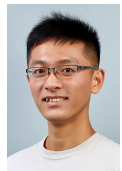
Matrix Denoising with Doubly Heteroscedastic Noise

Fundamental Limits and Optimal Spectral Methods

Yihan Zhang[*] Marco Mondelli[*]



[*]Institute of Science and Technology Austria

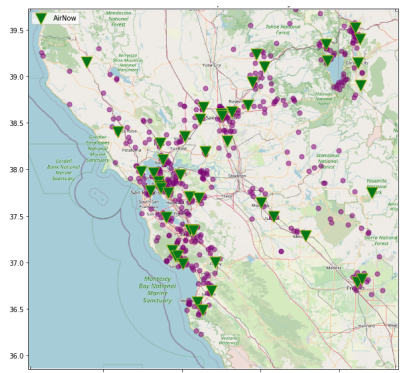
NeurIPS 2024

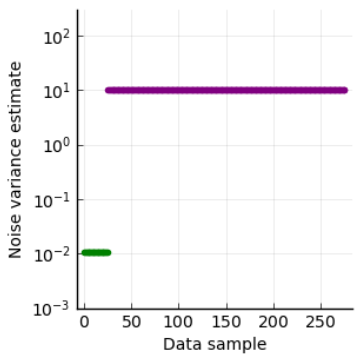
Institute of Science and Technology Austria

Example 1: spatial heteroscedasticity

Air quality [HGBF21]: hourly readings of $\mathrm{PM}_{2.5}$ particulate density ($\mu g/m^3$) in central California during Feb 9-13, 2021

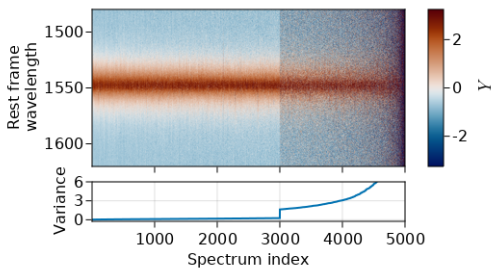
- AirNow: 25 high-precision sensors
- PurpleAir: 250 low-precision sensors

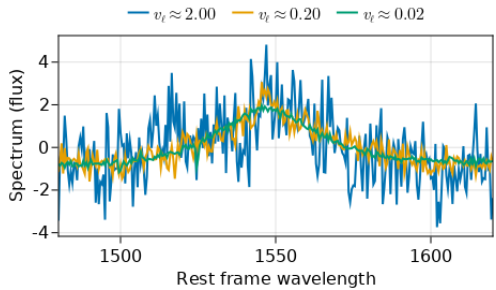




Example 2: temporal heteroscedasticity

Star spectra [APA+20, LHM+20, HYFB23]: $n = 5000$ spectra for a fixed star, each a $d = 281$-dimensional vector recording the flux at d wavelengths in $[1480, 1620]$ nm





Matrix denoising with doubly heteroscedastic noise

Yihan Zhang and Marco Mondelli. "Matrix Denoising with Doubly Heteroscedastic Noise: Fundamental Limits and Optimal Spectral Methods." arXiv:2405.13912

$$A = \frac{\lambda}{n} u^* v^{*\top} + W \in \mathbb{R}^{n \times d}$$

where

$$W = \Xi^{1/2} \widetilde{W} \Sigma^{1/2}, \qquad \widetilde{W}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$$

proportional regime: $\frac{n}{d} \to \delta$

Bayes-optimality

Theorem ("Information-theoretic limit" [ZM24])

$$\lim_{n\to\infty} \sup_{\widetilde{u}} \mathbb{E}\left[\frac{\left|\left\langle \Xi^{-1/2}u^*, \widetilde{u}\right\rangle\right|^2}{\left\|\Xi^{-1/2}u^*\right\|_2^2 \|\widetilde{u}\|_2^2}\right] = \frac{q_u^*}{\mathbb{E}\left[\overline{\Xi}^{-1}\right]},$$

$$\lim_{d\to\infty} \sup_{\widetilde{v}} \mathbb{E}\left[\frac{\left|\left\langle \Sigma^{-1/2}v^*, \widetilde{v}\right\rangle\right|^2}{\left\|\Sigma^{-1/2}v^*\right\|_2^2 \|\widetilde{v}\|_2^2}\right] = \frac{q_v^*}{\mathbb{E}\left[\overline{\Sigma}^{-1}\right]},$$

where q_u^*, q_v^* *solve*

$$q_u^* = \mathbb{E}\left[\frac{\lambda^2 q_v^* \overline{\Xi}^{-2}}{\delta + \lambda^2 q_v^* \overline{\Xi}^{-1}}\right], \quad q_v^* = \mathbb{E}\left[\frac{\lambda^2 q_u^* \overline{\Sigma}^{-2}}{1 + \lambda^2 q_u^* \overline{\Sigma}^{-1}}\right]$$

Weak recovery threshold

Corollary

Positive overlaps with u^*, v^* are achievable if and only if

$$\lambda > \left(\delta \mathbb{E}\left[\overline{\Xi}^{-2}\right]^{-1} \mathbb{E}\left[\overline{\Sigma}^{-2}\right]^{-1}\right)^{\frac{1}{4}}$$

Remark

BBP threshold [BBAP05] for whitened $\Xi^{-1/2} A \Sigma^{-1/2}$:

$$\lambda > \left(\delta \mathbb{E}\left[\overline{\Xi}^{-1}\right]^{-2} \mathbb{E}\left[\overline{\Sigma}^{-1}\right]^{-2}\right)^{\frac{1}{4}}$$

Weak recovery threshold

Corollary

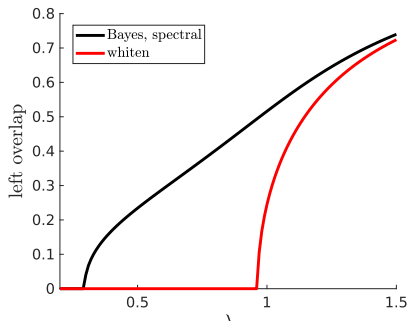
Positive overlaps with u^*, v^* are achievable if and only if

$$\lambda > \left(\delta \mathbb{E}\left[\overline{\Xi}^{-2}\right]^{-1} \mathbb{E}\left[\overline{\Sigma}^{-2}\right]^{-1}\right)^{\frac{1}{4}}$$

Remark

BBP threshold [BBAP05] for whitened $\Xi^{-1/2} A \Sigma^{-1/2}$:

$$\lambda > \left(\delta \mathbb{E}\left[\overline{\Xi}^{-1}\right]^{-2} \mathbb{E}\left[\overline{\Sigma}^{-1}\right]^{-2}\right)^{\frac{1}{4}}$$



A spectral estimator on steroids

Theorem (Efficient algorithm [ZM24])

The estimator

$$\widehat{u}^{\mathrm{spec}} = G_2(\Xi) u_1 \Big(\underbrace{G_1(\Xi) A F_1(\Sigma)}_{A^*} \Big)$$

- *achieves the Bayes-optimal weak recovery threshold;*

A spectral estimator on steroids

Theorem (Efficient algorithm [ZM24])

The estimator

$$\widehat{u}^{\mathrm{spec}} = G_2(\Xi) u_1 \Big(\underbrace{G_1(\Xi) A F_1(\Sigma)}_{A^*} \Big)$$

- *achieves the Bayes-optimal weak recovery threshold;*
- *achieves the Bayes-optimal overlap if* $\Xi = I_n$.

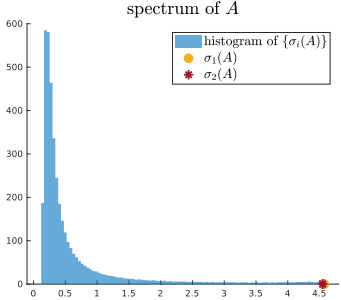
A spectral estimator on steroids

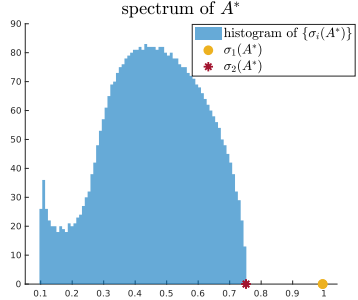
Theorem (Efficient algorithm [ZM24])

The estimator

$$\widehat{u}^{\mathrm{spec}} = G_2(\Xi) u_1 \Big(\underbrace{G_1(\Xi) A F_1(\Sigma)}_{A^*} \Big)$$

- *achieves the Bayes-optimal weak recovery threshold;*
- *achieves the Bayes-optimal overlap if* $\Xi = I_n$.





THANK YOU!

arXiv:2405.13912

zephyr.z798@gmail.com


Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, et al.
The 16th data release of the sloan digital sky surveys: First release from the apogee-2 southern survey and full release of eboss spectra.
The Astrophysical Journal Supplement Series, 249(1):3, jun 2020.


Jinho Baik, Gérard Ben Arous, and Sandrine Péché.
Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices.
Ann. Probab., 33(5):1643–1697, 2005.


David Hong, Kyle Gilman, Laura Balzano, and Jeffrey A. Fessler.
HePPCAT: probabilistic PCA for data with heteroscedastic noise.
IEEE Trans. Signal Process., 69:4819–4834, 2021.


David Hong, Fan Yang, Jeffrey A. Fessler, and Laura Balzano.
Optimally weighted PCA for high-dimensional heteroscedastic data.
SIAM J. Math. Data Sci., 5(1):222–250, 2023.

Brad W. Lyke, Alexandra N. Higley, J. N. McLane, Danielle P. Schurhammer, et al.
The sloan digital sky survey quasar catalog: Sixteenth data release.
The Astrophysical Journal Supplement Series, 250(1):8, aug 2020.

Yihan Zhang and Marco Mondelli.
Matrix denoising with doubly heteroscedastic noise: Fundamental limits and optimal spectral methods.
arXiv preprint arXiv:2405.13912, 2024.