_m(n)\epsilon_k + \epsilon_0 \lesssim \frac{\beta\epsilon'}{d^{3/2}}, \tag{H.5}$$

where we pick $\beta \asymp \eta^2\alpha^{m-1}/d(r-1)!\tilde{\gamma}_r$. If this is the case, we have have

$$d(\check{E} \oplus F_1, E_k \oplus F') = \|\mathbf{P}_{\check{E}} + \mathbf{P}_{F_1} - \mathbf{P}_{E_k} + \mathbf{P}_{F'}\|_F \leq \epsilon_k + \epsilon' =: \epsilon_{k+1}.$$

Suppose the algorithm terminates in l steps. Then $l \leq d$, and if we fix a desired $\epsilon_l < 1$, then iterating the inequalities (H.4) and (H.5) shows us that we just require

$$\epsilon_0 \leq \epsilon_l/\text{poly}_m(n)^d = \epsilon_l/\text{poly}_{m,d}(n).$$

By Theorem 19, this condition can be met with a sample size that grows according to $\text{poly}_{m,d}(n)$. ■

Definition 41 *Let $\tilde{\boldsymbol{X}}$ be an isotropic random vector in $\mathbb{R}^d$. For any positive integer m, $\gamma_r > \eta > 0$, we say that $\tilde{\boldsymbol{X}}$ is (m, μ)-norm-moment-identifiable if*

$$\left|\mathbb{E}\{\|\tilde{\boldsymbol{X}}\|_2^{2r}\} - \mathbb{E}\{\|\boldsymbol{g}\|_2^{2r}\}\right| \geq \mu$$

for some integer $r \leq m/2$. Similarly, we say that $\tilde{\boldsymbol{X}}$ is (m, μ)-product-moment-identifiable if

$$\left|\mathbb{E}\{\langle\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}'\rangle^r\} - \mathbb{E}\{\langle\boldsymbol{g}, \boldsymbol{g}'\rangle^r\}\right| \geq \mu$$

for some integer $r \leq m$.

Lemma 42 *In the NGCA model (1), suppose $\tilde{\boldsymbol{X}}$ is (m, η)-moment-identifiable along every direction $\boldsymbol{v} \in E$ for some $\eta \in (0, 1)$. Then for any proper subspace E_k of E, $\boldsymbol{P}_{E_k^\perp}\tilde{\boldsymbol{X}}$ is either $(m, c\eta^2/\tilde{\gamma}_r)$-norm-moment-identifiable or it is $(m, c\eta^2)$-product-moment-identifiable.*

Proof Note that $\mathbf{P}_{E_k^\perp}\tilde{\mathbf{X}}$ is still (m, η)-moment-identifiable along every direction in $E \cap E_k^\perp$. As such, we may apply Theorem 6 to conclude. ■

Lemma 43 *For any real numbers a and b, we have $|e^b - e^a| \leq (e^a + e^b)|b - a|$.*

Proof Use the fact that $e^x(x-1) + 1 \geq 0$ for all real x. ■

Appendix I. Proof of Corollary 10

Proof By symmetry, we know that $\boldsymbol{\Phi}_{\tilde{\mathbf{X}},\alpha} = c_0\mathbf{I}_d$ is a scalar matrix. To compute c_0, we write

$$c_0 = \frac{1}{d}\text{Tr}(\boldsymbol{\Phi}_{\tilde{\mathbf{X}},\alpha}) = \frac{1}{d}\frac{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\|\tilde{\mathbf{X}}\|_2^2\}}{\mathbb{E}\{e^{-\alpha\|\tilde{\mathbf{X}}\|_2^2}\}} = \frac{1}{d}\frac{e^{-\alpha d}d}{e^{-\alpha d}} = 1.$$

Combining this with Lemma 31 and 15 allows us to write

$$\boldsymbol{\Phi}_{\mathbf{X},\alpha} = \left(\begin{array}{c|c} \mathbf{I}_d & 0 \\ \hline 0 & (2\alpha+1)^{-1}\mathbf{I}_{n-d,} \end{array}\right).$$

By our choice of α, this gives an eigengap of

$$1 - \frac{1}{1+2\alpha} \geq \alpha = \frac{c}{n}.$$

Our assumption that $N \gtrsim dn^2(n + \log(1/\delta))/\epsilon^2$ together with Theorem 19 then guarantees that

$$\|\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha} - \mathbf{\Phi}_{\mathbf{X},\alpha}\| \leq \frac{c\epsilon}{\sqrt{d}n} \tag{I.1}$$

with high probability. We may now apply Theorem 40 to see that $d(F, E) \leq \epsilon$ where F is the subspace spanned by the top d eigenvectors of $\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}$.

It remains to see that F is discovered by the algorithm. But then (I.1) implies that

$$\lambda_i(\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} \geq \lambda_i(\mathbf{\Phi}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} - \frac{c\epsilon}{\sqrt{d}n} \geq \frac{\alpha}{2}$$

for $1 \leq i \leq d$, and similarly,

$$\lambda_i(\hat{\mathbf{\Phi}}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} \leq \lambda_i(\mathbf{\Phi}_{\mathbf{X},\alpha}) - \frac{1}{1+2\alpha} + \frac{c\epsilon}{\sqrt{d}n} \leq \frac{\alpha}{4}$$

for $d+1 \leq i \leq n$. The final inequality in both lines holds after choosing c to be small enough. We therefore see that the top d eigenvalues are indeed those that are identified by truncating at level $\beta = \alpha/3$. ■