# Edgeworth and confidence interval correction in spiked PCA

Iain Johnstone & Jeha Yang

Statistics & Biomedical Data Science, Stanford & Two Sigma

Shanghai, December 10, 2019

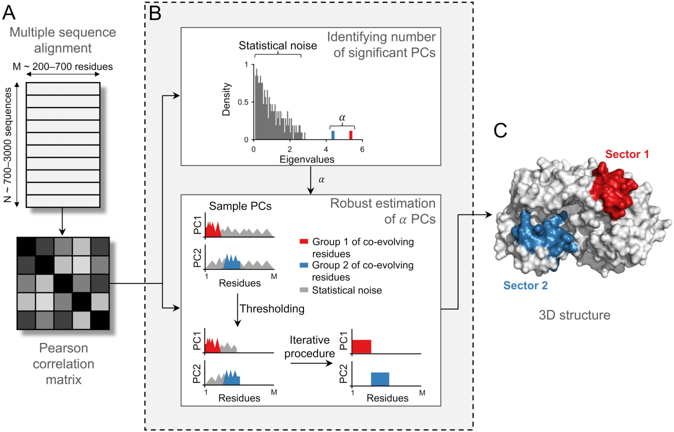
# Edgeworth and confidence interval correction in spiked PCA

Iain Johnstone & Jeha Yang

Statistics & Biomedical Data Science, Stanford & Two Sigma

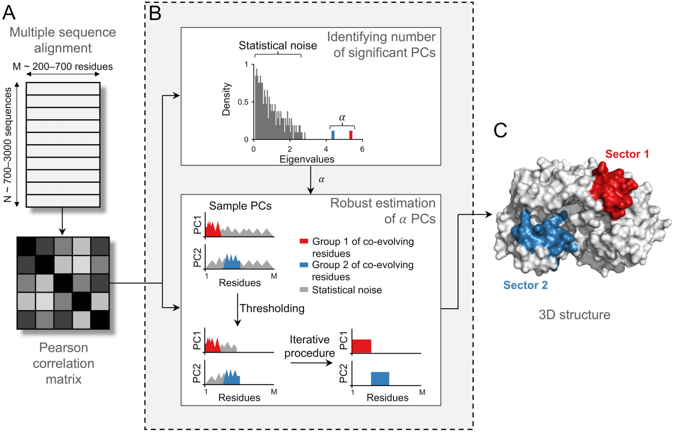
Shanghai, December 10, 2019

# Viral protein mutations and spiked models

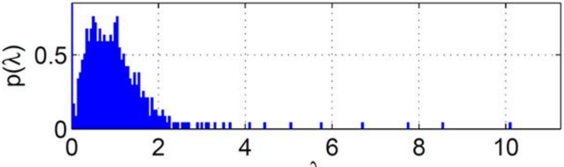



Quadeer et. al. PLOS Comp. Bio. 2018

# Viral protein mutations and spiked models




Quadeer et. al. PLOS Comp. Bio. 2018

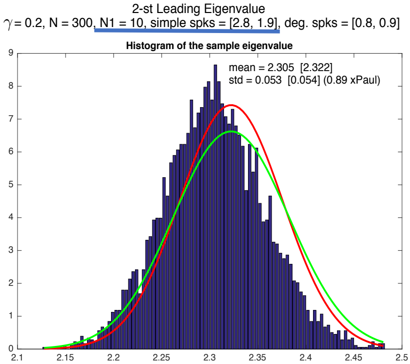


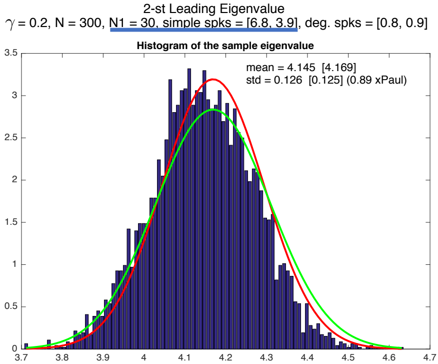
# A suggestive simulation on correlation matrices

[David Morales, Matt McKay]

$\rho_1 = 0.2$ ; $\rho_2 = 0.1$

2[nd] eigenvalue




Theoretical variance is pretty accurate, but there seems to be a shift in the mean (similar to what we've seen before in the eigenvector projections of sample covariance when spikes were close to each other)

# Outline

# High dimensional spiked PCA model

- Data : $X = [x_1 \cdots x_n]'$ with

$$x_1, \cdots, x_n \overset{i.i.d.}{\sim} N_{p+1}(0, \Sigma)$$

- Large dimensional asymptotic regime : as $n \to \infty$,

$$\gamma_n := p/n \to \gamma \in (0, \infty)$$

- Spiked eigenstructure of $\Sigma$ : for a fixed $r$,

$$\underbrace{\ell_1 > \cdots > \ell_r}_{Spikes} > 1 = \ell_{r+1} = \cdots = \ell_{p+1}$$

- Statistics : eigenvalues of sample covariance matrix $X'X/n$

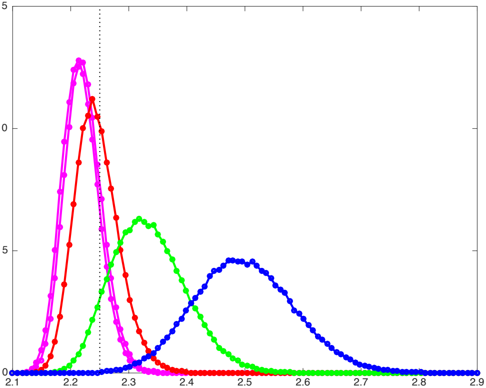
$$\hat{\rho}_1 \geq \cdots \geq \hat{\rho}_{p+1}$$

$\to$ w.l.o.g. $\Sigma$ is diagonal

# Largest Eigenvalue $\hat{\rho}_1$: Numerical illustration

$p = 200, n = 800 \qquad$ [i.e. $\gamma_n = p/n = 0.25$]

| | | subcritical | critical | supercritical |
|---|---|---|---|---|
| Spike | $h = \ell - 1$ : | 0, 0.25, | $h_+ = 0.5$, | 0.75, 1. |

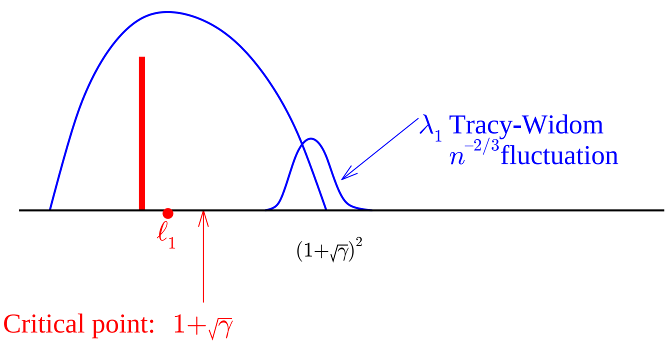


# Finite rank model, $K = 1$: phase transition

$\Sigma = \mathrm{diag}(\ell_1, 1, \ldots, 1)$ $\quad$ $\boxed{p/n \to \gamma}$.

Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$:

[Baik–Ben Arous–Peché,05]

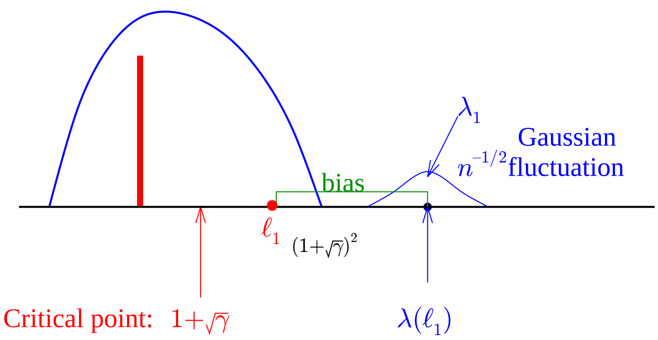


# Finite rank model, $K = 1$: phase transition

$\Sigma = \text{diag}(\ell_1, 1, \ldots, 1)$ $\boxed{p/n \to \gamma}$.

Interior point transition at $\ell_1 = 1 + \sqrt{\gamma}$:

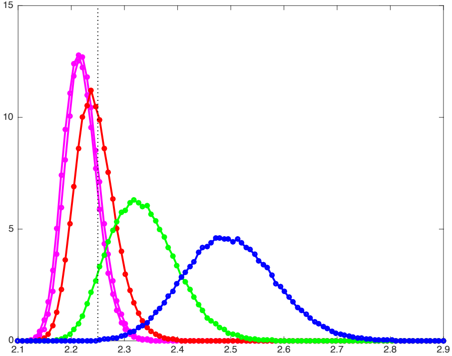
[Baik–Ben Arous–Peché,05]



# Largest Eigenvalue $\hat{\rho}_1$: Numerical illustration

$$p = 200, n = 800 \qquad [\text{i.e. } \gamma_n = p/n = 0.25]$$

| | | subcritical | critical | supercritical |
|---|---|---|---|---|
| Spike | $h =$ | 0, 0.25, | $h_+ = 0.5$, | 0.75, 1. |




Edge: $(1 + \sqrt{\gamma_n})^2 = 2.25$

# Largest eigenvalue: Phase transition

Different rates, limit distributions:

$$\text{For } h < \sqrt{\gamma}: \qquad n^{2/3}\left[\frac{\hat{\rho}_1 - \mu(\gamma_n)}{\tau(\gamma_n)}\right] \quad \stackrel{\mathcal{D}}{\Rightarrow} \quad TW_\beta,$$

$$\text{For } h > \sqrt{\gamma}: \qquad n^{1/2}\left[\frac{\hat{\rho}_1 - \rho(h, \gamma_n)}{\sigma(h, \gamma_n)}\right] \quad \stackrel{\mathcal{D}}{\Rightarrow} \quad N(0, 1)$$

# Largest eigenvalue: Phase transition

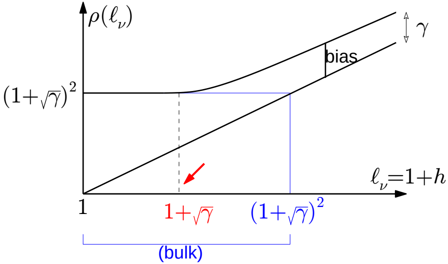
Different rates, limit distributions:

$$\text{For } h < \sqrt{\gamma}: \qquad n^{2/3}\left[\frac{\hat{\rho}_1 - \mu(\gamma_n)}{\tau(\gamma_n)}\right] \overset{\mathcal{D}}{\Rightarrow} TW_\beta,$$

$$\text{For } h > \sqrt{\gamma}: \qquad n^{1/2}\left[\frac{\hat{\rho}_1 - \rho(h,\gamma_n)}{\sigma(h,\gamma_n)}\right] \overset{\mathcal{D}}{\Rightarrow} N(0,1)$$

with

$$\rho(h,\gamma) = (1+h)\left(1+\frac{\gamma}{h}\right) \qquad \sigma^2(h,\gamma) = 2(1+h)^2\left(1-\frac{\gamma}{h^2}\right)$$




Statistical physics lit, 94- Baik-Ben Arous-Peche(05) , Paul (07) Baik-Silverstein (06), Bloemendal-Virag (11) Mo (11) , Wang (12) Benaych-Georges-Guionnet-Maida (11)

# Normal approximation – multiple spikes

- Assume that all spikes are simple, supercritical :

$$\ell_1 > \cdots > \ell_r > 1 + \sqrt{\gamma}$$

- Asymptotic mutual independence:

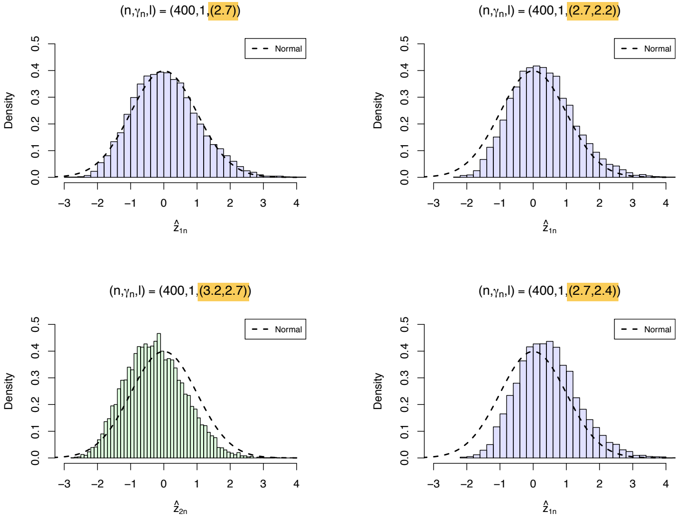
$$\text{with } \rho_{kn} := \rho(\ell_k, \gamma_n), \quad \sigma_{kn} := \sigma(\ell_k, \gamma_n),$$

$$(\hat{z}_{kn})_{k=1,\cdots,r} := \left( n^{1/2} \frac{(\hat{\rho}_k - \rho_{kn})}{\sigma_{kn}} \right)_{k=1,\cdots,r} \Rightarrow N(0, I_r)$$

Shi (2013)

# Edgeworth approximations

# Inaccuracy of approximations : $\hat{z}_{kn}$ associated with $\ell_k = 2.7$



# Traditional Edgeworth

(Smooth function of) means model: Petrov, 1975, Hall, 1992

$$S_n = \frac{1}{\sqrt{n\kappa_{2n}}} \sum_{i=1}^{n} X_{ni} \qquad \text{indep, mean 0, } \in \mathbb{R}^d, \quad d \textbf{ fixed}$$

$$\kappa_{jn} = \frac{1}{n} \sum_{1}^{n} \mathbb{E} X_{ni}^j \qquad \text{moments}$$

First order expansion:

$$\mathbb{P}(S_n \le x) = \Phi(x) + n^{-1/2} p(x) \phi(x) + o(n^{-1/2})$$

$$p(x) = \frac{-\kappa_{3n}}{\kappa_{2n}^{3/2}} \frac{H_2(x)}{6}, \qquad H_2(x) = x^2 - 1.$$

skewness correction

# Single spike, first order expansion for $\hat{\rho}_1$

$$\hat{z}_{1n} = n^{1/2}(\hat{\rho}_1 - \rho_{1n})/\sigma_{1n}$$

**Theorem** In spiked model, $h_1 = \ell_1 - 1 > \sqrt{\gamma}$, $\gamma_n = p/n$,

$$\mathbb{P}\left(\hat{z}_{1n} \leq x\right) = \Phi(x) + n^{-1/2} p_{1n}(x)\phi(x) + o(n^{-1/2}),$$

uniformly in $x \in \mathbb{R}$, with

$$\begin{aligned} p_{1n}(x) &= -\alpha_{2n} H_2(x) - \alpha_{0n} \\ \alpha_{2n} &= \alpha_2(h_1, \gamma_n) = \frac{\sqrt{2}}{3} \frac{h_1^3 + \gamma_n}{(h_1^2 - \gamma_n)^{3/2}}, \\ \alpha_{0n} &= \alpha_0(h_1, \gamma_n) = \frac{\gamma_n}{\sqrt{2}} \frac{h_1 + 1}{(h_1^2 - \gamma_n)^{3/2}} \end{aligned}$$

# Coefficients of Edgeworth expansion for single-spike

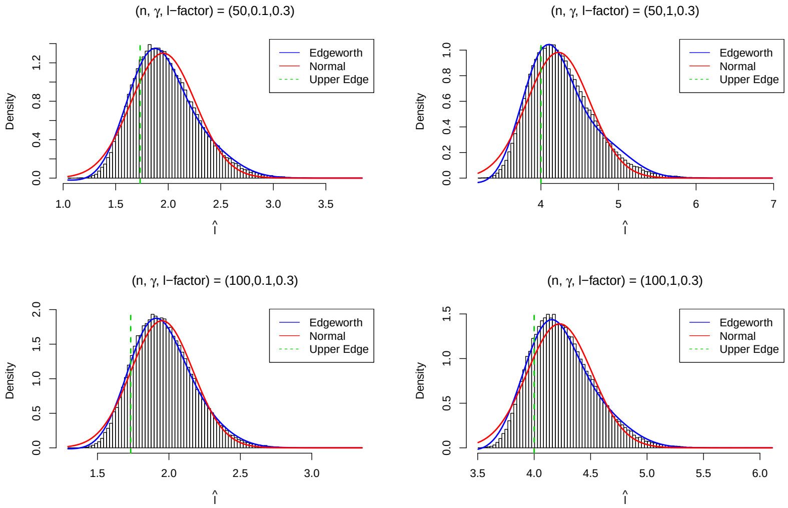
$$\alpha_2(h_1,\gamma_n) = \frac{\sqrt{2}}{3}\frac{h_1^3+\gamma_n}{(h_1^2-\gamma_n)^{3/2}}, \qquad \alpha_0(h_1,\gamma_n) = \frac{\gamma_n}{\sqrt{2}}\frac{h_1+1}{(h_1^2-\gamma_n)^{3/2}}$$

- Larger for "harder" cases i.e. larger $\gamma$ and smaller $h$ $(> \sqrt{\gamma})$
- Larger than the fixed $p$ case i.e. $\gamma = 0,\ \alpha_2 = \sqrt{2}/3,\ \alpha_0 = 0$

  Muirhead-Chikuse (1975)
- Empirically reasonable if

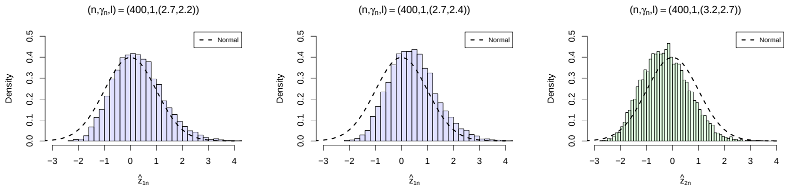
$$\frac{9}{2}\frac{\alpha_2^2}{n} = \frac{(h_1^3+\gamma)^2}{n(h_1^2-\gamma)^3} \le 0.2$$

# Single Spike Simulation



# Edgeworth for multiple spikes

# Eigenvalues are repulsive!




- joint density of $(\hat{\rho}_1, \cdots, \hat{\rho}_{n \wedge (p+1)})$ has a Jacobian factor

$$\prod_{i<j} |\hat{\rho}_i - \hat{\rho}_j|$$

$\rightarrow$ pushes eigenvalues apart

- **But,** not visible at leading order (for supercritical spikes:)

$$(\hat{z}_{kn})_{k=1,\cdots,r} \Rightarrow N(0, I_r)$$

# Multi spike, first order expansion for $\hat{\rho}_k$

$$\hat{z}_{kn} = n^{1/2}(\hat{\rho}_k - \rho_{kn})/\sigma_{kn}$$

**Theorem** In spiked model, $h_k = \ell_k - 1 > \sqrt{\gamma}$, $\gamma_n = p/n$,

$$\mathbb{P}\left(\hat{z}_{kn} \le x\right) = \Phi(x) + n^{-1/2} p_{kn}(x)\phi(x) + o(n^{-1/2}),$$

uniformly in $x \in \mathbb{R}$, with

$$p_{kn}(x) = -\alpha_2(h_k, \gamma_n) H_2(x) - \alpha_{0,k}(\boldsymbol{h}, \gamma_n)$$

$$\alpha_2(h_k, \gamma_n) = \frac{\sqrt{2}}{3} \frac{h_k^3 + \gamma_n}{(h_k^2 - \gamma_n)^{3/2}},$$

$$\alpha_{0,k}(\boldsymbol{h}, \gamma) = \frac{1}{\sqrt{2}} \frac{h_k + 1}{(h_k^2 - \gamma)^{1/2}} \left[ \frac{\gamma}{h_k^2 - \gamma} + \sum_{j \neq k} \frac{h_j}{h_k - h_j} \right]$$

# Interpretation

Edgeworth corrected density

$$\phi + n^{-1/2}(\alpha_2 H_3 + \alpha_0 H_1)\phi$$

Relative to single spike case: $\alpha_2$ unchanged, but

$$\Delta\alpha_0 = \alpha_{0,k}(\boldsymbol{h}, \gamma_n) - \alpha_0(h_k, \gamma_n) = \frac{1}{\sqrt{2}} \frac{h_k + 1}{(h_k^2 - \gamma_n)^{1/2}} \sum_{j \neq k} \frac{h_j}{h_k - h_j}$$

- $\Delta\alpha_0 > 0$, e.g. smaller spikes $h_j < h_k$, push density to right,

  conversely for $\Delta\alpha_0 < 0$
- closer spikes $\Rightarrow$ larger effect
- additive in $\ell_j$, $j \neq k$

# Repulsion example 1 : $\hat{z}_{kn}$ associated with $\ell_k = 2.7$

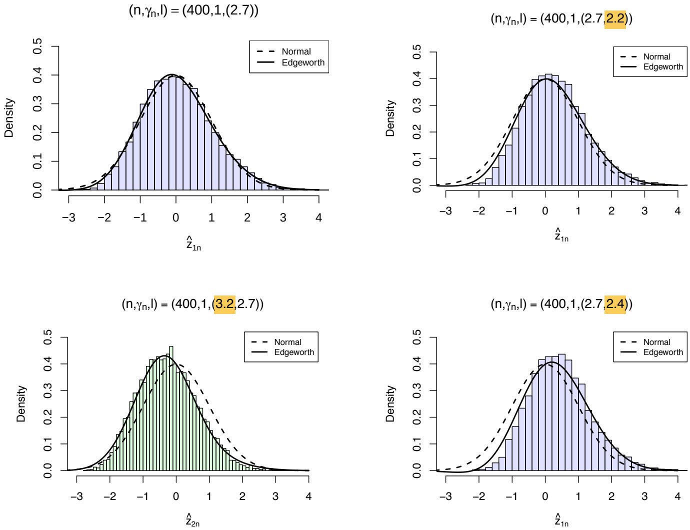
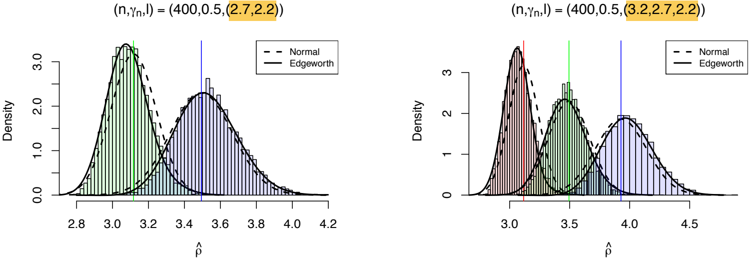



Figure: Density of $\hat{z}_{kn}$ associated with $\ell_k = 2.7$

# Repulsion example 2 : histograms of $(\hat{\rho}_k)_{k=1,\cdots,r}$ together




Blue, red, green vertical lines correspond to $\rho_{1n}, \rho_{2n}, \rho_{3n}$, respectively.

# Explaining the Repulsion Correction

# Perturbation setup

Recall $\quad \ell_1 > \cdots > \ell_r > 1 + \sqrt{\gamma} > 1 = \ell_{r+1} = \cdots = \ell_{p+1}$

Focus on $\ell_k$: $\quad n^{-1} X' X v_k = \hat{\rho}_k v_k,$

$$\hat{\rho}_k \to \rho_{kn} = \rho(\ell_k, \gamma_n) = \ell_k + \gamma \frac{\ell_k}{\ell_k - 1}$$

Permute columns:

$$X = [\sqrt{\ell_k} Z_1,\ Z_2 \Sigma_2^{1/2}] \qquad \Sigma_2 = \mathrm{diag}(\boldsymbol{\ell}_{(k)}, 1, \cdots, 1)$$

Population eigenvalues of $\Sigma_2$:

$$H_{(k)} = \left(1 - \frac{r-1}{p}\right) \delta_1 + \frac{1}{p} \sum_{j \neq k} \delta_{\ell_j} = \delta_1 + p^{-1} H^{\Delta}$$

# Standard first steps

$$n^{-1}X'Xv_k = \hat{\rho}_k v_k \qquad X = [\sqrt{\ell_k}Z_1,\ Z_2\Sigma_2^{1/2}]$$

$$n^{-1}Z_2\Sigma_2\Sigma_2' = U\Lambda U' \qquad U \in O(n),\ \Lambda = \mathrm{diag}(\lambda_1 \geq \cdots \lambda_n)$$

$$z = U'Z_1 \sim N(0, I_n) \quad z \perp\!\!\!\perp \Lambda \qquad \text{(Gaussian assumptions!)}$$

Schur complement, Woodbury formula, resolvent,..

$$R(x) = (\Lambda - xI_n)^{-1}$$

$\Rightarrow$ Key equation:

$$(\hat{\rho}_k - \rho_{kn})[1 + \ell_k n^{-1} z' \tilde{R}_{kn} z] = -\ell_k \rho_{kn}[n^{-1} z' R(\rho_{kn}) z + \ell_k^{-1}]$$

# The Forward Map $H \to \mathsf{F}_{\gamma,H}$

Silverstein equation: $H$ probability measure on $\mathbb{R}$, $\gamma > 0$,

$$z(\mathsf{m}) = -\frac{1}{\mathsf{m}} + \gamma \int \frac{t}{1 + t\mathsf{m}} dH(t), \qquad \mathsf{m} \in \mathbb{C}^+$$

$z(\mathsf{m}) = z$ has unique solution $\mathsf{m}(z)$ for $z \in \mathbb{C}^+$, and

$$\mathsf{m}(z) = \int \frac{1}{\lambda - z} d\mathsf{F}(\lambda) = m_{\mathsf{F}}(z)$$

defines (Stieltjes transform of) a probability distribution $\mathsf{F} = \mathsf{F}_{\gamma,H}$.

| | | |
|---|---|---|
| Population: | $\Sigma_p$ | $H_p = F^{\Sigma_p} = \frac{1}{p} \sum \delta_{\sigma_i}$ |
| Sample: | $\mathsf{B}_n = n^{-1} Z_p \Sigma_p Z_p'$ | $F^{\mathsf{B}_n} = \frac{1}{n} \sum \delta_{\lambda_i}$ |

If $H_p \Rightarrow H, \; p/n \to \gamma$ $\qquad F^{\mathsf{B}_n} \Rightarrow \mathsf{F}_{\gamma,H}$

(Marcenko-Pastur-Bai-Silverstein)

# Stochastic Decomposition

$$n^{-1} z' f(\Lambda) z = n^{-1} \sum f(\lambda_i) z_i^2 = n^{-1} \sum f(\lambda_i) + n^{-1/2} S_n(f)$$

$$S_n(f) = n^{-1/2} \sum f(\lambda_i)(z_i^2 - 1) \qquad (\Lambda \perp\!\!\!\perp z)$$

$$n^{-1} \sum f(\lambda_i) = \int f(\lambda_i)\, d\mathrm{F}_{\gamma_n, H_n}(\lambda) \quad + n^{-1} \left[ \sum f(\lambda_i) - n \int f\, d\mathrm{F}_{\gamma_n, H_n} \right]$$

$$= \mathrm{F}_{\gamma_n, H_n}(f) \qquad + n^{-1} G_n(f)$$

deterministic equiv. Bai-Silverstein CLT

$\Rightarrow$

$$n^{-1} z' f(\Lambda) z = \mathrm{F}_{\gamma_n, H_n}(f) + n^{-1/2} S_n(f) + n^{-1} G_n(f)$$

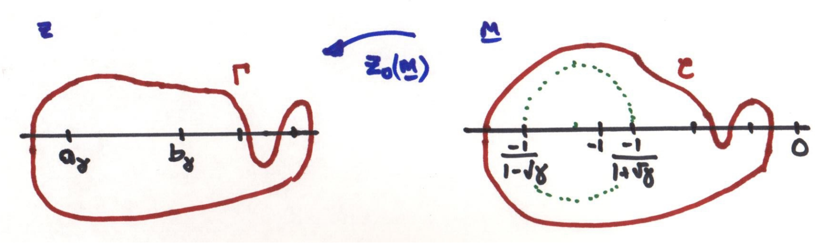
# Perturbing the centering

$H = \delta_1 + p^{-1} H^{\Delta}$ From Wang-Silverstein-Yao, 2014

$$\mathrm{F}_{\gamma,H}(f) = \mathrm{F}_{\gamma}(f) + n^{-1} A(f) + O(n^{-2})$$

$$A(f) = \frac{1}{2\pi i} \int_{\mathcal{C}} f(z_0(\mathrm{m}) w(\mathrm{m}) d\mathrm{m}$$

$$z_0(\mathrm{m}) = -\frac{1}{\mathrm{m}} + \frac{\gamma}{1+\mathrm{m}} \qquad w(\mathrm{m}) = \int \frac{t}{1+t\mathrm{m}} dH^{\Delta}(t)$$



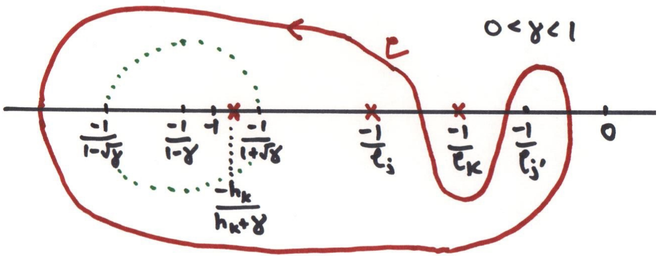
# Evaluating $A_n(g_{kn})$

In WSY 14, set $H \leftarrow H_{(k)n} = \delta_1 + \frac{1}{p}\sum_{j\neq k}(\delta_{\ell_j} - \delta_1)$

$f(z) \leftarrow g_{kn}(z) = (\rho_{kn} - z)^{-1}, \quad w(\mathrm{m}) = \sum_{j\neq k}\left(\frac{\ell_j}{1+\ell_j \mathrm{m}} - \frac{1}{1+\mathrm{m}}\right)$

$$A_n(g_{kn}) = \frac{1}{2\pi i}\int_{\mathcal{C}} \sum_{j\neq k} t_j(\mathrm{m})d\mathrm{m} = \frac{h_k}{(h_k^2 - \gamma)} \sum_{j\neq k} \frac{h_j}{h_k - h_j}$$

repulsion term



# Back to $n^{-1} z' R(\rho_{kn}) z$

$$-R(\rho_{kn}) = -(\Lambda - \rho_{kn} I_n)^{-1} = g_{kn}(\Lambda)$$

**Decomposition:**

$$-n^{-1} z' R(\rho_{kn}) z \approx \mathsf{F}_{\gamma_n}(g_{kn}) + n^{-1/2} S_n(g_{kn}) + n^{-1} D_n(g_{kn})$$

$$D_n(g_{kn}) = G_n(g_{kn}) + A_n(g_{kn}) + O(n^{-1})$$

$$= \tilde{\alpha}_{0,k}(\boldsymbol{h}, \gamma_n) + Z_{kn},$$

since, from Bai-Silverstein CLT

$$G_n(g_{kn}) = \mu_{\gamma_n}(g_{kn}) + Z_{kn}, \qquad \mu_{\gamma_n}(g_{kn}) = \frac{\gamma_n h_k}{(h_k^2 - \gamma_n)^2}$$

bulk term

# Key linearization

$$\hat{z}_{kn} = \frac{n^{1/2}(\hat{\rho}_k - \rho_{kn})}{\sigma_{kn}} \approx \frac{S_n(g_{kn}) + n^{-1/2} D_n(g_{kn})}{\sigma_{kn} \mathsf{F}_{\gamma_n}(g_{kn}^2) + h.o.t.}$$

Delta method for Edgeworth expansion, + conditioning

$$\mathbb{P}\{\hat{z}_{kn} \leq x\} = \mathbb{E}\Big\{\mathbb{P}\{S_n(g_{kn}) \leq y_n(x) | \Lambda\}\Big\} + o(n^{-1/2})$$

Final steps:

- Edgeworth expansion (conditional on $\Lambda$)
- uncondition; identify terms

# Edgeworth (conditional on $\Lambda$ )

$$S_n(g_{kn}) = n^{-1} \sum X_{ni} \qquad X_{ni} = c_{ni}(z_i^2 - 1) \qquad c_{ni} = g_{kn}(\lambda_i)$$

From e.g. Petrov 1975, n.i.d. case:

$$\mathbb{P}\Big\{ \frac{1}{\bar\kappa_{2n}\sqrt{n}} \sum X_{ni} \le y \Big| \Lambda \Big\} = \Phi(y) - \frac{\bar\kappa_{jn}}{\bar\kappa_{2n}^{3/2}\sqrt{n}} \frac{H_2(y)}{6} \phi(y) + o(n^{-1/2})$$

Cumulants: $\bar\kappa_{jn} = \kappa_j n^{-1} \sum_1^n c_{ni}^j = \kappa_j \mathsf{F}_{\gamma_n}(g_{kn}^j) + O(n^{-1/2})$

quadratic term:

$$\frac{\bar\kappa_{jn}}{\bar\kappa_{2n}^{3/2}} = \frac{\sqrt{2}}{3} \frac{h_k^3 + \gamma_n}{(h_k^2 - \gamma_n)^{3/2}} + O(n^{-1/2}) = \alpha_2(h_k, \gamma_n) + O(n^{-1/2})$$

# Assembling pieces

$$\mathbb{P}\{\hat{z}_{kn} \le x\} = \mathbb{E}\left[\Phi(y_n) - \frac{\alpha_2(h_k,\gamma_n)}{\sqrt{n}}\frac{H_2(y_n)}{6}\phi(y_n) + o(n^{-1/2})\right]$$

$$y_n = y_n(x) = x - \frac{1}{\bar{\kappa}_{2n}\sqrt{n}}D_n(g_{kn}) \qquad \text{repulsive shift}$$

$$= x - \frac{1}{\bar{\kappa}_{2n}\sqrt{n}}[\tilde{\alpha}_{0,k}(\boldsymbol{h},\gamma_n) + Z_{kn}]$$

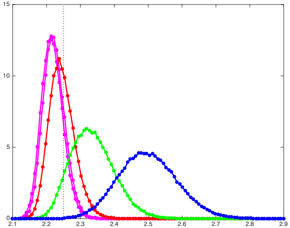
$$\mathbb{E}\Phi(y_n) \approx \Phi(x) - \frac{\alpha_{0k}(\boldsymbol{h},\gamma_n)}{\sqrt{n}}\phi(x)$$

Final result:

$$\mathbb{P}\{\hat{z}_{kn} \le x\} = \Phi(x) - \frac{1}{\sqrt{n}}\left[\alpha_2(h_k,\gamma_n)\frac{H_2(y)}{6} + \alpha_{0k}(\boldsymbol{h},\gamma_n)\right]\phi(x) + o(n^{-1/2})$$

# Confidence intervals after selection

# Inference for supercritical spikes




Below bulk edge:

Even for supercritical $\ell_k$,

$\mathbb{P}\{\hat{\rho}_k < b(\gamma)\}$ can be significant!

$\rightarrow$ **inference after selection of supercritical spikes**

Selection rule: select all $\hat{\rho}_k, k = 1, \cdots, \hat{r}$ such that

$$\hat{\rho}_k > \theta_n := b(\gamma_n) + n^{-1/3}\sqrt{\gamma_n}$$

Consistent: $\mathbb{P}\left(\hat{r} = r\right) = 1 - o(n^{-m}),\ m \in \mathbb{N}$

*Minimal conditioning*: Liu-Markovic-Tibshirani (2018)

$$\hat{\rho}_k \mid \hat{\rho}_k > \theta_n$$

# Pivots

Exact distribution of $\hat{\rho}_k$:

$$\overline{F}_{kn}(x, \ell) = \mathbb{P}_\ell(\hat{\rho}_k > x)$$

Exact pivot given $\hat{\rho}_k > \theta_n$:

$$u_{kn}(\hat{\rho}_k, \ell) := \frac{\overline{F}_{kn}(\hat{\rho}_k, \ell)}{\overline{F}_{kn}(\theta_n, \ell)} \sim U(0,1) \qquad \text{for all } \ell$$

Approach:

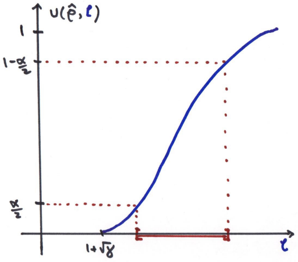
1. Approximate $\overline{F}_{kn}$ by Gaussian, Edgeworth, ...
2. Form approximate pivots $\quad u^A_{kn}(\hat{\rho}_k, \ell) \approx U(0,1)$
3. Confidence intervals: $\quad \{\ell_k > 1 + \sqrt{\gamma} \ : \ u^A_{kn}(\hat{\rho}_k, \ell) \in I\}$

# Pivots ctd.

$$\{\ell_k > 1 + \sqrt{\gamma} \;:\; u^A_{kn}(\hat{\rho}_k, \boldsymbol{\ell}) \in I\}$$

$$I = \begin{cases} [0, 1-\alpha] & \text{upper} \\ [\alpha/2, 1-\alpha/2] & \text{two-sided...} \end{cases}$$

Usually $\ell_k \to u^A_{kn}(\hat{\rho}_k, \boldsymbol{\ell})$ is monotone $\nearrow$

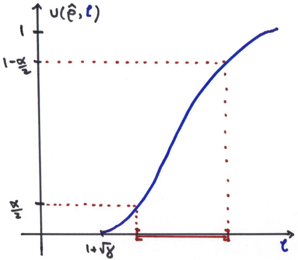


# Pivots ctd.

$$\{\ell_k > 1 + \sqrt{\gamma} \ : \ u^A_{kn}(\hat{\rho}_k, \boldsymbol{\ell}) \in I\}$$

$$I = \begin{cases} [0, 1-\alpha] & \text{upper} \\ [\alpha/2, 1-\alpha/2] & \text{two-sided...} \end{cases}$$

Usually $\ell_k \to u^A_{kn}(\hat{\rho}_k, \boldsymbol{\ell})$ is monotone $\nearrow$




Gaussian example:

$$\overline{F}_{kn}(x, \boldsymbol{\ell}) \approx \overline{\Phi}(z_n(x, \ell_k)), \qquad z_n(x, \ell) = n^{1/2} \frac{x - \rho(\ell, \gamma_n)}{\sigma(\ell, \gamma_n)}$$

$\to$ Selective $Z$ pivot:

$$u^z_n(\hat{\rho}_k, \ell_k) := \frac{\overline{\Phi}(z_n(\hat{\rho}_k, \ell_k))}{\overline{\Phi}(z_n(\theta_n, \ell_k))}$$

# Edgeworth pivots

Edgeworth approximation

$$\Phi^E_{kn}(x, \boldsymbol{\ell}) = \Phi(x) + n^{-1/2} p_k(x; \boldsymbol{\ell}, \gamma_n)\phi(x)$$

$\to$ Selective $E$ pivot: [estimated $\hat{\boldsymbol{\ell}}$ : $\hat{\ell}_j = \rho_n^{-1}(\hat{\rho}_j)$]

$$u^E_{kn}(\hat{\boldsymbol{\rho}}, \ell_k) := \frac{\overline{\Phi}^E_{kn}(z_n(\hat{\rho}_k, \ell_k), \hat{\boldsymbol{\ell}})}{\overline{\Phi}^E_{kn}(z_n(\theta_n, \ell_k), \hat{\boldsymbol{\ell}})}$$

Positive (E) pivot:

$$u^P_{kn}(\hat{\boldsymbol{\rho}}, \ell_k) := \begin{cases} u^E_{kn}(\hat{\boldsymbol{\rho}}, \ell_k) & \text{if } \overline{\Phi}^E_{kn}(z_n(\hat{\rho}_k, \ell_k), \hat{\boldsymbol{\ell}}) > 0 \\ u^z_n(\hat{\boldsymbol{\rho}}, \ell_k) & \text{otherwise} \end{cases}$$

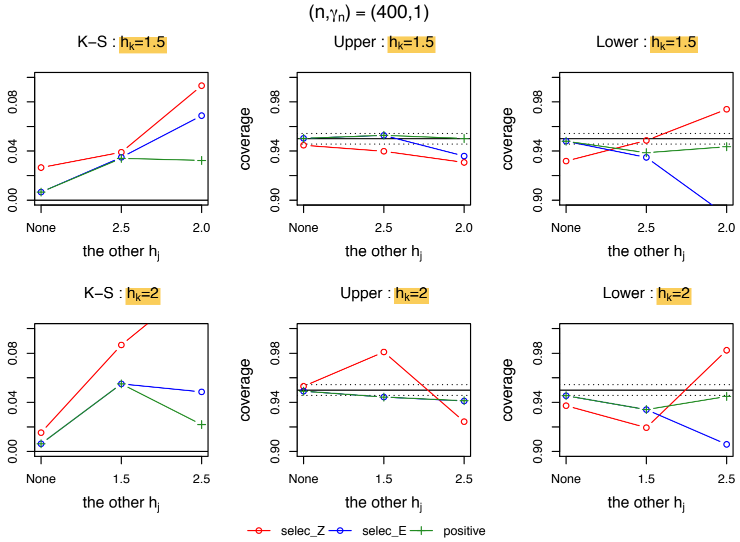
# Coverage accuracy

**Theorem:** Uniformly in $\alpha \in [0,1]$, for any $1 \leq k \leq r$,

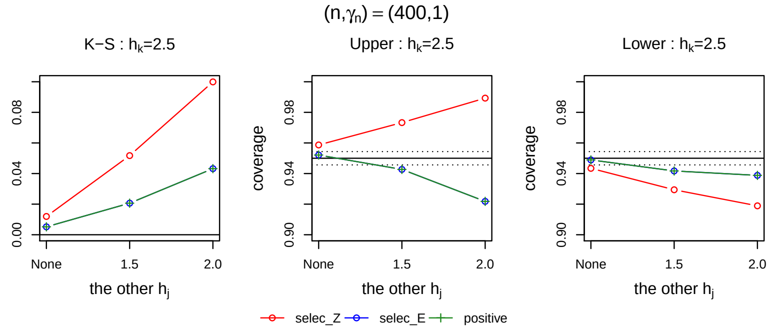
$$\mathbb{P}\{u(\hat{\boldsymbol{\rho}}) \leq \alpha \mid \hat{\rho}_k > \theta_n\} - \alpha$$
$$= \begin{cases} O(n^{-1/2}) & \text{for } u(\hat{\boldsymbol{\rho}}) = u_n^z(\hat{\rho}_k, \ell_k), \\ o(n^{-1/2}) & \text{for } u(\hat{\boldsymbol{\rho}}) = u_{kn}^E(\hat{\boldsymbol{\rho}}, \ell_k), u_{kn}^P(\hat{\boldsymbol{\rho}}, \ell_k) \end{cases}$$

- Consequence of the Edgeworth expansion
- also holds for clipped pivots $((u(\hat{\boldsymbol{\rho}}) \vee 0) \wedge 1)$

# Numerical coverage – 2 spikes



# Numerical coverage – 2 spikes




- Repulsion stronger for closer spikes $\rightarrow$ worse approximations
- selective E(o) has $\overline{\Phi}^E < 0$ with prob $> 5\%$ in tough cases:
$$\boldsymbol{h} = (2.0, 1.5), (2.5, 2.0)$$
- Positive pivot(+) usually fixes this!

# Future work

- Other models, e.g. low rank denoising

$$X = \sum_{k=1}^{r} \ell_k \boldsymbol{u}_k \boldsymbol{u}_k' + Z$$

- corrections for joint distributions
- non-Gaussian data
- second order expansions: LSS obstacle

Reference: (single spike) Yang & J., *Statistica Sinica* 2018.
(multispike) in preparation.

THANK YOU!