Spiked Covariance Model and BBP Phase Transition

Recall: For $Y \in \mathbb{R}^{n \times p}$ with independent rows of mean 0, covariance Σ, $\hat{\Sigma} = \frac{1}{n} Y^T Y \in \mathbb{R}^{p \times p}$ is the sample covariance matrix.

If $\Sigma = I$ and entries of Y are independent (with bounded moments of all orders), then as $n, p \to \infty$ with $p/n \to \gamma \in (0,\infty)$, the ESD of $\hat{\Sigma}$ converges weakly a.s. to the Marcenko-Pastur law μ_γ. This has density

$$d\mu_\gamma(x) = \frac{1}{2\pi\gamma x}\sqrt{(\gamma_+ - x)(x - \gamma_-)}\,\mathbb{1}\{x \in [\gamma_-, \gamma_+]\}$$

where $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$,

plus a point mass at 0 if $\gamma > 1$.

Def: Suppose $Y = X\Sigma^{1/2}$ ($\mathbb{E}[X_{ij}] = 0$, $\mathbb{E}[|X_{ij}|^2] = 1$, $\mathbb{E}[|X_{ij}|^k] \le C_k$) where

- A fixed number ℓ of eigenvalues $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_\ell > 1$ of Σ are larger than 1.
- All other eigenvalues $\lambda_{\ell+1} = \lambda_{\ell+2} = \dots = \lambda_p$ of Σ equal 1.

This is called the spiked covariance model (Johnstone '01).

At the population level, the covariance Σ has ℓ "signal" principal components. Each sample (row of Y) in $\mathbb{R}^p$ has significant variation along these ℓ directions, and isotropic noise in the remaining $p-\ell$ directions.

Q: What is the behavior of principal components of the sample covariance $\hat{\Sigma}$?

Thm (Baik, Ben Arous, Péché '05): Let X_{ij} be complex Gaussian: $X_{ij} = X_{ij}^{(1)} + i\, X_{ij}^{(2)}$ where $X_{ij}^{(1)}, X_{ij}^{(2)} \overset{iid}{\sim} N(0, \frac{1}{2})$. Define $\hat{\Sigma} = \frac{1}{n} Y^* Y$ where $Y = X\Sigma^{1/2}$. Suppose $n, p \to \infty$ with $p/n \to \gamma \in (0,\infty)$, and $\ell, \lambda_1, \dots, \lambda_\ell$ are fixed. Then:

(a) If all eigenvalues $\lambda_1, \ldots, \lambda_\ell < 1+\sqrt{\gamma}$, then the largest eigenvalue $\hat{\lambda}_1$ of $\hat{\Sigma}$ satisfies

$$\hat{\lambda}_1 \xrightarrow{a.s.} (1+\sqrt{\gamma})^2, \text{ the right endpoint of support of } \mu_\gamma.$$

$$n^{2/3} \cdot \tfrac{1}{\sqrt{\gamma}}(1+\tfrac{1}{\sqrt{\gamma}})^{-4/3}\left(\hat{\lambda}_1 - (1+\sqrt{\gamma})^2\right) \xrightarrow{d} F_{GUE} \quad \text{if } |\tfrac{p}{n}-\gamma| = o(n^{-2/3}),$$

where F_{GUE} is the GUE Tracy-Widom Law.

(b) If $\lambda_1 = \ldots = \lambda_k = 1+\sqrt{\gamma}$, and $\lambda_{k+1}, \ldots, \lambda_\ell < 1+\sqrt{\gamma}$, then

$$\hat{\lambda}_1 \xrightarrow{a.s.} (1+\sqrt{\gamma})^2$$

$$n^{2/3} \cdot \tfrac{1}{\sqrt{\gamma}}(1+\tfrac{1}{\sqrt{\gamma}})^{-4/3}\left(\hat{\lambda}_1 - (1+\sqrt{\gamma})^2\right) \xrightarrow{d} F_k, \quad \text{if } |\tfrac{p}{n}-\gamma| = o(n^{-2/3}),$$

where $F_1, F_2, F_3, \ldots$ are a certain generalization of F_{GUE}.

(c) If $\lambda_1 = \ldots = \lambda_k > 1+\sqrt{\gamma}$, and $\lambda_{k+1}, \ldots, \lambda_\ell < \lambda_1$, then

$$\hat{\lambda}_1 \xrightarrow{a.s.} \lambda_1\left(1+\frac{\gamma}{\lambda_1 - 1}\right)$$

$$\sqrt{n} \cdot \lambda_1^{-1}\left(1-\frac{\gamma}{(\lambda_1-1)^2}\right)^{-1/2}\left(\hat{\lambda}_1 - \lambda_1\left(1+\frac{\gamma}{\lambda_1-1}\right)\right) \xrightarrow{d} G_k \quad \text{if } |\tfrac{p}{n}-\gamma| = o(n^{-1/2}),$$

where G_1 is the standard Gaussian distribution, and G_k is the law of largest eigenvalue of a $k \times k$ GUE matrix (without rescaling).

Remark: (a-c) illustrate a phase transition phenomenon: If $\lambda_1 < 1+\sqrt{\gamma}$, the signal is covered by the spectral noise. If $\lambda_1 > 1+\sqrt{\gamma}$, an outlier eigenvalue separates from μ_γ. This is now called the "Baik-Ben Arous-Péché" (BBP) phase transition.

Remark: For real Gaussian X, as well as $X \in \mathbb{R}^{n \times p}$ having more general iid entries, the first-order (almost-sure) limits are the same. However the second-order (distributional) limits are different:

- If all $\lambda_1, \ldots, \lambda_\ell < 1+\sqrt{\gamma}$, and $|\tfrac{p}{n}-\gamma| = o(n^{-2/3})$, then

$$n^{2/3} \cdot \tfrac{1}{\sqrt{\gamma}}(1+\tfrac{1}{\sqrt{\gamma}})^{-4/3}\left(\hat{\lambda}_1 - (1+\sqrt{\gamma})^2\right) \xrightarrow{d} F_{GOE},$$

  where F_{GOE} is the GOE Tracy-Widom law. (This is different from F_{GUE}.) See Lee, Schnelli "Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population", 2016 for recent results.

- If X is real Gaussian, $X_{ij} \overset{iid}{\sim} N(0,1)$, and $\lambda_1 = \ldots = \lambda_k > 1+\sqrt{\gamma}$, $\lambda_{k+1}, \ldots, \lambda_\ell < 1+\sqrt{\gamma}$, and $|\tfrac{p}{n}-\gamma| = o(n^{-1/2})$, then

$$\sqrt{n} \cdot \lambda_1^{-1}\left(1-\frac{\gamma}{(\lambda_1-1)^2}\right)^{-1/2}\left(\hat{\lambda}_1 - \lambda_1\left(1+\frac{\gamma}{\lambda_1-1}\right)\right) \xrightarrow{d} \tilde{G}_k,$$

  where $\tilde{G}_1$ is $N(0,2)$, and $\tilde{G}_k$ is the law of largest eigenvalue of $k \times k$ GOE matrix (without rescaling).

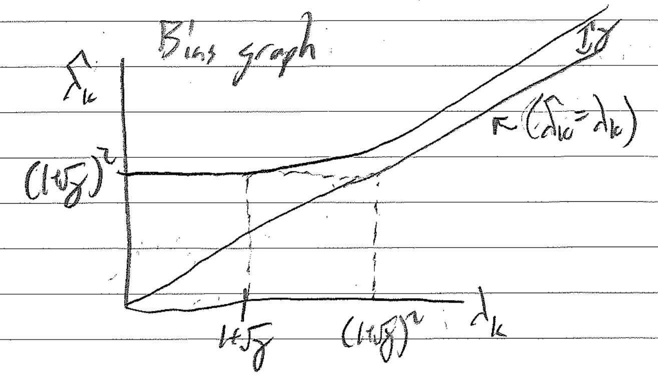
  For more general X, second-order behavior of $\hat{\lambda}_1$ depends on the kurtosis of X_{ij}. See Bai, Yao "Central limit theorems for eigenvalues in a spiked population model", 2008 and Benaych-Georges, Guionnet, Maida "Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices", 2011.

- For $\lambda_1 = \ldots = \lambda_k = 1+\sqrt{\gamma}$, see Bloemendal, Virag "Limits of spiked random matrices I + II", 2013 + 2016.

We'll focus today on the outlier eigenvalues and eigenvectors for real Gaussian X, and show the following:

Thm (Paul '07): Suppose $X_{ij} \overset{iid}{\sim} N(0,1)$, and Σ follows the spiked covariance model with fixed ℓ, $\lambda_1 > \lambda_2 > \ldots > \lambda_\ell$. Let $\tfrac{p}{n} \to \gamma \in (0,1)$, with $|\tfrac{p}{n}-\gamma| = o(n^{-1/2})$. Then the ℓ largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_\ell$ of $\hat{\Sigma}$ satisfy:

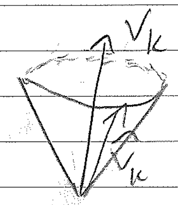
(a) $\hat{\lambda}_k \xrightarrow{a.s.} \begin{cases} (1+\sqrt{\gamma})^2 & \text{if } \lambda_k \in (1, 1+\sqrt{\gamma}] \\ \lambda_k\left(1+\frac{\gamma}{\lambda_k - 1}\right) & \text{if } \lambda_k > 1+\sqrt{\gamma} \end{cases}$

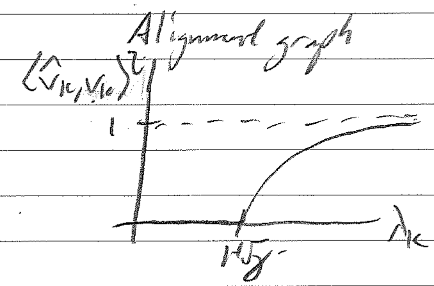


(b) If $\lambda_k > 1+\sqrt{\gamma}$, then

$$\sqrt{n}\left(\hat{\lambda}_k - \lambda_k\left(1+\frac{\gamma}{\lambda_k-1}\right)\right) \xrightarrow{d} \mathcal{N}\left(0, 2\lambda_k^2\left(1-\frac{\gamma}{(\lambda_k-1)^2}\right)\right) \quad (*)$$

(c) If $\lambda_k > 1+\sqrt{\gamma}$, then the corresponding unit eigenvectors $\hat{v}_k, v_k$ satisfy $\langle \hat{v}_k, v_k\rangle^2 \xrightarrow{a.s.} \left(1-\frac{\gamma}{(\lambda_k-1)^2}\right)/\left(1+\frac{\gamma}{\lambda_k-1}\right)$. Furthermore, for any deterministic unit vector $u \in \mathbb{R}^p$ orthogonal to v_k, $\langle \hat{v}_k, u\rangle^2 \xrightarrow{a.s.} 0$.





Implications for statistical inference:

(1) $\hat{\lambda}_k$ is upward biased for λ_k. However:

$z_k = \lambda_k\left(1+\frac{\gamma}{\lambda_k-1}\right) \Leftrightarrow \lambda_k = z_k m_\gamma(z_k)/(1+z_k m_\gamma(z_k))$ where m_γ is the Stieltjes transform of the M-P law. So $\tilde{\lambda}_k = \hat{\lambda}_k m_\gamma(\hat{\lambda}_k)/(1+\hat{\lambda}_k m_\gamma(\hat{\lambda}_k))$ is consistent for λ_k. Also, $\sqrt{n}(\tilde{\lambda}_k - \lambda_k)$ is asymptotically normal, with variance given by $(*)$ and the delta method. Note: Variance in $(*)$ relies on normality of X_{ij}.

(2) In general, it's not possible to consistently estimate v_k as $n, p \to \infty$ (if λ_k remains bounded). We know v_k should be near a certain cone, centered around $\hat{v}_k$, but the direction of error is unknown and (almost uniformly) random.

However, for a fixed unit vector $u \in \mathbb{R}^p$ (not depending on the data), we can consistently estimate $|v_k^T u|$. Note that $|\hat{v}_k^T u|$ is downward biased: Write $u = \alpha v_k + \sqrt{1-\alpha^2}\, w$, where w is orthogonal to v_k. Then $\langle \hat{v}_k, w\rangle \to 0$, so

$$|\langle \hat{v}_k, u\rangle| = \alpha \cdot |\langle \hat{v}_k, v_k\rangle| + o(1)$$
$$\to \alpha \cdot \frac{1-\frac{\gamma}{(\lambda_k-1)^2}}{1+\frac{\gamma}{\lambda_k-1}} = |\langle v_k, u\rangle| \cdot \underbrace{\frac{1-\frac{\gamma}{(\lambda_k-1)^2}}{1+\frac{\gamma}{\lambda_k-1}}}_{\rho_k}$$

Here $\rho_k \in (0,1)$ is unknown, but may be estimated by plugging in $\tilde{\lambda}_k$ for λ_k. Letting this plug-in estimate be $\hat{\rho}_k$, we see that $\hat{\rho}_k^{-1} \cdot |\langle \hat{v}_k, u\rangle|$ is consistent for $|\langle v_k, u\rangle|$.

(3) This is all assuming $\mathbb{E}X_{ij}^2 = 1$ for all i, j, i.e. $\Sigma = I +$ low-rank perturbation. It is more common to assume $\Sigma = \sigma^2 I +$ low-rank, and where I may further be restricted to a $\tilde{p}$-dimensional subspace of $\mathbb{R}^p$ for some "effective dimension" $\tilde{p}$. The parameters $\tilde{p}, \sigma^2$ may be estimated by matching the shape of the eigenvalue distribution to the Marcenko-Pastur law — see Patterson, Price, Reich "Population structure and eigenanalysis" 2006 for one implementation of this idea.

Proof of Thm (a): Recall $\hat{\Sigma} = \frac{1}{n} Y^T Y = \frac{1}{n} \Sigma^{1/2} X^T X \Sigma^{1/2}$. Note that

$\Sigma^{1/2} = V \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \ddots & & \\ & & \sqrt{\lambda_\ell} & \\ & & & 1 \\ & & & & \ddots \\ & & & & & 1 \end{pmatrix} V^T$ for the eigenvector basis V of Σ, and

- $\hat{\Sigma}$ has the same eigenvalues as $V^T \hat{\Sigma} V$
- $V^T X^T X V$ is equal in distribution to $X^T X$, by rotational invariance of $N(0, I)$.

So we may assume $\Sigma = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_\ell & \\ & & & 1 \\ & & & & \ddots \\ & & & & & 1 \end{pmatrix} = \begin{array}{cc} \ell & p-\ell \\ \begin{pmatrix} \Sigma_1 & 0 \\ 0 & I \end{pmatrix} \end{array}$. Write

$$\hat{\Sigma} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \qquad X = (\underbrace{X_1}_{\ell} \;\; \underbrace{X_2}_{p-\ell})$$

(columns of $\hat{\Sigma}$ of widths ℓ and $p-\ell$)

Then $S_{22} = \frac{1}{n} X_2^T X_2$ is a white sample covariance matrix. We'll assume:

Thm (Silverstein '85, Bai, Yin '93): All eigenvalues of S_{22} converge a.s. to $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$.

(This may be proven by either the moment method or Stieltjes transform method, as in the Wigner case.)

Then for any eigenvalue $\hat{\lambda}$ of $\hat{\Sigma}$ outside $((1-\sqrt{\gamma})^2 - \varepsilon, (1+\sqrt{\gamma})^2 + \varepsilon)$,

$$0 = \det(\hat{\Sigma} - \hat{\lambda} I) = \det(S_{22} - \hat{\lambda} I) \cdot \det(S_{11} - \hat{\lambda} I_\ell - S_{12}(S_{22} - \hat{\lambda} I_{p-\ell})^{-1} S_{21})$$

$$\Leftrightarrow 0 = \det(S_{11} - \hat{\lambda} I_\ell - S_{12} R(\hat{\lambda}) S_{21})$$

where $R(z) = (S_{22} - z I_{p-\ell})^{-1}$ is the resolvent of S_{22}.

We have: $S_{11} - S_{12} R(z) S_{21}$

$$= \frac{1}{n} \Sigma_1^{1/2} X_1^T X_1 \Sigma_1^{1/2} - \frac{1}{n^2} \Sigma_1^{1/2} X_1^T X_2 \left(\tfrac{1}{n} X_2^T X_2 - z I\right)^{-1} X_2^T X_1 \Sigma_1^{1/2}$$

$$= \frac{1}{n} \Sigma_1^{1/2} X_1^T \underbrace{\left(I - \frac{1}{n} X_2 \left(\tfrac{1}{n} X_2^T X_2 - z I\right)^{-1} X_2^T\right)}_{M} X_1 \Sigma_1^{1/2}$$

For any deterministic $M \in \mathbb{R}^{n \times n}$ with $\|M\| \leq C$,

$$\frac{1}{n} X_1^T M X_1 - \mathbb{E}\left[\frac{1}{n} X_1^T M X_1\right] \to 0 \text{ a.s. (entrywise) by Hanson-Wright.}$$

We have $\mathbb{E}\left[\frac{1}{n}(X_1^T M X_1)_{ij}\right] = \begin{cases} \frac{1}{n} \operatorname{Tr} M & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases}$

So $\frac{1}{n} X_1^T M X_1 = (\frac{1}{n} \operatorname{Tr} M) \cdot I_\ell + o(1)$.

Since X_1, X_2 are independent, apply this to M conditional on X_2.

Note that $\|M\| \leq 1 + \frac{1}{n} \|X_2\| \cdot \|R(z)\| \cdot \|X_2^T\| = 1 + \|\frac{1}{n} X_2^T X_2\| \cdot \|R(z)\| \leq C$ for $z \in \mathbb{C}$ s.t. $\operatorname{dist}(z, [(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]) > \varepsilon$, and

$$\frac{1}{n} \operatorname{Tr} M = 1 - \frac{1}{n} \operatorname{Tr}\left[\frac{1}{n} X_2^T X_2 \cdot \left(\tfrac{1}{n} X_2^T X_2 - z I_{p-\ell}\right)^{-1}\right]$$

$$= 1 - \frac{1}{n} \operatorname{Tr}\left[I_{p-\ell} + z\left(\tfrac{1}{n} X_2^T X_2 - z I_{p-\ell}\right)^{-1}\right] \qquad (A(A-z)^{-1} = I + z(A-z)^{-1})$$

$$\xrightarrow{a.s.} 1 - \gamma - z \cdot \gamma \cdot m(z)$$ where $m(z)$ is the Stieltjes transform of the Marcenko-Pastur law.

So letting $\hat{K}(z) = S_{11} - S_{12} R(z) S_{21} - z I_\ell$, we have

$$\hat{K}(z) \xrightarrow{a.s.} K(z) = (1 - \gamma - z\gamma m(z)) \cdot \Sigma_1 - z I_\ell \quad \text{(entrywise)}$$

By same argument as in Wigner case, outlier eigenvalues $\hat{\lambda}$ converge to roots of $0 = \det K(z) \Leftrightarrow 0 = (1 - \gamma - z\gamma m(z)) \cdot \lambda_k - 1$,

$$\Leftrightarrow \lambda_k = z \cdot (1 - \gamma - z\gamma m(z))^{-1} \text{ for some } k.$$

Recall $m(z)$ satisfies $\gamma z m(z)^2 + (z + \gamma - 1) m(z) + 1 = 0$, so

$1 + z m(z) = (1 - \gamma - z\gamma m(z)) \cdot m(z)$. Then outliers $\hat{\lambda}$ converge to z

satisfying $\lambda_k = \frac{zm(z)}{1+zm(z)}$.

We have $zm(z) = \int \frac{z}{x-z}\, d\mu(x) = \int \left(-1 + \frac{x}{x-z}\right) d\mu(x) = -1 + \int \frac{x}{x-z}\, d\mu(x)$.

This is

- Strictly increasing in z over $(-\infty, (1-\sqrt{\gamma})^2)$ and $((1+\sqrt{\gamma})^2, \infty)$.
- Positive for $z \in (-\infty, (1-\sqrt{\gamma})^2)$, less than -1 for $z \in ((1+\sqrt{\gamma})^2, \infty)$.
- Approaches $-(1+\frac{1}{\sqrt{\gamma}})$ as $z \downarrow (1+\sqrt{\gamma})^2$

Then $z \mapsto \frac{zm(z)}{1+zm(z)}$ is

- Also increasing in z over $(-\infty, (1-\sqrt{\gamma})^2)$ and $((1+\sqrt{\gamma})^2, \infty)$.
- Less than 1 for $z \in (-\infty, (1-\sqrt{\gamma})^2)$, greater than 1 for $z \in ((1+\sqrt{\gamma})^2, \infty)$.
- Approaches $1+\sqrt{\gamma}$ as $z \downarrow (1+\sqrt{\gamma})^2$.

Since $\lambda_1 > \dots > \lambda_\ell > 1$, $\lambda_k = \frac{zm(z)}{1+zm(z)}$ has a root if and only if $\lambda_k > 1+\sqrt{\gamma}$. In this case, can check that the root is

$$z_k \equiv \lambda_k \left(1 + \frac{\gamma}{\lambda_k - 1}\right) \text{ as desired.}$$

Proof of Thm (c): Again, it suffices to consider $\Sigma = \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\ell & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix}$. Then the population eigenvectors $v_1, \dots, v_\ell$ are the standard basis vectors $e_1, \dots, e_\ell$.

For $\lambda_k > 1+\sqrt{\gamma}$: Write the eigenvector $\hat{v}$ for $\hat{\lambda}_k \equiv \hat{\lambda}$ as $(\hat{v}_1, \hat{v}_2)$, and

$$\hat{\lambda}\hat{v} = \hat{\Sigma}\hat{v} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \end{pmatrix}.$$

$$\Rightarrow 0 = (S_{11} - \hat{\lambda} I)\hat{v}_1 + S_{12}\hat{v}_2$$

$$0 = S_{21}\hat{v}_1 + (S_{22} - \hat{\lambda} I)\hat{v}_2$$

Since $S_{22} - \hat{\lambda} I$ is invertible a.s. for all large n, letting $R(z) = (S_{22} - zI)^{-1}$,

$\hat{v}_2 = -R(\hat{\lambda}) \cdot S_{21}\hat{v}_1$.

We deduce from this:

(i) $1 = \|\hat{v}\|^2 = \|\hat{v}_1\|^2 + \|\hat{v}_2\|^2 = \hat{v}_1^T (I + S_{12} R(\hat{\lambda})^2 S_{21}) \hat{v}_1$

(ii) $0 = (S_{11} - \hat{\lambda} I - S_{12} R(\hat{\lambda}) S_{21}) \hat{v}_1$.

Recall $\underbrace{S_{11} - zI - S_{12}R(z)S_{21}}_{\hat{K}(z)} \xrightarrow{a.s.} K(z) = (1-\gamma-z\gamma m(z))\Sigma_1 - zI_\ell$

$= \mathrm{diag}\left((1-\gamma-z\gamma m(z))\lambda_k - z\right)$

and $\hat{\lambda} \equiv \hat{\lambda}_k$ converges to the root z_k where $(1-\gamma-z\gamma m(z))\lambda_k - z = 0$.

Then $S_{11} - \hat{\lambda} I - S_{12}R(\hat{\lambda})S_{21} \xrightarrow{a.s.} \begin{pmatrix} d_1 & & & & \\ & \ddots & & & \\ & & d_{k-1} & & \\ & & & 0 & \\ & & & & d_{k+1} \\ & & & & & \ddots \\ & & & & & & d_\ell \end{pmatrix}$ for some non-zero $d_1, \dots, d_{k-1}, d_{k+1}, \dots, d_\ell$.

This and (ii) imply $e_j^T \hat{v}_1 \to 0$ for all $j \neq k$, so $\hat{v}_1 = \pm\|\hat{v}_1\| e_k + o(1)$.

Note that in (i), $I + S_{12}R(z)^2 S_{21} = -\hat{K}'(z)$. By a Cauchy integral argument (as in the Wigner case), $\hat{K}'(z) \xrightarrow{a.s.} K'(z)$ entrywise.

Let $\lambda(z) = \frac{zm(z)}{1+zm(z)} \Rightarrow zm(z) = \frac{\lambda(z)}{1-\lambda(z)} = -1 + \frac{1}{1-\lambda(z)}$

Then $\frac{d}{dz}[zm(z)] = \lambda'(z) \cdot \frac{1}{(1-\lambda(z))^2}$. Also differentiating $z = \lambda(z)\left(1 + \frac{\gamma}{\lambda(z)-1}\right)$,

$1 = \left(1 + \frac{\gamma}{\lambda(z)-1} - \frac{\gamma\lambda(z)}{(\lambda(z)-1)^2}\right)\lambda'(z) = \left(1 - \frac{\gamma}{(\lambda(z)-1)^2}\right)\lambda'(z)$. So

$\frac{d}{dz}[zm(z)] = \left(1 - \frac{\gamma}{(\lambda(z)-1)^2}\right)^{-1} \cdot \frac{1}{(\lambda(z)-1)^2}$. Applying $\lambda(z_k) = \lambda_k$,

$$K'(z_k)_{kk} = -\frac{\gamma\lambda_k}{(\lambda_k-1)^2}\left(1 - \frac{\gamma}{(\lambda_k-1)^2}\right)^{-1} - 1 = -\left(1 + \frac{\gamma}{\lambda_k-1}\right) \Big/ \left(1 - \frac{\gamma}{(\lambda_k-1)^2}\right).$$

From $\hat{v}_1 = \pm\|\hat{v}_1\|e_k + o(1)$ and (i),

$$1 = \|\hat{v}_1\|^2 (-\hat{K}'(\hat{\lambda}))_{kk} + o(1) = \|\hat{v}_1\|^2 \cdot \left(1 + \frac{\gamma}{\lambda_k-1}\right) \Big/ \left(1 - \frac{\gamma}{(\lambda_k-1)^2}\right) + o(1)$$

$\Rightarrow \|\hat{v}_1\|^2 \to \left(1-\frac{\gamma}{(\lambda_k-1)^2}\right) / \left(1+\frac{\gamma}{\lambda_k-1}\right).$

Thus: $|e_j^T \hat{v}|^2 \to \begin{cases} 0 & j \in \{1,\dots,\ell\} \setminus \{k\} \\ \dfrac{1-\frac{\gamma}{(\lambda_k-1)^2}}{1+\frac{\gamma}{\lambda_k-1}} & j=k. \end{cases}$

Finally, observe that the distribution of $\hat{v}_2$ is rotationally invariant on $\mathbb{R}^{p-\ell}$, because the law of $\hat{\Sigma}$ doesn't change under the transformation $\hat{\Sigma} \mapsto \begin{pmatrix} I & 0 \\ 0 & Q \end{pmatrix} \hat{\Sigma} \begin{pmatrix} I & 0 \\ 0 & Q^T \end{pmatrix}$ for any orthogonal $Q \in \mathbb{R}^{(p-\ell)\times(p-\ell)}$.

Then for any unit vector $u \in \operatorname{span}(e_{\ell+1},\dots,e_p)$, $\mathbb{E}[|u^T \hat{v}|^2] = \mathbb{E}[\|\hat{v}_2\|^2] \cdot \frac{1}{p-\ell} \leq \frac{1}{p-\ell}$.

By a concentration of measure argument, $|u^T \hat{v}|^2 \xrightarrow{a.s.} 0$.

Combining with $|e_j^T \hat{v}|^2 \to 0$ for all $j \neq k$, we get $|u^T \hat{v}|^2 \xrightarrow{a.s.} 0$ for any deterministic unit vector u orthogonal to $v_k \equiv e_k$.

Proof sketch of Thm (b): We first make the preceding arguments more quantitative:

Let $\hat{K}(z) = S_{11} - zI_\ell - S_{12}(S_{22}-zI)^{-1}S_{21}$

$$= \frac{1}{n}\Sigma_1^{1/2} X_1^T \underbrace{\left(I - \frac{1}{n}X_2\left(\frac{1}{n}X_2^T X_2 - zI\right)^{-1} X_2^T\right)}_{M} X_1 \Sigma_1^{1/2} - zI_\ell$$

$$K(z) = \Sigma_1(1-\gamma-z\gamma m(z)) - zI_\ell.$$

Then $$\hat{K}(z) - K(z) = \underbrace{\left(\frac{1}{n}\Sigma_1^{1/2}X_1^T M X_1 \Sigma_1^{1/2} - \Sigma_1 \cdot \frac{1}{n}\operatorname{Tr} M\right)}_{A(z)} + \underbrace{\left(\Sigma_1 \frac{1}{n}\operatorname{Tr} M - \Sigma_1 \cdot (1-\gamma-z\gamma m(z))\right)}_{B(z)}.$$

Hanson-Wright in fact gives $A(z) \prec \frac{1}{\sqrt{n}}$.

For z such that $\operatorname{dist}(z, [(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]) > \delta$, a more detailed analysis of the Stieltjes transform shows $|m_n(z) - m(z)| \prec \frac{1}{n}$, so $B_n \prec \frac{1}{n} + o(\frac{p}{n} - \gamma)$.

Then $\|\hat{v}_1/\|\hat{v}_1\| - e_k\| \prec \frac{1}{\sqrt{n}}$, $|\hat{\lambda}_k - z_k| \prec \frac{1}{\sqrt{n}}$ where $z_k = \lambda_k(1+\frac{\gamma}{\lambda_k-1})$.

Note $\hat{K}(\hat{\lambda}_k)\hat{v}_1 = 0$, and $K(z_k)e_k = 0$. Then on the one hand,

$$e_k^T \hat{K}(\hat{\lambda}_k)e_k = e_k^T(\hat{K}(\hat{\lambda}_k) - \hat{K}(z_k))e_k + \underbrace{e_k^T(\hat{K}(z_k) - K(z_k))e_k}_{e_k^T A(z_k) e_k + O_\prec(\frac{1}{n})}.$$

On the other hand,

$$e_k^T \hat{K}(\hat{\lambda}_k) e_k = \left(\frac{\hat{v}_1}{\|\hat{v}_1\|} - e_k\right)^T \hat{K}(\hat{\lambda}_k)\left(\frac{\hat{v}_1}{\|\hat{v}_1\|} - e_k\right) \prec \frac{1}{n}.$$

So $(\hat{K}(\hat{\lambda}_k) - \hat{K}(z_k))_{kk} + A(z_k)_{kk} \prec \frac{1}{n}$.

Apply Taylor expansion: $\hat{K}(\hat{\lambda}_k) - \hat{K}(z_k) = (\hat{\lambda}_k - z_k)\cdot \hat{K}'(z_k) + O_\prec(\frac{1}{n})$

$$= (\hat{\lambda}_k - z_k)\cdot K'(z_k) + O_\prec(\tfrac{1}{n}).$$

So: $(\hat{\lambda}_k - z_k)\cdot K'(z_k)_{kk} + A(z_k)_{kk} \prec \frac{1}{n}$, and

$$\sqrt{n}(\hat{\lambda}_k - z_k) = \sqrt{n}\,A(z_k)_{kk} / (-K'(z_k)_{kk}) + O_\prec(\tfrac{1}{\sqrt{n}}).$$

From part (a), $-K'(z_k)_{kk} = \left(1+\frac{\gamma}{\lambda_k-1}\right)/\left(1-\frac{\gamma}{(\lambda_k-1)^2}\right)$.

We apply a CLT for $A(z)_{kk}$:

Lemma: If $w \in \mathbb{R}^n$ has iid $N(0,1)$ entries, and $A \in \mathbb{R}^{n\times n}$ satisfies $\|A\|/\|A\|_F \to 0$, then

$$(w^T A w - \operatorname{Tr} A)\cdot \|A\|_F^{-1} \to N(0,2).$$

Proof: By diagonalizing A and rotational invariance of w, we may assume A is diagonal. Then $w^T A w - \operatorname{Tr} A = \sum_{i=1}^n A_{ii}(w_i^2 - 1)$.

This has variance $\sum_{i=1}^n A_{ii}^2 \operatorname{Var}[w_i^2 - 1] = 2\|A\|_F^2$.

Then the result follows from the Lindeberg/Lyapunov CLT.

Note that $A(z)_{kk} = \lambda_k \frac{1}{n}\left(w^T M w - \operatorname{Tr} M\right)$, where w is column k of X_1.

Applying the Lemma to $A = \frac{\lambda_k M}{\sqrt{n}}$, we get

$$\sqrt{n}\, A(z)_{kk} \cdot \left\|\frac{\lambda_k M}{\sqrt{n}}\right\|_F^{-1} \to N(0,2).$$

Finally, $\frac{1}{n}\|M\|_F^2 = \frac{1}{n}\operatorname{Tr} M^2 \to \frac{\left(1+\frac{\gamma}{\lambda_k-1}\right)^2}{1-\frac{\gamma}{(\lambda_k-1)^2}}$ by a similar computation as for $\frac{1}{n}\operatorname{Tr} M$. The additional ingredient is

$$\frac{1}{n}\operatorname{Tr}\left(S_{22}(S_{22}-zI)^{-1}S_{22}(S_{22}-zI)^{-1}\right)$$

$$\xrightarrow{a.s.} \gamma\int \frac{x^2}{(x-z)^2}\,d\mu(x) = \gamma\int\left(1+\frac{2z}{x-z}+\frac{z^2}{(x-z)^2}\right)d\mu(x)$$

$$= \gamma\left(1+2z\,m(z)+z^2 m'(z)\right),$$

with $z_k m(z_k) = \frac{\lambda_k}{1-\lambda_k}$ and a computation for $m'(z_k)$ as before.

Combining these, $\sqrt{n}(\hat{\lambda}_k - z_k) \xrightarrow{L} N(0,\sigma^2)$ where

$$\sigma^2 = 2\cdot\lambda_k^2\cdot\frac{\left(1+\frac{\gamma}{\lambda_k-1}\right)^2}{1-\frac{\gamma}{(\lambda_k-1)^2}}\cdot\left(\frac{1-\frac{\gamma}{(\lambda_k-1)^2}}{1+\frac{\gamma}{\lambda_k-1}}\right)^2 = 2\lambda_k^2\left(1-\frac{\gamma}{(\lambda_k-1)^2}\right).$$

Problem 7. (a) Implement a method that estimates principal eigenvalues of Σ in the spiked covariance model, based on observing X, and also provides 95% confidence intervals in the setting of real Gaussian data.

Note that $m(z)$ may be computed from its analytic expression, or estimated as $\frac{1}{p}\sum_i \frac{1}{\hat{\lambda}_i - z}$ where the sum is over the "bulk" eigenvalues $\hat{\lambda}_i$ of $\hat{\Sigma}$.

Are these the same? Which leads to more accurate estimates/confidence intervals in practice?

(b) How does the coverage of the confidence intervals change with the number of spike eigenvalues and their spacings? With p/n?

(c) Extend your implementation to data where X_{ij} are not Gaussian. (You will need to find the expressions for the limiting normal distribution of the outlier eigenvalues in the literature.) This should include a method of estimating unknown parameters in the normal limit, e.g. the kurtosis of X_{ij}.