Recover the spectrum of covariance matrix: a non-asymptotic iterative method

Juntao Duan*, Ionel Popescu†, Heinrich Matzinger‡

January 4, 2022

Abstract

It is well known the sample covariance has a consistent bias in the spectrum, for example spectrum of Wishart matrix follows the Marchenko-Pastur law. We in this work introduce an iterative algorithm 'Concent' that actively eliminate this bias and recover the true spectrum for small and moderate dimensions.

Keywords: covariance spectrum; covariance; eigenvalues;

1 Introduction

In life when we observe some object, we often collect information from multiple perspectives. For example, we measure differences and similarities of certain animal species by color, sex, height, weight etc, which we call features (or explanatory, independent variables). This unavoidably will end up with a vector collecting those (say p) features $(X_1, X_2, \cdots, X_p)$. As we collect sample instances with those p features, we will find each feature X_i follows some probability distribution. For example the animal species' biological sex follows a Bernoulli distribution with parameter $p \approx 0.5$.

It is fundamental to understand the relations between the p features. In principle, we know everything if we know the joint distribution of the p features, for example cumulative distribution

$$F_{X_1,X_2,\cdots,X_p}(t_1, t_2, \cdots, t_p) = \mathbb{P}(X_1 \le t_1, X_2 \le t_2, \cdots, X_p \le t_p)$$

However, this is unlikely to happen in real world. And estimation of the joint distribution become impossible as p increases due to curse of dimensionality. Simply put, the number of samples required to estimate the joint distribution will grow at k^p where k is the number of samples required to estimate any marginal X_i within requested accuracy.

Since moments (mean, variance, correlation) determines many useful information of random variables, the most natural simplification of the problem is to estimate the moments of those features. Namely,

$$\mathbb{E}\, X_1^{a_1} X_2^{a_2} \cdots X_p^{a_p}, \quad a_1, \cdots a_p \ge 0$$

*School of Mathematics, Georgia Institute of Technology, juntaoduan@gmail.com

†University of Bucharest, Faculty of Mathematics and Computer Science, Institute of Mathematics of the Romanian Academy,

‡School of Mathematics, Georgia Institute of Technology,

In particular, collecting first order moments will obtain mean vector μ, and second order centered moments will be covariance matrix Σ.

$$\vec{\mu} = \begin{bmatrix} \mathbb{E}\, X_1 \\ \vdots \\ \mathbb{E}\, X_p \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \operatorname{Cov} X_1 & \cdots & \operatorname{Cov}(X_1, X_p) \\ & \ddots & \\ \operatorname{Cov}(X_p, X_1) & \cdots & \operatorname{Cov} X_p \end{bmatrix}$$

Estimating each element in μ and Σ is not difficult. Since moments can be estimated by taking average statistics from samples. Let k-th samples of X_i be $X_i^{\omega_k}$.

$$\mu_i := \mathbb{E}\, X_i \approx \hat{\mu}_i := \frac{1}{n} \sum_{k=1}^{n} X_i^{\omega_k}, \quad \Sigma_{i,j} := \operatorname{Cov}(X_i, X_j) \approx \hat{\Sigma}_{i,j} := \frac{1}{n} \sum_{k=1}^{n} X_i^{\omega_k} X_j^{\omega_k} - \hat{\mu}_i \hat{\mu}_j$$

Then for each entry as a random variable, with the law of large number we can conclude the error goes to zero in the limit and central limit theorem will control the fluctuation of the error is of order $O(\frac{1}{\sqrt{n}})$ where n is the sample size.

Let us first fix some notations. Let X be the data matrix with n samples (rows) and p columns (features). Then the sample covariance matrix is

$$\frac{1}{n} X^T X - \frac{1}{n^2} X^T \mathbb{1} (X^T \mathbb{1})^T, \quad \mathbb{1} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$$

For simplification of the analysis, we will assume all random variables have mean 0, so that $\vec{\mu} = 0$. Then the sample covariance matrix is simplified as

$$\hat{\Sigma} = \frac{1}{n} X^T X$$

However, as a high dimensional vector or matrix, entry-wise behavior is usually misleading. The matrix $\hat{\Sigma}$ usually exhibits fundamentally different behavior even entries behave as expected. Specifically, the spectrum (eigenvalues) of $\hat{\Sigma}$ has a consistent bias compared with Σ. For example if we fix $p/n \to c$, spectrum of sample covariance matrix for X with i.i.d. entries mean 0 and variance 1, converges to the Marchenko-Pastur law (see [13]) as $n \to \infty$.

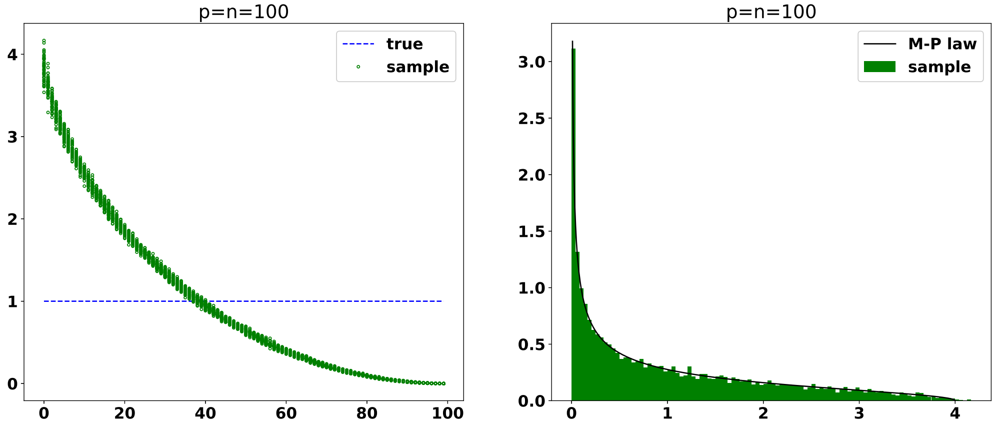

Figure 1: We take $n = p = 100$. The sample covariance matrix ($\hat{\Sigma}$) of standard Gaussian of dimension 100. The true spectrum is $\lambda = 1$ since $\Sigma = I$. On the left we see sample spectrum is concentrated around a biased curve. On the right the histogram of the sample spectrum can be fitted to Marchenko-Pastur distribution closely.

In the example as shown in Figure 1, the true covariance matrix is identity, $\Sigma = I$. Since largest eigenvalue of $\hat{\Sigma}$ is very close to 4, we see largest eigenvalues of the error matrix $\lambda_{max}(\hat{\Sigma} - \Sigma) \approx 4 - 1 = 3$. Therefore $\|\hat{\Sigma} - \Sigma\| \approx 3$. Similarly, for arbitrary true covariance Σ, the spectral norm of the error matrix $\|\hat{\Sigma} - \Sigma\|$ behave similar to a constant (depend on $\frac{p}{n}$) multiple of $\|\Sigma\|$ (see [8]). Any attempt using sample covariance in a matrix fashion will yield a significant error, for example principle component analysis, MANOVA, factor analysis and linear discriminant analysis etc. in multivariate analysis.

In many practical applications, the spectrum of the true covariance matrix contains essential information about the structure of the data at hand. Therefore, recovering the true spectrum is critical to understand the behavior of various models we use for the data. In the case $\Sigma = I$ as shown in figure 1, we can expect a reverting process that may recover the true spectrum from Marchenko-Pastur distribution. In the case of general covariance matrix Σ, a series of results (see Silverstein [15], Bai and Yin [1], Yin, Bai and Krishnaiah [18]) have shown the spectrum of the sample covariance $\hat{\Sigma}$ converge to the free product of spectrum of Σ with a Marchenko-Pastur distribution, provided the spectrum of the true covariance Σ converges. The main result is summarized as follows.

Theorem 1. *Assume the following.*

1. *The entries of $X_p = (X_{i,j})_{n \times p}$ are i.i.d. real random variables for all p.*
2. *$E[X_{1,1}] = 0$, $E[|X_{1,1}|^2] = 1$.*
3. *Let $p/n \to c > 0$ as $p \to \infty$.*
4. *Let Σ_p $(p \times p)$ be non-negative definite symmetric random matrix with spectrum distribution F^{Σ_p} (If $\{\lambda_i\}_{1 \le i \le p}$ are the eigenvalues of Σ_p, then $F^{\Sigma_p} = \sum_1^p \frac{1}{p}\delta_{\lambda_i}(x)$) such that F^{Σ_p} almost surely converges weakly to F^{Σ} on $[0, \infty)$.*
5. *X_p and Σ_p are independent.*

Then the spectrum distribution of $W_p = \frac{1}{n}\Sigma_p^{1/2} X_p^T X_p \Sigma_p^{1/2}$, denoted as F^{W_p} almost surely converges weakly to F^W. F^W is the unique probability measure whose Stieltjes transform $m(z) = \int \frac{dF^W(x)}{x - z}$, $z \in \mathbb{C}^+$ satisfies the equation

$$-\frac{1}{m} = z - c\int \frac{t}{1 + tm} dF^{\Sigma}(t) \quad \forall z \in \mathbb{C}^+ \tag{1.1}$$

In theory, from the limiting distribution of the estimated covariance matrix, one could retrieve Σ using the Stieltjes-transform from free probability. This idea, pioneered by EI Karoui [5], attempts to discretize the Marchenko-Pastur equation 1.1 then estimate spectrum by minimizing the residuals. Recently there is a series of follow-up work attempt to improve the discretization and the optimization for example using different discretization and quantization strategy in [10, 11] and moving from complex plane to real line in [12].

Another type of approach is based on an explicit formula to relate the moments of true limiting spectrum and moments of sample limiting spectrum distribution see for example [2, 9]. This formula approximates the finite dimensional relation with an asymptotic normal error. However, this approach is rather restrictive computationally since it has to solve polynomial equations and invert moments back to distribution. Mostly, it can only deal with the case that true spectrum has only a small number of unique values.

However, using limiting result from random matrix theory does not guarantee good performance in finite dimensional case. Even though the estimators are proven consistent in the limit, the rate of convergence was not well understood and in fact very often to be slow. From Figure 2, the random matrix approaches ‘Quest’ [10] and ‘Moment’ from [9] usually falls short. Instead, we introduce an iterative algorithm, ‘Concent’ method (black in Figure 2), which solves a random optimization problem based on concentration of sample spectrum. Moreover it also exploited sample covariance eigenvectors to correct sample spectrum. This method exhibits remarkable reconstruction due to the sub-Gaussian concentration behavior of sample spectrum which become effective even with very small n and p.

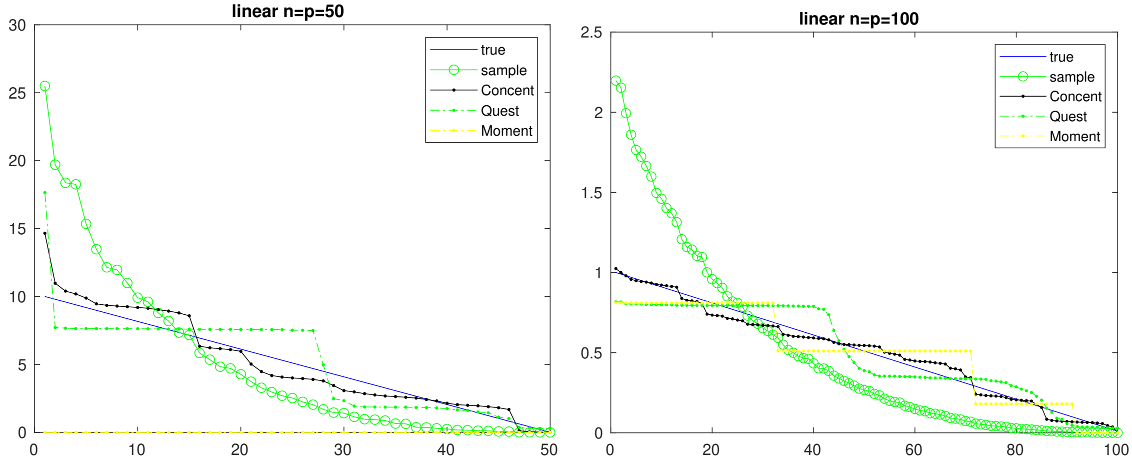

Figure 2: We take $n = p = 50$ on the left, and $n = p = 100$ on the right. The true covariance matrix Σ has diagonalization $Q\Lambda Q^T$ where the diagonal matrix Λ has eigenvalues spaced from 0 to 10 on the left and 0 to 1 on the right uniformly. The sample spectrum (in green) has a convex shape which is significantly different from the true spectrum. The ‘Quest’ estimator form Ledoit [10] performs poorly because of its discontinuity from discretization. The ‘Moment’ estimator from [9] is not gaining useful information at all on the left due to true spectrum has values > 1, which will result higher moments overflow in computer. ‘Moment’ On the right still perform poorly even we restrict spectrum ≤ 1.

There is also many other work trying to find better estimator than sample covariance matrix which do not necessarily estimate true spectrum. For example, under sparsity or low rank condition of the true covariance, shrinkage method on sample covariance exhibits appealing performance, see Stein [16], Bickel and Levina [3] and Donoho [7]. See [6] for a detailed review.

The remaining of this paper is structured as follows. First, we show by various simulations the concentration of the sample spectrum in section 2. This shows the bias in sample spectrum is consistent. Any sample spectrum can be used as a biased baseline. Then in section 2.2, we propose an optimization problem and its approximations based on this concentration property. Then in 2.3 we outline an iterative eigenvector correction algorithm which actively improves approximated optimization solution. At the end, we show some simulations to demonstrate it works well in various settings in section 3.

2 From concentration of sample spectrum to recovery by optimization

2.1 Concentration of sample spectrum

Let $\hat{\Lambda}$ (green in Figure 1) be spectrum of sample covariance matrix $\hat{\Sigma}$. The mean of sample spectrum $\mathbb{E}\,\hat{\Lambda}$ is very far from the true spectrum Λ (red in Figure 1). However $\hat{\Lambda}$ are very concentrated together around $\mathbb{E}\,\hat{\Lambda}$. It is easily observed but not necessarily easy to prove. We formulate the simplest case (assuming Gaussian) below.

Theorem 2. *Let Λ be the diagonal matrix with the true spectrum, i.e. eigenvalues of $p \times p$ true covariance matrix Σ (Σ and Λ are unknown in practice). Assume all spectrum are sorted decreasingly. Let $\mathcal{N}$ be a random $n \times p$ matrix with i.i.d. Gaussian random variables with mean 0 and variance 1, which is unknown in practice as well. Suppose we observe data matrix $X = \mathcal{N}^T \Lambda^{1/2}$. Then denote $\hat{\Lambda}$ as the spectrum of the sample covariance $W = \frac{1}{n} X^T X = \frac{1}{n}\Lambda^{1/2}\mathcal{N}^T\mathcal{N}\Lambda^{1/2}$ with eigenvalues decreasingly ordered $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. Then we have the sample spectrum is concentrated around its mean,*

$$\max_{1\leq k \leq p} \mathbb{P}(\|\hat{\lambda}_k - \mathbb{E}\,\hat{\lambda}_k\|_\infty > \|\Lambda\|_2 t) < Ce^{-cnt} \tag{2.1}$$

The proof is based on the Lipschitz continuity of eigenvalue function and a Gaussian concentration inequality, which we present later. Essentially, this concerns the local statistics for the eigenvalues of finite dimensional sample covariance matrix. For the general case $\mathcal{N}$ is not Gaussian entries, it is much more complicated and we would expect a simple sub-exponential tail bound.

For the simplest case $\Sigma = I$ (with general random variables) , the universality [17] would imply the behavior is close to Wishart matrix as long as first four moments are matched for the entries. For the case of $\Sigma \neq I$ (with general random variables), there is no concentration or universality result available. We suspect there is a sub-exponential tail if first four moments are matched and sub-exponential tail if not.

Conjecture 3. *Fixing all assumptions in Theorem 2, except the random matrix $\mathcal{N}$ have different entries. If entries of $\mathcal{N}$ is not Gaussian but has first four moments matched with Gaussian then we still have sub-exponential tail e^{-cnt}.*

Proof: The difficulty in proving such concentration result mainly due to the complexity of eigenvalue function. We will show the eigenvalue function is Lipschitz. Let $diag(\Lambda) = (\lambda_1, \cdots, \lambda_p)$ a vector of diagonal eigenvalues sorted in decreasing order. Define the diagonalization map $f : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ that

$$diag(\hat{\Lambda}) := f(\mathcal{N}) = Q^T\left(\frac{1}{n}\Lambda^{1/2}\mathcal{N}^T\mathcal{N}\Lambda^{1/2}\right)Q$$

where Q is orthogonal eigenvector matrix of $\frac{1}{n}\Lambda^{1/2}\mathcal{N}^T\mathcal{N}\Lambda^{1/2}$. Then we compute a perturbation

$$\begin{aligned}\|f(\mathcal{N}) - f(\mathcal{N}')\|_\infty &= \|\hat{\Lambda} - \hat{\Lambda}'\|_\infty \\ &= \|\hat{\Lambda} - \hat{\Lambda}'\|_2 \\ &= \|Q^T\left(\frac{1}{n}\Lambda^{1/2}(\mathcal{N}^T\mathcal{N} - \mathcal{N}'^T\mathcal{N}')\Lambda^{1/2}\right)Q\|_2 \\ &\leq \frac{1}{n}\|\Lambda\|_2\|\mathcal{N}^T\mathcal{N} - \mathcal{N}'^T\mathcal{N}'\|_2\end{aligned}$$

$$\leq \frac{1}{n}\|\Lambda\|_2(\|\mathcal{N}^T(\mathcal{N}-\mathcal{N}')\|_2 + \|(\mathcal{N}-\mathcal{N}')^T\mathcal{N}'\|_2)$$

Notice for rectangular matrices A, B, $\|AB\|_2 \leq \|A\|_2\|B\|_F$. Then we conclude

$$\|\mathcal{N}^T(\mathcal{N}-\mathcal{N}')\|_2 \leq \|\mathcal{N}^T\|_2\|\mathcal{N}-\mathcal{N}'\|_F = \|\mathcal{N}\|_2\|\mathcal{N}-\mathcal{N}'\|_F$$

and similarly

$$\|(\mathcal{N}-\mathcal{N}')^T\mathcal{N}'\|_2 = \|\mathcal{N}'^T(\mathcal{N}-\mathcal{N}')\|_2 \leq \|\mathcal{N}'\|_2\|\mathcal{N}-\mathcal{N}'\|_F$$

This leads to the bound on the variation of eigenvalues

$$\|\hat{\Lambda}-\hat{\Lambda}'\|_\infty \leq \frac{1}{n}\|\Lambda\|_2(\|\mathcal{N}\|_2 + \|\mathcal{N}'\|_2)\|\mathcal{N}-\mathcal{N}'\|_F$$

This means f has Lipschitz constant $L \leq \frac{1}{n}\|\Lambda\|_2(\|\mathcal{N}\|_2 + \|\mathcal{N}'\|_2)$, which of course is a random variable since $\mathcal{N}, \mathcal{N}'$ are random matrix. However, we can bound this Lipschitz constant with high probability. Recall for Gaussian random matrix $\mathcal{N}$ we have the concentration of the norm (or singular value [14])

$$\mathbb{P}(\|\mathcal{N}\|_2 > C_0(\sqrt{n}+\sqrt{p}) + t) \leq 2e^{-c_0 t^2}$$

With overwhelming probability, at least $1 - 2e^{-c_0 S^2(n+p)}$, the function f has Lipschitz constant $L \leq \frac{1}{n}\|\Lambda\|_2 2(C_0 + S)(\sqrt{n}+\sqrt{p})$. Then recall Gaussian concentration inequality (can be found in many textbooks for example [4]) states that for $g : \mathbb{R}^N \to \mathbb{R}$ with Lipschitz constant L then for a Gaussian random vector, we have

$$\mathbb{P}\left(|g(X) - \mathbb{E}\, g(X)| > t\right) < 2e^{-\frac{t^2}{2L^2}}$$

Taking $N = np$ and set $g = f|_k$ which restricts f on the k-th coordinate of its codomain. Therefore apply the Gaussian concentration inequality with $L \leq \frac{1}{n}\|\Lambda\|_2 2(C_0 + S)(\sqrt{n}+\sqrt{p}) = (C_1 + c_1 S)\|\Lambda\|_2/\sqrt{n}$ (we assumed $p \leq O(n)$), we obtain concentration for each sample eigenvalue,

$$\begin{aligned}\mathbb{P}\left(|\hat{\lambda}_k - \mathbb{E}\,\hat{\lambda}_k| > \|\Lambda\|_2 t\right) &= \mathbb{P}\left(|\hat{\lambda}_k - \mathbb{E}\,\hat{\lambda}_k| > \|\Lambda\|_2 t, L \leq (C_1 + c_1 S)\|\Lambda\|_2/\sqrt{n}\right)\\ &\quad + \mathbb{P}\left(|\hat{\lambda}_k - \mathbb{E}\,\hat{\lambda}_k| > \|\Lambda\|_2 t, L > (C_1 + c_1 S)\|\Lambda\|_2/\sqrt{n}\right)\\ &< 2e^{-\frac{nt^2}{2(C_1+c_1S)^2}} + 2e^{-c_0 n S^2}\\ &< 4e^{-cnt}\end{aligned}$$

where the last step we selected $S = O(\sqrt{t})$. Then notice this holds for any eigenvalue, we conclude.

$$\max_k \mathbb{P}\left(\|\hat{\lambda}_k - \mathbb{E}\,\hat{\lambda}_k\| > \|\Lambda\|_2 t\right) < Ce^{-cnt}$$

□

From Figure 3, clearly sample spectrum are concentrated around a convex curve which is significantly different from the true spectrum ($\Sigma \neq I$).

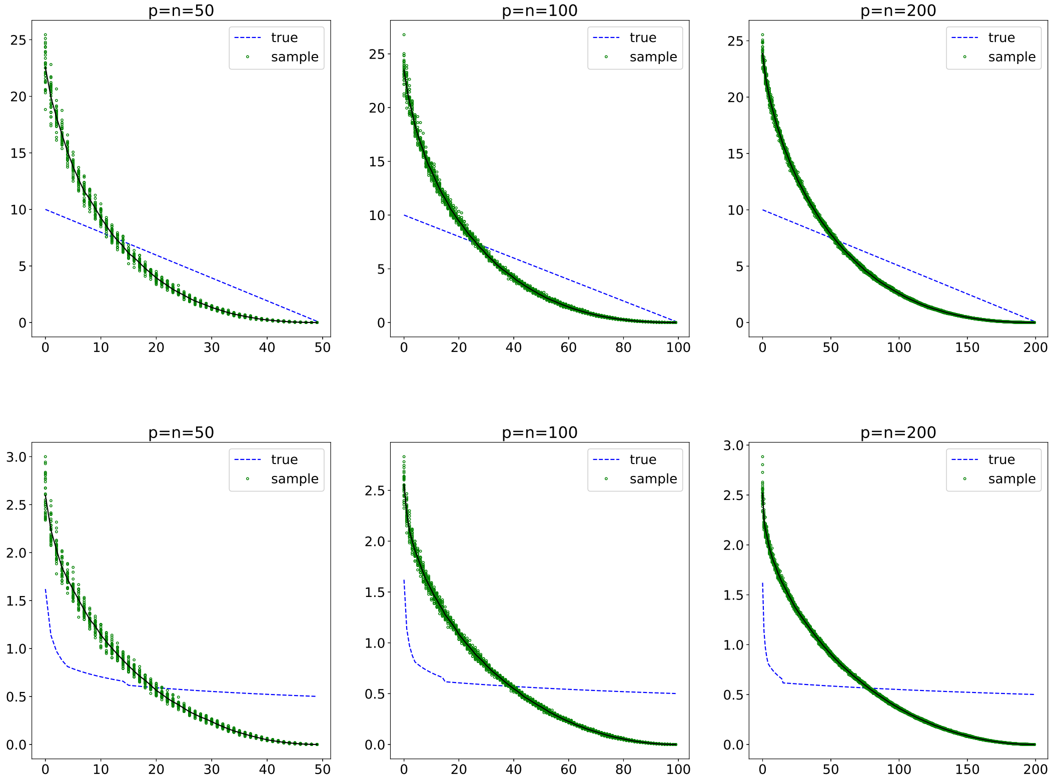

Figure 3: The pictures in the first row has true covariance Σ with eigenvalues ranging from 0 to 10 evenly of step size $1/p$. The graph in the second row has true spectrum Σ with eigenvalues ranging from 0 to 10 evenly of step size $1/p$.

2.2 Random optimization

Let Λ be the true spectrum, and $\hat{\Lambda}$ be the sample covariance spectrum. Due to the concentration in previous section, we propose the following optimization,

$$\min_{D\geq 0}\sum_{k=1}^{K}\left\|\hat{\Lambda}-\mathrm{eig}(D^{1/2}\mathcal{N}_k^T\mathcal{N}_k D^{1/2})\right\| \tag{2.2}$$

where $\mathcal{N}_i$ is $n\times p$ random matrix with i.i.d. standard normal random variables, and $\mathrm{eig}(\cdot)$ is computing the eigenvalues and sort in descending order. One note that the norm could be ℓ_1,ℓ_2 or any vector norm. However, by analyzing the performance on simulation, we did not observe too much difference for various norms. for computational reason ℓ_2 is taken in the following discussion.

One caveat of this formulation is the optimization problem does not have convex structure so that it can not be solved by fast algorithms available in convex optimization literature. The reason for the complexity in the objective function is due to $\mathrm{eig}(\cdot)$ need to compute eigenvalues and sort afterwards. One could use a generic global optimizer search engine (e.g. Genetic algorithm) but it will be extremely slow and essentially not applicable for large dimension. So we replace it with a approximation of the problem. First, we translate the problem into linear function in D in side the norm if we knew the true spectrum Λ. That is we want the variable D be outside of the eigenvalue computation.

$$\Lambda \approx \underset{D\geq 0}{\mathrm{argmin}} \sum_{k=1}^{K} \left\| \hat{\Lambda} - R_k D \right\| \tag{2.3}$$

where R_k is a vector obtained by element-wise division,

$$R_k = \frac{\mathrm{eig}(\Lambda^{1/2}\mathcal{N}_k^T\mathcal{N}_k\Lambda^{1/2})}{\Lambda}$$

Of course, R_k is not obtainable, so we replace it with an estimator

$$\hat{R}_k = \frac{\mathrm{eig}(\tilde{\Lambda}^{1/2}\mathcal{N}_k^T\mathcal{N}_k\tilde{\Lambda}^{1/2})}{\tilde{\Lambda}}$$

where $\tilde{\Lambda}$ is any reasonable estimator of the covariance spectrum. In principle, one can iteratively find a sequence of such estimators $\tilde{\Lambda}_k$. We use the sample spectrum to start the iteration $\tilde{\Lambda}_0 = \hat{\Lambda}$. Thus we arrived at

$$\Lambda_{concent} = \underset{D\geq 0}{\mathrm{argmin}} \sum_{k} \left\| \hat{\Lambda} - \hat{R}_k D \right\| \tag{2.4}$$

Now let's derive an explicit formula for ℓ_2 minimization. The objective function can be rewritten as

$$f = \sum_k \left\| \hat{\Lambda} - \hat{R}_k D \right\|^2 = \sum_{k=1}^{K}\sum_{j=1}^{p}(\hat{\lambda}_j - \hat{R}_{kj}d_j)^2$$

Then set partial derivatives $\partial f/\partial d_j$ to be zero, we found

$$d_j = \frac{\hat{\lambda}_j \sum_{k=1}^{K} \hat{R}_{kj}}{\sum_{k=1}^{K} \hat{R}_{kj}^2}$$

Here d_j will serve as an estimator of the true spectrum λ_j. And $\hat{R}_{kj}$'s are approximates of $\hat{\lambda}_j/\lambda_j$. In principle, the simplest approximation would be taking average of such ratio to get a naive estimator $\hat{\lambda}_j \sum_{k=1}^{K} \hat{R}_{kj}/K$. But instead our ℓ_2 minimization give a second order correction of this naive approach.

This procedure can be repeated many times to obtian improved ratios However, this approach has many parts replaced by estimators instead of the true, thus we will propose a follow up eigenvector correction procedure.

2.3 An eigenvector correction

We start with any estimator, say with $\Lambda_0 = \Lambda_{concent}$. Then in $k+1$-th step, we simulate a sample covariance matrix and diagonalize it

$$W_k = \Lambda_k^{1/2}\mathcal{N}^T\mathcal{N}\Lambda_k^{1/2}/n \quad \rightarrow \quad W_k = V_k D_k V_k^T$$

Then we obtain the next estimator by the diagonal elements of the matrix $V_k\hat{\Lambda}V_k^T$.

$$\Lambda_{k+1} = diag(V_k\hat{\Lambda}V_k^T)$$

We give a heuristic argument here to explain its effectiveness. Let $\hat{W} = Q^T\Lambda^{1/2}\mathcal{N}^T\mathcal{N}\Lambda^{1/2}Q/n$ be the given sample covariance matrix, then the true covariance matrix is $W = Q\Lambda Q^T$, then diagonal elements of $Q^T\hat{W}Q$ will be a good estimator of true spectrum. Since

$$diag(Q^T\hat{W}Q) = q_k^T\hat{W}q_k = \lambda_k\frac{1}{n}\sum_{i=1}^{n}\mathcal{N}_{i,k}^2 \to \lambda_k$$

where $\mathcal{N}_{i,k}$ is i,k-th entry of $\mathcal{N}$. In our procedure, the V_k will play a similar role as Q. One limitations about this procedure is that it does not apply to high dimensional setting, for example $p \geq 10^4$. The computation would be too demanding due to the eigen-decomposition used in the iteration. However for small or moderate dimensions (for example $p = 10^3$), the iteration converges fast and usually less than 10 iterations would be sufficient.

Combining the two approach, we propose the following 'Concent' algorithm for spectrum recovery.

Algorithm 1: 'Concent': Eigenvector corrected random optimization

```
Data: $X \in \mathbb{R}^{n\times p}$, $n, p > 0$
Initialization: $l \geq 10$, $K \geq 10$                              // $l$ total iterations
$\hat{\Lambda} = \text{eig}(X^TX/n)$
Initialization: $\Lambda_1 = \hat{\Lambda}$
for $i \leftarrow 1$ to $l$ do
    /* 1.  Approximated random optimization                                    */
    Generate $\mathcal{N}_1, \cdots, \mathcal{N}_K$              // random $n \times p$ standard normal matrix
    for $k = 1$ to $K$ do
        $\Lambda_{temp} = \text{eig}({\Lambda_i}^{1/2}\mathcal{N}_k^T\mathcal{N}_k{\Lambda_i}^{1/2}/n)$
        Create ratio vector $\hat{R}_k = diag(\Lambda_{temp})/diag(\Lambda_i)$
    end
    Create diagonal matrix $S = diag(\frac{\sum_{k=1}^K \hat{R}_{kj}}{\sum_{k=1}^K \hat{R}_{kj}^2}, \cdots \frac{\sum_{k=1}^K \hat{R}_{kj}}{\sum_{k=1}^K \hat{R}_{kj}^2})$
    $\Lambda_i = \hat{\Lambda}S$                                       // approximated estimator

    /* 2.  Eigenvector correction                                              */
    Generate $\mathcal{N}$                                       // random $n \times p$ standard normal matrix
    Compute $W = \Lambda_i^{1/2}\mathcal{N}^T\mathcal{N}\Lambda_i^{1/2}/n$
    $VDV^T \leftarrow Diagonalization(W)$
    $\Lambda_{i+1} \leftarrow diag(V\hat{\Lambda}V^T)$                          // approximated estimator
end
Output: $\Lambda_{l+1}$
```

In simulations, we found the *'Approximated random optimization'* and *'Eigenvector correction'* are interchangeable and both serve as approximated iterative algorithm to solve the random optimization 2.2. The only difference is that using eigenvectors will produce a smoother spectrum.

3 Simulation

Here we show the simulations on various settings of our method. We also compare with 'Quest' estimator form Ledoit [10] and 'Moment' estimator from [9].

3.1 Simulated spectrum

We have shown our 'Concent' method performing well for linear spectrum in Figure 2. We next exam the case that true spectrum has a convex or concave shape.

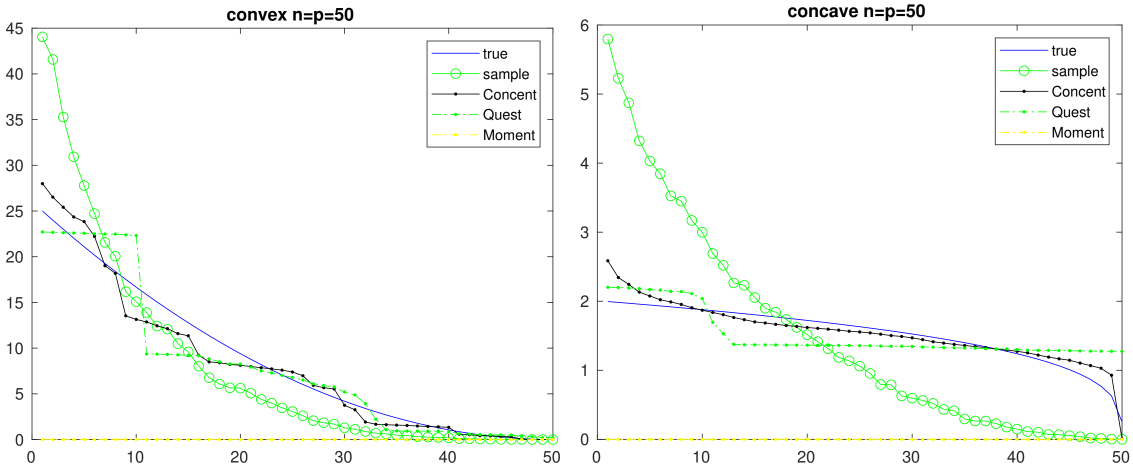

Figure 4: The true spectrum on the left takes x^2 where x is evenly spaced in $(0, 5]$. On the right the true spectrum is taking $x^{0.3}$.

When we deal with spectrum of special unknown structure, it's still possible to recover the smoothing approximation of the true spectrum using our 'Concent' algorithm. Here is a simulation with spectrum of step shape and sparse shape.

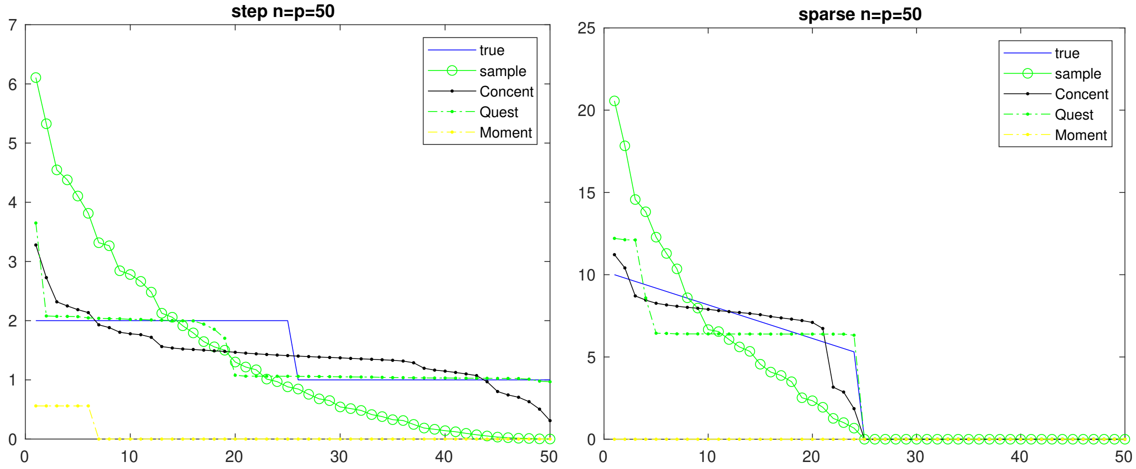

Figure 5: On the left, the true spectrum are half 2's and half 1's. On the right the true spectrum is taking by zero out last half of the linear spectrum.

3.2 Real world data

We compare the result with the true spectrum generated from large sample size real stock data. The 'true' spectrum is taken from 50 stocks with 1000 days. $n/p = 20$ which means the 'true' spectrum is relatively close to the spectrum of real stock covariance matrix.

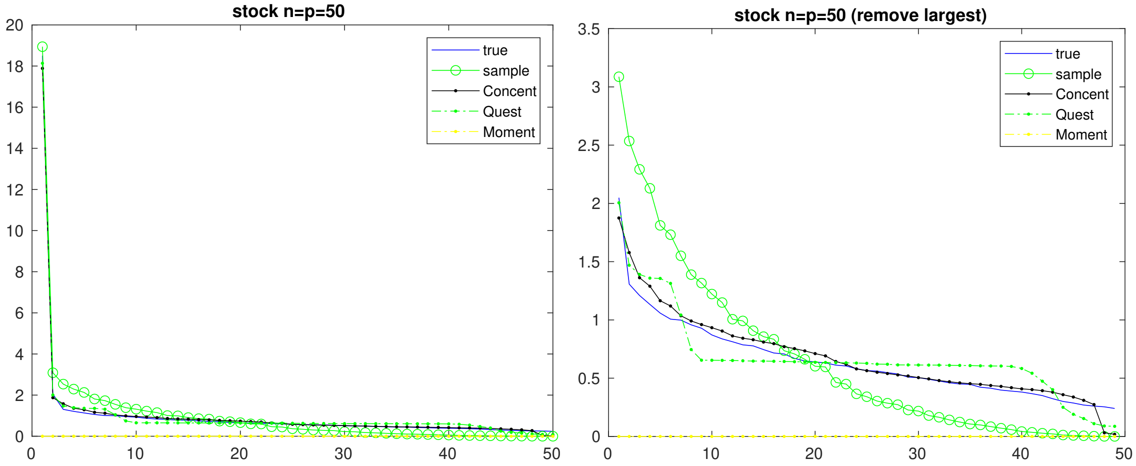

Figure 6: On the right, we removed the largest eigenvalue to make it easy to see the difference.

Another example we study is amazon reviews dataset. We take 50 products with 4082 reviews for each. Therefore $n/p \approx 80$, we are confident the sample spectrum from this data is close to the true spectrum.

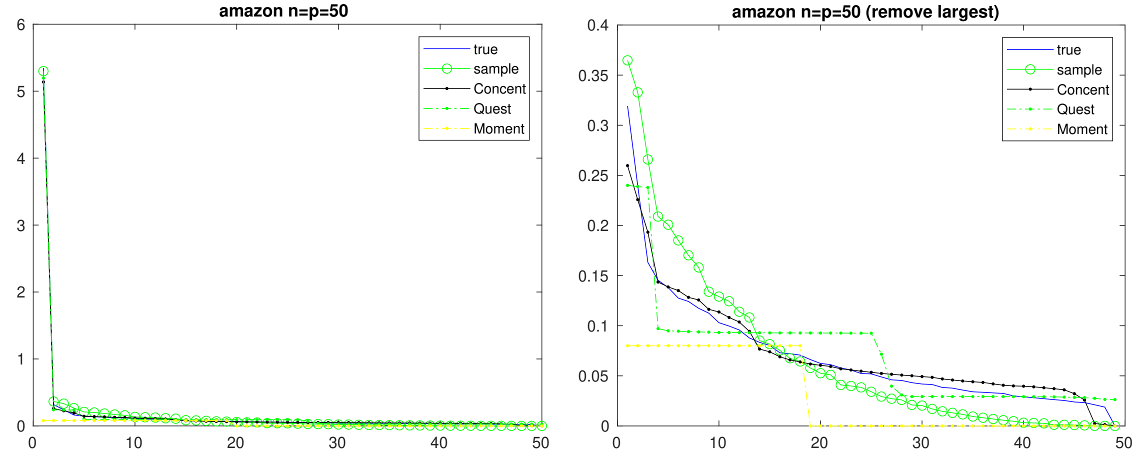

Figure 7: On the right, we removed the largest eigenvalue to show the difference in bulk eigenvalues.

In the majority of those cases, ‘Concent’ outperforms others. There is another significant advantage of ‘Concent’. That is its robustness against to the random generated samples. In other words, taken any sample spectrum, ‘Concent’ would be able to use it to recover the true spectrum due to powerful finite dimensional concentration of sample spectrum. On the other hand, in Figure 8, ‘Quest’ (in red) varies quite significantly even when sample spectrum (in green) changes very little.

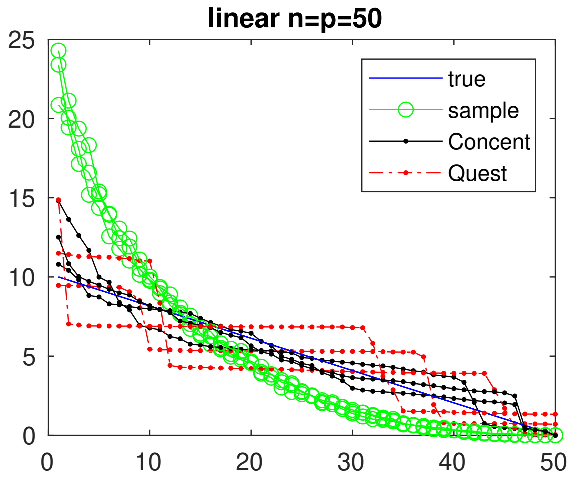

Figure 8: We put 3 simulated recovery together. Even the sample spectrum is very concentrated but the Quest varies significantly.

Moments method generally does not produce relevant result. The total number of moments can be computed are too few (around 10). When eigenvalues are large than 1, large moment will blow up. In the case of eigenvalues are less than 1, then higher order moments will be close to zero and produce no information.

4 Conclusion

We derive a concentration of sample spectrum result and propose a random optimization to recover the true spectrum. Our method of recovering the spectrum is based on finite dimensional concentration of measure behavior so that it provides a competitive performance for small and moderate dimensional covariance matrix. It is much more stable compared with 'Quest' and 'Moment' method which are based on properties of the limiting random matrix behavior. From simulations we showed our algorithms overcome several weakness of 'Quest' and 'Moment' method. 'Quest' method is very sensitive to small changes in sample spectrum and usually produce a discontinuous estimator. 'Moment' method does not work properly for small or moderate dimensions and will blow up for eigenvalues larger than 1.

There are several limitations of our method. First, it has expensive diagonalization procedure which will be hard to implement for large dimensions. Second, for discontinuous true spectrum, the recovery is only possible if the structure is known otherwise it produces smoothing approximations of the true spectrum.

References

[1] Z. D. Bai and Y. Q. Yin. Convergence to the semicircle law. *Ann. Probab.*, 16(2):863–875, 04 1988.

[2] Zhidong Bai, Jiaqi Chen, and Jianfeng Yao. On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Australian and New Zealand Journal of Statistics*, 52(4):423–437, 2010.

[3] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[5] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.*, 36(6):2757–2790, 12 2008.

[6] Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

[7] Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Trans. Inform. Theory*, 63(4):2137–2152, 2017.

[8] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

[9] W. Kong and G. Valiant. Spectrum estimation from samples. *Annals of Statistics*, 45(5):2218–2247, 2017.

[10] Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139(Supplement C):360 – 384, 2015.

[11] Olivier Ledoit and Michael Wolf. Numerical implementation of the quest function. *Computational Statistics & Data Analysis*, 115:199–223, 2017.

[12] Weiming Li, Jiaqi Chen, Yingli Qin, Zhidong Bai, and Jianfeng Yao. Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *Journal of Statistical Planning and Inference*, 143(11):1887 – 1897, 2013.

[13] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[14] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

[15] J.W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331 – 339, 1995.

[16] C Stein. Estimation of a covariance matrix. *Ritz lecture at annual IMS meeting in Atlanta 1975*, 1975.

[17] Terence Tao and Van Vu. Random covariance matrices: Universality of local statistics of eigenvalues. *The Annals of Probability*, 40(3):1285–1315, 2012.

[18] Y.Q Yin, Z.D Bai, and P.R Krishnaiah. Limiting behavior of the eigenvalues of a multivariate f matrix. *Journal of Multivariate Analysis*, 13(4):508 – 516, 1983.