Weak Detection of Signal in the Spiked Wigner Model

Hye Won Chung and Ji Oon Lee

Korea Advanced Institute of Science and Technology (KAIST)

2019 ICML

Model and Applications: Spiked Wigner Model

- Signal: $\boldsymbol{x} \in \mathbb{R}^N$
- Noise: H is an $N \times N$ real symmetric random matrix (Wigner matrix)
- Data: Signal-plus-noise

$$M = \sqrt{\lambda}\boldsymbol{x}\boldsymbol{x}^T + H$$

(λ: Signal-to-Noise Ratio (SNR))

- Applications
  - Community detection: recover $\boldsymbol{x} \in \{1, -1\}^N$ indicating communities each node belongs to where noise H is Bernoulli distribution
  - Submatrix localization: recover small blocks of M with atypical mean where H follows Gaussian distribution

Reliable (Strong) Detection vs. Weak Detection

$$M = \sqrt{\lambda} \boldsymbol{x}\boldsymbol{x}^T + H \qquad (\|\boldsymbol{x}\|_2 = 1, \quad \|H\| \to 2)$$

- If $\lambda > 1$, signal can be detected and recovered by principal component analysis (PCA)
  - BBP transition: the largest eigenvalue converges to $\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}} > 2$
- If $\lambda < 1$ and the noise is Gaussian,
  - signal cannot be detected by PCA
  - no tests based on the eigenvalues can reliably detect the signal (Montanari, Reichman, Zeitouni '17)
  - no tests can reliably detect the signal (El Alaoui, Krzakala, Jordan '18)
- For $\lambda < 1$, consider hypothesis testing:

$$\boldsymbol{H}_0 : \lambda = 0, \qquad \boldsymbol{H}_1 : \lambda > 0$$

- Error: $\mathrm{err}(\lambda) = \mathbb{P}(\hat{\boldsymbol{H}} = \boldsymbol{H}_1 | \boldsymbol{H}_0) + \mathbb{P}(\hat{\boldsymbol{H}} = \boldsymbol{H}_0 | \boldsymbol{H}_1)$

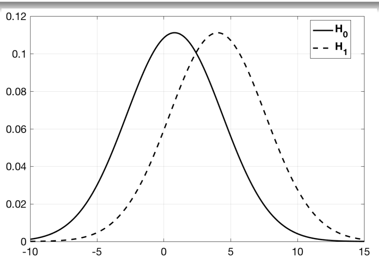
Proposed Test Based on Linear Spectral Statistics (LSS)

- Denoting by $\mu_1, \ldots, \mu_N$ the eigenvalues of M, LSS is defined as

$$L_N(f) = \sum_{i=1}^{N} f(\mu_i) \text{ for analytic } f \text{ on an open interval containing } [-2, 2]$$

Theorem (Central Limit Theorem (CLT) for LSS 1 (Chung and Lee'19))

$$\left(\sum_{i=1}^{N} f(\mu_i) - N \int_{-2}^{2} \frac{\sqrt{4 - z^2}}{2\pi} f(z)\, \mathrm{d}z \right) \Rightarrow \mathcal{N}(m_M(f), V_M(f))$$



Goal: find $f = f_\lambda^*$ that maximizes

$$\left| \frac{m_M(f) - m_H(f)}{\sqrt{V_M(f)}} \right|$$

Hypothesis Testing - Algorithm 1

- Proposed test (Algorithm 1):
  - Compute the test statistic $L_\lambda := \sum_{i=1}^{N} f_\lambda^*(\mu_i) - N \int_{-2}^{2} \frac{\sqrt{4-z^2}}{2\pi} f_\lambda^*(z)\,\mathrm{d}z$ (L_λ is a simple function of M)
  - Set $m_\lambda = \frac{1}{2}(m_M(f_\lambda^*) + m_H(f_\lambda^*))$
  - Accept $\boldsymbol{H}_0$ if $L_\lambda \le m_\lambda$; Accept $\boldsymbol{H}_1$ if $L_\lambda > m_\lambda$
- Universality:
  - For any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 = 1$, the proposed test and its error do not change, and thus the test **does not need any prior information on $\boldsymbol{x}$**
  - The proposed test **does not depend on the distribution of the noise** H except on $\mathbb{E}[H_{ii}^2]$ and $\mathbb{E}[H_{ij}^4]$
- Optimality:
  - The proposed test is with the lowest error among all tests based on LSS
  - For Gaussian noise, the proposed test achieves the optimal error (LLR)

Entrywise Transformation

Assume that $\sqrt{N}H_{ij}$ is drawn from a distribution with a density g

- Fisher information

$$F^H = \int_{-\infty}^{\infty} \frac{g'(w)^2}{g(w)} \mathrm{d}w$$

  ($F^H \geq 1$ with equality if and only if g is Gaussian)

- Entrywise transformation

$$h(w) := -\frac{g'(w)}{g(w)}, \qquad \widetilde{M}_{ij} = \frac{1}{F^H N} h(\sqrt{N} M_{ij})$$

- If $\lambda > \frac{1}{F^H}$, signal can be reliably detected by PCA after the entrywise transformation (Perry, Wein, Bandeira, Moitra, '18)
- If $\lambda < \frac{1}{F^H}$, will our test be improved by the entrywise transformation?

Hypothesis Testing - Algorithm 2

Theorem (CLT for LSS 2 (Chung and Lee'19))

Denoting by $\widetilde{\mu}_1, \dots, \widetilde{\mu}_N$ the eigenvalues of $\widetilde{M}$

$$\left(\sum_{i=1}^{N} f(\widetilde{\mu}_i) - N \int_{-2}^{2} \frac{\sqrt{4-z^2}}{2\pi} f(z)\,\mathrm{d}z \right) \Rightarrow \mathcal{N}(m_{\widetilde{M}}(f), V_{\widetilde{M}}(f))$$

- Find $f = \widetilde{f}_\lambda^*$ that maximizes $\left| \frac{m_{\widetilde{M}}(f) - m_{\widetilde{M}_0}(f)}{\sqrt{V_{\widetilde{M}}(f)}} \right|$
- Proposed test (Algorithm 2):
  - Compute the test statistic $\widetilde{L}_\lambda := \sum_{i=1}^{N} \widetilde{f}_\lambda^*(\widetilde{\mu}_i) - N \int_{-2}^{2} \frac{\sqrt{4-z^2}}{2\pi} \widetilde{f}_\lambda^*(z)\,\mathrm{d}z$
  - Set $\widetilde{m}_\lambda = \frac{1}{2}(m_{\widetilde{M}}(\widetilde{f}_\lambda^*) + m_{\widetilde{M}_0}(\widetilde{f}_\lambda^*))$
  - Accept $\boldsymbol{H}_0$ if $\widetilde{L}_\lambda \leq \widetilde{m}_\lambda$; Accept $\boldsymbol{H}_1$ if $\widetilde{L}_\lambda > \widetilde{m}_\lambda$

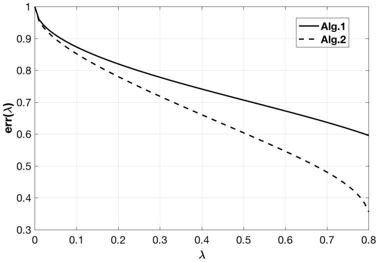
Example: Limiting Errors of the Two Tests

- Suppose that the density function of the noise matrix is given by

$$g(x) = \frac{1}{2\cosh(\pi x/2)} = \frac{1}{e^{\pi x} + e^{-\pi x}}$$

- Apply the entrywise transformation

$$h(x) = -\frac{g'(x)}{g(x)} = \frac{\pi}{2}\tanh\frac{\pi x}{2}, \quad \widetilde{M}_{ij} = \sqrt{\frac{2}{N}}\tanh\left(\frac{\pi\sqrt{N}}{2}M_{ij}\right)$$



Conclusion / Future Works

- We proposed a hypothesis test for a signal detection problem in a rank-one spiked Wigner model
- The proposed test is based on LSS for which we proved the CLT
- University and optimality of the proposed test
- When the density of the noise matrix is known and it is not Gaussian, the test error can be improved by the entrywise transformation
- Future work: further generalization to the model with higher-rank and/or non-Wigner noise

Poster Number: 206