The Annals of Statistics
2017, Vol. 45, No. 5, 2218–2247
DOI: 10.1214/16-AOS1525

SPECTRUM ESTIMATION FROM SAMPLES

BY WEIHAO KONG AND GREGORY VALIANT[1]

Stanford University

We consider the problem of approximating the set of eigenvalues of the covariance matrix of a multivariate distribution (equivalently, the problem of approximating the "population spectrum"), given access to samples drawn from the distribution. We consider this recovery problem in the regime where the sample size is comparable to, or even sublinear in the dimensionality of the distribution. First, we propose a theoretically optimal and computationally efficient algorithm for recovering the moments of the eigenvalues of the population covariance matrix. We then leverage this accurate moment recovery, via a Wasserstein distance argument, to accurately reconstruct the vector of eigenvalues. Together, this yields an eigenvalue reconstruction algorithm that is asymptotically consistent as the dimensionality of the distribution and sample size tend toward infinity, even in the sublinear sample regime where the ratio of the sample size to the dimensionality tends to zero. In addition to our theoretical results, we show that our approach performs well in practice for a broad range of distributions and sample sizes.

1. Introduction. One of the most insightful properties of a multivariate distribution (or dataset) is the vector of eigenvalues of the covariance of the distribution or dataset. This vector of eigenvalues—the "spectrum"—contains important information about the structure and geometry of the distribution. Indeed, the first step in understanding many high-dimensional distributions is to compute the eigenvalues of the covariance of the data, often with the aim of understanding whether there exist lower dimensional subspaces that accurately capture the majority of the structure of the high-dimensional distribution (e.g., as a first step in performing Principal Component Analysis).

Given independent samples drawn from a multivariate distribution over $\mathbb{R}^d$, when can this vector of eigenvalues of the (distribution/"population") covariance be accurately computed? In the regime in which the number of samples, n, is significantly larger than the dimension d, the empirical covariance matrix of the samples will be an accurate approximation of the true distribution covariance (assuming some modest moment bounds), and hence the empirical spectrum will accurately reflect the true population spectrum. In the linear or sublinear regime in which n is comparable to, or significantly smaller than d, the empirical covariance

Received February 2016; revised October 2016.

[1]Supported by NSF CAREER Award CCF-1351108 and a Sloan fellowship.

MSC2010 subject classifications. 62H12, 62H10.

Key words and phrases. Spectrum estimation, eigenvalues of covariance matrices, sublinear sample size, method of moments, random matrix theory, high-dimensional inference.

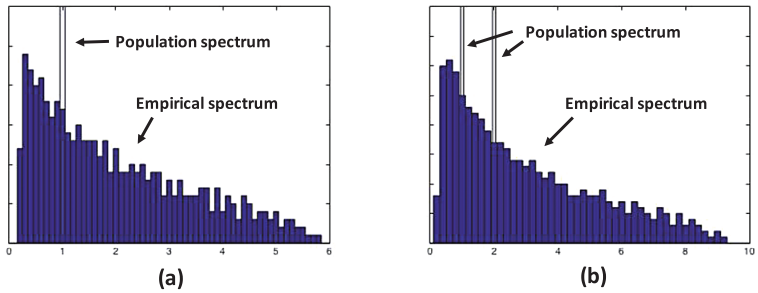

FIG. 1. *Plots of the nonzero eigenvalues of the empirical covariance, corresponding to* $n = 500$ *samples from*: (a) *the* $d = 1000$ *dimensional Gaussian with identity covariance, and* (b) *the* $d = 1000$ *dimensional "*2*-spike" Gaussian distribution whose covariance has* 500 *eigenvalues with value* 1, *and* 500 *eigenvalues of value* 2. *Note that the empirical spectra are poor approximations of the true population spectra. To what extent can the eigenvalues of the true distribution* (*the population spectrum*) *be accurately recovered from the samples, particularly in the "sublinear" data regime where* $n \ll d$*?*

will be significantly different from the population covariance of the distribution. Both the eigenvalues, and corresponding eigenvectors (principal components) of this empirical covariance matrix may be misleading. (See Figure 1 for an illustration of this fact.) The basic question we consider and answer affirmatively is: *In this linear or sublinear sample regime in which the eigenvalues of the empirical covariance are misleading, is it possible to recover accurate estimates of the eigenvalues of the underlying population covariance?*

This question of understanding the relationship between the empirical spectrum and the population spectrum of the underlying distribution has a long history of study, both from the perspective of characterizing the empirical spectrum, and with the goal of correcting for its biases. In the former vein, the seminal work of Anderson [2] considered the joint distribution of the empirical eigenvalues in the asymptotic regime as n tends toward infinity, for fixed dimension, d. The work of Marcenko and Pastur [29, 33] and more recent advances in random matrix theory have enabled analysis of the empirical spectrum, particularly in the asymptotic regime where the dimension and sample size scale linearly with each other (see, e.g., Bai and Silverstein's recent book [4]). We provide a more detailed discussion of the relationship between this characterization of the empirical spectrum and the problem of recovering the population spectrum in Section 1.2.

In the latter vein, works have considered both the end objective of recovering the population spectrum, as well as the objective of estimating the population covariance matrix. In their seminal work [17], James and Stein's proposed a shrinkage estimator for covariance estimation which uses the empirical eigenvectors, but "shrinks" the empirical eigenvalues to reduce the overall error due to the differences between the empirical and population spectra. Takemura [36] and Dey and Srinivasan [10] extended this work of James and Stein, obtaining orthogonally

invariant minimax covariance estimators under Stein's loss. There are many other approaches to eigenvalue shrinkage in this early line of work, for example, [12, 14, 15, 34, 35]. More recently, there has been a significant effort to develop optimal covariance estimators in the asymptotic regime where both n and d tend to infinity. This includes work of Ledoit and Wolf [21–23], Schäfer and Strimmer [32], and more recent work of Donoho et al. [11] who considered shrinkage estimators in the spiked covariance model (when all population eigenvalues take a constant number of different values).

While both the problems of covariance estimation and spectrum estimation face the common challenge that the empirical spectrum might differ significantly from the population spectrum, the problems are different. It is not clear whether optimal estimators for one of the problems can be leveraged to yield optimal or near-optimal estimators for the other problem. After formally defining the specific problem that we tackle—estimating the population spectrum in the linear and sublinear data regimes—in Section 1.2, we provide a technical discussion of the more modern related work on spectrum estimation, beginning with the seminal work of Karoui [13] and Burda [6, 7].

1.1. *Setup and definition.* We focus on the general setting where the multivariate distribution over $\mathbb{R}^d$ in question is defined by a real-valued distribution X, with zero mean, variance 1, and fourth moment β and a real $d \times d$ matrix $\mathbf{S}$. A sample of n vectors, viewed as a $n \times d$ data matrix $\mathbf{Y}$ consisting of n vectors drawn independently from the distribution corresponding to the pair $(X, \mathbf{S})$ is given by $\mathbf{Y} = \mathbf{XS}$ where $\mathbf{X} \in \mathbb{R}^{n \times d}$ has i.i.d. entries drawn according to distribution X. Note that this setting encompasses the case where the data is drawn from the uniform distribution over a d-dimensional unit cube, and the case of a multivariate Gaussian (corresponding to the distribution X being the standard Gaussian and the covariance of the corresponding multivariate Gaussian given by $\mathbf{S}^T\mathbf{S}$).

Throughout, we denote the corresponding *population covariance* matrix $\Sigma = \mathbf{S}^T\mathbf{S}$, and its eigenvalues by $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_d$, with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$. Our objective will be to recover an accurate approximation to this sorted vector of eigenvalues, $\boldsymbol{\lambda}$, given a data matrix $\mathbf{Y}$ as defined above. It is also convenient to regard the vector of eigenvalues as a distribution over $\mathbb{R}$, consisting of d equally-weighted point masses at locations $\lambda_1, \ldots, \lambda_d$; we refer to this distribution as the *population spectral distribution* D_Σ. We note that the task of learning the sorted vector $\boldsymbol{\lambda}$ in L_1 distance is closely related to the task of learning the spectral distribution in *Wasserstein* distance (i.e., "earthmover distance"): the L_1 distance between two sorted vectors of length d is exactly d times the Wasserstein-1 distance between the corresponding point-mass distributions. Similarly, given a distribution, Q, that is close to the true spectral distribution D_Σ in Wasserstein distance, the length d

vector whose ith element is given by the ith $(d+1)$-quantile[2] of Q will be close, in L_1 distance, to the sorted vector of population eigenvalues.

1.2. *Related work.* Before formally stating our main result of accurate spectrum estimation in the sublinear data regime, we discuss the context of our results and the connections to existing related work on spectrum estimation.

Population and sample spectra: *The Marcenko–Pastur law.* Given the setting described above, where we observe an $n \times d$ data matrix $\mathbf{Y} = \mathbf{XS}$ with population covariance given by $\mathbf{S}^T\mathbf{S}$, in the regime in which the dimensionality of the data, d, is linear in the number of samples, n, much is known about the mapping from the population spectral distribution D_Σ, to the empirical spectral distribution of the samples. Specifically, provided the ratio of the number of samples, n, to the dimensionality of the samples, d, is bounded below by some constant $\gamma > 0$, for sufficiently large n, d, the *expected* empirical spectral distribution will be well approximated by a deterministic function of the population spectral distribution. This deterministic function characterizing the correspondence between the empirical and population spectra is known as the Marcenko–Pastur law, which is defined in terms of the Stieltjes' transform (also referred to as the *Cauchy transform*) of the spectral distribution [29, 33]. At least in the linear regime in which n and d scale together, it is not hard to show that the empirical spectral distribution will be close to the *expected* empirical spectral distribution, and hence, asymptotically, the Marcenko–Pastur law will give an accurate characterization of the empirical spectral distribution. We refer the reader to Chapter 3 of Bai and Silverstein's book [4] for a thorough treatment of the Stieltjes' transform and Marcenko–Pastur law.

Inverting the Marcenko–Pastur law. Perhaps the most natural approach to recovering the population spectrum from the data matrix $\mathbf{Y}$ is to attempt to invert the mapping between population spectrum and expected empirical spectrum given by the Marcenko–Pastur law. The seminal work of Karoui [13] shows that this inversion can be represented via a linear program, and that, in the linear regime where $n/d \to \gamma \in (0, \infty)$, the reconstruction will be asymptotically consistent. This work also demonstrates the practical viability of this approach on a series of synthetic data, for the setting $d = 100$, $n = 500$. Building off the work of Karoui, Li et al. [25] considered applying this approach to a parametric model where the population spectral distribution has a constant (finite) support. They also suggested extending the Marcenko–Pastur law to the real line, allowing the optimization to be conducted over the reals, which makes the optimization procedure both easier to implement and more computationally efficient. Another approach to invert

[2]For $i = 1, \ldots, d$, the ith $(d+1)$-quantile of a distribution Q is defined to be the minimum value, x, with the property that the cumulative distribution function of Q at x is at least $i/(d+1)$.

the Marcenko–Pastur law directly, proposed by Ledoit and Wolf [23, 24], exploits the natural discreteness of the population spectrum in finite dimensions, and optimized over the Marcenko–Pastur law on the real line. Simulations demonstrated that this approach yields significant improvements in the accuracy of the recovered spectrum, versus the earlier approach of Karoui [13].

In a similar spirit, Mestre [30] considered the task of recovering the population spectrum in the setting where the population spectral distribution has a constant support (say of size r) and where the weights (but not the values) of each point mass are known a priori. Mestre proposed an algorithm for recovering the values of the support of the population spectrum via inverting the Marcenko–Pastur law, which is successful provided the empirical spectral distribution consists of r clusters of values, corresponding to the r point masses of the population spectrum. Provided sufficient separation between the point masses of the population spectrum, in the linear regime where n and d have constant ratio, the requirements of the algorithm are satisfied.

There seem to be two limitations to this general approach of "inverting" the Marcenko–Pastur law. The first is that the Marcenko–Pastur law, in general, is poorly equipped to deal with the sublinear-sample regime where $n \ll d$. In this sublinear sample-size regime, for example, with $n = d^{2/3}$, even if the population spectrum has a specific limiting distribution, the expected empirical spectral distribution may not converge. The second drawback is the difficulty of obtaining theoretical bounds on the accuracy of the recovered spectral distribution. This seems mainly due to the difficulty of analyzing the robustness of inverting the Marcenko–Pastur law. Specifically, given a dimension and sample size, it seems difficult to characterize the set of population spectral distributions that map, via the Marcenko–Pastur law, to a given neighborhood of a specific empirical spectral distribution. This analysis is further complicated in the sublinear sample regime by the potential lack of concentration of the empirical spectrum.

Method of moments. There have been several works that approach the spectrum recovery problem via the method of moments [3, 26, 31]. Rao et al. [31] observed the fact that the moments of the empirical spectral distribution have a limiting Gaussian distribution whose mean and variance are functions of the population spectrum. Given these moment distributions, they proposed a maximum likelihood approach to recover the parameters of the population spectrum in the setting where the spectrum consists of a constant number of point masses. In a similar fashion, Bai et al. [3] directly estimate the moments of the population spectrum from the empirical moments, via a system of polynomial equations that is derived from the Marcenko–Pastur law. In the linear sample-size setting, Bai et al. show that their recovery is consistent. We note that an immediate consequence of our accurate moment estimation (Theorem 1), together with the fact that a distribution supported on at most r values is robustly determined by its first $2r - 1$ moments

(see, e.g., [3]), yields the fact that for such spectral distributions, consistent estimation is possible in the sublinear data regime as long as $\frac{n}{d^{1-\frac{2}{2r-1}}} \to \infty$.

The recent work of Li and Yao [26] essentially interpolates between the approach of Mestre [30] and Bai et al. [3] to tackle the setting where the spectrum consists of a constant, r, number of point masses, but where the empirical spectrum cannot be partitioned into r corresponding clusters. For these "mixed" clusters, they employ the moment-based approach of Bai et al., and show consistency in the linear sample-size regime.

Finally, the work of Burda et al. [6] from the physics community employs a method of moment approach to recovering specific classes of population spectra, for example, the 3-spike case. This work is essentially a method-of-moments approach to inverting the Marcenko–Pastur law in specific cases, although this work seems to be unaware of the Marcenko–Pastur law and the related literature relating the empirical and population spectra.

Sketching bi-linear forms. In a recent work [27], Li et al. consider a seemingly unrelated problem, the problem of *sketching* matrix norms. Namely, suppose one wishes to approximate the kth moment of the spectrum of a $d \times d$ matrix, Σ, $\|\Sigma\|_k^k = \sum_{i=1}^d \lambda_i^k$, but rather than working directly with the matrix Σ, one only has access to a much smaller matrix that is a bilinear *sketch* of Σ. The question is how to design this sketch: for some $r, s \ll d$, can one design distributions $\mathcal{A}$ and $\mathcal{B}$ over $r \times d$ and $d \times s$ matrices, respectively, such that for any Σ, with high probability, given matrices $\mathbf{A}$ and $\mathbf{B}$ drawn respectively from $\mathcal{A}$ and $\mathcal{B}$, $\|\Sigma\|_k$ can be approximated based on the $r \times s$ matrix $\mathbf{A}\Sigma\mathbf{B}$? The authors consider setting $\mathcal{A}$ and $\mathcal{B}$ to have i.i.d. Gaussian entries, and show that such sketches are information theoretically optimal, to constant factors.

The connection between sketching matrix norms and recovering moments of the population covariance is that the matrix $\mathbf{Y}\mathbf{Y}^T = \mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{X}^T$ can be viewed as a bilinear sketch of the matrix $\mathbf{S}\mathbf{S}^T$. While $\mathbf{S}\mathbf{S}^T$ is not the population covariance matrix, it has the same eigenvalues (and hence same moments) as the population covariance.

The main difference between our work and the work of Li et al. is the conceptual difference in focus: we are focussed on recovering the population spectrum from limited data; they are focussed on defining small sketching matrices for matrix norms. The approach to moment recovery of [27] and our work both leverage a simple unbiased moment estimator (see Fact 1), though our techniques differ in two ways: first, [27] is concerned with establishing the minimum sketch size from an information theoretic perspective, and the proposed algorithm is not computationally efficient; second, from a technical perspective, the proof of correctness in [27] focusses on the Gaussian setting, and it seems difficult to extend their analysis techniques to the more general setting that we consider. In particular, to prove the variance bound of the moment estimator in the more general setting, we take

a rather different route and employ a variant of the approach of Yin and Krishnaiah [38].

Other works on spectrum reconstruction. There are also several other works on the population spectrum recovery problem for specific classes of population covariance. These include the paper of Bickel and Levina [5] who obtain accurate reconstruction in the sublinear-sample setting for the class of population covariance matrices whose off-diagonal entries decrease quickly with their distance to the diagonal (e.g., as in the class of Toeplitz matrices).

1.3. *Summary of approach and results*. Our approach to recovering the population spectral distribution from a given data matrix is via the method of moments, and is motivated by the observation (also leveraged in [27]) that there is a natural unbiased estimator for the kth moment of the population spectral distribution.

FACT 1. Fix a list of k distinct integers $\sigma = (\sigma_1, \ldots, \sigma_k)$ with $\sigma_i \in \{1, \ldots, n\}$. Let $\mathbf{Y} = \mathbf{XS}$ where $\mathbf{X} \in \mathbb{R}^{n \times d}$ consists of entries drawn i.i.d. from a distribution of zero mean and variance 1, and $\mathbf{S} \in \mathbf{R}^{d \times d}$. Letting $\mathbf{A} = \mathbf{YY}^T$, we have

$$\mathbf{E}\left[\prod_{i=1}^{k} \mathbf{A}_{\sigma_i, \sigma_{(i \bmod k)+1}}\right] = \sum_{i=1}^{d} \lambda_i^k,$$

where the expectation is over the randomness of the entries of $\mathbf{X}$, and λ_i is the ith eigenvalue of the population covariance matrix $\mathbf{S}^T\mathbf{S}$.

The above fact, whose simple proof is given in Section 2, suggests that a good algorithm for estimating the kth moment of the spectral distribution would be to compute the sum of the above quantity over *all* sets σ of distinct indices. The naive algorithm for computing such a sum would take time $O(n^k)$ to evaluate, and it seems unlikely that a significantly faster algorithm exists.[3] Fortunately, as we show, there is a simple algorithm that computes the sum over all *increasing* lists of k indices; additionally, such a sum results in an estimator with comparable variance to the computationally intractable estimator corresponding to the sum over all lists. This algorithm, together with a careful analysis of the variance of the corresponding estimator, yields the following theorem.

THEOREM 1 (Efficient moment estimation). *There is an algorithm that takes* $\mathbf{Y} = \mathbf{XS}$ *and an integer* $k \geq 1$ *as input, runs in time* $\mathrm{poly}(n, d, k)$ *and with probability at least* $1 - \delta$, *outputs an estimate of* $\frac{1}{d}\|\mathbf{S}^T\mathbf{S}\|_k^k$ (*i.e.*, $\frac{1}{d}\sum_i \lambda_i^k$) *with multiplicative*

[3]The ability to efficiently compute this sum would imply an efficient algorithm for counting the number of k-cycles in a graph, which is NP-hard, for general k [1].

error at most

$$\frac{f(k)}{\sqrt{\delta}} \max\left(\frac{d^{k/2-1}}{n^{k/2}}, \frac{d^{\frac{1}{4}-\frac{1}{2k}}}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right),$$

where the function $f(k) = 2^{6k}k^{3k}\beta^{k/2}$.

Restated slightly, the above theorem shows that the kth moment of the population spectrum can be accurately computed in the sublinear data regime, provided $n \geq c_k d^{1-\frac{2}{k}}$, for some constant c_k dependent on k. In the asymptotic regime as $d \to \infty$, this theorem implies that the multiplicative error of the estimate of the kth moment goes to zero provided $\frac{n}{d^{1-\frac{2}{k}}} \to \infty$. This moment recovery is useful in its own right, as these moments of the spectral distribution (also referred to as the *Schatten* matrix norms of the population covariance) provide insights into the population distribution (see, e.g., [16] and the survey [28]).

The recovery guarantees of Theorem 1 are optimal to constant factors: to accurately estimate the kth moment of the population spectrum to within a constant multiplicative error, the sample size n must scale at least as $d^{1-2/k}$, as is formalized by the following lower bound, which is a corollary to the lower bound in [27].

COROLLARY 1. *Fix a constant integer* k, *and suppose there exists an algorithm that, for any* $d \times d$ *matrix* $\mathbf{S}$, *when given an* $n \times d$ *data matrix* $\mathbf{Y} = \mathbf{XS}$ *with entries of* $\mathbf{X}$ *chosen i.i.d. as above, outputs an estimate* y *satisfying the following with probability at least* $3/4$: $0.9\|\mathbf{SS}^T\|_k^k \leq y \leq 1.1\|\mathbf{SS}^T\|_k^k$; *then* $n \geq cd^{1-2/k}$, *for an absolute constant* c *independent of* n, d *and* k.

Given accurate estimates of the low-order moments of the population spectral distribution, an accurate approximation of the list of population eigenvalues can be recovered by first solving the moment inverse problem—namely finding a distribution D whose moments are close to the recovered moments, for example, via linear programming—and then returning the vector of length d whose ith element is given by the ith $(d+1)$-quantile of distribution D. Altogether, this yields a practically viable polynomial-time algorithm with the following theoretical guarantees for recovering the population spectrum.

THEOREM 2 (Main theorem). *Consider an* $n \times d$ *data matrix* $\mathbf{Y} = \mathbf{XS}$, *where* $\mathbf{X} \in R^{n \times d}$ *has i.i.d. entries with mean* 0, *variance* 1, *and fourth moment* β *and* $\mathbf{S}$ *is a real* $d \times d$ *matrix s.t. the eigenvalues of the population covariance* $\Sigma = \mathbf{S}^T\mathbf{S}$, $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_d$, *are upper bounded by a constant* b. *There is an algorithm that takes* $\mathbf{Y}$ *as input and for any integer* $k \geq 1$ *runs in time* $\mathrm{poly}(n, d, k)$ *and outputs* $\hat{\boldsymbol{\lambda}} = \hat{\lambda}_1, \ldots, \hat{\lambda}_d$ *with expected* L_1 *error satisfying*

$$\mathbf{E}\left[\sum_{i=1}^{d} |\lambda_i - \hat{\lambda_i}|\right] \leq bd\left(f(k)\left(\frac{d^{k/2-1}}{n^{k/2}} + \frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}}\right) + \frac{C}{k} + \frac{1}{d}\right),$$

where C is an absolute constant and $f(k) = C'(6k)^{3k+1}\beta^{k/2}$ for an absolute constant C'.

This theorem implies that our population spectrum estimator is asymptotically consistent in terms of Wasserstein distance, even in the *sublinear* sample-size regime where $\frac{d}{n} \to \infty$:

COROLLARY 2 (Consistent sublinear sample-size estimation). *Fix a limiting spectral distribution p_∞ that is absolutely bounded by a constant, and a sequence of absolutely bounded population spectral distributions, $p_1, p_2, \ldots$ and corresponding population covariance matrices $\Sigma_1, \Sigma_2, \ldots$, such that p_d is the spectral distribution of Σ_d, and p_d converges weakly to p_∞ as $d \to \infty$. Given a sequence of data matrices, with the dth matrix $\mathbf{Y}_d = \mathbf{X}_d\mathbf{S}_d$ being $n_d \times d$ with $\mathbf{S}_d^T\mathbf{S}_d = \Sigma_d$ and entries of $\mathbf{X}_d$ chosen i.i.d. with zero mean, variance 1, and bounded fourth moment, then our algorithm outputs a distribution q_d on input $\mathbf{Y}_d$ such that q_d converges weakly to p_∞, provided $\frac{n_d}{d^{1-\varepsilon}} \to \infty$ for every positive constant ε. (For example, taking $n_d = \frac{d}{\log d}$ yields asymptotically consistent sublinear sample spectrum estimation.)*

The proof of Theorem 2 follows from combining Theorem 1 with the following proposition that bounds the Wasserstein distance between two distributions in terms of their discrepancies in low-order moments.

PROPOSITION 1. *Given two distributions with respective density functions p, q supported on $[-1, 1]$ whose first k moments are $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, respectively, the Wasserstein distance, $W_1(p, q)$, between p and q is bounded by*

$$W_1(p, q) \le \frac{C}{k} + g(k)\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2,$$

where C is an absolute constant, and $g(k) = C'3^k$ for an absolute constant C'.

The proof of the above proposition proceeds by leveraging the dual definition of Wasserstein distance: $W_1(p, q) = \sup_{f \in Lip} \int_{-\infty}^{\infty} f(x)(p(x) - q(x))\,dx$, where *Lip* denotes the set of all Lipschitz-1 functions. Our proof argues that for any Lipschitz function f, after convolving it with a special "bump" function, $\hat{b}$, which is a scaled Fourier transform of the bump function used in [18], the resulting function $f * \hat{b}$ has small high-order derivatives and is close to f in L_∞ norm. Given the small high-order derivates of $f * \hat{b}$, there exists a good degree-k polynomial interpolation of this function, P_k: the closeness of the first k moments of p and q implies a bound on the integral $\int P_k(x)(p(x) - q(x))\,dx$, from which we derive a bound on the original Wasserstein distance. Our approach to approximating a Lipschitz-1

function with a degree n polynomial can also be seen as a constructive proof of a special case of Jackson's theorem (see, e.g., Theorem 7.4 in [8]).

We also show, via a Chebyshev polynomial construction, that the inverse linear dependence of Proposition 1 between the number of moments k, and the Wasserstein distance between the distributions, is tight in the case that the moments exactly match.

PROPOSITION 2. *For any even k, there exits a pair of distributions p, q, each consisting of $k/2$ point masses, supported within the unit interval $[-1, 1]$, s.t. p and q have identical first $k-2$ moments, and Wasserstein distance $W_1(p, q) > 1/2k$.*

1.4. *Organization of paper.* In Section 2, we motivate and state our algorithms for accurately recovering the moments of the population spectrum, and prove Theorem 1. The most cumbersome component of this proof of correctness of our algorithm is the proof of a bound on the variance of our moment estimator; we defer this proof to the online supplementary material [20]. In Section 3, we state our algorithm for leveraging accurate moment reconstruction to recover the population spectrum, and describe the connection between the Wasserstein distance between spectral distributions, and L_1 distance between the vectors. In Section 4, we establish Propositions 1 and 2, which bound the Wasserstein distance between two distributions in terms of their discrepancies in low-order moments, completing our proof of Theorem 2. Section 5 contains some results illustrating the empirical performance of our approach.

2. Estimating the spectral moments. The core of our approach to recovering the moments of the population spectral distribution is a convenient *unbiased* estimator for these moments, first proposed in the recent word of Li et al. [27] on sketching matrix norms. This estimator is defined via the notion of a *cycle*.

DEFINITION 1. Given integers n and k, a *k-cycle* is a sequence of k distinct integers, $\sigma = (\sigma_1, \ldots, \sigma_k)$ with $\sigma_i \in [n]$. Given an $n \times n$ matrix A, each cycle, σ, defines a product:

$$\mathbf{A}_\sigma = \prod_{i=1}^{k} \mathbf{A}_{\sigma_i, \sigma_{i+1}},$$

with the convention that $\sigma_{k+1} = \sigma_1$, for ease of notation.

The following observation demonstrates the utility of the above definition.

FACT 2. For any k-cycle σ, a symmetric $d \times d$ real matrix $\mathbf{T}$, and a random $n \times d$ matrix X with i.i.d. entries with mean 0 and variance 1,

$$\mathbf{E}\left[\left(\mathbf{X}^T \mathbf{T} \mathbf{X}\right)_\sigma\right] = \operatorname{trace}\left(\mathbf{T}^k\right),$$

where the expectation is over the randomness of X.

PROOF. We can expand $\mathbf{E}[(\mathbf{X}^T\mathbf{T}\mathbf{X})_\sigma]$ as follows, where γ_{k+1} is shorthand for γ_1:

$$\sum_{\delta_1,\ldots,\delta_k,\gamma_1,\ldots,\gamma_k\in[d]} \mathbf{E}\left[\prod_{i=1}^{k} \mathbf{X}_{\delta_i,\sigma_i}\mathbf{T}_{\delta_i,\gamma_{i+1}}\mathbf{X}_{\gamma_{i+1},\sigma_{i+1}}\right] = \sum_{\delta_1,\ldots,\delta_k}\prod_{i=1}^{k}\mathbf{T}_{\delta_i,\delta_{i+1}} = \operatorname{trace}(\mathbf{T}^k).$$

The first equality holds since, for every term of the expression, the expectation of that term is zero unless each of the entries of $\mathbf{X}$ appears at least twice. Because the σ_i are distinct, in every nonzero term, each of the entries of $\mathbf{X}$ will appear exactly twice and $\delta_i = \gamma_i$. □

The above fact shows that each k-cycle yields an unbiased estimator for the kth spectral moment of T. While each estimator is unbiased, the variance will be extremely large. Perhaps the most natural approach to reducing this variance, would be to compute the average over *all* k-cycles. Unfortunately, such an estimator seems intractable, from a computational standpoint. The naive algorithm for computing this average—simply iterating over the $\binom{n}{k}$ different k-cycles—would take time $O(n^k)$ to evaluate. It seems unlikely that a significantly faster algorithm exists, assuming that $P \neq NP$, as an efficient algorithm to compute this average over k-cycles would imply an efficient algorithm for counting the number of simple k-cycles in a graph (i.e., loops of length k with no repetition of vertices), which is known to be NP-hard for general k (see, e.g., [1]).

One computationally tractable variant of this average over all k-cycles, would be to relax the condition that the k elements of each cycle be distinct. This quantity is simply the trace of the matrix $(\mathbf{X}^T\mathbf{T}\mathbf{X})^k$, which is trivial to compute! Unfortunately, this *exactly* corresponds to the kth moment of the empirical spectrum, which is a significantly biased approximation of the population spectral moment (e.g., as illustrated in Figure 1).

Our algorithm proceeds by computing the average of all *increasing* k-cycles.

DEFINITION 2. An *increasing* k-cycle $\sigma = (\sigma_1, \ldots, \sigma_k)$ is a k-cycle with the additional property that $\sigma_1 < \sigma_2 < \cdots < \sigma_k$.

We observe that, perhaps surprisingly, there is a simple and computationally tractable algorithm for computing the average over all increasing cycles. Given $\mathbf{Y} = \mathbf{X}\mathbf{S}$, instead of computing the trace of $(\mathbf{Y}^T\mathbf{Y})^k$, which would correspond to the empirical kth moment, we instead zero out the diagonal and lower-triangular entries of $\mathbf{Y}^T\mathbf{Y}$ in the "first" $k-1$ copies of $\mathbf{Y}^T\mathbf{Y}$ in the product $(\mathbf{Y}^T\mathbf{Y})^k$. It is not hard to see that this exactly corresponds to preserving the set of increasing cycles, as the contribution to a diagonal entry of the product corresponding to a nonincreasing cycle will include a lower-triangular entry of one of the terms, and hence will be zero (see Lemma 1 for a formal proof). This motivates Algorithm 1 for estimating the kth moment of the population spectrum.

Algorithm 1 [Estimating the kth moment]

Input: $Y \in R^{n \times d}$

Set, $A = YY^T$, and let $G = A_{up}$ be the matrix A with the diagonal and lower triangular entries set to zero.

Output: $\frac{\mathrm{tr}(G^{k-1}A)}{d \cdot \binom{n}{k}}$

Our main moment estimation theorem characterizes the performance of Algorithm 1.

THEOREM 1. *Given a data matrix* $\mathbf{Y} = \mathbf{XS}$ *where the entries of* X *are chosen i.i.d. with mean* 0, *variance* 1 *and fourth moment* β, *Algorithm* 1 *runs in time* $\mathrm{poly}(n, d, k)$ *and with probability at least* $1 - \delta$, *outputs an estimate of* $\frac{1}{d}\|\mathbf{S}^T\mathbf{S}\|_k^k$ (*i.e.*, $\frac{1}{d}\sum_i \lambda_i^k$) *with multiplicative error at most*

$$\frac{f(k)}{\sqrt{\delta}} \max\left(\frac{d^{k/2-1}}{n^{k/2}}, \frac{d^{\frac{1}{4}-\frac{1}{2k}}}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right),$$

where the function $f(k) = 2^{6k}k^{3k}\beta^{k/2}$.

The following restatement of the above theorem emphasizes the fact that accurate estimation of the population spectral moments is possible in the sublinear sample regime where $n = o(d)$.

COROLLARY 3. *Suppose* X *is a random* $n \times d$ *matrix whose entries are chosen i.i.d as described above. For any constant* $c > 1$, *there exists a function* $f_c(k)$ *such that, given* $n = f_c(k)d^{1-2/k}$, *for any* $d \times d$ *real matrix* S, *Algorithm* 1 *takes data matrix* $\mathbf{Y} = \mathbf{XS}$ *as input, runs in time* $\mathrm{poly}(d, k)$ *and with probability at least* 3/4 (*over the randomness of* $\mathbf{X}$), *outputs an estimate,* y, *of the* k*th population spectral moment that is a multiplicative approximation in the following sense*:

$$\left(1 - \frac{c^{2k}-1}{c^{2k}+1}\right)\|\mathbf{S}^T\mathbf{S}\|_k^k \le y \le \left(1 + \frac{c^{2k}-1}{c^{2k}+1}\right)\|\mathbf{S}^T\mathbf{S}\|_k^k.$$

As the above corollary shows, for any constant integer $k \ge 1$ there is a constant c_k such that taking $n = c_k d^{1-2/k}$ is sufficient to estimate the kth spectral moment accurately. For any constant k, this sublinear dependence between n and d is information theoretically optimal in the following sense (which is a stronger version of Corollary 1).

PROPOSITION 3. *Given the setting of Theorem* 1, *for any* $k > 1$, *suppose that an algorithm takes* $\mathbf{XS}$ *and with probability at least* 3/4 *computes* y *with* $(1 - \varepsilon)\|\mathbf{SS}^T\|_k^k \le y \le (1 + \varepsilon)\|\mathbf{SS}^T\|_k^k$ *for* $\varepsilon = (1.2^{2k} - 1)/(1.2^{2k} + 1)$. *Then* $\mathbf{X}$ *must be* $n \times d$ *for* $n \ge cd^{1-2/k}$ *for an absolute constant* c.

The above proposition follows as an immediate corollary from Theorem 3.2 of [27] by plugging in $S = X$, $T = I_d$, $A = S$ and $p = 2k$.

2.1. *Proof of Theorem* 1. The proof of this theorem follows from the following three components: Lemma 1 (below) shows that the efficient Algorithm 1 does in fact compute the average over all increasing k-cycles, $(\mathbf{Y}\mathbf{Y}^T)_\sigma$; Fact 2 guarantees that the average over such cycles is an unbiased estimator for the claimed quantity; and Proposition 4 bounds the variance of this estimator, which by Chebyshev's inequality, guarantees the claimed accuracy of Theorem 1. Our proof of this variance bound follows a similar approach as in [38].

The following lemma shows that Algorithm 1 computes the average over all increasing k-cycles, σ, of $(\mathbf{Y}\mathbf{Y}^T)_\sigma$; for an informal argument, see the discussion before the statement of Algorithm 1.

LEMMA 1. *Given an $n \times d$ data matrix $\mathbf{Y} = \mathbf{X}\mathbf{S}$, Algorithm* 1 *returns the average of $(\mathbf{Y}\mathbf{Y}^T)_\sigma$ taken over all increasing k-cycles, σ.*

PROOF. Let $A = \mathbf{Y}\mathbf{Y}^T = \mathbf{X}\mathbf{S}\mathbf{S}^T\mathbf{X}^T$; let $U_{i,j,m}$ denote the set of increasing m-cycles σ such that $\sigma_1 = i$ and $\sigma_m = j$, and define

$$F_{i,j,m} = \sum_{\sigma \in U_{i,j,m}} \prod_{\ell=1}^{m-1} A_{\sigma_\ell, \sigma_{\ell+1}}.$$

There is a simple recursive formula of $F_{i,j,m}$, given by

$$F_{i,j,m} = \sum_{\ell=i}^{j-1} F_{i,\ell,m-1} A_{\ell,j}. \tag{1}$$

Let G be the strictly upper triangular matrix of A, as in Algorithm 1, and let $F^{(m)}$ denote the matrix whose (i, j)th entry is $F_{i,j,m}$. The recursive formula 1 can be rewritten as $F^{(m)} = F^{(m-1)}G$. Given this, the sum over all increasing k-cycles is

$$\sum_{i,j} F_{i,j,k-1} A_{j,i} = \operatorname{tr}\big(F^{(k)}A\big) = \operatorname{tr}\big(G^{k-1}A\big),$$

as claimed. □

The main technical challenge in establishing the performance guarantees of our moment recovery, is bounding the variance of our (unbiased) estimator.

PROPOSITION 4. *Given the setup of Theorem* 1, *let U be the set of all increasing cycles of length k, then the following variance bound holds, where the function $f(k) = 2^{12k}k^{6k}\beta^k$:*

$$\mathbf{Var}\left[\frac{1}{|U|}\sum_{\sigma \in U}\left(\mathbf{X}^T\mathbf{T}\mathbf{X}\right)_\sigma\right] \le f(k)\max\left(\frac{d^{k-2}}{n^k}, \frac{d^{\frac{1}{2}-\frac{1}{k}}}{n}, \frac{1}{n}\right)\operatorname{tr}(\mathbf{T}^k)^2.$$

To see the high level approach to our proof of this proposition, consider the following: given lists of indices $\delta = (\delta_1, \ldots, \delta_k)$ and $\gamma = (\gamma_1, \ldots, \gamma_k)$, with $\delta_i, \gamma_i \in [d]$, we have the following equality:

$$(\mathbf{X}^T\mathbf{T}\mathbf{X})_\sigma = \sum_{\delta,\gamma\in[d]^k} \prod_{i=1}^{k} \mathbf{X}_{\sigma_i\delta_i}\mathbf{T}_{\delta_i,\gamma_{i+1}}\mathbf{X}_{\sigma_{i+1},\gamma_{i+1}}.$$

We now seek to bound each cross-term in the expansion of the total variance $\mathbf{Var}[\sum_{\sigma\in U}(\mathbf{X}^T\mathbf{T}\mathbf{X})_\sigma]$: namely, for a pair of increasing k-cycles, σ, π consider their contribution to the variance $\mathbf{E}[(\mathbf{X}^T\mathbf{T}\mathbf{X})_\sigma(\mathbf{X}^T\mathbf{T}\mathbf{X})_\pi]$ being

$$\sum_{\delta,\delta',\gamma,\gamma'\in[d]^k} \prod_{i=1}^{k} \mathbf{T}_{\delta_i,\gamma_{i+1}}\mathbf{T}_{\delta'_i,\gamma'_{i+1}} \prod_{i=1}^{k} \mathbf{X}_{\sigma_i,\delta_i}\mathbf{X}_{\sigma_i,\gamma_i}\mathbf{X}_{\pi_i,\delta'_i}\mathbf{X}_{\pi_i,\gamma'_i}.$$

We bound this sum by partitioning the set of summands, $\{(\delta, \delta', \gamma, \gamma')\}$ into classes. To motivate the role of these classes, consider the task of computing the expectation of the "$\mathbf{X}$" part of the expression, namely

$$\mathbf{E}\left[\prod_{i=1}^{k} \mathbf{X}_{\sigma_i,\delta_i}\mathbf{X}_{\sigma_i,\gamma_i}\mathbf{X}_{\pi_i,\delta'_i}\mathbf{X}_{\pi_i,\gamma'_i}\right],$$

for a given $\delta, \delta', \gamma, \gamma' \in [d]^k$. Thanks to the i.i.d. and zero mean properties of each entry $X_{i,j}$, most of terms are zero. The idea is to partition the set of summands that give rise to nonzero terms, via the creation of a list of constraints, $L = \{L_1, \ldots, L_m\}$, where each L_i contains only equalities and inequalities ("$\neq$") involving the indices of δ, δ', γ, γ'. For example, in the case $k = 2$, one such constraint could be $L_1 = \{\delta_1 = \delta'_1, \gamma_1 = \gamma'_1, \delta_2 = \gamma'_2, \gamma_2 = \delta'_2\}$, which specifies a subset of $\{(\delta, \delta', \gamma, \gamma')\}$ that satisfy each of the four specified equalities. We will design a set of these constraints, $L = \{L_1, \ldots, L_m\}$ satisfying the following useful properties:

1. Any lists of indices $\delta, \delta', \gamma, \gamma' \in [d]^k$ with the property that the expectation of the $\mathbf{X}$ "part" is zero, namely $\mathbf{E}[\prod_{i=1}^{k} \mathbf{X}_{\sigma_i,\delta_i}\mathbf{X}_{\sigma_i,\gamma_i}\mathbf{X}_{\pi_i,\delta'_i}\mathbf{X}_{\pi_i,\gamma'_i}] = 0$, does not satisfy any constraint $L_i \in L$.
2. Any lists of indices $\delta, \delta', \gamma, \gamma' \in [d]^k$ whose expectation of the $\mathbf{X}$ "part" is nonzero must satisfy exactly one of the constraint.
3. For any constraint $L_i \in L$, all lists of indices $\delta, \delta', \gamma, \gamma' \in [d]^k$ satisfying L_i have the same expected value of the $\mathbf{X}$ "part," namely

$$\mathbf{E}\left[\prod_{i=1}^{k} \mathbf{X}_{\sigma_i,\delta_i}\mathbf{X}_{\sigma_i,\gamma_i}\mathbf{X}_{\pi_i,\delta'_i}\mathbf{X}_{\pi_i,\gamma'_i}\right].$$

Given a set of constraints, L, satisfying the above, the set of summands $\{(\delta, \delta', \gamma, \gamma')\}$ corresponding to a constraint $L_i \in L$ have the same value of $\mathbf{E}[\prod_{i=1}^{k} \mathbf{X}_{\sigma_i,\delta_i}\mathbf{X}_{\sigma_i,\gamma_i}\mathbf{X}_{\pi_i,\delta'_i}\mathbf{X}_{\pi_i,\gamma'_i}]$, which will not be too difficult to bound. What remains is to deal with the $\mathbf{T}$ component of the expression. Our set of constraints

is also useful for this purpose. For example, consider the following sum over all $(\delta, \delta', \gamma, \gamma')$ that satisfy a constraint L_i:

$$\sum_{\delta,\delta',\gamma,\gamma' \text{ s.t. } L_i} \prod_{i=1}^{k} \mathbf{T}_{\delta_i,\gamma_{i+1}} \mathbf{T}_{\delta_i',\gamma_{i+1}'},$$

the equalities in L_i can be leveraged to simplify the calculation; revisiting the example above with $k = 2$, for instance, for the constraint $L_1 = \{\delta_1 = \delta_1', \gamma_1 = \gamma_1', \delta_2 = \gamma_2', \gamma_2 = \delta_2'\}$, the above expression simply becomes $\operatorname{tr}(T^4)$.

The full details of this partitioning scheme are rather involved, and are given in the online supplementary material [20].

3. From moments to spectrum. Given the accurate recovery of the moments of the population spectral distribution, as described in the previous section, we now describe the algorithm for recovering the population spectrum from these moments. We proceed via the natural approach to this moment inverse problem. The proposed algorithm has two parts: first, we recover a distribution whose moments closely match the estimated moments of the population spectrum (recovered via Algorithm 1); this recovery is performed via the standard linear programming approach. Given this recovered distribution, p^+, to obtain the vector of estimated population eigenvalues (the spectrum), one simply returns the length d vectors consisting of the $(d+1)$st-quantiles of distribution p^+—specifically, this is the vector whose ith component is the minimum value, x, with the property that the cumulative distribution function of p^+ at x is at least $i/(d+1)$. These two steps are formalized in Algorithm 2.

The following restatement of Theorem 2 quantifies the performance of Algorithm 2.

THEOREM 2. *Consider an $n \times d$ data matrix $\mathbf{Y} = \mathbf{XS}$, where $\mathbf{X} \in R^{n \times d}$ has i.i.d. entries with mean* 0, *variance* 1, *and fourth moment β, and $\mathbf{S}$ is a real $d \times d$ matrix s.t. the eigenvalues of the population covariance $\Sigma = \mathbf{S}^T\mathbf{S}$, $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_d$ are upper bounded by a constant $b \geq 1$. There is an algorithm that takes $\mathbf{Y}$ as input and for any integer $k \geq 1$ runs in time* $\operatorname{poly}(n, d, k)$ *and outputs $\hat{\boldsymbol{\lambda}} = \hat{\lambda}_1, \ldots, \hat{\lambda}_d$ with expected L_1 error satisfying*

$$\mathbf{E}\left[\sum_{i=1}^{d} |\lambda_i - \hat{\lambda}_i|\right] \leq bd\left(f(k)\left(\frac{d^{k/2-1}}{n^{k/2}} + \frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}}\right) + \frac{C}{k} + \frac{1}{d}\right),$$

where C is an absolute constant and $f(k) = C'(6k)^{3k+1}\beta^{k/2}$ for some absolute constant C'.

At the highest level, the proof of the above theorem has two main parts: the first part argues that if two distributions have similar first k moments, then the

Algorithm 2 [Moments to spectrum]

Input: Approximation to first k moments of population spectrum, $\hat{\boldsymbol{\alpha}}$, dimensionality d, and fine mesh of values $\mathbf{x} = x_0, \ldots, x_t$ that cover the range $[0, b]$ where b is an upper bound on the maximum population eigenvalue. Taking $x_i = i\varepsilon$ for $\varepsilon \leq 1/\max(d, n)$ is sufficient.

Output: Estimated population spectrum, $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$.

1. Let $\mathbf{p}^+$ be the solution to the following linear program, which we will regard as a distribution consisting of point masses at values $\mathbf{x}$:

$$\begin{aligned} &\underset{\mathbf{p}}{\text{minimize}} \ |\mathbf{V}\mathbf{p} - \hat{\boldsymbol{\alpha}}|_1 \\ &\text{subject to } \mathbf{1}^T\mathbf{p} = 1 \\ &\mathbf{p} > 0, \end{aligned} \tag{2}$$

   where the matrix $\mathbf{V}$ is defined to have entries $\mathbf{V}_{i,j} = x_j^i$.

2. Return the vector $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ where $\hat{\lambda}_i$ is the ith $(d+1)$st-quantile of distribution corresponding to $\mathbf{p}^+$, namely set $\hat{\lambda}_i = \min\{x_j : \sum_{\ell \leq j} p_\ell^+ \geq \frac{i}{d+1}\}$.

two distributions are "close" (in a sense that we will formalize soon). As applied to our setting, the guarantees of Algorithm 1 ensures that, with high probability, the distribution returned by Algorithm 2 will have similar first k moments to the true population spectral distribution, and hence these two distributions are "close." The second and straightforward part of the proof will then argue that if two distributions, p and p', are "close," and distribution p, consists of d equally weighted point masses (such as the true population spectral distribution), then the vectors given by the $(d+1)$-quantiles of distribution p' will be close, in L_1 distance, to the vector consisting of the locations of the point masses of distribution p. As we show, the right notion of "closeness" of distributions to formalize the above proof approach is the *Wasserstein* distance, also known as "earthmover" distance.

DEFINITION 3. Given two real-valued distributions p, q, with respective density functions $p(x)$, $q(x)$, the *Wasserstein* distance between them, denoted by $W_1(p, q)$ is defined to be the cost of the minimum cost scheme of moving the probability mass in distribution p to make distribution q, where the per-unit-mass cost of moving probability mass from value a to value b is $|a - b|$.

One can also define Wasserstein distance via a dual formulation (given by the Kantorovich–Rubinstein theorem [19] which yields exactly what one would expect from linear programming duality):

$$W_1(p, q) = \sup_{f : Lip(f) \leq 1} \int f(x) \cdot \big(p(x) - q(x)\big)\, dx,$$

where the supremum is taken over all functions with Lipschitz constant 1.

Two convenient properties of Wasserstein distance are summarized in the following easily verified facts. The first states that in the case of distributions consisting of d equally-weighted point masses, the Wasserstein distance *exactly* equals the L_1 distance between the sorted vectors of the locations of the point masses. The second fact states that given any distribution p, supported on a subset of the interval $[a, b]$, the distribution p' defined to place weight $1/d$ at each of the d $(d+1)$st-quantiles of distribution p, will satisfy $W_1(p, p') \leq \frac{b-a}{d}$. For our purposes, these two facts establish that, provided the distribution $\mathbf{p}^+$ returned by the linear programming portion of Algorithm 2 is close to the true population spectral distance, in Wasserstein distance, then the final step of the algorithm—the rounding of the distribution to the point masses at the quantiles—will yield a close L_1 approximation to the vector of the population spectrum.

FACT 3. Given two vectors $\vec{a} = (a_1, \ldots, a_d)$, and $\vec{b} = (b_1, \ldots, b_d)$ that have been sorted, that is, for all i, $a_i \leq a_{i+1}$ and $b_i \leq b_{i+1}$,

$$|\vec{a} - \vec{b}|_1 = d \cdot W_1(p_{\vec{a}}, p_{\vec{b}}),$$

where $p_{\vec{a}}$ denotes the distribution that puts probability mass $1/d$ on each value a_i, and $p_{\vec{b}}$ is defined analogously.

FACT 4. Given a distribution p supported on $[a, b]$, let distribution p' be defined to have probability mass $1/d$ at each of the d $(d+1)$st-quantiles of distribution p. Then $W_1(p, p') \leq \frac{b-a}{d}$.

The remaining component of our proof of Theorem 2 is to establish that the accurate moment recovery of Algorithm 1 as guaranteed by Theorem 1 is sufficient to guarantee that, with high probability, the distribution $\mathbf{p}^+$ returned by the linear program in the first step of Algorithm 2 is close, in Wasserstein distance, to the population spectral distribution. We establish this general robust connection between accurate moment estimation, and accurate distribution recovery in Wasserstein distance, via the following proposition, which we prove in Section 4.

PROPOSITION 1. *Given two distributions with respective density functions p, q supported on $[-1, 1]$ whose first k moments are $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, respectively, the Wasserstein distance, $W_1(p, q)$, between p and q is bounded by*

$$W_1(p, q) \leq \frac{C}{k} + g(k)\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2,$$

where C is an absolute constant, and $g(k) = C' 3^k$ for an absolute constant C'.

We conclude this section by assembling the pieces—the accurate moment estimation of Theorem 1, and the guarantees of the above proposition and Facts 3 and 4—to prove Theorem 2.

PROOF OF THEOREM 2. Given the data matrix Y and an upperbound on the population eigenvalues, b, we first divide each sample by $\sqrt{b}$ thereby reducing the problem to the setting where the eigenvalues are bounded by 1. We then run Algorithm 1 on the scaled data matrix to recover the (scaled) moments. Given these recovered moments, we then run Algorithm 2 to recover the scaled spectrum, and then return this recovered spectrum scaled by the factor of b. We now prove the correctness of this algorithm.

Let $\boldsymbol{\lambda}$ denote the vector of population eigenvalues, and let $\mathbf{p}$ denote the scaled spectral distribution obtained by dividing each entry of $\boldsymbol{\lambda}$ by b. Hence $\mathbf{p}$ is supported on $[0, 1]$. Denote the vector of the first k moments of this distribution as $\boldsymbol{\alpha}$. Consider the distribution $\mathbf{p}^+$ recovered by the linear programming step of Algorithm 2, and denote its first k moments with the vector $\boldsymbol{\alpha}^+$. We first argue that $\|\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}\|_2$ is small, and then will apply the Wasserstein distance bound of Proposition 1.

By Proposition 4 and the inequality $\mathbf{E}[X]^2 \le \mathbf{E}[X^2]$, the estimated moment vector $\hat{\boldsymbol{\alpha}}$, given as input to Algorithm 2, satisfies $\mathbf{E}\left[\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_1\right] \le \sum_{i=1}^{k} f(i) \times \max\left(\frac{d^{i/2-1}}{n^{i/2}}, \frac{d^{\frac{1}{4}-\frac{1}{2i}}}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$ where $f(k) = 2^{6k}k^{3k}\beta^{k/2}$. We now argue that there is a distribution that is a feasible point for the linear program that also has accurate moments. Specifically, consider taking the mesh $\mathbf{x} = x_1, \ldots, x_t$ of the linear program grid points to be an ε-mesh with $\varepsilon \le \frac{1}{\max(n,d)}$, and consider the feasible point of the linear program that corresponds to the true population spectral distribution whose support has been rounded to the nearest multiple of ε. This rounding changes the ith moment by at most $1 - (1-\varepsilon)^i$.

Hence, by the triangle inequality, this rounded population spectral distribution is a feasible point of the linear program, $\mathbf{p}^*$, with objective value at most $\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_1 + \sum_{i=1}^{k}(1-(1-\varepsilon)^i)$. Hence, also by the triangle inequality, the moments $\boldsymbol{\alpha}^+$ of the distribution $\mathbf{p}^+$ returned by the linear program will satisfy

$$\begin{aligned}
\mathbf{E}\left[\left\|\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}\right\|_2\right] &\le \mathbf{E}\left[\left\|\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}\right\|_1\right] \\
&\le 2\sum_{i=1}^{k}\left(f(i)\max\left(\frac{d^{i/2-1}}{n^{i/2}}, \frac{d^{\frac{1}{4}-\frac{1}{2i}}}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right) + 1 - (1-\varepsilon)^i\right) \\
&\le 2k\left(f(k)\left(\frac{d^{k/2-1}}{n^{k/2}} + \frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}}\right) + k\varepsilon\right).
\end{aligned}$$

Letting $\mathbf{p}^+_{\text{quant}}$ denote the distribution $\mathbf{p}^+$ that has been quantized so as to consist of d equally weighted point masses (according to the second step of Algorithm 2), by Fact 4 and Proposition 1, we have the following:

$$\begin{aligned}
W_1\left(\mathbf{p}^+_{\text{quant}}, \mathbf{p}\right) &\le W_1\left(\mathbf{p}^+_{\text{quant}}, \mathbf{p}^+\right) + W_1\left(\mathbf{p}^+, \mathbf{p}\right) \\
&\le \frac{1}{d} + \left(\frac{C}{k} + g(k)\left\|\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}\right\|_2\right).
\end{aligned}$$

Plugging in $\varepsilon \le 1/\max(n,d)$ and our bound on the moment discrepancy $\|\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}\|_2$ we get

$$W_1(\mathbf{p}^+_{\text{quant}}, \mathbf{p}) \le \frac{1}{d} + \frac{C}{k} + \widetilde{f}(k)\left(\frac{d^{k/2-1}}{n^{k/2}} + \frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}}\right),$$

where $\widetilde{f}(k) = C'(6k)^{3k+1}\beta^{k/2}$. Let $\hat{\boldsymbol{\lambda}}$ be the vector corresponds to distribution $\mathbf{p}^+_{\text{quant}}$ after multiplication by b. By Fact 3 we have

$$\mathbf{E}\left[\sum_{i=1}^{d} |\lambda_i - \hat{\lambda}_i|\right] \le bd\left(\widetilde{f}(k)\left(\frac{d^{k/2-1}}{n^{k/2}} + \frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}}\right) + \frac{C}{k} + \frac{1}{d}\right). \qquad \square$$

To yield Corollary 2 from Theorem 2, it suffices to show that, under the assumptions of the corollary, in the limit as $d \to \infty$, the number of moments that we can accurately estimate with our sublinear sample size, n_d, also goes to infinity (as $d \to \infty$). By assumption, $\frac{n_d}{d^{1-\varepsilon}} \to \infty$ for every constant $\varepsilon > 0$, and hence there is some function $\alpha(d)$ such that $\frac{n_d}{d^{1-\alpha(d)}} \to \infty$ with $\alpha(d) \to 0$; additionally, we may assume that $\alpha(d) \ge \frac{1}{\log\log d}$. By setting $k_d = \lfloor \frac{1}{\alpha(d)} \rfloor$, from Theorem 2, we examine the expected Wasserstein error of our reconstruction term by term. The first term satisfies

$$\begin{aligned}
\widetilde{f}(k_d)\frac{d^{k_d/2-1}}{n^{k_d 2}} &\le \big((ck_d)^{3k_d+1}\big)\frac{d^{k_d/2-1}}{n_d^{k_d/2}} \\
&\le \big((ck_d)^{3k_d+1}\big)\frac{d^{\frac{k_d}{2}-1}}{d^{\frac{k_d}{2}-\frac{k_d\alpha(d)}{2}}} \\
&\le \big((ck_d)^{3k_d+1}\big)d^{-1/2} \\
&\le \frac{(c\log\log d)^{1+3\log\log d}}{\sqrt{d}},
\end{aligned}$$

which tends to 0 as $d \to \infty$. The second term satisfies

$$\begin{aligned}
\widetilde{f}(k)\frac{1 + d^{\frac{1}{4}-\frac{1}{2k}}}{n^{1/2}} &\le (ck_d)^{3k_d+1}\frac{1 + d^{\frac{1}{4}-\frac{1}{2k_d}}}{d^{\frac{1}{2}-\frac{\alpha(d)}{2}}} \\
&\le (ck_d)^{3k_d+1}\frac{1}{d^{1/4}}
\end{aligned}$$

which tends to 0 as $d \to \infty$. C/k_d and $1/d$ also go to 0 as $d \to \infty$. Combining these four terms establishes Corollary 2.

4. Moments and Wasserstein distance. In this section, we prove Proposition 1, which establishes a general robust relationship between the disparity between the low-order moments of two univariate distributions, and the Wasserstein

distance (see Definition 3) between the distributions. This relatively straightforward proof proceeds via a constructive version of Jackson's theorem (see, e.g., Theorem 7.4 in [8]) which shows that Lipschitz functions can be well approximated by polynomials. For convenience, we restate Proposition 1, and the lower bound establishing its tightness.

PROPOSITION 1. *Given two distribution with respective density functions* p, q *supported on* $[-1, 1]$ *whose first* k *moments are* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ *and* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, *respectively, the Wasserstein distance,* $W_1(p, q)$, *between* p *and* q *is bounded by*

$$W_1(p, q) \leq \frac{C}{k} + g(k)\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2,$$

where C *is an absolute constant, and* $g(k) = C'3^k$ *for an absolute constant* C'.

The following lower bound shows that the inverse linear dependence in the above bound on the number of matching moments, k, is tight in the case where the moments exactly match.

PROPOSITION 2. *For any even* k, *there exits a pair of distributions* p, q, *each consisting of* $k/2$ *point masses, supported within the unit interval* $[-1, 1]$, *s.t.* p *and* q *have identical first* $k - 2$ *moments, and Wasserstein distance* $W_1(p, q) > 1/2k$.

4.1. *Proof of Proposition* 1. For clarity, we give an intuitive overview of the proof of Proposition 1 in the case where the first k moments of the two distributions in question match exactly. Consider a pair of distributions, p and q, whose first k moments match. Because p and q have the same first k moments, for *any* polynomial P of degree at most k, the inner product between P and $p - q$ is zero: $\int P(x)(p(x) - q(x))\,dx = 0$. The natural approach to bounding the Wasserstein distance, $\sup_{f \in Lip} \int f(x)(p(x) - q(x))\,dx$, is to argue that for any Lipschitz function, f, there is a polynomial P_f of degree at most k that closely approximates f. Indeed,

$$\begin{aligned}
&\int f(x)\big(p(x) - q(x)\big)\,dx \\
&\quad \leq \int |P_f(x) - f(x)|\big(p(x) - q(x)\big)\,dx + \int P_f(x)\big(p(x) - q(x)\big)\,dx \\
&\quad \leq 2\|f - P_f\|_\infty.
\end{aligned}$$

Hence, all that remains is to argue that there is a good degree k polynomial approximation of any Lipschitz function f. As the following standard fact shows, the approximation error of f by a degree-k polynomial is typically determined by the $k + 1$th order derivative of f.

FACT 5 (Polynomial interpolation; e.g., Theorem 2.2.4 of [9]). For a given function $g \in C^{k+1}[a,b]$, there exists a degree k polynomial P_g such that

$$\|g(x) - P_g(x)\|_\infty \leq \left(\frac{b-a}{2}\right)^{k+1} \frac{\max_{x\in[a,b]} |g^{(k+1)}(x)|}{2^k(k+1)!}.$$

While our function f is Lipschitz, its higher derivatives do not necessarily exist, or might be extremely large. Hence, before applying the above interpolation fact, we define a "smooth" version of f, which we denote f_s. This smooth function will have the property that $\|f - f_s\|_\infty$ is small, and that the derivatives of f_s are small. We will accomplish this by defining f_s to be the convolution of f with a special "bump" function $\hat{b}$ that we will define shortly. To motivate our choice of $\hat{b}$, consider the convolution of f with an arbitrary function, h: $f_s = f * h$. From the definition of convolution, the derivates of f_s satisfy the following property:

$$(f_s)^{(k+1)}(x) = \big(f * h^{(k+1)}\big)(x).$$

Hence, we can bound the derivatives of f_s by choosing h with small derivatives. Additionally, since we require that f_s is close to f, we also want h to be concentrated around 0 so the convolution will not change f too much in infinity norm.

We define f_s to be the convolution of function f with a scaled version of a special "bump" function $\hat{b}$ defined as the Fourier transform of the function $b(y)$ defined as

$$b(y) = \begin{cases} \exp\left(-\dfrac{y^2}{1-y^2}\right) & |y| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

This function was leveraged in a recent paper by Kane et al. [18], to smooth the indicator function while maintaining small higher derivates. As they show, the derivates of $\hat{b}$ are extremely well behaved: $\|\hat{b}^{(k)}\|_1 = O(\frac{1}{k})$ and $\|\hat{b}^{(k)}\|_\infty = O(1)$. The actual function that we convolve f with to obtain f_s will be a scaled version of this bump function $\hat{b}_c = c \cdot \hat{b}(cx)$ for an appropriate choice of c.

We note that if, instead of convolving by $\hat{b}_c$, we had convolved by a scaled Gaussian [or a scaled version of the function $b(x)$ rather its Fourier transform] the $O(1/k)$ dependence on the Wasserstein distance that we show in Proposition 1 would, instead, be $O(1/\sqrt{k})$.[4]

We now give the proof of Proposition 1 in the special case where the first k moments of p and q match exactly. The proof of the robust version is given in the online supplementary material [20], and is similar, though requires bounds on

[4]In the Gaussian case, this is because the kth derivative of a standard Gaussian, $G(x)$ is given by $G^{(k)}(x) = (-1)^k H_k(x) e^{-x^2}$ where $H_k(x)$ is the kth Hermite polynomial, and for even k the value of $H_k(0)$ is $\frac{k!}{(k/2)!}$ which is already too large to obtain better than an $O(\frac{1}{\sqrt{k}})$ dependence.

the coefficients of the interpolation polynomial P that approximates the smoothed function f_s.

PROOF OF "NONROBUST" PROPOSITION 1. Consider distributions p, q supported on the interval $[-1, 1]$ whose first k moments match. Given a Lipschitz function f, let $f_s = f * \hat{b}_c(x)$ where the scaled bump function $\hat{b}_c$ is as defined above, for a choice of c to be determined at the end of the proof. Letting P denote the degree k polynomial approximation of f_s we have the following:

$$\int f(x)\big(p(x) - q(x)\big)\,dx \le 2\|f - P\|_\infty \le 2\|f - f_s\|_\infty + 2\|f_s - P\|_\infty.$$

We bound each of these two terms. For the first term, $\|f - f_s\|_\infty$, we have that for any x:

$$\begin{aligned}
|f(x) - f_s(x)| &= \left| f(x) - \int_{-\infty}^{\infty} f(x-t)\hat{b}_c(t)\,dt \right| \\
&= \left| f(x)\left(1 - \int_{-\infty}^{\infty} \hat{b}_c(t)\,dt\right) + \int_{-\infty}^{\infty} \big(f(x) - f(x-t)\big)\hat{b}_c(t)\,dt \right| \\
&\le \int_{-\infty}^{\infty} |\hat{b}_c(t)t|\,dt.
\end{aligned}$$

Note that the last inequality holds since $\int \hat{b}_c(t)\,dt = b(0) = 1$ and f has Lipschitz constant at most 1, by assumption. To bound the above quantity, applying Lemma A.2 from [18] with $l = 0$, $n = 1$ yields $|\hat{b}_c(t)t| = O(1)$, with $l = 0$, $n = 3$ yields $|\hat{b}_c(t)t| = O(c^{-2}t^{-2})$. Splitting the integral into two parts, we have

$$\int_{-\infty}^{\infty} |\hat{b}_c(t)t|\,dt \le 2\left(\int_0^{1/c} |\hat{b}_c(t)t|\,dt + \int_{1/c}^{\infty} |\hat{b}_c(t)t|\,dt\right) = O\left(\frac{1}{c}\right).$$

We now bound the second term (the polynomial approximation error term) $\|f_s - P\|_\infty$, and then will specify the choice of c. From Fact 5, this term is controlled by the $(k+1)$st derivative of f_s:

$$\begin{aligned}
\big|(f_s)^{(k+1)}\big|_\infty &= c^{k+1}\big|\big(f * (\hat{b}^{(k+1)})_c\big)(x)\big|_\infty \\
&\le c^{k+1}|f|_\infty\big|\hat{b}^{(k+1)}\big|_1 \\
&= O\big(c^{k+1}\big),
\end{aligned}$$

where the inequality holds by the definition of convolution and the last equality applies Lemma A.3 from [18]. Hence, we have the following bound on the polynomial approximation error term:

$$\begin{aligned}
\|f_s - P\|_\infty &\le \frac{\max_{x\in[-1,1]} |f_s^{(k+1)}(x)|}{2^k(k+1)!} \\
&= O\left(\frac{c^{k+1}}{2^k(k+1)!}\right).
\end{aligned}$$

Setting $c = \Theta(k)$ balances the contribution from the two terms, $\|f - f_s\|_\infty$ and $\|f_s - P\|_\infty$, yielding the proposition in the nonrobust case. □

4.2. *Proof of Proposition* 2: *Wasserstein lower bound.* We now prove Proposition 2, showing that the $O(1/k)$ dependence of Proposition 1 is optimal up to constant factors, by constructing a sequence of distribution pairs p_k, q_k with the same first k moments but that have $O(\frac{1}{k})$ Wasserstein distance between them. The proof follows from leveraging a Chebyshev polynomial construction via the following general lemma.

LEMMA 2. *Given a polynomial P of degree j with j real roots $\{x_1, \ldots, x_j\}$, then letting P' denote the derivative of P, then for all $\ell \leq j-2$, $\sum_{i=1}^{j} x_i^\ell \cdot \frac{1}{P'(x_i)} = 0$.*

See Fact 14 in [37] for the very short proof of the above lemma.

Lemma 2 provides a very natural construction for a pair of distributions whose low-order moments match: simply begin with any polynomial P of degree k with k distinct (real) roots $x_1, \ldots, x_k$, and define the signed measure m, supported at the roots of P, with $m(x_i) = \frac{1}{P'(x_i)}$. Define distribution p^+ to be the positive portion of m, normalized so as to be a distribution and define p^- to be the negative portion of m scaled so as to be a distribution. Note that provided $k \geq 2$, the scaling factor for p^+ and p^- will be identical, as Lemma 2 guarantees that $\sum_{i=1}^{j} \frac{1}{P'(x_i)} = 0$, and hence the first $k-2$ moments of p^+ and p^- will agree.

Proposition 2 will follow from setting the polynomial P of the above construction to be the kth Chebyshev polynomial (of the first kind) T_k. We require the following properties of the Chebyshev polynomials, which can be easily verified by leveraging the trigonometric definition of the Chebyshev polynomials: $T_k(\cos(t)) = \cos(kt)$, and the fact that the derivative satisfies $T_k'(x) = k \cdot U_{k-1}(x)$ where U_j is the jth Chebyshev polynomial of the second kind, satisfying $U_j(\cos(t)) = \frac{\sin((j+1)t)}{\sin t}$.

FACT 6. Let $x_1, \ldots, x_k$ denote the roots of T_k, with $x_i = -\cos(\frac{(1+2(i-1))\pi}{2k})$, and set $y_i = 1/T_k'(x_i) = \frac{1}{kU_{k-1}(x_i)}$:

1. For $i \leq n/2$, $\frac{i}{k^2} \leq |y_i| \leq i\frac{\pi}{k^2}$.
2. For $i \leq n/2$, $\frac{5i}{k^2} \leq |x_{i+1} - x_i| \leq \frac{10i}{k^2}$.
3. $\sum_{i=1}^{n/2} y_{2i-1} = -\sum_{i=1}^{n/2} y_{2i} \in [\frac{1}{4}, \frac{1}{2}]$. (Hence the scaling factor required to make the distributions from the signed measure is at least 2.)

We now put the pieces together to complete the proof of Proposition 2.

PROOF OF PROPOSITION 2. By construction, and Lemma 2, letting p^+ denote the distribution corresponding to the positive portion of the signed measure corresponding to T_k, and p^- corresponding to the negative portion of the signed measure, we have that p^+ and p^- each consist of $k/2$ point masses, located at values in the interval $[-1, 1]$, and the first $k-2$ moments of p^+ and p^- are identical.

To lower bound the Wasserstein distance between p^+ and p^-, note that all the mass in p^+ must be moved to the support of p^-. Hence, the distance is lower bounded by the sum

$$\sum_{i=1}^{k/4} 2y_{2i}|x_{2i} - x_{2i-1}| \geq \sum_{i=1}^{k/4} 2\frac{2i}{k^2} \cdot \frac{10i}{k^2} = \frac{40}{k^4}\sum_{i=1}^{k/4} i^2 \geq \frac{40}{64k}.$$

□

5. Empirical performance. We evaluated the performance of our population spectrum recovery algorithm on a variety of synthetic distributions, for a range of dimensions and sample sizes. Recall that our algorithm consists of first applying Algorithm 1 to estimate the first k moments of the population spectral distribution, and then applying Algorithm 2 to recover a distribution whose moments closely match the estimated moments. Our matlab implementation is available from our websites.

5.1. *Implementation discussion.* Our estimates of higher-order spectral moments have larger variance than our estimates of the lower-order spectral moments, hence when solving the moment inverse problem, we should be more forgiving of discrepancies in higher moments. For example, we should require that the distribution we return match the estimated 1st and 2nd moments extremely accurately, while tolerating larger discrepancies between the 5th moment of the distribution that we return and the estimated 5th population spectral moment. We implemented this intuition as follows: in the linear program of Algorithm 2 that reconstructs a distribution from the moment estimates, the objective function of the linear program weighs the discrepancy between the ith moment of the returned distribution and the ith estimated moment by a coefficient $1/(c_i\hat{\alpha}_i)$, where $\hat{\alpha}_i$ is the ith estimated moment, and c_i is a scaling factor designed to capture the (multiplicative) standard deviation of the estimate. In our experiments, we set c_i to correspond to our bound on the standard deviation of the error in the ith recovered moment, implied by Theorem 1. This corresponds to setting

$$c_i = (2i)^{2i} \cdot \frac{\max(d^{i/2-1}, 1)}{n^{i/2}}.$$

This scaling is theoretically justified, and we made no effort to optimize it: it seems likely that the empirical performance can be improved with a more careful weighting function.

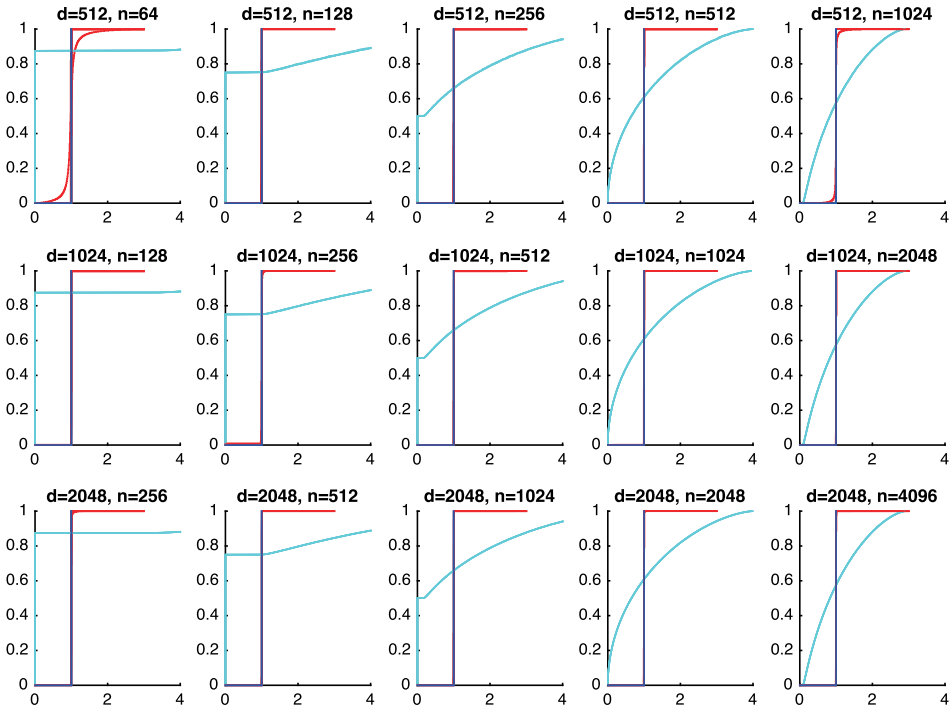

FIG. 2. *Empirical results for reconstructing the population spectrum for covariance* $\Sigma = I_d$ *for a range of sample sizes and dimensions. Red lines depict the cdf of the distribution recovered by our algorithm over five independent trials, the blue line depicts the cdf of the true population spectral distribution, and the cyan line depicts the cdf of the empirical spectral distribution* (*in one of the trials*).

In all runs of our algorithm, we estimated and matched the first 7 spectral moments (i.e., we set the parameter $k = 7$ in Algorithm 2). Considering higher moments beyond the 7th did not significantly improve the results for the dimension and sample sizes that we considered.

We would expect that the empirical performance could be improved by adaptively setting the number of moments to consider, based on the values of the lower order moments. Specifically, it would be natural to only consider higher moments if the lower order moments fail to robustly characterize the distribution. More generally, a variety of other approaches to the general moment inverse problem could be substituted in place of Algorithm 2, and might improve the empirical performance, though such directions are beyond the focus of this work.

5.2. *Experimental setup and results*. We evaluated our algorithm on four different types of population spectral distributions:

1. Identity covariance: $\Sigma_d = \mathbf{I}_d$. (Figure 2.)

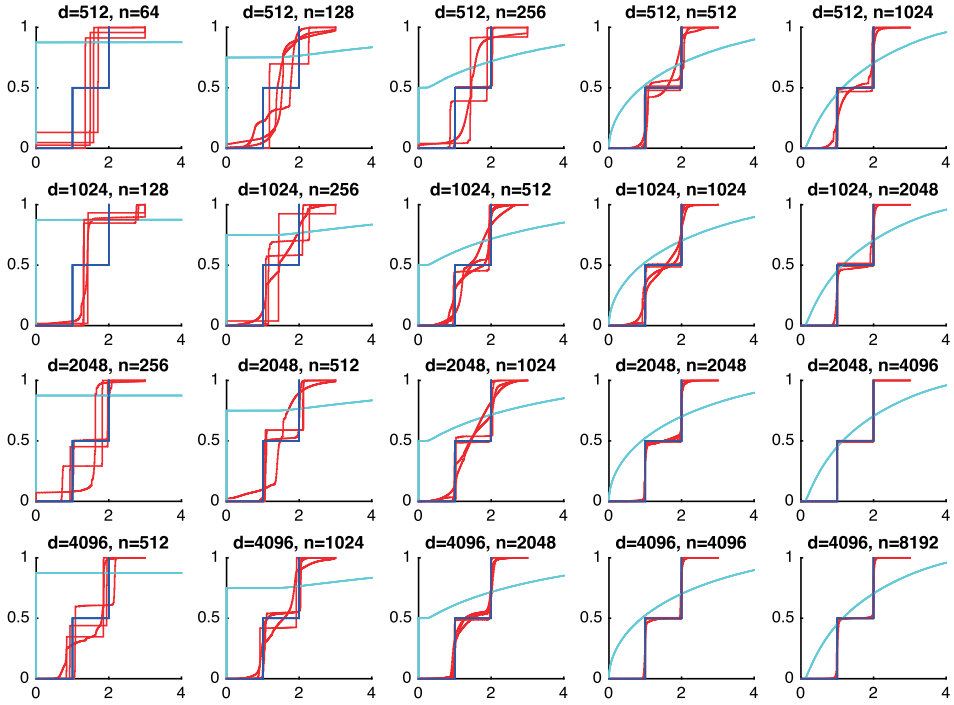

FIG. 3. *Empirical results for reconstructing the population spectrum for covariance matrices that have $d/2$ eigenvalues equal to* 1, *and $d/2$ eigenvalues equal to* 2. *Red lines depict the cdf of the distribution recovered by our algorithm over five independent trials, the blue line depicts the cdf of the true population spectral distribution, and the cyan line depicts the cdf of the empirical spectral distribution* (*in one of the trials*).

2. "Two spike" spectrum: Σ_d has $d/2$ eigenvalues equal to 1 and $d/2$ eigenvalues equal to 2. (Figure 3.)

3. Uniform spectrum: the eigenvalues of Σ_d are $\{2/d, 4/d, 6/d, \ldots, 2\}$, corresponding to a (discretized) uniform distribution over the range $[0, 2]$. (Figure 4.)

4. Toeplitz[5] covariance: $\Sigma_d(i, j) = 0.3^{|i-j|}$. (Figure 5.)

For each of the four types of population spectral distributions, we evaluated our algorithm for a variety of dimensions and sample sizes, taking $d =$

[5]Toeplitz matrices arise in numerous application areas, particularly in settings where each datapoint is a time-series and the correlation between two measurements decreases exponentially as a function of the chronological separation between the measurements.

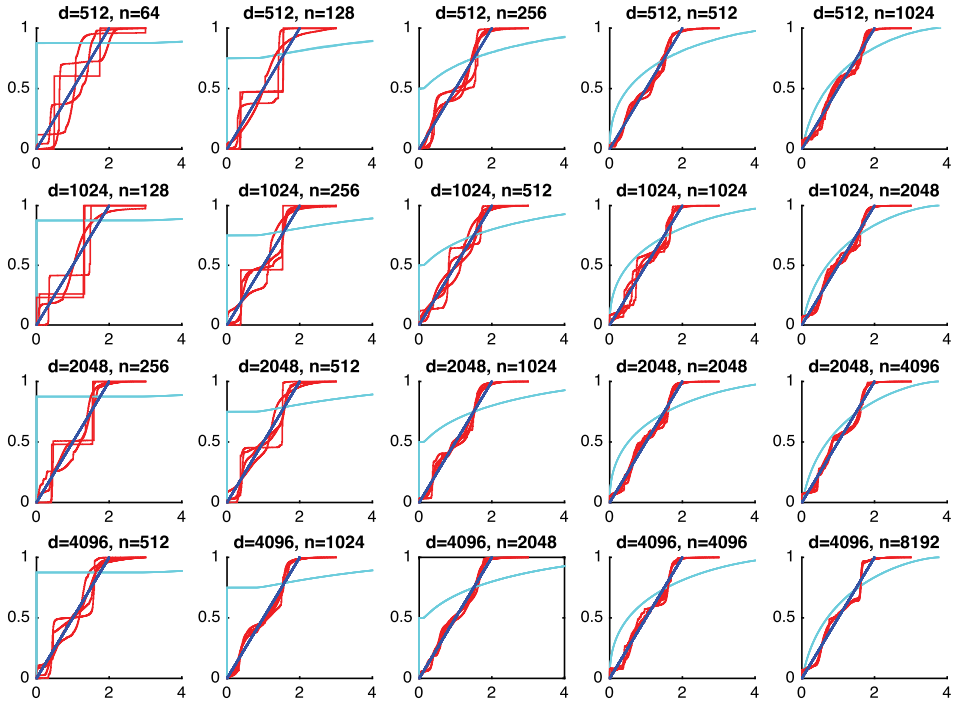

FIG. 4. *Empirical results for reconstructing the population spectrum for a covariance matrices whose eigenvalues correspond to the discretized uniform distribution on the interval* $[0, 2]$. *Red lines depict the cdf of the distribution recovered by our algorithm over five independent trials, the blue line depicts the cdf of the true population spectral distribution and the cyan line depicts the cdf of the empirical spectral distribution* (*in one of the trials*).

$512, 1024, 2048, 4096$, and for each value of d, we considered sample sizes $n = d/8, d/4, d/2, d, 2d$. For each setting, we ran our algorithm five times on independently drawn data. Figures 2–5 show the results of each run, showing the cdf of the estimated spectral distribution (red), together with the cdf of the population spectral distribution (blue), and the cdf of the empirical spectral distribution (cyan).

We observe that in general, for a fixed ratio of d/n, the results improve with larger d, as is implied by our theoretical sublinear sample size asymptotic consistency results (in spite of the daunting constant factors that appear in the analysis). Additionally, our approach has good performance for the more difficult distributions—the uniform and Toeplitz distributions—even in the $n \leq d$ regime.

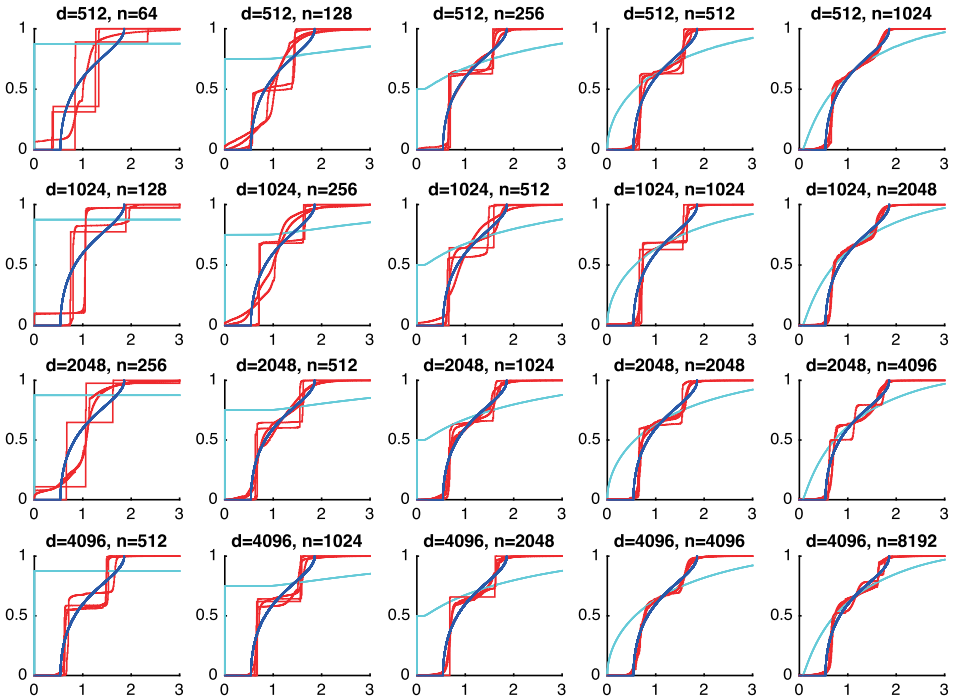

FIG. 5. *Empirical results for reconstructing the population spectrum for covariance* $\Sigma = T$ *where* $T_{i,j} = 0.3^{|i-j|}$ *is a* $d \times d$ *Toeplitz matrix. Red lines depict the cdf of the distribution recovered by our algorithm over five independent trials, the blue line depicts the cdf of the true population spectral distribution and the cyan line depicts the cdf of the empirical spectral distribution* (*in one of the trials*).

SUPPLEMENTARY MATERIAL

Supplement to "Spectrum estimation from samples" (DOI: 10.1214/16-AOS1525SUPP; .pdf). The supplement contains the technical details of the proofs of Propositions 1 and 4.

REFERENCES

[1] ALON, N., YUSTER, R. and ZWICK, U. (1997). Finding and counting given length cycles. *Algorithmica* **17** 209–223. MR1425734

[2] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Stat.* **34** 122–148. MR0145620

[3] BAI, Z., CHEN, J. and YAO, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.* **52** 423–437. MR2791528

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

[5] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[6] BURDA, Z., GÖRLICH, A., JAROSZ, A. and JURKIEWICZ, J. (2004). Signal and noise in correlation matrix. *Phys. A* **343** 295–310. MR2094415
[7] BURDA, Z., JURKIEWICZ, J. and WACŁAW, B. (2005). Spectral moments of correlated Wishart matrices. *Phys. Rev. E* (3) **71** 026111. MR2139960
[8] CAROTHERS, N. (2000). A short course on approximation theory.
[9] DE VILLIERS, J. (2012). *Mathematics of Approximation. Mathematics Textbooks for Science and Engineering* **1**. Atlantis Press, Paris. MR2962227
[10] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511
[11] DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. Preprint. Available at arXiv:1311.0851.
[12] EFRON, B. and MORRIS, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960
[13] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012
[14] HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** 531–544.
[15] HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** 586–597. MR0568722
[16] HARDT, M., LIGETT, K. and MCSHERRY, F. (2012). A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems* 2339–2347.
[17] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, Calif.. MR0133191
[18] KANE, D. M., NELSON, J. and WOODRUFF, D. P. (2010). On the exact space complexity of sketching and streaming small norms. In *Proceedings of the Twenty-First Annual ACM–SIAM Symposium on Discrete Algorithms* 1161–1178. SIAM, Philadelphia, PA. MR2809734
[19] KANTOROVIČ, L. V. and RUBINŠTEĬN, G. Š. (1957). On a functional space and certain extremum problems. *Dokl. Akad. Nauk SSSR* (*N.S.*) **115** 1058–1061. MR0094707
[20] KONG, W. and VALIANT, G. (2017). Supplement to "Spectrum estimation from samples." DOI:10.1214/16-AOS1525SUPP.
[21] LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718
[22] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411.
[23] LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942
[24] LEDOIT, O. and WOLF, M. (2013). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. Available at SSRN 2198287.
[25] LI, W., CHEN, J., QIN, Y., BAI, Z. and YAO, J. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *J. Statist. Plann. Inference* **143** 1887–1897.
[26] LI, W. and YAO, J. (2014). A local moment estimator of the spectrum of a large dimensional covariance matrix. *Statist. Sinica* **24** 919–936. MR3235405
[27] LI, Y., NGUYEN, H. L. and WOODRUFF, D. P. (2014). On sketching matrix norms and the top singular vector. In *Proceedings of the Twenty-Fifth Annual ACM–SIAM Symposium on Discrete Algorithms* 1562–1581. ACM, New York. MR3376474
[28] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.* **3** 123–224.

[29] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **1** 457–483.

[30] MESTRE, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54** 5113–5129.

[31] RAO, N. R., MINGO, J. A., SPEICHER, R. and EDELMAN, A. (2008). Statistical eigeninference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885. MR2485015

[32] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. MR2183942

[33] SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivariate Anal.* **55** 331–339.

[34] STEIN, C. (1975). Estimation of a covariance matrix. *Rietz Lecture, 39th Annual IMS Meeting, Atlanta, GA*.

[35] STEIN, C. (1977). Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation I. Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division* **74** 4–65.

[36] TAKEMURA, A. (1984). An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.* **8** 367–376. MR0767967

[37] VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing* 685–694. ACM, New York.

[38] YIN, Y. Q. and KRISHNAIAH, P. R. (1983). A limit theorem for the eigenvalues of product of two random matrices. *J. Multivariate Anal.* **13** 489–507.

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
353 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: whkong@stanford.edu
valiant@stanford.edu