The Annals of Statistics
2018, Vol. 46, No. 5, 2416–2451
https://doi.org/10.1214/17-AOS1625

OPTIMALITY AND SUB-OPTIMALITY OF PCA I: SPIKED RANDOM MATRIX MODELS

BY AMELIA PERRY[*,1,2], ALEXANDER S. WEIN[*,1,3], AFONSO S. BANDEIRA[†,4] AND ANKUR MOITRA[*,5]

Massachusetts Institute of Technology[*] and New York University[†]

A central problem of random matrix theory is to understand the eigenvalues of spiked random matrix models, introduced by Johnstone, in which a prominent eigenvector (or "spike") is planted into a random matrix. These distributions form natural statistical models for principal component analysis (PCA) problems throughout the sciences. Baik, Ben Arous and Péché showed that the spiked Wishart ensemble exhibits a sharp phase transition asymptotically: when the spike strength is above a critical threshold, it is possible to detect the presence of a spike based on the top eigenvalue, and below the threshold the top eigenvalue provides no information. Such results form the basis of our understanding of when PCA can detect a low-rank signal in the presence of noise. However, under structural assumptions on the spike, not all information is necessarily contained in the spectrum. We study the statistical limits of tests for the presence of a spike, including nonspectral tests. Our results leverage Le Cam's notion of contiguity and include:

(i) For the Gaussian Wigner ensemble, we show that PCA achieves the optimal detection threshold for certain natural priors for the spike.

(ii) For any non-Gaussian Wigner ensemble, PCA is sub-optimal for detection. However, an efficient variant of PCA achieves the optimal threshold (for natural priors) by pre-transforming the matrix entries.

(iii) For the Gaussian Wishart ensemble, the PCA threshold is optimal for positive spikes (for natural priors) but this is not always the case for negative spikes.

1. Introduction.

One of the most common ways to analyze a collection of data is to extract top eigenvectors of a sample covariance matrix that represent

Received April 2017; revised July 2017.

[1]Equal contributors.

[2]Supported in part by NSF CAREER Award CCF-1453261 and a grant from the MIT NEC Corporation.

[3]Conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

[4]Supported by NSF Grants DMS-1317308, DMS-1712730 and DMS-1719545. Part of this work was done while with the MIT Department of Mathematics.

[5]Supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship and an Alfred P. Sloan Fellowship.

MSC2010 subject classifications. 62H15, 62B15.

Key words and phrases. Random matrix, principal component analysis, hypothesis testing, deformed Wigner, spiked covariance, contiguity, power envelope, phase transition.

directions of largest variance, often referred to as principal component analysis (PCA). Starting from the work of Karl Pearson, this technique has been a mainstay in statistics and throughout the sciences for more than a century. For instance, genome-wide association studies construct a correlation matrix of expression levels, whereby PCA is able to identify collections of genes that work together. PCA is also used in economics to extract macroeconomic trends and to predict yields and volatility [Litterman and Scheinkman (1991), Forni et al. (2000), Stock and Watson (2002), Egloff, Leippold and Wu (2010)], and in network science to find well-connected communities [McSherry (2001)]. More broadly, it underlies much of exploratory data analysis, dimensionality reduction and visualization.

Classical random matrix theory provides a suite of tools to characterize the behavior of the eigenvalues of various random matrix models in high-dimensional settings. Nevertheless, most of these works can be thought of as focusing on a pure noise model [Anderson, Guionnet and Zeitouni (2010), Bai and Silverstein (2010), Tao (2012)] where there is not necessarily any low-rank structure to extract. A direction initiated by Johnstone (2001) has brought this powerful theory closer to statistical questions by introducing *spiked models* that are of the form "signal + noise." Such models have yielded fundamental new insights on the behaviors of several methods such as principal component analysis (PCA) [Johnstone and Lu (2004), Paul (2007), Nadler (2008)], sparse PCA [Amini and Wainwright (2008), Vu and Lei (2012), Berthet and Rigollet (2013a), Ma (2013), Shen, Shen and Marron (2013), Cai, Ma and Wu (2013), Birnbaum et al. (2013), Deshpande and Montanari (2014a), Krauthgamer, Nadler and Vilenchik (2015)], and synchronization algorithms [Singer (2011), Boumal et al. (2014), Bandeira, Boumal and Singer (2014), Boumal (2016)]. More precisely, given a true signal in the form of an n-dimensional unit vector x called the *spike*, we can define two natural spiked random matrix ensembles as follows:

- Spiked (Gaussian) Wishart: observe the sample covariance $Y = \frac{1}{N} XX^\top$, where X is an $n \times N$ matrix with columns drawn i.i.d. from $\mathcal{N}(0, I_n + \beta xx^\top)$, in the high-dimensional setting where the sample count N and dimension n scale proportionally as $n/N \to \gamma$. We allow $\beta \in [-1, \infty)$.
- Spiked Wigner: observe $Y = \lambda xx^\top + \frac{1}{\sqrt{n}} W$, where W is an $n \times n$ random symmetric matrix with entries drawn i.i.d. (up to symmetry) from a fixed distribution of mean 0 and variance 1.

We adopt a Bayesian viewpoint, taking the spike x to be drawn from an arbitrary but known prior. This enables our approach to address structural assumptions on the spike, such as sparsity or an entrywise constraint to $\{\pm 1/\sqrt{n}\}$, to model variants of sparse PCA or community detection [Deshpande, Abbe and Montanari (2016)].

The Wishart model describes the sample covariance of high-dimensional data. The Gaussian Wigner distribution arises from the Wishart as a particular small-γ limit [Johnstone and Onatski (2015)]. The spiked Wigner model also describes

various inference problems where pairwise measurements are observed between n entities; this captures, for instance, Gaussian variants of community detection [Deshpande, Abbe and Montanari (2016)] and $\mathbb{Z}/2$ synchronization [Javanmard, Montanari and Ricci-Tersenghi (2016)].

We will refer to the parameter β or λ as the signal-to-noise ratio (SNR). In each of the above models, we study the following statistical questions:

- *Detection*: For what values of the SNR is it possible to consistently test [with probability $1 - o(1)$ as $n \to \infty$] between a random matrix drawn from the spiked distribution and one from the unspiked distribution?
- *Recovery*: For what values of the SNR can an estimator $\hat{x}$ achieve correlation with the true spike x that is bounded above zero as $n \to \infty$?

We primarily study the detection problem, which has previously been explored in various statistical models [Donoho and Jin (2004), Cai, Jin and Low (2007), Sun and Nobel (2008), Ingster, Tsybakov and Verzelen (2010), Arias-Castro, Candès and Durand (2011), Arias-Castro, Candès and Plan (2011), Arias-Castro, Bubeck and Lugosi (2012), Butucea and Ingster (2013), Sun and Nobel (2013), Arias-Castro and Verzelen (2014), Verzelen and Arias-Castro (2015)].

The spiked random matrix models above all enjoy a sharp characterization of the performance of PCA through random matrix theory. In the complex Wishart case, the seminal work of Baik, Ben Arous and Péché (2005) showed that when $\beta > \sqrt{\gamma}$ an isolated eigenvalue emerges from the Marchenko–Pastur-distributed bulk. Later Baik and Silverstein (2006) established this result in the real Wishart case. In the Wigner case, the top eigenvalue separates from the semicircular bulk when $\lambda > 1$ [Péché (2006), Féral and Péché (2007), Capitaine, Donati-Martin and Féral (2009), Pizzo, Renfrew and Soshnikov (2013)]. Each result establishes a sharp *spectral threshold* at which PCA (top eigenvalue) is able to solve the detection problem for the respective spiked random matrix model. Moreover, it is known that above this threshold, the top eigenvector correlates nontrivially with x, while the correlation concentrates about zero below the threshold. Despite detailed research on the spectral properties of spiked random matrix models, much less is known about the more general statistical question: can any hypothesis test consistently detect the presence of a spike below the threshold where PCA succeeds? Our main goal in this paper is to address this question in each of the models above, and as we will see, the answer varies considerably across them. Our results shed new light on how much of the accessible information about x is *not* captured by the top eigenvalue, or even by the full spectrum.

Several recent works have examined this question. Onatski, Moreira and Hallin (2013) study the spiked Wishart model where x is an arbitrary unknown unit vector (which, by rotational symmetry, is equivalent to drawing x from the uniform prior on the unit sphere). They identify the optimal hypothesis testing power (between spiked and unspiked) and in particular show that there is no test to consistently detect the presence of a spike below the spectral threshold. Even more recent work

[Onatski, Moreira and Hallin (2014), Dobriban (2017), Ke (2016)] elaborates on this point in other spiked models. In the Gaussian Wigner model, it has been established by Montanari, Reichman and Zeitouni (2015) and Johnstone and Onatski (2015) that detection is impossible below the spectral threshold, and the former used techniques similar to those of the present paper, which are not fundamentally limited to spherically symmetric models; indeed, these techniques were applied to sparse PCA by Banks et al. (2017).

In another line of work, several papers have studied recovery in structured spiked random matrix models through approximate message passing [Donoho, Maleki and Montanari (2009), Bayati and Montanari (2011), Javanmard and Montanari (2013)], Guerra interpolation [Guerra (2003)], and other tools originating from statistical physics. These results span sparse PCA [Deshpande and Montanari (2014b), Lesieur, Krzakala and Zdeborová (2015a)], nonnegative PCA [Montanari and Richard (2016)], cone-constrained PCA [Deshpande, Montanari and Richard (2014)] and general structured PCA [Rangan and Fletcher (2012), Lesieur, Krzakala and Zdeborová (2015b), Deshpande, Abbe and Montanari (2016), Krzakala, Xu and Zdeborová (2016), Barbier et al. (2016), Lelarge and Miolane (2016)]. Methods based on approximate message passing typically exhibit the same threshold as PCA, but above the threshold they obtain better (and often optimal) estimates of the spike. In many cases, the above techniques give the asymptotic minimum mean square error (MMSE) and, in particular, identify the threshold for nontrivial recovery. However, they do not typically address the detection problem (although we expect the detection and recovery thresholds to match), and they tend to be restricted to i.i.d. priors.

We develop a number of general-purpose tools for proving both upper and lower bounds on detection. We defer the precise statement of our results in each model to their respective sections, but for now we highlight some of our main results:

- In the Gaussian Wigner model, detection is impossible below the spectral threshold ($\lambda = 1$) for priors such as the spherical prior[6] (Corollary 3.14), the Rademacher prior[7] (Corollary 3.12) and any sufficiently sub-Gaussian prior (Theorem 3.10). We also study sparse Rademacher priors,[8] where we see that the spectral threshold is sometimes optimal and sometimes sub-optimal depending on the sparsity level (Section 3.7).
- In the Wigner model with non-Gaussian noise, the spectral threshold is never optimal (subject to mild conditions): there is an entrywise pre-transformation on the observed matrix that exploits the non-Gaussianity of the noise and strictly im-

[6] x is uniform on the unit sphere in $\mathbb{R}^n$.

[7] x is i.i.d. uniform on $\{\pm 1/\sqrt{n}\}$.

[8] x is i.i.d. where each entry is 0 with probability $1-\rho$ and otherwise uniform on $\{\pm\frac{1}{\sqrt{\rho n}}\}$.

proves the performance of PCA (Theorem 4.8). This method was first described by Lesieur, Krzakala and Zdeborová (2015b) and we give a rigorous analysis. Moreover, we provide a lower bound (Theorem 4.4), which often matches this upper bound.

- In the Wishart model, the PCA threshold is optimal for the spherical prior, both for positive and negative β. For the Rademacher prior, PCA is optimal for all positive β; however, in the less-studied case of negative β, an inefficient algorithm succeeds below the spectral threshold when γ is sufficiently large. This exposes a new statistical phase transition that seems to be previously unexplored. For the sparse Rademacher prior, PCA can be sub-optimal in both the positive and negative β regimes, but it is always optimal for sufficiently large positive β.

We emphasize that when we say PCA is optimal, we refer only to the threshold for consistent detection. In essentially all cases we consider (except the spherical prior), the top eigenvector has sub-optimal estimation error above the threshold; optimal error is often given by an approximate message passing algorithm such as that of Deshpande, Abbe and Montanari (2016). Furthermore, PCA does not achieve optimal hypothesis testing power below the threshold, and in fact no method based on a finite number of top eigenvalues can be optimal in this sense [Onatski, Moreira and Hallin (2013, 2014), Johnstone and Onatski (2015), Dobriban (2017)].

All our lower bounds follow a similar pattern and are based on the notion of *contiguity* introduced by Le Cam (1960). On a technical level, we show that a particular second moment is bounded which (as is standard in contiguity arguments) implies that the spiked distribution cannot be consistently distinguished [with $o(1)$ error as $n \to \infty$] from the corresponding unspiked distribution. We develop general tools for controlling the second moment based on sub-Gaussianity and large deviations theory that apply across a range of models and a range of different priors on x.

While bounds on the second moment do not *a priori* imply anything about the recovery problem, it follows from results of Banks et al. (2017) that many of our nondetection results imply the corresponding nonrecovery results. The value of the second moment also yields bounds on hypothesis testing power (see Proposition 2.5).

Our work fits into an emerging theme in statistics: we indicate several scenarios when PCA is sub-optimal but the only known tests that beat it are computationally inefficient. Such computational versus statistical gaps have received considerable recent attention [e.g., Berthet and Rigollet (2013b), Ma and Wu (2015)], often in connection with sparsity. We provide evidence for a new such gap in the negatively-spiked Wishart model with the Rademacher prior, offering an example where sparsity is not present.

Outline. In Section 2, we give preliminaries on contiguity and the second moment method. In Section 3, we study the spiked Gaussian Wigner model, in Section 4 we study the spiked non-Gaussian Wigner model and in Section 5 we study the spiked Wishart model. Some proofs are deferred to the Appendices in Perry et al. (2017).

2. Contiguity and the second moment method. Contiguity and related ideas will play a crucial role in this paper. First introduced by Le Cam (1960), contiguity is a central concept in the asymptotic theory of statistical experiments, and has found many applications throughout probability and statistics. Our work builds on a history of using contiguity and related tools such as the *small subgraph conditioning method* to establish fundamental results about random graphs [e.g., Robinson and Wormald (1994), Janson (1995), Molloy et al. (1997); see Wormald (1999) for a survey] and impossibility results for detecting community structure in the sparse stochastic block model [Mossel, Neeman and Sly (2015), Banks et al. (2016)]. Contiguity is formally defined as follows.

DEFINITION 2.1 [Le Cam (1960)]. Let distributions P_n, Q_n be defined on the measurable space $(\Omega_n, \mathcal{F}_n)$. We say that the sequence Q_n is *contiguous* to P_n, and write $Q_n \lhd P_n$, if for any sequence A_n of events,

$$\lim_{n\to\infty} P_n(A_n) = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} Q_n(A_n) = 0.$$

Contiguity readily implies that the distributions P_n and Q_n cannot be consistently distinguished (given a single sample) in the following sense.

OBSERVATION 2.2. *If $Q_n \lhd P_n$, then there is no hypothesis test of the alternative Q_n against the null P_n with* $\Pr[\textit{type I error}] + \Pr[\textit{type II error}] = o(1)$.

Note that $Q_n \lhd P_n$ and $P_n \lhd Q_n$ are not equivalent, but either of them implies nondistinguishability. Also, showing that two (sequences of) distributions are contiguous does not rule out the existence of a test that distinguishes between them with constant error probability (better than random guessing). In fact, such tests do exist for the spiked Wigner and Wishart models, for instance by thresholding the trace of the matrix; optimal tests are discussed by Onatski, Moreira and Hallin (2013) and Johnstone and Onatski (2015).

Our goal in this paper is to show thresholds below which spiked and unspiked random matrix models are contiguous. We will do this through computing a particular second moment, related to the χ^2-divergence as $1 + \chi^2(Q_n \parallel P_n)$, through a classical form of the second moment method.

LEMMA 2.3. *Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of distributions on $(\Omega_n, \mathcal{F}_n)$. If the second moment*

$$\mathop{\mathbb{E}}_{P_n}\left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2\right]$$

exists and remains bounded as $n \to \infty$, then $Q_n \lhd P_n$.

All of the contiguity results in this paper will follow through Lemma 2.3 and its conditional variant below. The roles of P_n and Q_n are not symmetric, and we will always take P_n to be the unspiked distribution and take Q_n to be the spiked distribution, as the second moment is more tractable to compute in this direction. We include the proof of Lemma 2.3 here for completeness.

PROOF. Let $\{A_n\}$ be a sequence of events. Using Cauchy–Schwarz,

$$Q_n(A_n) = \int_{A_n} \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\,\mathrm{d}P_n \le \sqrt{\int_{A_n}\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 \mathrm{d}P_n}\cdot\sqrt{\int_{A_n}\mathrm{d}P_n}.$$

The first factor on the right-hand side is bounded; so if $P_n(A_n) \to 0$ then also $Q_n(A_n) \to 0$. □

There will be times when the above second moment is unbounded but we are still able to prove contiguity using a modified second moment that conditions away from rare "bad" events that would otherwise dominate the second moment. This idea has appeared previously [Arias-Castro and Verzelen (2014), Verzelen and Arias-Castro (2015), Banks et al. (2016, 2017)].

LEMMA 2.4. *Let ω_n be an event that occurs with probability $1 - o(1)$ under Q_n. Let $\tilde{Q}_n$ be the conditional distribution of Q_n given ω_n. If the modified second moment $\mathbb{E}_{P_n}[(\mathrm{d}\tilde{Q}_n/\mathrm{d}P_n)^2]$ remains bounded as $n \to \infty$, then $Q_n \lhd P_n$.*

PROOF. By Lemma 2.3, we have $\tilde{Q}_n \lhd P_n$. As $Q_n \lhd \tilde{Q}_n$ we have $Q_n \lhd P_n$. □

Moreover, given a value of the second moment, we are able to obtain bounds on the tradeoff between type I and type II error in hypothesis testing, which are valid nonasymptotically.

PROPOSITION 2.5. *Consider a hypothesis test of a simple alternative Q against a simple null P. Let α be the probability of type I error, and β the probability of type II error. Regardless of the test, we must have*

$$\frac{(1-\beta)^2}{\alpha} + \frac{\beta^2}{(1-\alpha)} \le \mathop{\mathbb{E}}_{P}\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^2,$$

assuming the right-hand side is defined and finite. Furthermore, this bound is tight: for any $\alpha, \beta \in (0,1)$ *there exist* P, Q *and a test for which equality holds.*

PROOF. Let A be the event that the test selects the alternative Q, and let $\bar{A}$ be its complement.

$$\mathop{\mathbb{E}}_{P}\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^2 = \int \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q = \int_A \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q + \int_{\bar{A}} \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q$$
$$\geq \frac{(\int_A \mathrm{d}Q)^2}{\int_A (\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}Q} + \frac{(\int_{\bar{A}} \mathrm{d}Q)^2}{\int_{\bar{A}} (\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}Q} = \frac{(1-\beta)^2}{\alpha} + \frac{\beta^2}{(1-\alpha)},$$

where the inequality follows from Cauchy–Schwarz. The following example shows tightness: let $P = \text{Bernoulli}(\alpha)$ and let $Q = \text{Bernoulli}(1-\beta)$. On input 0, the test chooses P, and on input 1, it chooses Q. □

Although contiguity is a statement about nondetection rather than nonrecovery, our results also have implications for nonrecovery. In general, the detection problem and recovery problem can have different thresholds, but such counterexamples are often unnatural. For a wide class of problems with additive Gaussian noise, the results of Banks et al. (2017) imply that if the second moment from above is bounded then nontrivial recovery is impossible. This result applies to the Gaussian Wigner model and the positively-spiked ($\beta > 0$) Wishart model,[9] and so our nondetection results immediately imply nonrecovery results in those settings.

3. Gaussian Wigner models.

3.1. *Main results.*

We define the spiked Gaussian Wigner model.

DEFINITION 3.1. A *spike prior* is a family of distributions $\mathcal{X} = \{\mathcal{X}_n\}$, where $\mathcal{X}_n$ is a distribution over $\mathbb{R}^n$. We require our priors to be normalized so that $x^{(n)}$ drawn from $\mathcal{X}_n$ has $\|x^{(n)}\| \to 1$ (in probability) as $n \to \infty$.

DEFINITION 3.2. For $\lambda \geq 0$ and a spike prior $\mathcal{X}$, we define the spiked Gaussian Wigner model $\text{GWig}(\lambda, \mathcal{X})$ as follows. We first draw a spike $x \in \mathbb{R}^n$ from the prior $\mathcal{X}_n$. Then we reveal

$$Y = \lambda x x^\top + \frac{1}{\sqrt{n}} W,$$

where W is drawn from the $n \times n$ GOE (Gaussian orthogonal ensemble), that is, W is a random symmetric matrix with off-diagonal entries $\mathcal{N}(0,1)$, diagonal entries $\mathcal{N}(0,2)$ and all entries independent (except for symmetry $W_{ij} = W_{ji}$). We denote the unspiked model ($\lambda = 0$) by $\text{GWig}(0)$.

[9] For the Wishart case, consider the asymmetric $n \times N$ matrix of samples, which can be equivalently written as $\sqrt{\beta} x u^\top + W$ where $u \sim \mathcal{N}(0, I_N)$ and W is i.i.d. $\mathcal{N}(0,1)$.

It is well known that this model admits the following spectral behavior.

THEOREM 3.3 [Féral and Péché (2007), Benaych-Georges and Nadakuditi (2011)]. *Let Y be drawn from* $\mathrm{GWig}(\lambda, \mathcal{X})$ *with any spike prior $\mathcal{X}$ supported on unit vectors* ($\|x\| = 1$):

- *If $\lambda \le 1$, the top eigenvalue of Y converges almost surely to 2 as $n \to \infty$, and the top (unit-norm) eigenvector v has trivial correlation with the spike: $\langle v, x\rangle^2 \to 0$ almost surely.*
- *If $\lambda > 1$, the top eigenvalue converges almost surely to $\lambda + 1/\lambda > 2$, and v estimates the spike nontrivially: $\langle v, x\rangle^2 \to 1 - 1/\lambda^2$ almost surely.*

It follows that if $\|x\| \to 1$ in probability then the above convergence holds in probability (instead of almost surely). Thus PCA solves the detection and recovery problems precisely when $\lambda > 1$. In the critical case $\lambda = 1$ or near-critical case $\lambda \to 1$, there is also a test to consistently distinguish the spiked and unspiked models based on their spectra [Johnstone and Onatski (2015)]; see Appendix A (in the supplement [Perry et al. (2017)]) for details. Our goal is now to investigate whether detection is possible when $\lambda < 1$.

As a starting point, we compute the second moment of Lemma 2.3.

PROPOSITION 3.4. *Let $\lambda \ge 0$ and let $\mathcal{X}$ be a spike prior. Let $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$ and $P_n = \mathrm{GWig}_n(0)$. Let x and x' be independently drawn from $\mathcal{X}_n$. Then*

$$\mathop{\mathbb{E}}_{P_n}\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 = \mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right).$$

We defer the proof of this proposition until Section 3.2. For specific choices of the prior $\mathcal{X}$, our goal will be to show that if λ is below some critical $\lambda^*_{\mathcal{X}}$, this second moment is bounded as $n \to \infty$ (implying that detection is impossible). We will specifically consider the following types of priors.

DEFINITION 3.5. Let $\mathcal{X}_{\mathrm{sph}}$ denote the spherical prior: x is a uniformly random unit vector in $\mathbb{R}^n$.

By spherical symmetry, the spherical prior is equivalent to asking for a test that works for *any* unit-norm spike (i.e., no prior). Without loss of generality, any test for the spherical prior depends only on the spectrum.

DEFINITION 3.6. If π is a distribution on $\mathbb{R}$ with $\mathbb{E}[\pi] = 0$ and $\mathrm{Var}[\pi] = 1$, let $\mathrm{i.i.d.}(\pi/\sqrt{n})$ denote the spike prior that samples each coordinate of x independently from $\pi/\sqrt{n}$.

We will give two general techniques for showing contiguity for various priors. We call the first method the *sub-Gaussian method*, and it is presented in Section 3.3. The idea is that if the correlation $\langle x, x' \rangle$ between two independent draws from the prior is sufficiently sub-Gaussian, this implies strong tail bounds on $\langle x, x' \rangle$ which can be integrated to show that the second moment is bounded. For instance, this gives results in the case of an i.i.d. prior where the entrywise distribution π is sub-Gaussian.

In Section 3.6, we present our second method, the *conditioning method*, which uses the conditional second moment method and can improve upon the sub-Gaussian method in some cases. It only applies to finitely-supported i.i.d. priors and is based on a result from Banks et al. (2016).

For certain natural priors, we are able to show contiguity for all $\lambda < 1$, matching the spectral threshold. In particular, this holds for the spherical prior $\mathcal{X}_{\text{sph}}$ (Corollary 3.14), the i.i.d. Gaussian prior $\text{i.i.d.}(\mathcal{N}(0, 1/n))$ (Corollary 3.11), the i.i.d. Rademacher prior $\text{i.i.d.}(\pm 1/\sqrt{n})$ (Corollary 3.12), and more generally for $\text{i.i.d.}(\pi/\sqrt{n})$ where π is strictly sub-Gaussian (Theorem 3.10).

Not all priors are as well behaved as those above. In Section 3.7, we discuss the sparse Rademacher prior, where we see that the PCA threshold is not always optimal.

In Section 3.5, we show that (in some sense) similar priors have the same detection threshold (Proposition 3.13). One corollary (Corollary 3.15) is that regardless of the prior, no test based only on the eigenvalues can succeed below the $\lambda = 1$ threshold.

Our results often yield the limit value of the second moment and, therefore, imply asymptotic bounds on hypothesis testing via Proposition 2.5; see Appendix B (in the supplement) for details.

3.2. *Second moment computation.* We begin by computing the second moment $\mathbb{E}_{P_n}[(\mathrm{d}Q_n/\mathrm{d}P_n)^2]$ where $Q_n = \text{GWig}_n(\lambda, \mathcal{X})$ and $P_n = \text{GWig}_n(0)$. First, we simplify the likelihood ratio:

$$\frac{\mathrm{d}Q_n}{\mathrm{d}P_n} = \frac{\mathbb{E}_{x\sim\mathcal{X}_n} \exp(-\frac{n}{4}\langle Y - \lambda xx^\top, Y - \lambda xx^\top\rangle)}{\exp(-\frac{n}{4}\langle Y, Y\rangle)}$$

$$= \mathop{\mathbb{E}}_{x\sim\mathcal{X}_n} \exp\left(\frac{\lambda n}{2}\langle Y, xx^\top\rangle - \frac{n\lambda^2}{4}\langle xx^\top, xx^\top\rangle\right).$$

Now passing to the second moment:

$$\mathop{\mathbb{E}}_{P_n}\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2 = \mathop{\mathbb{E}}_{x,x'\sim\mathcal{X}_n} \mathop{\mathbb{E}}_{Y\sim P_n} \exp\left(\frac{\lambda n}{2}\langle Y, xx^\top + x'x'^\top\rangle \right.$$

$$\left. - \frac{n\lambda^2}{4}\left(\langle xx^\top, xx^\top\rangle + \langle x'x'^\top, x'x'^\top\rangle\right)\right),$$

where x and x' are drawn independently from $\mathcal{X}_n$. We now simplify the Gaussian moment-generating function over the randomness of Y, and cancel terms, to arrive at the expression

$$= \mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right),$$

which proves Proposition 3.4.

3.3. *The sub-Gaussian method.* In this section, we give a general method for controlling the second moment $\mathbb{E}_{x,x'} \exp(\frac{n\lambda^2}{2}\langle x, x'\rangle^2)$. We will need the concept of a sub-Gaussian random variable.

DEFINITION 3.7. A $\mathbb{R}^n$-valued random variable X is σ^2-*sub-Gaussian* if $\mathbb{E}[X] = 0$ and, for all $v \in \mathbb{R}^n$, $\mathbb{E}\exp(\langle v, X\rangle) \le \exp(\sigma^2\|v\|^2/2)$.

The most general form of the sub-Gaussian method is the following.

PROPOSITION 3.8. *Let $\mathcal{X}$ be any spike prior. Let $P_n = \mathrm{GWig}_n(0)$ and $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$. With x and x' drawn independently from $\mathcal{X}_n$, suppose $\langle x, x'\rangle$ is (σ^2/n)-sub-Gaussian for some constant σ. If $\lambda < 1/\sigma$, then $\mathbb{E}_{x,x'} \exp(\frac{n\lambda^2}{2}\langle x, x'\rangle^2)$ is bounded and so $Q_n \lhd P_n$.*

PROOF. Using the well-known sub-Gaussian tail bound $\Pr[|\langle x, x'\rangle| \ge t] \le 2\exp(-nt^2/2\sigma^2)$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,x'} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) &= \int_0^\infty \Pr\left[\exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) \ge u\right] \mathrm{d}u \\
&= \int_0^\infty \Pr\left[|\langle x, x'\rangle| \ge \sqrt{\frac{2\log u}{n\lambda^2}}\right] \mathrm{d}u \\
&\le \int_0^\infty 2u^{-1/\sigma^2\lambda^2} \,\mathrm{d}u
\end{aligned}$$

which is finite (uniformly in n) provided $\lambda < 1/\sigma$. □

We next show that it is sufficient for the prior itself to be (multivariate) sub-Gaussian.

PROPOSITION 3.9. *Let $P_n = \mathrm{GWig}_n(0)$ and $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$. Suppose $\mathcal{X}_n$ is (σ^2/n)-sub-Gaussian. If $\lambda < 1/\sigma$, then $Q_n \lhd P_n$.*

PROOF. Let $\delta > 0$. We use the conditional second moment method (Lemma 2.4), taking $\tilde{\mathcal{X}}_n$ to be the conditional distribution of $\mathcal{X}_n$ given the $(1 - o(1))$-probability event $\|x\| \le 1 + \delta$. With $\tilde{Q}_n = \mathrm{GWig}_n(\lambda, \tilde{\mathcal{X}})$, the conditional second moment $\mathbb{E}_{P_n}(\mathrm{d}\tilde{Q}_n/\mathrm{d}P_n)^2$ is (by Proposition 3.4)

$$\mathop{\mathbb{E}}_{x,x'\sim\tilde{\mathcal{X}}} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) \le (1 + o(1)) \mathop{\mathbb{E}}_{x\sim\mathcal{X},x'\sim\tilde{\mathcal{X}}} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right).$$

With $x \sim \mathcal{X}$ and $x' \sim \tilde{\mathcal{X}}$, we have that $\langle x, x'\rangle$ is $(\sigma^2(1+\delta)^2/n)$-sub-Gaussian because for any $v \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{X},x'\sim\tilde{\mathcal{X}}} \exp\big(v\langle x, x'\rangle\big) \le \mathop{\mathbb{E}}_{x'\sim\tilde{\mathcal{X}}} \exp\big(\sigma^2 v^2 \|x'\|^2/2n\big) \le \exp\big(\sigma^2 v^2 (1+\delta)^2/2n\big).$$

Choosing δ small enough so that $\lambda < 1/(\sigma(1+\delta))$, the result now follows from Proposition 3.8. □

Specializing to i.i.d. priors, it is sufficient for the distribution of each entry to be sub-Gaussian. In this case, we can also compute the limit value of the (conditional) second moment.

THEOREM 3.10 (Sub-Gaussian method for i.i.d. priors). *Let π be a mean-zero unit-variance distribution on $\mathbb{R}$ and let $\mathcal{X} = \mathrm{i.i.d.}(\pi/\sqrt{n})$. Let $P_n = \mathrm{GWig}_n(0)$, $Q_n = \mathrm{GWig}_n(\lambda, \mathcal{X})$, and $\tilde{Q}_n$ as in the proof of Proposition 3.9. Suppose π is σ^2-sub-Gaussian. If $\lambda < \frac{1}{\sigma}$ then $\lim_{n\to\infty} \mathbb{E}_{P_n}(\mathrm{d}\tilde{Q}_n/\mathrm{d}P_n)^2 = (1-\lambda^2)^{-1/2} < \infty$ and so $Q_n \lhd P_n$.*

PROOF. Since π is σ^2-sub-Gaussian, it follows easily from the definition that $\mathcal{X}_n$ is (σ^2/n)-sub-Gaussian and so contiguity follows from Proposition 3.9. To compute the limit value, by the central limit theorem we have that for $x, x' \sim \mathcal{X}$, $\sqrt{n}\langle x, x'\rangle$ converges in distribution to $\mathcal{N}(0,1)$. The same holds for $x, x' \sim \tilde{\mathcal{X}}$. By the continuous mapping theorem applied to $g(z) = \exp(\lambda^2 z^2/2)$, we also get convergence in distribution $\exp(n\lambda^2\langle x, x'\rangle^2/2) \xrightarrow{d} \exp(\lambda^2\chi_1^2/2)$. The convergence in expectation $\mathbb{E}_{x,x'\sim\tilde{\mathcal{X}}} \exp(n\lambda^2\langle x, x'\rangle^2/2) \to \mathbb{E}\exp(\lambda^2\chi_1^2/2) = (1-\lambda^2)^{-1/2}$ follows since the sequence $\exp(n\lambda^2\langle x, x'\rangle^2/2)$ is uniformly integrable; this is clear from the final step of the proof of Proposition 3.8 (which has no dependence on n). □

Since $\mathrm{Var}[\pi] = 1$, π cannot be σ^2-sub-Gaussian with $\sigma < 1$. If π is 1-sub-Gaussian ("strictly sub-Gaussian") then Theorem 3.10 gives a tight result, matching the spectral threshold. For instance, the standard Gaussian distribution is 1-sub-Gaussian, so we have the following.

COROLLARY 3.11. *If $\lambda < 1$, then* $\mathrm{GWig}(\lambda, \mathrm{i.i.d.}(\mathcal{N}(0, 1/n))) \lhd \mathrm{GWig}(0)$.

Note that the i.i.d. Gaussian prior is very similar to the spherical prior; in Section 3.5 we show how to transfer the proof to the spherical prior.

3.4. *Application*: *The Rademacher prior.* If π is a Rademacher random variable (uniform on $\{\pm 1\}$), then $\mathrm{i.i.d.}(\pi/\sqrt{n})$ is the *Rademacher prior*, which we abbreviate as $\mathrm{i.i.d.}(\pm 1/\sqrt{n})$. This case of the Gaussian Wigner model has been studied by Deshpande, Abbe and Montanari (2016) and Javanmard, Montanari and Ricci-Tersenghi (2016) as a Gaussian model for community detection and $\mathbb{Z}/2$ synchronization. The former proves that the spectral threshold $\lambda = 1$ is precisely the threshold above which nontrivial recovery of the signal is possible. We further show contiguity below this $\lambda = 1$ threshold (which, recall, is not implied by nonrecovery).

COROLLARY 3.12. *If* $\lambda < 1$, *then* $\mathrm{GWig}(\lambda, \mathrm{i.i.d.}(\pm 1/\sqrt{n})) \lhd \mathrm{GWig}(0)$.

PROOF. The Rademacher distribution is 1-sub-Gaussian by Hoeffding's lemma, so the proof follows from Theorem 3.10. □

Perhaps it is surprising that the spectral threshold is optimal for the Rademacher prior because it suggests that there is no way to exploit the ± 1 structure. However, PCA is only optimal in terms of the threshold and not in terms of error in recovering the spike once $\lambda > 1$. An efficient estimator that asymptotically minimizes the mean squared error is the approximate message passing algorithm of Deshpande, Abbe and Montanari (2016).

3.5. *Comparison of similar priors.* We show that two similar priors have the same contiguity threshold, in the following sense.

PROPOSITION 3.13. *Let* $\lambda^* \geq 0$. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be spike priors. Suppose that* $x \sim \mathcal{X}_n$ *and* $y \sim \mathcal{Y}_n$ *can be coupled such that* $y = \alpha x$ *where* $\alpha = \alpha_n$ *is a random variable with* $\alpha_n \to 1$ *in probability as* $n \to \infty$. *Suppose that for each* $\lambda < \lambda^*$, *the second moment* $\mathbb{E}_{x,x' \sim \mathcal{X}} \exp(\frac{n\lambda^2}{2}\langle x, x'\rangle^2)$ *remains bounded as* $n \to \infty$. *Then for any* $\lambda < \lambda^*$, $\mathrm{GWig}(\lambda, \mathcal{Y}) \lhd \mathrm{GWig}(0)$.

PROOF. Let $\lambda < \lambda^*$ and $\delta > 0$. Let $\tilde{\mathcal{Y}}$ be the conditional distribution of $\mathcal{Y}$ given the $(1 - o(1))$-probability event $\alpha \leq 1 + \delta$. Letting $\tilde{Q}_n = \mathrm{GWig}(\lambda, \tilde{\mathcal{Y}})$ and $P_n = \mathrm{GWig}(0)$, we have

$$\begin{aligned}
\frac{\mathrm{d}\tilde{Q}_n}{\mathrm{d}P_n} &= \mathop{\mathbb{E}}_{y,y' \sim \tilde{\mathcal{Y}}} \exp\left(\frac{n\lambda^2}{2}\langle y, y'\rangle^2\right) \\
&= (1 + o(1)) \mathop{\mathbb{E}}_{x,x' \sim \mathcal{X}} \mathbb{1}_{\alpha \leq 1+\delta} \mathbb{1}_{\alpha' \leq 1+\delta} \exp\left(\frac{n\lambda^2}{2}(\alpha\alpha')^2\langle x, x'\rangle^2\right) \\
&\leq (1 + o(1)) \mathop{\mathbb{E}}_{x,x' \sim \mathcal{X}} \exp\left(\frac{n\lambda^2}{2}(1+\delta)^4\langle x, x'\rangle^2\right)
\end{aligned}$$

which is bounded provided we choose δ small enough so that $\lambda(1+\delta)^2 < \lambda^*$. The result now follows from the conditional second moment method (Lemma 2.4). □

We can now show that the spectral threshold is optimal for the spherical prior (uniform on the unit sphere) by comparison to the i.i.d. Gaussian prior; this result was obtained previously by Montanari, Reichman and Zeitouni (2015), Johnstone and Onatski (2015).

COROLLARY 3.14. *If* $\lambda < 1$, *then* $\mathrm{GWig}(\lambda, \mathcal{X}_{\mathrm{sph}}) \lhd \mathrm{GWig}(0)$.

PROOF. We have shown that for any $\lambda < 1$, the second moment is bounded for a conditioned version of the i.i.d. Gaussian prior (conditioning on $\|x\| \le 1+\delta$); see Corollary 3.11. This conditioned Gaussian prior can be coupled to the spherical prior as required by Proposition 3.13, due to Gaussian spherical symmetry. The result follows from Proposition 3.13. □

A more direct proof for the spherical prior is possible using known properties of the confluent hypergeometric function; see Appendix C.

Another corollary is that *any* prior $\mathcal{X}$ (with $\|x\| \to 1$ in probability) and for any $\lambda < 1$, contiguity holds on the level of spectra; this implies that no test depending only on the eigenvalues can succeed below the $\lambda = 1$ threshold, even though other tests can in some cases (e.g., the sparse Rademacher prior of Section 3.7).

COROLLARY 3.15. *Let* $\mathcal{X}$ *be any spike prior* (*with* $\|x\| \to 1$ *in probability*). *Let* Q_n *be the joint distribution of eigenvalues of* $\mathrm{GWig}_n(\lambda, \mathcal{X})$ *and let* P_n *be the joint distribution of eigenvalues of* $\mathrm{GWig}_n(0)$. *If* $\lambda < 1$, *then* $Q_n \lhd P_n$.

PROOF. Due to Gaussian spherical symmetry, the distribution of eigenvalues of the spiked matrix depends only on the norm of the spike and not its direction; thus without loss of generality, $\mathcal{X}$ is a mixture of spherical priors, over a norm distribution converging in probability to 1. The result now follows from Proposition 3.13 and Corollary 3.14. □

3.6. *The conditioning method.* In this section, we give an alternative to the sub-Gaussian method that can give tighter results in some cases. Here, we give an overview, with the full details deferred to Appendix D in the supplement. Throughout this section, we require the prior to be $\mathcal{X} = \mathrm{i.i.d.}(\pi/\sqrt{n})$ where π has finite support.

The main idea is that the second moment takes a particular form involving a multinomial random variable; it turns out that this exact form has been studied by Banks et al. (2016) in the context of contiguity in the stochastic block model. Following their work, we apply the conditional second moment method (Lemma 2.4), conditioning on a high-probability "good" event where the empirical distribution

of x is close to $\pi/\sqrt{n}$. Proposition 5 in Banks et al. (2016) provides an exact condition (involving an optimization problem over matrices) for boundedness of the conditional second moment. This method improves upon the sub-Gaussian method in some cases (see, e.g., Section 3.7).

Let $\Delta_{s^2}(\pi)$ denote the set of nonnegative vectors $\alpha \in \mathbb{R}^{s^2}$ with row- and column-sums prescribed by π, that is, treating α as an $s \times s$ matrix, we have (for all i) that row i and column i of α each sum to π_i. Let $D(u, v)$ denote the KL divergence between two vectors: $D(u, v) = \sum_i u_i \log(u_i/v_i)$.

THEOREM 3.16 (Conditioning method). *Let $\mathcal{X} = \text{i.i.d.}(\pi)$ where π has mean zero, unit variance and finite support $\Sigma \subseteq \mathbb{R}$ with $|\Sigma| = s$. Let $Q_n = \text{GWig}_n(\lambda, \mathcal{X})$ and $P_n = \text{GWig}_n(0)$. Define the $s \times s$ matrix $\beta_{ab} = ab$ for $a, b \in \Sigma$. Identify π with the vector of probabilities $\pi \in \mathbb{R}^{\Sigma}$, and define $\bar{\alpha} = \pi\pi^{\top}$. Let*

$$\overline{\lambda}_{\mathcal{X}} = \left[\sup_{\alpha \in \Delta_{s^2}(\pi)} \frac{\langle \alpha, \beta \rangle^2}{2D(\alpha, \bar{\alpha})}\right]^{-1/2}.$$

If $\lambda < \overline{\lambda}_{\mathcal{X}}$, then $Q_n \lhd P_n$.

In Appendix D, we give the full proof and also compute that the limit value of the conditional second moment is $(1 - \lambda^2)^{-1/2}$ (the same as in Theorem 3.10). We also explain the intuition behind the matrix optimization problem.

3.7. *Application*: *The sparse Rademacher prior.* Now consider the case where $\pi = \sqrt{1/\rho}\mathcal{R}(\rho)$ where $\mathcal{R}(\rho)$ is the sparse Rademacher distribution with sparsity $\rho \in (0, 1]$: $\mathcal{R}(\rho)$ is 0 with probability $1 - \rho$, and otherwise uniform on $\{\pm 1\}$. Here, we give a summary of our results, with full details deferred to Appendix E.

We know from Corollary 3.12 that when $\rho = 1$, detection is impossible below the spectral threshold. However, for sufficiently small ρ (roughly 0.054), an exhaustive search procedure is known to perform detection for some range of λ values below the spectral threshold [Banks et al. (2017)]. Toward a matching lower bound, we would like to find ρ^* as small as possible such that PCA is optimal for all $\rho \geq \rho^*$.

Using the sub-Gaussian method (Theorem 3.10), it follows that PCA is optimal for all $\rho \geq 1/3$. The conditioning method (Theorem 3.16) improves this constant substantially, to roughly 0.184. Using a more sophisticated method that conditions on an event depending jointly on the signal and noise, Perry, Wein and Bandeira (2016) improve the constant further, to roughly 0.138. Similar (but quantitatively weaker) results have been obtained by Banks et al. (2017).

Based on heuristics from statistical physics, Lesieur, Krzakala and Zdeborová (2015b) predicted that the exact ρ value at which PCA becomes sub-optimal is given by the replica-symmetric (RS) formula, which yields $\rho_{\text{RS}} \approx 0.09$. It was

later proven rigorously that ρ_{RS} is the exact threshold for nontrivial *recovery* below $\lambda = 1$, and that if $\rho < \rho^*$ then detection below $\lambda = 1$ is possible (by thresholding the free energy) [Krzakala, Xu and Zdeborová (2016), Barbier et al. (2016), Lelarge and Miolane (2016)]. It remains open to show that detection is impossible below $\lambda = 1$ for all $\rho \geq \rho_{\mathrm{RS}}$. Lesieur, Krzakala and Zdeborová (2015b) also conjecture a computational gap: when $\lambda < 1$, no polynomial-time algorithm can perform detection or recovery (regardless of ρ).

4. Non-Gaussian Wigner models.

4.1. *Main results.* We first define the spiked non-Gaussian Wigner model.

DEFINITION 4.1. In the *general spiked Wigner model* $\mathrm{Wig}(\lambda, \mathcal{P}, \mathcal{P}_d, \mathcal{X})$, one observes a matrix

$$Y = \lambda x x^\top + \frac{1}{\sqrt{n}} W,$$

with the spike x drawn from a spike prior $\mathcal{X}$, and the entries of noise matrix W drawn independently up to symmetry, with the off-diagonal entries drawn from a distribution $\mathcal{P}$ and the diagonal entries drawn from a second distribution $\mathcal{P}_d$. For the sake of normalization, we assume that $\mathcal{P}$ has mean 0 and variance 1.

Recall that the prior $\mathcal{X}$ is required to obey the normalization $\|x\| \to 1$ in probability (see Definition 3.1).

The spectral behavior of this model is well understood[10] [see, e.g., Féral and Péché (2007), Capitaine, Donati-Martin and Féral (2009), Pizzo, Renfrew and Soshnikov (2013), Benaych-Georges and Nadakuditi (2011)]. In fact it exhibits *universality* [see, e.g., Tao and Vu (2014)]: regardless of the choice of the noise distributions $\mathcal{P}, \mathcal{P}_d$ (with sufficiently many finite moments), many properties of the spectrum behave the same as if $\mathcal{P}$ were a standard Gaussian distribution. In particular, for $\lambda \leq 1$, the spectrum bulk has a semicircular distribution and the maximum eigenvalue converges almost surely to 2. For $\lambda > 1$, an isolated eigenvalue emerges from the bulk with value converging to $\lambda + 1/\lambda$, and (under suitable assumptions) the top eigenvector has squared correlation $1 - 1/\lambda^2$ with the truth.

In stark contrast, we will show that from a statistical standpoint, universality breaks down entirely: the detection problem becomes easier when the noise is non-Gaussian. Let $\mathcal{X}$ be a spike prior, and suppose that through the second moment

[10]Many of the results cited here assume $\|x\| = 1$ and show almost-sure convergence of various quantities. Since we assume only $\|x\| \to 1$ in probability, the same convergence is true only in probability (which is enough for our purposes).

method, we can establish contiguity between the *Gaussian* spiked and unspiked models whenever λ lies below some critical value

$$\lambda^*_{\mathcal{X}} \triangleq \sup\left\{\lambda \;\middle|\; \mathbb{E}_{x,x'\sim\mathcal{X}} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) \text{ is bounded as } n \to \infty\right\}.$$

The detection threshold for the non-Gaussian Wigner model depends on $\lambda^*_{\mathcal{X}}$ as well as a parameter $F_{\mathcal{P}}$ (defined below) that depends on the noise distribution $\mathcal{P}$.

THEOREM (Informal; see Theorems 4.4 and 4.8). *Under suitable conditions (see Assumptions 4.3 and 4.7), the spiked model is contiguous to the unspiked model for all $\lambda < \lambda^*_{\mathcal{X}}/\sqrt{F_{\mathcal{P}}}$; but when $\lambda > 1/\sqrt{F_{\mathcal{P}}}$, there exists an entrywise transformation f such that the spiked and unspiked models can be consistently distinguished via the top eigenvalue of $f(\sqrt{n}Y)$.*

Recall that if we take the spike prior to be, for example, spherical or Rademacher, we have $\lambda^*_{\mathcal{X}} = 1$, implying that our upper and lower bounds match, and thus our pre-transformed PCA procedure achieves the optimal threshold for *any* noise distribution (subject to regularity assumptions). For reasons discussed later (see Appendix G in the supplement), we require $\mathcal{P}$ to be a continuous distribution with a density function $p(w)$. The parameter $F_{\mathcal{P}}$, which quantifies its difficulty, is the Fisher information of $\mathcal{P}$ under translation:

$$F_{\mathcal{P}} = \mathop{\mathbb{E}}_{w\sim\mathcal{P}}\left[\left(\frac{p'(w)}{p(w)}\right)^2\right] = \int_{-\infty}^{\infty} \frac{p'(w)^2}{p(w)}\,\mathrm{d}w.$$

Gaussian noise enjoys an extremal value of this Fisher information, qualifying it as the unique hardest noise distribution (among a large class).

PROPOSITION 4.2 [Pitman (1979), page 37]. *Let $\mathcal{P}$ be a real distribution with a C^1, nonvanishing density function $p(w)$. Suppose* $\mathrm{Var}[\mathcal{P}] = 1$. *Then $F_{\mathcal{P}} \geq 1$, with equality if and only if $\mathcal{P}$ is a standard Gaussian.*

This is effectively a form of the Cramér–Rao inequality, and can be exploited for a proof of the central limit theorem [Brown (1982), Barron (1986)].

Our upper bound proceeds by a pre-transformed PCA procedure. Define $f(w) = -p'(w)/p(w)$, where p is the probability density function of the noise $\mathcal{P}$. Given the observed matrix Y, we apply f entrywise to $\sqrt{n}Y$, and examine the largest eigenvalue. This entrywise transformation approximately yields another spiked Wigner model, but with improved signal-to-noise ratio. One can derive the transformation $-p'(w)/p(w)$ by using calculus of variations to optimize the signal-to-noise ratio of this new spiked Wigner model. This phenomenon is illustrated in Figures 1 and 2.

To intuitively understand why non-Gaussian noise makes the detection problem easier, consider the extreme case where the noise distributions $\mathcal{P}$, $\mathcal{P}_d$ are uniform on $\{\pm 1\}$, with mean 0 and variance 1. Since the noise contribution $\frac{1}{\sqrt{n}}W$

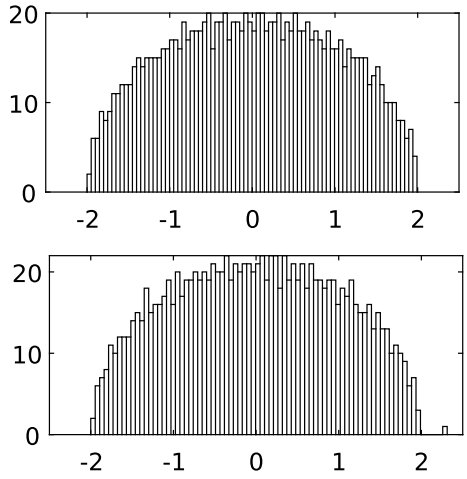

FIG. 1. *Spectrum of a spiked Wigner matrix* ($\lambda = 0.9, n = 1200$) *with bimodal noise*, *before* (*above*) *and after* (*below*) *the entrywise transformation. An isolated eigenvalue is evident only in the latter.*

is entrywise exactly $\pm 1/\sqrt{n}$, it is very easy to detect and identify the small signal perturbation $\lambda x x^\top$, which is entrywise $O(1/n)$. If there is no spike, all the entries will be $\pm 1/\sqrt{n}$ (exactly). If there is a spike, each entry will be $\pm 1/\sqrt{n}$ plus a much smaller offset. One can therefore subtract off the noise and recover the signal exactly. In fact, if we let the noise be a smoothed version of $\{\pm 1\}$ (so that the derivative p' exists), the entrywise transformation $-p'(w)/p(w)$ is precisely implementing this noise-subtraction procedure. This justifies the restriction to continuous noise distributions because any distribution with a point mass admits a similar trivial recovery procedure and we will not have contiguity for *any* $\lambda > 0$; see Appendix G for details.

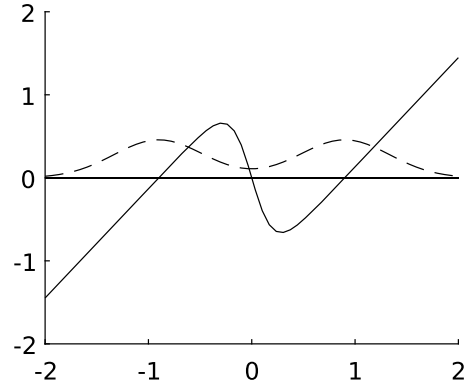

FIG. 2. *The noise density* p (*dashed*) *and entrywise transformation* $-p'/p$ (*solid*). *The bimodal noise is a convolution of Rademacher and Gaussian random variables.*

The above results on non-Gaussian noise parallel a *channel universality* phenomenon for mutual information, due to Krzakala, Xu and Zdeborová (2016) (shown for finitely-supported i.i.d. priors). The pre-transformed PCA procedure we use for our upper bound was previously suggested by Lesieur, Krzakala and Zdeborová (2015b) based on linearizing an approximate message passing algorithm, but to our knowledge, no rigorous results have been previously established about its performance in general. Other entrywise pre-transformations have been shown to improve spectral approaches to various structured PCA problems [Deshpande and Montanari (2014a), Kannan and Vempala (2016)].

4.2. *Lower bound.* In this section, we state our main statistical lower bound that establishes contiguity in the non-Gaussian Wigner setting. Given a noise distribution $\mathcal{P}$, define the *translation function*

$$\tau(a,b) = \log \mathop{\mathbb{E}}_{\mathcal{P}}\left[\frac{\mathrm{d}T_a\mathcal{P}}{\mathrm{d}\mathcal{P}}\frac{\mathrm{d}T_b\mathcal{P}}{\mathrm{d}\mathcal{P}}\right] = \log \mathop{\mathbb{E}}_{z\sim\mathcal{P}}\left[\frac{p(z-a)}{p(z)}\frac{p(z-b)}{p(z)}\right],$$

where $T_a\mathcal{P}$ denotes the translation of distribution $\mathcal{P}$ by a. For instance, the translation function of standard Gaussian noise is $\tau(a,b) = ab$.

ASSUMPTION 4.3. (i) The prior $\mathcal{X}$ satisfies (as usual) $\|x\| \to 1$ in probability, and furthermore $\mathcal{X}$ is (σ^2/n)-sub-Gaussian for some constant σ^2 (see Definition 3.7).

(ii) The prior $\mathcal{X}$ satisfies high-probability norm bounds: for $q = 2, 4, 6, 8$, there exists a constant α_q for which, with probability $1 - o(1)$ over $x \sim \mathcal{X}$, we have $\|x\|_q \le \alpha_q n^{\frac{1}{q}-\frac{1}{2}}$.

(iii) We assume the distributions $\mathcal{P}, \mathcal{P}_d$ have nonvanishing density functions $p(w)$, $p_d(w)$ and translation functions τ, τ_d that are C^4 in a neighborhood of $(0,0)$.

Our main lower bound result is the following.

THEOREM 4.4. *Under Assumption* 4.3, $\mathrm{Wig}(\lambda, \mathcal{P}, \mathcal{P}_d, \mathcal{X})$ *is contiguous to* $\mathrm{Wig}(0, \mathcal{P}, \mathcal{P}_d)$ *for all* $\lambda < \lambda^*_{\mathcal{X}}/\sqrt{F_{\mathcal{P}}}$.

We defer the proof to Appendix F. In Appendix F, we also show that the assumptions on $\mathcal{X}$ are satisfied for the spherical prior and for reasonable i.i.d. priors; see Propositions 4.5 and 4.6 below. The assumptions on $\mathcal{P}, \mathcal{P}_d$ are satisfied by any mixture of Gaussians of positive variance, for example.

PROPOSITION 4.5. *Conditions* (i) *and* (ii) *in Assumption* 4.3 *are satisfied for the spherical prior* $\mathcal{X}_{\mathrm{sph}}$.

PROPOSITION 4.6. *Consider an i.i.d. prior* $\mathcal{X} = \mathrm{i.i.d.}(\pi/\sqrt{n})$ *where* π *is zero-mean, unit-variance and sub-Gaussian with some constant* σ^2. *Then conditions* (i) *and* (ii) *in Assumption* 4.3 *are satisfied.*

4.3. *Pre-transformed PCA.* In this section, we analyze a pre-transformed PCA procedure for the non-Gaussian spiked Wigner model. We need the following regularity assumptions.

ASSUMPTION 4.7. Of the prior $\mathcal{X}$ we require (as usual) $\|x\| \to 1$ in probability, and we also assume that with probability $1 - o(1)$, all entries of x are small: $|x_i| \le n^{-1/2+\alpha}$ for some fixed $\alpha < 1/8$. Of the noise $\mathcal{P}$, we assume the following:

(i) $\mathcal{P}$ has a nonvanishing C^3 density function $p(w) > 0$,

(ii) Letting $f(w) = -p'(w)/p(w)$, we have that f and its first two derivatives are polynomially-bounded: there exists $C > 0$ and an even integer $m \ge 2$ such that $|f^{(\ell)}(w)| \le C + w^m$ for all $0 \le \ell \le 2$.

(iii) With m as in (ii), $\mathcal{P}$ has finite moments up to $5m$: $\mathbb{E}|\mathcal{P}|^k < \infty$ for all $1 \le k \le 5m$.

The main theorem of this section is the following.

THEOREM 4.8. *Let $\lambda \ge 0$ and let $\mathcal{X}, \mathcal{P}$ satisfy Assumption 4.7. Let $\hat{Y} = \sqrt{n}\, Y$ where Y is drawn from* $\mathrm{Wig}(\lambda, \mathcal{P}, \mathcal{P}_d, \mathcal{X})$. *Let $f(\hat{Y})$ denote entrywise application of the function $f(w) = -p'(w)/p(w)$ to $\hat{Y}$, except we define the diagonal entries of $f(\hat{Y})$ to be zero*:

- *If $\lambda \le 1/\sqrt{F_{\mathcal{P}}}$, then $\frac{1}{\sqrt{n}}\lambda_{\max}(f(\hat{Y})) \to 2\sqrt{F_{\mathcal{P}}}$ as $n \to \infty$.*
- *If $\lambda > 1/\sqrt{F_{\mathcal{P}}}$, then $\frac{1}{\sqrt{n}}\lambda_{\max}(f(\hat{Y})) \to \lambda F_{\mathcal{P}} + \frac{1}{\lambda} > 2\sqrt{F_{\mathcal{P}}}$ as $n \to \infty$ and furthermore the top (unit-norm) eigenvector v of $f(\hat{Y})$ correlates with the spike: $\langle v, x\rangle^2 \ge (\lambda - 1/\sqrt{F_{\mathcal{P}}})^2/\lambda^2 - o(1)$ with probability $1 - o(1)$.*

Convergence is in probability. Here, $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue.

The proof is deferred to Appendix H in the supplement, but the main idea is that the entrywise transformation f approximately produces another spiked (non-Gaussian) Wigner matrix with a different signal-to-noise ratio λ, and we can choose f to optimize this.

We have set the diagonal entries to zero for convenience, but this is not essential: so long as we define the diagonals of $f(\hat{Y})$ so that the largest (in absolute value) diagonal entry is $o(\sqrt{n})$, the diagonal entries can only change the spectral norm of $f(\hat{Y})$ by $o(\sqrt{n})$ and so the result still holds.

5. Spiked Wishart models.

5.1. *Main results.* We first formally define the spiked Wishart model.

DEFINITION 5.1. Let $\gamma > 0$ and $\beta \in [-1, \infty)$. Let $\mathcal{X} = \{\mathcal{X}_n\}$ be a spike prior. The spiked (Gaussian) Wishart model $\mathrm{Wish}(\gamma, \beta, \mathcal{X})$ on $n \times n$ matrices is defined as follows: we first draw a hidden spike $x \sim \mathcal{X}_n$, and then reveal $Y = \frac{1}{N} XX^\top$, where X is an $n \times N$ matrix whose columns are sampled independently from $\mathcal{N}(0, I + \beta xx^\top)$; the parameters N and n scale proportionally with $n/N \to \gamma$ as $n \to \infty$. If $\beta < 0$ and $|\beta| \cdot \|x\|^2 > 1$ (so that the covariance matrix is not positive semidefinite), output a failure event $\perp$.

Recall that spike priors are required to satisfy $\|x\| \to 1$ in probability (Definition 3.1). Our contiguity results will apply even to the case when the sample matrix X is revealed.

The spiked Wishart model admits the following spectral behavior. In this high-dimensional setting, the spectrum bulk of Y converges to the Marchenko–Pastur distribution with shape parameter γ. By results of Baik, Ben Arous and Péché (2005) and Baik and Silverstein (2006), it is known that the top eigenvalue consistently distinguishes the spiked and unspiked models when $\beta > \sqrt{\gamma}$. In fact, matching lower bounds are known in the absence of a prior (equivalently, for the spherical prior) due to Onatski, Moreira and Hallin (2013): for $0 \le \beta < \sqrt{\gamma}$, no hypothesis test distinguishes this spiked model from the unspiked model with $o(1)$ error. In the case of $-1 \le \beta < 0$, a corresponding PCA threshold exists: the minimum eigenvalue exits the bulk when $\beta < -\sqrt{\gamma}$ [Baik and Silverstein (2006)], but we are not aware of lower bounds in the literature. The case of $\beta < -1$ is of course invalid, as the covariance matrix must be positive semidefinite. As in the Wigner model, consistent detection is possible in the critical case $|\beta| = \sqrt{\gamma}$, at least when $\beta > 0$; see Appendix A in the supplement.

Our goal in this section will be to give lower and upper bounds on the statistical threshold for γ (as a function of β) for various priors on the spike. We begin with a crude lower bound that allows us to transfer any lower bound for the Gaussian Wigner model into a lower bound for the Wishart model. Recall that $\lambda^*_{\mathcal{X}}$ denotes the threshold for boundedness of the Gaussian *Wigner* second moment:

$$\lambda^*_{\mathcal{X}} \triangleq \sup\left\{\lambda \,\middle|\, \mathop{\mathbb{E}}_{x,x' \sim \mathcal{X}} \exp\left(\frac{n\lambda^2}{2}\langle x, x'\rangle^2\right) \text{ is bounded as } n \to \infty\right\}. \tag{1}$$

PROPOSITION 5.2. *Let $\mathcal{X}$ be a spike prior. If $\beta^2 < 1 - e^{-\gamma(\lambda^*_{\mathcal{X}})^2}$ then* $\mathrm{Wish}(\gamma, \beta, \mathcal{X})$ *is contiguous to* $\mathrm{Wish}(\gamma)$.

The proof can be found in Section 5.5.2. A consequence of the above is that if $\lambda^*_{\mathcal{X}} = 1$, so that the spectral method is optimal in the Wigner setting, it follows that the ratio between the above Wishart lower bound $(1 - e^{-\gamma(\lambda^*_{\mathcal{X}})^2})$ and the spectral upper bound (γ) tends to 1 as $\gamma \to 0$. This reflects the fact that the Wigner model is a particular $\gamma \to 0$ limit of the Wishart model [Johnstone and Onatski (2015)].

For $\beta > 0$, we will later give an even stronger implication from Wigner to Wishart lower bounds (Corollary 5.9).

Although Proposition 5.2 is a strong bound for small γ, it is rather weak for large γ (and in particular does not cover the case $\beta \geq 1$). In Section 5.3, we will remedy this by giving a much tighter lower bound (Theorem 5.7) which depends on the rate function of the large deviations of the prior. The proof involves an application of the conditional second moment method whereby we condition away from certain "bad" events depending on interactions between the signal and noise [similar to Perry, Wein and Bandeira (2016)]. One consequence (Corollary 5.9) of our lower bound roughly states that if detection is impossible below the spectral threshold ($\lambda = 1$) in the *Wigner* model, then it is also impossible below the spectral threshold ($|\beta| = \sqrt{\gamma}$) in the Wishart model for all positive β. (This is not true for negative β.)

We complement our lower bounds with the following upper bound.

THEOREM 5.3. *Let $\beta \in (-1, \infty)$. Let $\mathcal{X}_n$ be a spike prior supported on at most c^n points, for some fixed $c > 0$. If*

$$2\gamma \log c < \beta - \log(1 + \beta)$$

then there is a (inefficient) test that consistently distinguishes between the spiked Wishart model $\mathrm{Wish}(\gamma, \beta, \mathcal{X})$ *and the unspiked model* $\mathrm{Wish}(\gamma)$.

The test that gives this upper bound is based on the maximum likelihood estimator (MLE), computed by exhaustive search over all possible spikes. The proof, which can be found in Appendix I, is a simple application of the Chernoff bound and the union bound. For some priors (such as i.i.d. sparse Rademacher), we can get the most mileage out of this theorem by first conditioning on a $(1 - o(1))$-probability event (e.g., x has a typical number of nonzeros) in order to decrease the value of c.

We will typically not consider the boundary case $\beta = -1$. Note, however, that if $\beta = -1$ and the prior is finitely-supported (for each n), with $\|x\| = 1$ almost surely, then detection is possible for any γ: in the spiked model, the spike is orthogonal to all of the samples; but in the unspiked model, with probability 1 there will not exist a vector in the support of the prior that is orthogonal to all of the samples.

We now summarize the implications of our lower and upper bounds for some specific priors.

- *Spherical*: For the spherical prior (x is drawn uniformly from the unit sphere), it was known previously that the PCA threshold $|\beta| = \sqrt{\gamma}$ is optimal for all positive β [Onatski, Moreira and Hallin (2013)]. We show that the PCA threshold is also optimal for all $-1 < \beta < 0$.
- *Rademacher*: For the Rademacher prior $\mathrm{i.i.d.}(\pm 1/\sqrt{n})$, we show that the PCA threshold is optimal for all $\beta > 0$. However, when β is negative and sufficiently close to -1, the MLE of Theorem 5.3 succeeds below the PCA threshold.

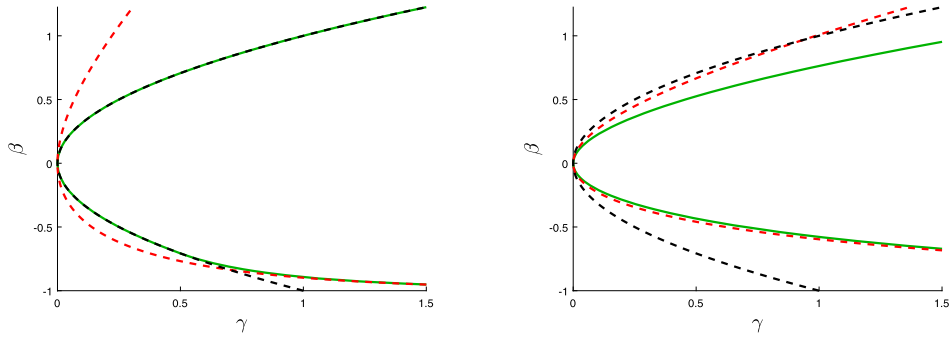

FIG. 3. *Upper and lower bounds for the spiked Wishart model with Rademacher prior* (*left panel*) *and sparse Rademacher prior with* $\rho = 0.03$ (*right panel*). *PCA succeeds to the left of the dashed black curve* $\beta^2 = \gamma$. *To the right of the solid green curve, detection is impossible* (*by Theorem* 5.7; *see Appendix N for details*). *To the left of the dashed red curve, detection is possible via the inefficient MLE algorithm of Theorem* 5.3. (*The red curve is not a tight analysis of the MLE and is sometimes weaker than the PCA bound.*) *For the Rademacher prior, the lower bound matches PCA for all* $\beta > -0.7$, *but the MLE succeeds below the PCA threshold for all* $\beta < -0.84$. *For the sparse Rademacher prior with any* ρ, *the lower bound matches PCA for sufficiently large positive* β (*not shown*); *see Proposition* 5.10.

- *Sparse Rademacher* (defined in Section 3.7): If the sparsity ρ is sufficiently small, the MLE beats PCA in both the positive- and negative-β regimes. However, for any fixed ρ, if β is sufficiently large (and positive) then the PCA threshold is optimal.

See Appendix N for details on the above results, including how they follow from our general upper and lower bounds (Theorems 5.3 and 5.7). Figure 3 depicts our upper and lower bounds for the Rademacher and sparse Rademacher priors.

As in the Wigner model, our methods often yield the limit value of the (conditional) second moment, and thus imply asymptotic bounds on hypothesis testing power via Proposition 2.5; see Appendix B for details.

5.2. *Rate functions.* Our main lower bound will depend on the prior through tail probabilities of the correlation $\langle x, x' \rangle$ of two spikes x, x' drawn independently from the prior $\mathcal{X}$. These tail probabilities are encapsulated by the rate function $f_{\mathcal{X}} : [0, 1) \to [0, \infty)$ of the large deviations of $\mathcal{X}$, which is intuitively defined by $\Pr[|\langle x, x' \rangle| \geq t] \approx \exp(-n f_{\mathcal{X}}(t))$. Formally, we define $f_{\mathcal{X}}$ as follows.

DEFINITION 5.4. Let $\mathcal{X} = \{\mathcal{X}_n\}$ be a spike prior. For x, x' drawn independently from $\mathcal{X}_n$ and $t \in [0, 1)$, let

$$f_{n,\mathcal{X}}(t) = -\frac{1}{n} \log \Pr\big[|\langle x, x' \rangle| \geq t\big].$$

Suppose we have $f_{n,\mathcal{X}}(t) \geq b_{n,\mathcal{X}}(t)$ for some sequence of functions $b_{n,\mathcal{X}}$ that converges uniformly on $[0, 1)$ to $f_{\mathcal{X}}$ as $n \to \infty$. Then we call such $f_{\mathcal{X}}$ the *rate function* of the prior $\mathcal{X}$.

Without loss of generality, $f_{\mathcal{X}}(0) = 0$ and $f_{\mathcal{X}}(t)$ is nondecreasing. Note that a tail bound of the form $\Pr[|\langle x, x'\rangle| \geq t] \leq \text{poly}(n) \exp(-n f_{\mathcal{X}}(t))$ is sufficient to establish that $f_{\mathcal{X}}$ is a rate function.

We now state the rate functions for some priors of interest. It is proven by Perry, Wein and Bandeira (2016) that these indeed satisfy the definition of rate function.

PROPOSITION 5.5 [Perry, Wein and Bandeira (2016)]. *We have the following rate functions for the spherical, Rademacher and sparse Rademacher priors*:

- *Spherical*: $f_{\text{sph}}(t) = -\frac{1}{2}\log(1 - t^2)$.
- *Rademacher*: $f_{\text{Rad}}(t) = \log 2 - H(\frac{1+t}{2})$.
- *Sparse Rademacher*[11] *with sparsity* ρ:

$$f_\rho(t) = \min_{\zeta \in [\max(\rho t, 1-2\rho), \rho]} G_\rho(\zeta) + \zeta f_{\text{Rad}}\left(\frac{\rho t}{\zeta}\right),$$

where

$$G_\rho(\zeta) = -H\big(\{\zeta, \rho - \zeta, \rho - \zeta, 1 - 2\rho + \zeta\}\big) + 2H(\rho).$$

Here, $H(p) = -p \log p - (1 - p)\log(1 - p)$ *is the binary entropy, and* $H(\{p_i\}) = -\sum_i p_i \log p_i$.

Note that rate functions for general i.i.d. priors can be easily derived from large deviations theory (Cramér's theorem) since $\langle x, x'\rangle$ is the sum of n i.i.d. random variables; this is how the Rademacher rate function is derived. However, to obtain stronger results in some cases, one may use a variant of the prior that conditions on typical outcomes [similar to our conditioning method for the Wigner model (Section 3.6) or Appendix A of Banks et al. (2016)]; this is how the sparse Rademacher rate function is derived.

We will need the following strengthening of the notion of rate function.

DEFINITION 5.6. We say that a rate function $f_{\mathcal{X}}$ for a prior $\mathcal{X}$ admits a *local Chernoff bound* if there exists $T > 0$ and $C > 0$ such that for any n,

$$\Pr\big[|\langle x, x'\rangle| \geq t\big] \leq C \exp\big(-n f_{\mathcal{X}}(t)\big) \qquad \forall t \in [0, T],$$

where x and x' are drawn independently from $\mathcal{X}_n$.

The Rademacher and sparse Rademacher rate functions in Proposition 5.5 each admit a local Chernoff bound; see Perry, Wein and Bandeira (2016).

[11] This is for a variant of the sparse Rademacher prior where the sparsity is exactly ρn. See Appendix N for details on how this extends to our variant.

5.3. *Main lower bound result.* We are now ready to state our main lower bound result. Recall that $\lambda^*_{\mathcal{X}}$ denotes the Wigner threshold (1).

THEOREM 5.7. *Let $\mathcal{X}$ be a spike prior with rate function $f_{\mathcal{X}}$. Let $\beta > -1$ and $\gamma^* > 0$. Suppose that either*:

(i) $\beta^2/\gamma^* \le (\lambda^*_{\mathcal{X}})^2$, *or*
(ii) $f_{\mathcal{X}}$ *admits a local Chernoff bound* (*Definition* 5.6).

If

$$\gamma^* f_{\mathcal{X}}(t) \ge (1+\beta)\frac{t(w-t)}{1-t^2} + \frac{1}{2}\log\left(\frac{1-w^2}{1-t^2}\right) \qquad \forall t \in (0,1), \tag{2}$$

where

$$w = \sqrt{A^2+1} - A \qquad \text{with } A = \frac{1-t^2}{2t(\beta+1)},$$

then $\mathrm{Wish}(\gamma, \beta, \mathcal{X})$ *is contiguous to* $\mathrm{Wish}(\gamma)$ *for all $\gamma > \gamma^*$.*

We expect condition (ii) to hold for all reasonable priors; condition (i) yields a weaker result in some cases but is sometimes more convenient. Some basic properties of (2) are discussed in Appendix J. In Appendix M, we establish the following monotonicity.

PROPOSITION 5.8. *Let $\mathcal{X}$ be a spike prior. Fix $\lambda > 0$ and $\bar{\beta} \in (-1,\infty)\setminus\{0\}$. If* (2) *holds for $\bar{\beta}$ and $\gamma^* = \bar{\beta}^2/\lambda^2$, then it also holds for any $\beta > \bar{\beta}$ and $\gamma^* = \beta^2/\lambda^2$.*

In particular, if $\lambda = 1$ (so that $\gamma^* = \beta^2$, corresponding to the spectral threshold) we have that if Theorem 5.7 shows that the PCA threshold is optimal for some $\bar{\beta} \in (-1,\infty)\setminus\{0\}$, then the PCA threshold is also optimal for all $\beta > \bar{\beta}$.

The following connection to the Wigner model is also proved in Appendix M, corresponding to the $\beta \to 0$ limit of the monotonicity property above.

COROLLARY 5.9. *Suppose $\langle x, x'\rangle$ is (σ^2/n)-sub-Gaussian* (*Definition* 3.7), *where x and x' are drawn independently from $\mathcal{X}_n$. Then for any $\beta > 0$ and any $\gamma > \beta^2\sigma^2$ we have* $\mathrm{Wish}(\gamma, \beta, \mathcal{X}) \lhd \mathrm{Wish}(\gamma)$.

Recall that the sub-Gaussian condition above implies a *Wigner* lower bound for all $\lambda < 1/\sigma$ (Proposition 3.8). This means whenever Proposition 3.8 implies that the PCA threshold is optimal for the Wigner model, we also have that the PCA threshold is optimal for the Wishart model for any positive β. Conversely, if Theorem 5.7 shows that PCA is optimal for all $\beta > 0$ then it is also optimal for the Wigner model (see Proposition M.2). In light of the above monotonicity

(Proposition 5.8), these results make sense because the Wigner model corresponds to the $\gamma \to 0$ limit of the Wishart model [Johnstone and Onatski (2015)].

We also show (in Appendix M) that for a wide range of priors, the PCA threshold becomes optimal for sufficiently large β.

PROPOSITION 5.10. *Suppose* $\mathcal{X} = \text{i.i.d.}(\pi/\sqrt{n})$ *where* π *is a mean-zero unit-variance distribution for which* $\pi\pi'$ *(product of two independent copies of* π*) has a moment-generating function* $M(\theta) \triangleq \mathbb{E}\exp(\theta\pi\pi')$ *which is finite on an open interval containing zero. Then there exists* $\bar{\beta}$ *such that for any* $\beta \geq \bar{\beta}$ *and any* $\gamma > \beta^2$ *we have* $\text{Wish}(\gamma, \beta, \mathcal{X}) \lhd \text{Wish}(\gamma)$.

A final property of Theorem 5.7 is that it gives similar thresholds for similar priors in the sense of Proposition 3.13 for the Wigner model; see Proposition L.1 for details.

5.4. *Lower bound proof summary.* The full proof of Theorem 5.7 will be completed in the next section, but we now describe the proof outline and give some preliminary results. We approach contiguity for the spiked Wishart model through the second moment method outlined in Section 2. Note that detection can only become easier when given the original sample matrix X (instead of $\frac{1}{N}XX^\top$), so we establish the stronger statement that the spiked distribution on X is contiguous to the unspiked distribution. We first simplify the second moment in high generality.

PROPOSITION 5.11. *For any* $|\beta| < 1$, *there exists* $\delta > 0$ *such that the following holds. Let* $\mathcal{X}$ *be a spike prior supported on vectors* x *with* $1 - \delta \leq \|x\| \leq 1 + \delta$. *In distribution* Q_n, *let a hidden spike* x *be drawn from* $\mathcal{X}_n$, *and let* N *independent samples* y_i, $1 \leq i \leq N$, *be revealed from the normal distribution* $\mathcal{N}(0, I_{n\times n} + \beta xx^\top)$. *In distribution* P_n, *let* N *independent samples* y_i, $1 \leq i \leq N$, *be revealed from* $\mathcal{N}(0, I_{n\times n})$. *Then we have*

$$\mathop{\mathbb{E}}_{P_n}\left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2\right] = \mathop{\mathbb{E}}_{x,x'\sim\mathcal{X}}\left[(1-\beta^2\langle x, x'\rangle^2)^{-N/2}\right].$$

This result has appeared in higher generality [Cai, Ma and Wu (2015)]; for completeness, we give the proof in Section 5.5.1. The condition $1 - \delta \leq \|x\| \leq 1 + \delta$ will not be an issue because we can always consider a modified prior that conditions on this $(1 - o(1))$-probability event (see Lemma 2.4). Note that the above second moment has the curious property of symmetry under replacing β with $-\beta$. In contrast, the original Wishart model does not, since for instance $\beta > 1$ is allowed while $\beta < -1$ is not. As a result, the second moment method gives good results for negative β but substantially sub-optimal results for positive β. To remedy this, we will apply the conditional second moment method (Lemma 2.4), conditioning on an event that depends jointly on the signal and noise (we previously only conditioned on the signal).

The proof of Theorem 5.7 has two parts. In Section 5.5.2, we control the *small deviations* of the second moment, that is, the contribution from $\langle x, x'\rangle^2$ values at most some small $\varepsilon > 0$. Here, we use either the Wigner lower bound (i) or the local Chernoff bound (ii) [combined with (2)], whichever is provided. This step uses the basic second moment of Proposition 5.11 without conditioning. In Section 5.6, we complete the proof by controlling the remaining *large deviations* of the conditional second moment. Here, we use the condition (2) on the rate function of the prior.

We remark that conditions (i) and (ii) in Theorem 5.7 are related because using the sub-Ggaussian method of Section 3.3, a Chernoff-type bound on $\langle x, x'\rangle$ implies a Wigner lower bound; note however that a local Chernoff bound only needs to hold near $t = 0$.

5.5. *Proof of lower bound.* This section is devoted to proving Theorem 5.7. Along the way, we will also prove Propositions 5.11 and 5.2.

5.5.1. *Second moment computation*: *Proof of Proposition* 5.11. We first compute

$$\begin{aligned}
&\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}(y_1, \dots, y_N) \\
&\quad = \mathop{\mathbb{E}}_{x' \sim \mathcal{X}} \left[\prod_{i=1}^{n} \frac{\exp(-\frac{1}{2} y_i^\top (I + \beta x' x'^\top)^{-1} y_i)}{\sqrt{\det(I + \beta x' x'^\top)} \exp(-\frac{1}{2} y_i^\top y_i)} \right] \\
&\quad = \mathop{\mathbb{E}}_{x'} \left[\det(I + \beta x' x'^\top)^{-N/2} \prod_{i=1}^{N} \exp\left(-\frac{1}{2} y_i^\top \left((I + \beta x' x'^\top)^{-1} - I \right) y_i \right) \right].
\end{aligned}$$

Note that $(I + \beta x' x'^\top)^{-1}$ has eigenvalue $(1 + \beta \|x'\|^2)^{-1}$ on x' and eigenvalue 1 on the orthogonal complement of x'. Thus $(I + \beta x' x'^\top)^{-1} - I = \frac{-\beta}{1+\beta\|x'\|^2} x' x'^\top$, and we have

$$= \mathop{\mathbb{E}}_{x'} \left[(1 + \beta \|x'\|^2)^{-N/2} \prod_{i=1}^{N} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|x'\|^2} \langle y_i, x' \rangle^2 \right) \right].$$

Passing to the second moment, we compute

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P_n} \left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n} \right)^2 \right] \\
&\quad = \mathop{\mathbb{E}}_{Q_n} \left[\frac{\mathrm{d}Q_n}{\mathrm{d}P_n} \right] \\
&\quad = \mathop{\mathbb{E}}_{x, x'} \left[(1 + \beta \|x'\|^2)^{-N/2} \prod_{i=1}^{N} \mathop{\mathbb{E}}_{y_i \sim \mathcal{N}(0, I + \beta x x^\top)} \exp\left(\frac{1}{2} \frac{\beta}{1 + \beta \|x'\|^2} \langle y_i, x' \rangle^2 \right) \right].
\end{aligned}$$

Over the randomness of y_i, we have $\langle y_i, x'\rangle \sim \mathcal{N}(0, \|x'\|^2 + \beta\langle x, x'\rangle^2)$, so that the inner expectation can be simplified using the moment-generating function (MGF) of the χ_1^2 distribution:

$$= \mathop{\mathbb{E}}_{x,x'}\left[(1+\beta\|x'\|^2)^{-N/2}\prod_{i=1}^{N}\left(1-\frac{\beta}{1+\beta\|x'\|^2}(\|x'\|^2+\beta\langle x,x'\rangle^2)\right)^{-1/2}\right]$$
$$= \mathop{\mathbb{E}}_{x,x'}\left[(1-\beta^2\langle x,x'\rangle^2)^{-N/2}\right]$$

as desired. Here, the MGF step requires

$$\frac{\beta}{1+\beta\|x'\|^2}(\|x'\|^2+\beta\langle x,x'\rangle^2) < 1. \tag{3}$$

Provided that $\|x\|$ and $\|x'\|$ are sufficiently close to 1, this is true so long as either $|\beta| < 1$ (as assumed by Proposition 5.11) or $\langle x, x'\rangle^2$ is sufficiently small (as in the small deviations of the next section).

5.5.2. *Small deviations and proof of Proposition 5.2.* We now show how to bound the *small deviations*

$$S(\varepsilon) \triangleq \mathop{\mathbb{E}}_{x,x'\sim\mathcal{X}}(1-\beta^2\langle x,x'\rangle^2)^{-N/2}\mathbb{1}_{\langle x,x'\rangle^2\le\varepsilon}$$

of the Wishart second moment in terms of the Wigner second moment. [Assume $\|x\|, \|x'\|$ are sufficiently close to 1 and $\varepsilon > 0$ is a sufficiently small constant so that (3) holds.] Letting $\hat{\gamma} = n/N$ so that $\hat{\gamma} \to \gamma$, we have

$$S(\varepsilon) = \mathop{\mathbb{E}}_{x,x'\sim\mathcal{X}}\exp\left(\frac{-n}{2\hat{\gamma}}\log(1-\beta^2\langle x,x'\rangle^2)\right)\mathbb{1}_{\langle x,x'\rangle^2\le\varepsilon}$$
$$\le \mathop{\mathbb{E}}_{x,x'\sim\mathcal{X}}\exp\left(\frac{-n}{2\hat{\gamma}\varepsilon^2}\log(1-\varepsilon^2\beta^2)\langle x,x'\rangle^2\right)$$

using the convexity of $t \mapsto -\log(1-\beta^2 t)$. Note that this resembles the Wigner second moment and so (by definition of $\lambda_{\mathcal{X}}^*$) it is bounded as $n \to \infty$ so long as

$$\frac{-1}{\gamma\varepsilon^2}\log(1-\varepsilon^2\beta^2) < (\lambda_{\mathcal{X}}^*)^2. \tag{4}$$

Proposition 5.2 now follows by setting $\varepsilon = 1+\delta$ for small $\delta > 0$ and conditioning the prior on $\|x\|^2 \le 1+\delta$. [See Section 3.5 for similar arguments; note that the conditioning can only increase the Wigner second moment by a $1+o(1)$ factor.] Furthermore, using the bound $\log t \ge 1 - 1/t$ we have the following fact that will be used in the proof of Theorem 5.7.

LEMMA 5.12. *If $\beta^2/\gamma < (\lambda_{\mathcal{X}}^*)^2$, then there exists $\varepsilon > 0$ such that $S(\varepsilon)$ is bounded as $n \to \infty$.*

Note that $\beta^2/\gamma < (\lambda^*_{\mathcal{X}})^2$ is precisely condition (i) in the statement of Theorem 5.7. If instead condition (ii) holds, we can control the small deviations using the following lemma, deferred to Appendix K.

LEMMA 5.13. *If* (2) *holds and* $f_{\mathcal{X}}$ *admits a local Chernoff bound, then there exists* $\varepsilon > 0$ *such that* $S(\varepsilon)$ *is bounded as* $n \to \infty$.

5.6. *Proof of Theorem* 5.7. We now prove our main lower bound result using the conditional second moment method. Define Q_n and P_n as in Proposition 5.11. For a vector $x \in \mathbb{R}^n$ and an $n \times n$ matrix Y, define the "good" event $\Omega(x, Y)$ by

$$x^\top Y x/\|x\|^2 \in \big[(1+\beta\|x\|^2)(1-\eta), (1+\beta\|x\|^2)(1+\eta)\big],$$

where $\eta = \frac{\log n}{\sqrt{n}}$. Note that under Q_n (where x is the spike and Y is the Wishart matrix: $Y = \frac{1}{N} X X^\top$ where the columns of X are the samples y_i), $x^\top Y x/\|x\|^2 \sim (1+\beta\|x\|^2)\chi^2_N/N$ and so $\Omega(x, Y)$ occurs with probability $1 - o(1)$. Let $\tilde{Q}_n$ be the conditional distribution of Q_n given $\Omega(x, Y)$.

For simplicity, we now specialize to the case where $\mathcal{X}$ is supported on unit vectors $\|x\| = 1$; see Appendix L for the general case. Similar to the proof of Proposition 5.11, we compute the conditional second moment as follows:

$$\frac{\mathrm{d}\tilde{Q}_n}{\mathrm{d}P_n} = (1+o(1)) \mathop{\mathbb{E}}_{x' \sim \mathcal{X}} \left[\mathbb{1}_{\Omega(x', Y)} (1+\beta)^{-N/2} \prod_{i=1}^N \exp\left(\frac{1}{2} \frac{\beta}{1+\beta} \langle y_i, x' \rangle^2 \right) \right]$$

and so $\mathbb{E}_{P_n} (\frac{\mathrm{d}\tilde{Q}_n}{\mathrm{d}P_n})^2 = (1+o(1)) \mathbb{E}_{x, x' \sim \mathcal{X}} m(\langle x, x' \rangle)$ where

$$\begin{aligned} m(\langle x, x' \rangle) &= \mathop{\mathbb{E}}_{Y \sim P_n} (1+\beta)^{-N} \\ &\quad \times \exp\left(\frac{N}{2} \frac{\beta}{1+\beta} (x^\top Y x + x'^\top Y x') \right) \mathbb{1}_{\Omega(x, Y)} \mathbb{1}_{\Omega(x', Y)} \qquad (5) \\ &= \mathop{\mathbb{E}}_{Y \sim P_n} (1+\beta)^{-N} \exp\left(N\beta \left(1 + \frac{\Delta}{2} + \frac{\Delta'}{2} \right) \right) \mathbb{1}_{|\Delta| \le \eta} \mathbb{1}_{|\Delta'| \le \eta}, \end{aligned}$$

where Δ, Δ' are defined by $x^\top Y x = (1+\beta)(1+\Delta)$ and $x'^\top Y x' = (1+\beta)(1+\Delta')$. We will see below that m is indeed only a function of $\langle x, x' \rangle$.

5.6.1. *Interval* $|\alpha| \in [\varepsilon, 1-\varepsilon]$. Let $\alpha = \langle x, x' \rangle$. Let $\varepsilon > 0$ be a small constant (not depending on n), to be chosen later. First, let us focus on the contribution from $|\alpha| \in [\varepsilon, 1-\varepsilon]$, that is, we want to bound

$$M_1 \triangleq \mathop{\mathbb{E}}_{\alpha} \big[\mathbb{1}_{|\alpha| \in [\varepsilon, 1-\varepsilon]} m(\alpha) \big].$$

For $Y \sim P_n$ and with x, x' fixed unit vectors, the matrix

$$\begin{pmatrix} Nx^\top Yx & Nx^\top Yx' \\ Nx^\top Yx' & Nx'^\top Yx' \end{pmatrix}$$

follows the 2×2 Wishart distribution with N degrees of freedom and shape matrix

$$\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}, \qquad \alpha = \langle x, x' \rangle.$$

By integrating over $c = x^\top Yx'$ and using the PDF of the Wishart distribution, we have

$$\begin{aligned} m(\alpha) = \iiint (1+\beta)^2 \exp\Big\{ N\Big[&-\log(1+\beta) + \beta\Big(1 + \frac{\Delta}{2} + \frac{\Delta'}{2}\Big) \\ &+ \Big(\frac{1}{2} - \frac{3}{N}\Big)\log\big((1+\beta)^2(1+\Delta)(1+\Delta') - c^2\big) \\ &- \frac{1}{1-\alpha^2}\Big((1+\beta)\Big(1 + \frac{\Delta}{2} + \frac{\Delta'}{2}\Big) - \alpha c\Big) - \frac{1}{2}\log(1-\alpha^2) \\ &+ \log(N/2) - \frac{1}{N}\log\Gamma_2(N/2)\Big]\Big\}\, \mathrm{d}c\, \mathrm{d}\Delta\, \mathrm{d}\Delta', \end{aligned}$$

where the integration is over the domain $|\Delta| \le \eta$, $|\Delta'| \le \eta$, and $|c| \le (1+\beta) \times \sqrt{(1+\Delta)(1+\Delta')}$, and Γ_2 denotes the multivariate gamma function.

Using $\eta = o(1)$ and applying Stirling's approximation to Γ_2, we have for $|\alpha| \in [\varepsilon, 1-\varepsilon]$,

$$\begin{aligned} m(\alpha) \le \max_{|c| \le 1+\beta} (1+\beta)^2 \exp\Big\{ N\Big[&-\log(1+\beta) + \beta + \frac{1}{2}\log\big((1+\beta)^2 - c^2\big) \\ &- \frac{1+\beta-\alpha c}{1-\alpha^2} - \frac{1}{2}\log\big(1-\alpha^2\big) + 1 + o(1)\Big]\Big\}, \end{aligned}$$

where the $o(1)$ is uniform in α. Letting $w = c/(1+\beta)$ and solving explicitly for the optimal w,

$$\begin{aligned} m(\alpha) &\le m_1(\alpha) \\ &\triangleq (1+\beta)^2 \exp\Big\{ N\Big[(1+\beta)\frac{\alpha(w-\alpha)}{1-\alpha^2} + \frac{1}{2}\log\Big(\frac{1-w^2}{1-\alpha^2}\Big) + o(1)\Big]\Big\}, \end{aligned}$$

where

$$w = w(\alpha) = \pm\sqrt{A^2+1} - A \qquad \text{with } A = \frac{1-\alpha^2}{2\alpha(\beta+1)}$$

and $\pm$ has the same sign as α.

We now show how to bound the contribution to M_1 from positive α; the proof for negative α is similar. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\alpha}\left[\mathbb{1}_{\alpha\in[\varepsilon,1-\varepsilon]}m_1(\alpha)\right]\\
&\quad=\int_0^\infty \Pr\left[\mathbb{1}_{\alpha\in[\varepsilon,1-\varepsilon]}m_1(\alpha)\geq u\right]\mathrm{d}u\\
&\quad=\int_0^\infty \Pr\left[\alpha\in[\varepsilon,1-\varepsilon]\text{ and }m_1(\alpha)\geq u\right]\mathrm{d}u\\
&\quad=m_1(\varepsilon)\Pr\left[\alpha\in[\varepsilon,1-\varepsilon]\right]\\
&\qquad+\int_{m_1(\varepsilon)}^{m_1(1-\varepsilon)}\Pr\left[\alpha\in[\varepsilon,1-\varepsilon]\text{ and }m_1(\alpha)\geq u\right]\mathrm{d}u.
\end{aligned}$$

Since $m_1(\alpha)$ is strictly increasing on $[0,1]$ (see Appendix J), we can apply the change of variables $u=m_1(t)$ to obtain

$$\begin{aligned}
&=m_1(\varepsilon)\Pr\left[\alpha\in[\varepsilon,1-\varepsilon]\right]\\
&\qquad+\int_\varepsilon^{1-\varepsilon}\Pr\left[\alpha\in[\varepsilon,1-\varepsilon]\text{ and }\alpha\geq t\right]m_1(t)\,O(N)\,\mathrm{d}t\\
&\leq m_1(\varepsilon)\Pr[\alpha\geq\varepsilon]+O(N)\int_\varepsilon^{1-\varepsilon}\Pr[\alpha\geq t]m_1(t)\,\mathrm{d}t.
\end{aligned}$$

Plugging in the rate function to bound $\Pr[\alpha\geq\varepsilon]$ and $\Pr[\alpha\geq t]$, we obtain $M_1=o(1)$ provided that (2) holds. The contribution from negative α yields the same condition (2) due to the symmetry $w(-\alpha)=-w(\alpha)$ and $m_1(-\alpha)=m_1(\alpha)$.

5.6.2. *Interval* $|\alpha|\in[0,\varepsilon)$. This case needs special consideration because both sides of (2) approach 0 as $t\to 0$ and so the last step above requires α to be bounded away from 0. Since [up to a factor of $1+o(1)$] conditioning Q_n on $\Omega(x,Y)$ only decreases the second moment (for each value of α), we can revert back to the basic second moment: the contribution $M_2\triangleq\mathbb{E}_\alpha[\mathbb{1}_{|\alpha|\in[0,\varepsilon)}m(\alpha)]$ is bounded by the small deviations $S(\varepsilon^2)$ from Section 5.5.2. It therefore follows from either Lemma 5.12 or Lemma 5.13 that provided ε is small enough, M_2 is bounded as $n\to\infty$.

5.6.3. *Interval* $|\alpha|\in(1-\varepsilon,1]$. This case needs special consideration because in the calculations for the $[\varepsilon,1-\varepsilon]$ interval, certain terms in the exponent blow up at $|\alpha|=1$, which prevents us from replacing Δ,Δ' by an error term that is $o(1)$ uniformly in α. To deal with this case, we will bound $m(\alpha)$ by its worst-case value $m(1)$.

To see that $m(1)$ is the worst case, notice from (5) that up to an $\exp(o(N))$ factor (which will turn out to be negligible), $m(\alpha)$ is proportional to

$\Pr[|\Delta| \le \eta \text{ and } |\Delta'| \le \eta]$. Since $Nx^\top Yx$ and $Nx'^\top Yx'$ each follow at χ^2_N distribution (with correlation that increases with $|\alpha|$), this probability is maximized when they are perfectly correlated at $|\alpha| = 1$.

We now proceed to bound $m(1)$. Let $Y \sim P_n$, and let x, x' be fixed unit vectors with $|\alpha| = 1$. We have that $Nx^\top Yx$ follows a χ^2_N distribution, with $Nx'^\top Yx' = Nx^\top Yx$. Similar to the computation for $[\varepsilon, 1-\varepsilon]$, we obtain

$$m(1) \le m_3 \triangleq (1+\beta)\exp\left\{N\left[-\frac{1}{2}\log(1+\beta) - \frac{1}{2}(1-\beta) + \frac{1}{2} + o(1)\right]\right\}$$

and

$$M_3 \triangleq \mathop{\mathbb{E}}_{\alpha}\left[\mathbb{1}_{|\alpha|\in(1-\varepsilon,1]} m(\alpha)\right] \le \exp\big(o(N)\big)\Pr\big[|\alpha| \ge 1-\varepsilon\big] m_3.$$

Plugging in the rate function, M_3 is $o(1)$ provided that $\gamma f_{\mathcal{X}}(1-\varepsilon) > -\frac{1}{2} \times \log(1+\beta) - \frac{1}{2}(1-\beta) + \frac{1}{2}$. This follows from (2) (near $t = 1$) provided ε is small enough (since $f_{\mathcal{X}}$ is an increasing function of t).

Acknowledgments. The authors are indebted to Philippe Rigollet for helpful discussions and for many comments on a draft. We thank the anonymous reviewers for many helpful, detailed comments.

SUPPLEMENTARY MATERIAL

Optimality and sub-optimality of PCA in spiked random matrix models: Supplementary proofs (DOI: 10.1214/17-AOS1625SUPP; .pdf). Contains proofs omitted from this paper for the sake of length.

REFERENCES

AMINI, A. A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory* 2454–2458.

ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices. Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2012). Detection of correlations. *Ann. Statist.* **40** 412–435. MR3014312

ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

ARIAS-CASTRO, E. and VERZELEN, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42** 940–969.

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

BANDEIRA, A. S., BOUMAL, N. and SINGER, A. (2014). Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. Available at arXiv:1411.3272.

BANKS, J., MOORE, C., NEEMAN, J. and NETRAPALLI, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *29th Annual Conference on Learning Theory* 383–416.

BANKS, J., MOORE, C., VERZELEN, N., VERSHYNIN, R. and XU, J. (2017). Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. Available at arXiv:1607.05222.

BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVA, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. Available at arXiv:1606.04142.

BARRON, A. R. (1986). Entropy and the central limit theorem. *Ann. Probab.* **14** 336–342. MR0815975

BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. MR2810285

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201

BERTHET, Q. and RIGOLLET, P. (2013a). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BERTHET, Q. and RIGOLLET, P. (2013b). Complexity theoretic lower bounds for sparse principal component detection. In *COLT* 1046–1066.

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

BOUMAL, N. (2016). Nonconvex phase synchronization. *SIAM J. Optim.* **26** 2355–2377. MR3566919

BOUMAL, N., SINGER, A., ABSIL, P.-A. and BLONDEL, V. D. (2014). Cramér–Rao bounds for synchronization of rotations. *Inf. Inference* **3** 1–39.

BROWN, L. D. (1982). A proof of the central limit theorem motivated by the Cramér–Rao inequality. In *Statistics and Probability*: *Essays in Honor of C. R. Rao* (G. Kallianpur, P. Krishnaiah and J. Ghosh, eds.) 141–148. North-Holland, Amsterdam. MR659464

BUTUCEA, C. and INGSTER, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688.

CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449.

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110.

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815.

CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37** 1–47. MR2489158

DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2016). Asymptotic mutual information for the binary stochastic block model. In 2016 *IEEE International Symposium on Information Theory* 185–189.

DESHPANDE, Y. and MONTANARI, A. (2014a). Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems* 334–342.

DESHPANDE, Y. and MONTANARI, A. (2014b). Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory* 2197–2201.

DESHPANDE, Y., MONTANARI, A. and RICHARD, E. (2014). Cone-constrained principal component analysis. In *Advances in Neural Information Processing Systems* 2717–2725.

DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. MR3670197

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

EGLOFF, D., LEIPPOLD, M. and WU, L. (2010). The term structure of variance swap rates and optimal variance swap investments. *J. Financ. Quant. Anal.* **45** 1279–1310.

FÉRAL, D. and PÉCHÉ, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.* **272** 185–228. MR2291807

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econ. Stat.* **82** 540–554.

GUERRA, F. (2003). Broken replica symmetry bounds in the mean field spin glass model. *Comm. Math. Phys.* **233** 1–12. MR1957729

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526.

JANSON, S. (1995). Random regular graphs: Asymptotic distributions and contiguity. *Combin. Probab. Comput.* **4** 369–405. MR1377557

JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference* **2** 115–144. MR3311445

JAVANMARD, A., MONTANARI, A. and RICCI-TERSENGHI, F. (2016). Phase transitions in semidefinite relaxations. *Proc. Natl. Acad. Sci. USA* **113** E2218–E2223.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

JOHNSTONE, I. M. and LU, A. Y. (2004). Sparse principal components analysis. Unpublished manuscript 7.

JOHNSTONE, I. M. and ONATSKI, A. (2015). Testing in high-dimensional spiked models. Available at arXiv:1509.07269.

KANNAN, R. and VEMPALA, S. (2016). Beyond spectral: Tight bounds for planted Gaussians. Available at arXiv:1608.03643.

KE, Z. T. (2016). Detecting rare and weak spikes in large covariance matrices. Available at arXiv:1609.00883.

KRAUTHGAMER, R., NADLER, B. and VILENCHIK, D. (2015). Do semidefinite relaxations solve sparse PCA up to the information limit? *Ann. Statist.* **43** 1300–1322. MR3346704

KRZAKALA, F., XU, J. and ZDEBOROVÁ, L. (2016). Mutual information in rank-one matrix estimation. Available at arXiv:1603.08447.

LELARGE, M. and MIOLANE, L. (2016). Fundamental limits of symmetric low-rank matrix estimation. Available at arXiv:1611.03888.

LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015a). Phase transitions in sparse PCA. In *IEEE International Symposium on Information Theory* 1635–1639.

LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015b). MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In 53*rd Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 680–687.

LE CAM, L. (1960). *Locally Asymptotically Normal Families of Distributions*: *Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. Univ. California Press, Berkeley, CA.

LITTERMAN, R. B. and SCHEINKMAN, J. (1991). Common factors affecting bond returns. *J. Fixed Income* **1** 54–61.

MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698

MCSHERRY, F. (2001). Spectral partitioning of random graphs. In *Proceedings of the* 42*nd IEEE Symposium on Foundations of Computer Science* 529–537.

MOLLOY, M. S. O., ROBALEWSKA, H., ROBINSON, R. W. and WORMALD, N. C. (1997). 1-factorizations of random regular graphs. *Random Structures Algorithms* **10** 305–321. MR1606218

MONTANARI, A., REICHMAN, D. and ZEITOUNI, O. (2015). On the limitation of spectral methods: From the Gaussian hidden clique problem to rank-one perturbations of Gaussian tensors. In *Advances in Neural Information Processing Systems* 217–225.

MONTANARI, A. and RICHARD, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inform. Theory* **62** 1458–1484.

MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461.

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. MR3189485

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173.

PERRY, A., WEIN, A. S. and BANDEIRA, A. S. (2016). Statistical limits of spiked tensor models. Available at arXiv:1612.07728.

PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2017). Supplement to "Optimality and sub-optimality of PCA I: Spiked random matrix models." DOI:10.1214/17-AOS1625SUPP.

PITMAN, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. Chapman & Hall, London; A Halsted Press Book, Wiley, New York. MR0549771

PIZZO, A., RENFREW, D. and SOSHNIKOV, A. (2013). On finite rank deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 64–94. MR3060148

RANGAN, S. and FLETCHER, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In *IEEE International Symposium on Information Theory* 1246–1250.

ROBINSON, R. W. and WORMALD, N. C. (1994). Almost all regular graphs are Hamiltonian. *Random Structures Algorithms* **5** 363–374.

SHEN, D., SHEN, H. and MARRON, J. S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115** 317–333.

SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179.

SUN, X. and NOBEL, A. B. (2008). On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res.* **9** 2431–2453.

SUN, X. and NOBEL, A. B. (2013). On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix. *Bernoulli* **19** 275–294. MR3019495

TAO, T. (2012). *Topics in Random Matrix Theory*. *Graduate Studies in Mathematics* **132**. Amer. Math. Soc., Providence, RI. MR2906465

TAO, T. and VU, V. (2014). Random matrices: The universality phenomenon for Wigner ensembles. In *Modern Aspects of Random Matrix Theory*. *Proc. Sympos. Appl. Math.* **72** 121–172. Amer. Math. Soc., Providence, RI. MR3288230

VERZELEN, N. and ARIAS-CASTRO, E. (2015). Community detection in sparse random networks. *Ann. Appl. Probab.* **25** 3465–3510. MR3404642

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. In *Proceedings of the* 15*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 1278–1286.

WORMALD, N. C. (1999). Models of random regular graphs. In *Surveys in Combinatorics*, 1999 (*Canterbury*). *London Mathematical Society Lecture Note Series* **267** 239–298. Cambridge Univ. Press, Cambridge. MR1725006

A. PERRY
A. S. WEIN
DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: ameliaperry@mit.edu
awein@mit.edu

A. S. BANDEIRA
DEPARTMENT OF MATHEMATICS
AND
CENTER FOR DATA SCIENCE
COURANT INSTITUTE OF
MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
251 MERCER STREET
NEW YORK, NEW YORK 10012
USA
E-MAIL: bandeira@cims.nyu.edu

A. MOITRA
DEPARTMENT OF MATHEMATICS
AND
COMPUTER SCIENCE AND ARTIFICIAL
INTELLIGENCE LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: moitra@mit.edu