Matrix Denoising with Weighted Loss

William Leeb
University of Minnesota, Twin Cities

UC Davis MADDD Seminar
March 10$^{\text{th}}$, 2020

The observation model:

- We observe a "signal plus noise" matrix $\mathbf{Y}$ of the form
$$\mathbf{Y} = \mathbf{X} + \mathbf{G}$$

- $\mathbf{Y}$ is of size p-by-n, where p and n are large.

- $\mathbf{X}$ is a rank r signal matrix, where $r \ll p, n$.

- $\mathbf{G}$ is a matrix of additive Gaussian noise.

The goal: Estimate $\mathbf{X}$ from $\mathbf{Y}$.

A more detailed look:

- Write the SVD of $\mathbf{X}$:

$$\mathbf{X} = \sum_{k=1}^{r} t_k \mathbf{u}_k \mathbf{v}_k^T,$$

  where $t_k > 0$, and the $\mathbf{u}_k$, $\mathbf{v}_k$ are orthonormal vectors.

- Write the SVD of $\mathbf{Y}$:

$$\mathbf{Y} = \sum_{k=1}^{\min\{p,n\}} \lambda_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T,$$

  where $\lambda_k > 0$, and the $\hat{\mathbf{u}}_k$, $\hat{\mathbf{v}}_k$ are orthonormal vectors.

- The entries of $\mathbf{G}$ have distribution $G_{ij} \sim N(0, 1/n)$.

- We study the problem as p, n grow to infinity, and r stays fixed:

$$\lim_{n\to\infty} \frac{p}{n} = \gamma < \infty$$

The loss function:

- We measure the error between $\mathbf{X}$ and $\widehat{\mathbf{X}}$ with a *weighted loss*.

- Specificically, we use a loss function of the form:

$$\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2$$

- Here, Ω and Π are matrices with p columns and n columns, respectively.

Why weighted loss? We consider three applications:

- Submatrix denoising

- Heteroscedastic noise

- Missing data

Submatrix denoising:

- Suppose we are only interested in estimating a submatrix $\mathbf{X}_0$ of $\mathbf{X}$.

- We use information from the entire matrix $\mathbf{X}$, but only penalize errors on $\mathbf{X}_0$.

- Let Ω and Π project onto the rows and columns of $\mathbf{X}_0$; then the natural loss is

$$\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2$$

- We can show that denoising the full $\mathbf{X}$ and projecting onto $\mathbf{X}_0$ is typically better than denoising $\mathbf{X}_0$ directly.

Heteroscedastic noise:

- Observe $\mathbf{Y}' = \mathbf{X}' + \mathbf{N}$, where $\mathbf{N}$ has rank 1 variance structure:

$$\mathbf{N} = \mathbf{A}^{1/2}\mathbf{G}\mathbf{B}^{1/2}.$$

  Our goal is to estimate $\mathbf{X}'$.

- *Whiten* the noise:

$$\mathbf{Y} = \mathbf{A}^{-1/2}\mathbf{Y}'\mathbf{B}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{X}'\mathbf{B}^{-1/2} + \mathbf{G} \equiv \mathbf{X} + \mathbf{G}$$

- Estimate $\mathbf{X} = \mathbf{A}^{-1/2}\mathbf{X}'\mathbf{B}^{-1/2}$ with a method tailored for white noise, and then unwhiten:

$$\widehat{\mathbf{X}}' = \mathbf{A}^{1/2}\widehat{\mathbf{X}}\mathbf{B}^{1/2}$$

- The mean squared error is then

$$\|\widehat{\mathbf{X}}' - \mathbf{X}'\|_F^2 = \|\mathbf{A}^{1/2}\widehat{\mathbf{X}}\mathbf{B}^{1/2} - \mathbf{A}^{1/2}\mathbf{X}\mathbf{B}^{1/2}\|_F^2 = \|\mathbf{A}^{1/2}(\widehat{\mathbf{X}} - \mathbf{X})\mathbf{B}^{1/2}\|_F^2,$$

  which is a weighted loss.

- It can be shown (later in this talk) that whitening improves the signal-to-noise ratio.

Missing data:

- We observe $\mathcal{F}(\mathbf{Y}')$, M random entries of $\mathbf{Y}' = \mathbf{X}' + \mathbf{G}$. Our goal is to estimate $\mathbf{X}'$.

- Assume rank 1 sampling, with row and column sampling probabilities $\mathbf{P}$ and $\mathbf{Q}$.

- It can be shown that

$$\mathbf{Y} \equiv \mathbf{P}^{-1/2}\mathcal{F}^*(\mathcal{F}(\mathbf{Y}'))\mathbf{Q}^{-1/2} \sim \mathbf{X} + \text{white noise}$$

  where $\mathbf{X} = \mathbf{P}^{-1/2}\mathbf{X}'\mathbf{Q}^{-1/2}$.

- We denoise $\mathbf{Y}$ to get $\widehat{\mathbf{X}}$, then estimate $\mathbf{X}'$ by $\widehat{\mathbf{X}}' = \mathbf{P}^{1/2}\widehat{\mathbf{X}}\mathbf{Q}^{1/2}$.

- The mean squared error is then:

$$\|\widehat{\mathbf{X}}' - \mathbf{X}'\|_F^2 = \|\mathbf{P}^{1/2}\widehat{\mathbf{X}}\mathbf{Q}^{1/2} - \mathbf{P}^{1/2}\mathbf{X}\mathbf{Q}^{1/2}\|_F^2 = \|\mathbf{P}^{1/2}(\widehat{\mathbf{X}} - \mathbf{X})\mathbf{Q}^{1/2}\|_F^2,$$

  which is a weighted loss.

Return to $\mathbf{Y} = \mathbf{X} + \mathbf{G}$, $G_{ij} \sim N(0, 1/n)$, $\mathbf{X}$ rank r.

- A standard approach to estimating $\mathbf{X}$ is *singular value shrinkage.*

- Singular value shrinkage performs an SVD of $\mathbf{Y}$:

$$\mathbf{Y} = \sum_{k=1}^{\min(n,p)} \lambda_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$$

- We then replace the observed singular values λ_j with new singular values t_k, leaving the observed singular vectors fixed:

$$\widehat{\mathbf{X}} = \sum_{k=1}^{r} \hat{t}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$$

- Since $\mathbf{X}$ has rank r, we set all but the top r components of $\widehat{\mathbf{X}}$ to 0.

- With unweighted Frobenius loss, singular value shrinkage is known to be an optimal procedure (Shabalin and Nobel, 2013; Donoho and Gavish, 2014).

- Furthermore, there are explicit formulas for the asymptotically optimal singular values $\hat{t}_1, \ldots, \hat{t}_r$ of $\widehat{\mathbf{X}}$.

- Computing the optimal singular values $\hat{t}_k$ requires knowing two things:

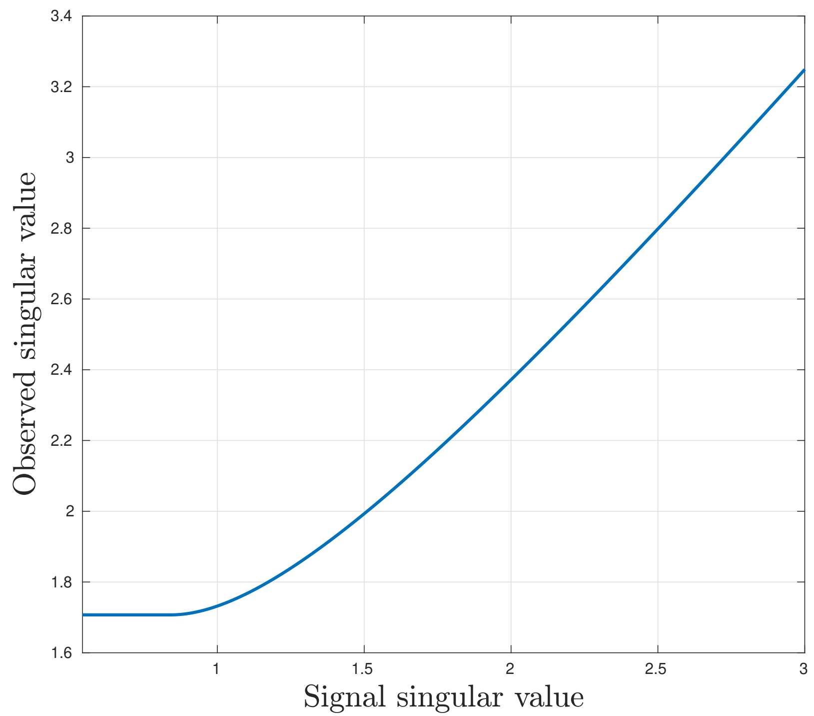
    1. The angles between the *population singular vectors* $\mathbf{u}_k$ and $\mathbf{v}_k$ and the *empirical singular vectors* $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{v}}_k$.

    2. The relation between the *population singular values* t_k and the *empirical eigenvalues* λ_k of $\mathbf{Y}$.

- These are derived by Paul (2007).

- The top r singular values of $\mathbf{Y}$ converge almost surely to the following expression:

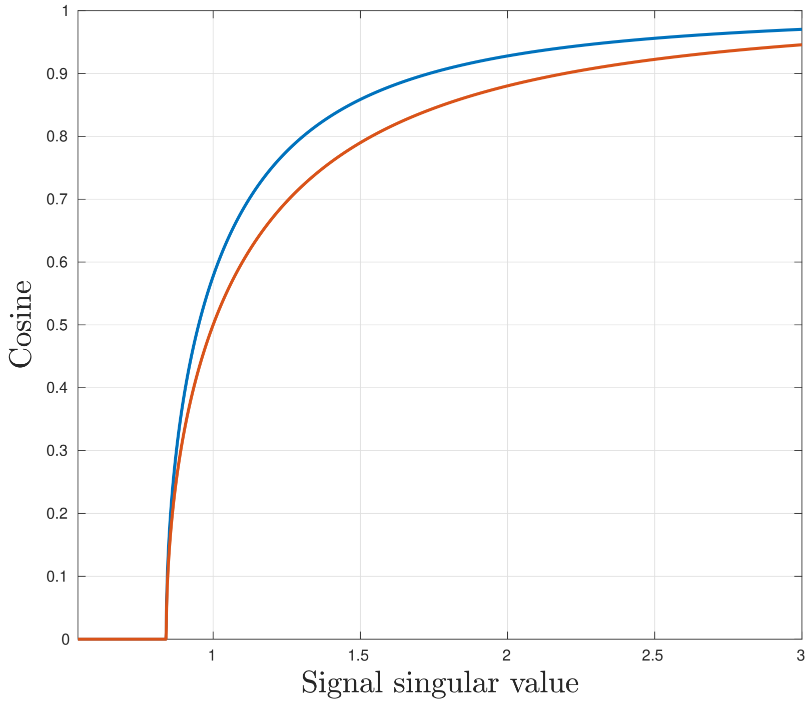
$$\lambda_k = \begin{cases} \sqrt{(t_k^2+1)(1+\gamma/t_k^2)}, & \text{if } t_k^2 > \sqrt{\gamma} \\ 1+\sqrt{\gamma}, & \text{otherwise} \end{cases}$$



- The cosines between empirical and population singular vectors converge almost surely:

$$\langle \mathbf{u}_j, \hat{\mathbf{u}}_k \rangle^2 \longrightarrow c_{j,k}^2 = \begin{cases} \frac{1-\gamma/t_k^4}{1+\gamma/t_k^2}, & \text{if} \quad j = k \ \text{ and } \ t_k^2 > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}$$

$$\langle \mathbf{v}_j, \hat{\mathbf{v}}_k \rangle^2 \longrightarrow \tilde{c}_{j,k}^2 = \begin{cases} \frac{1-\gamma/t_k^4}{1+1/t_k^2}, & \text{if} \quad j = k \ \text{ and } \ t_k^2 > \sqrt{\gamma} \\ 0, & \text{otherwise} \end{cases}$$



- The asymptotic mean squared error (AMSE) is:

$$\|\widehat{\mathbf{X}} - \mathbf{X}\|_F^2 = \sum_{k=1}^{r} (t_k^2 + \hat{t}_k^2 - 2 t_k \hat{t}_k c_k \tilde{c}_k).$$

- This is minimized by:

$$\hat{t}_k = t_k c_k \tilde{c}_k,$$

  with error

$$\text{AMSE} = \sum_{k=1}^{r} t_k^2 (1 - c_k^2 \tilde{c}_k^2).$$

- So long as $t_k^2 > \sqrt{\gamma}$, $\hat{t}_k$ is estimable from the observed data. Otherwise, the k^{th} component of X is lost in the noise.

What about *weighted* Frobenius loss, $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2$?

- We generalize singular value shrinkage to the class of *spectral estimators*, of the form:

$$\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T.$$

- $\widehat{\mathbf{U}} \in \mathbb{R}^{p \times r}$ and $\widehat{\mathbf{V}} \in \mathbb{R}^{n \times r}$ are the top singular vectors of $\mathbf{Y}$.

- $\widehat{\mathbf{B}}$ is an r-by-r matrix, to be optimized over:

$$\widehat{\mathbf{B}} = \operatorname*{argmin}_{\widehat{\mathbf{B}}'} \mathcal{L}(\widehat{\mathbf{U}}\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T, \mathbf{X})$$

Optimal spectral denoising:

- Solving for the optimal $\widehat{\mathbf{B}}$ is easy in principle:

$$\widehat{\mathbf{B}} = \mathbf{D}^{-1}\mathbf{C}\mathrm{diag}(\mathbf{t})\widetilde{\mathbf{C}}^T\widetilde{\mathbf{D}}^{-1},$$

where $\mathbf{t} = (t_1, \dots, t_r)$, and

$$\mathbf{D} = \widehat{\mathbf{U}}^T\Omega^T\Omega\widehat{\mathbf{U}}$$

$$\widetilde{\mathbf{D}} = \widehat{\mathbf{V}}^T\Pi^T\Pi\widehat{\mathbf{V}}$$

$$\mathbf{C} = \widehat{\mathbf{U}}^T\Omega^T\Omega\mathbf{U}$$

and

$$\widetilde{\mathbf{C}} = \widehat{\mathbf{V}}^T\Pi^T\Pi\mathbf{V}$$

- These are the matrices of *weighted* inner products between singular vectors of $\mathbf{X}$ and $\mathbf{Y}$.

Estimating $\widehat{\mathbf{B}} = \mathbf{D}^{-1}\mathbf{C}\mathrm{diag}(\mathbf{t})\widetilde{\mathbf{C}}^T\widetilde{\mathbf{D}}^{-1}$:

- The singular values $t_1, \ldots, t_r$ are estimable, as we've seen.
- The matrices $\mathbf{D} = \widehat{\mathbf{U}}^T\Omega^T\Omega\mathbf{U}$ and $\widetilde{\mathbf{D}} = \widehat{\mathbf{V}}^T\Pi^T\Pi\mathbf{V}$ are observed.
- We must estimate $\mathbf{C}$ and $\widetilde{\mathbf{C}}$, or all inner products of the form

$$\hat{\mathbf{u}}_k^T\Omega^T\Omega\mathbf{u}_l$$

  and

$$\hat{\mathbf{v}}_k^T\Pi^T\Pi\mathbf{v}_l$$

  for $1 \leq k, l \leq r$.
- We will show the formulas on the next slide.

Estimating the weighted inner products:

- When $t_j, t_k > \gamma^{1/4}$,

$$\hat{\mathbf{u}}_j \Omega^T \Omega \mathbf{u}_k \to \begin{cases} (d_k - s_k^2 \mu)/c_k, & \text{if } j = k \\ d_{jk}/c_k, & \text{if } j \neq k \end{cases}$$

  and

$$\hat{\mathbf{v}}_j \Pi^T \Pi \mathbf{v}_k \to \begin{cases} (\tilde{d}_k - \tilde{s}_k^2 \nu)/\tilde{c}_k, & \text{if } j = k \\ \tilde{d}_{jk}/\tilde{c}_k, & \text{if } j \neq k \end{cases}$$

  where

$$\mu = \lim_{p \to \infty} \frac{1}{p} \mathrm{tr}(\Omega^T \Omega)$$

  and

$$\nu = \lim_{n \to \infty} \frac{1}{n} \mathrm{tr}(\Pi^T \Pi)$$

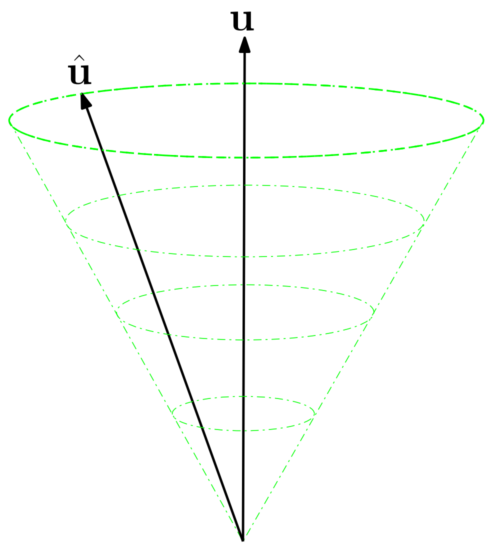
- Note that c_k and $\tilde{c}_k$ are estimable, as we've seen already.

Sketch of the derivation:

- Decompose $\hat{\mathbf{u}}_k$ into signal and noise components:

$$\hat{\mathbf{u}}_k = c_k \mathbf{u}_k + s_k \tilde{\mathbf{u}}_k$$

- The unit vector $\tilde{\mathbf{u}}_k$ is orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_r$, and uniformly random.



- The $\tilde{\mathbf{u}}_k$ also satisfy the *Hanson-Wright* formula. For any bounded A:

$$\tilde{\mathbf{u}}_k^T A \tilde{\mathbf{u}}_k \sim \frac{1}{p}\mathrm{tr}(A).$$

Sketch of the derivation:

- Applying Ω gives:

$$\Omega\hat{\mathbf{u}}_k = c_k\Omega\mathbf{u}_k + s_k\Omega\tilde{\mathbf{u}}_k$$

- Taking inner products with certain vectors, we can read off parameters.

- For example, the squared norm of each side is:

$$\|\Omega\hat{\mathbf{u}}_k\|^2 = c_k^2\|\Omega\mathbf{u}_k\|^2 + s_k^2\mu$$

  from which we can solve for $\|\Omega\mathbf{u}_k\|^2$.

Sketch of the derivation:

- Next, take inner products of $\Omega \mathbf{u}_k$ with each side of

$$\Omega \hat{\mathbf{u}}_k = c_k \Omega \mathbf{u}_k + s_k \Omega \tilde{\mathbf{u}}_k$$

- This gives:

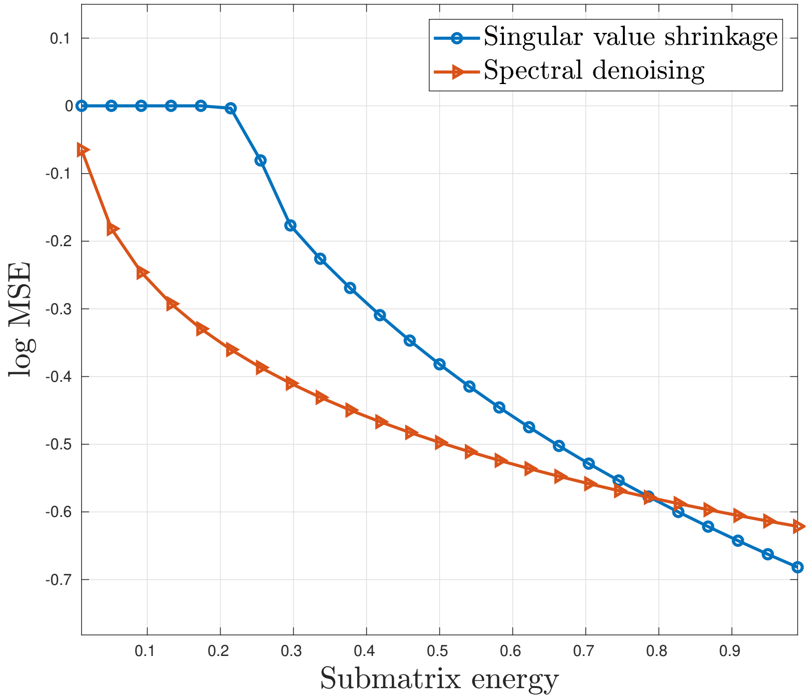
$$\hat{\mathbf{u}}_k^T \Omega^T \Omega \mathbf{u}_k = c_k \|\Omega \mathbf{u}_k\|^2$$

- This is known, since we already know $\|\Omega \mathbf{u}_k\|^2$.

The derivation of the cross terms $\hat{\mathbf{u}}_k^T \Omega^T \Omega \mathbf{u}_l$, $k \neq l$, proceeds similarly.

Submatrix estimation:

- We estimate a submatrix $\mathbf{X}_0 = \Omega \mathbf{X} \Pi^T$ of $\mathbf{X}$ by estimating $\mathbf{X}$ using spectral denoising with loss $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2$, and taking $\widehat{\mathbf{X}}_0 = \Omega \widehat{\mathbf{X}} \Pi^T$.
- We compare this approach with optimal singular value shrinkage applied to $\mathbf{Y}_0 = \Omega \mathbf{Y} \Pi^T$.



- Errors are plotted against the fraction of $\mathbf{X}$'s energy contained in $\mathbf{X}_0$.
- We prove that unless $\mathbf{X}_0$ contains an overwhelming fraction of $\mathbf{X}$'s energy, using the full matrix outperforms denoising $\mathbf{Y}_0$ directly.

Heteroscedastic noise:

- Observe $\mathbf{Y}' = \mathbf{X}' + \mathbf{N}$, where $\mathbf{N}$ has rank 1 variance structure:

$$\mathbf{N} = \mathbf{A}^{1/2}\mathbf{G}\mathbf{B}^{1/2}$$

- *Whiten* the noise:

$$\mathbf{Y} = \mathbf{A}^{-1/2}\mathbf{Y}'\mathbf{B}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{X}'\mathbf{B}^{-1/2} + \mathbf{G} \equiv \mathbf{X} + \mathbf{G}$$

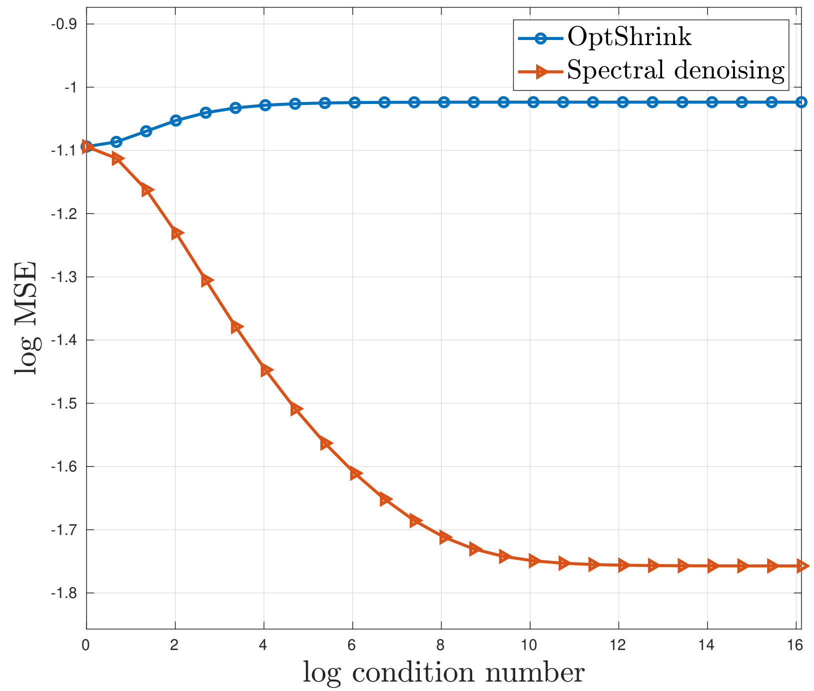
- Estimate $\mathbf{X} = \mathbf{A}^{-1/2}\mathbf{X}'\mathbf{B}^{-1/2}$ with optimal spectral denoising with weighted loss

$$\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\mathbf{A}^{1/2}(\widehat{\mathbf{X}} - \mathbf{X})\mathbf{B}^{1/2}\|_F^2,$$

- Finally, unwhiten: $\widehat{\mathbf{X}}' = \mathbf{A}^{1/2}\widehat{\mathbf{X}}\mathbf{B}^{1/2}$.

Comparison with OptShrink:

- We compare with optimal singular value shrinkage, without whitening (Nadakuditi, 2014):

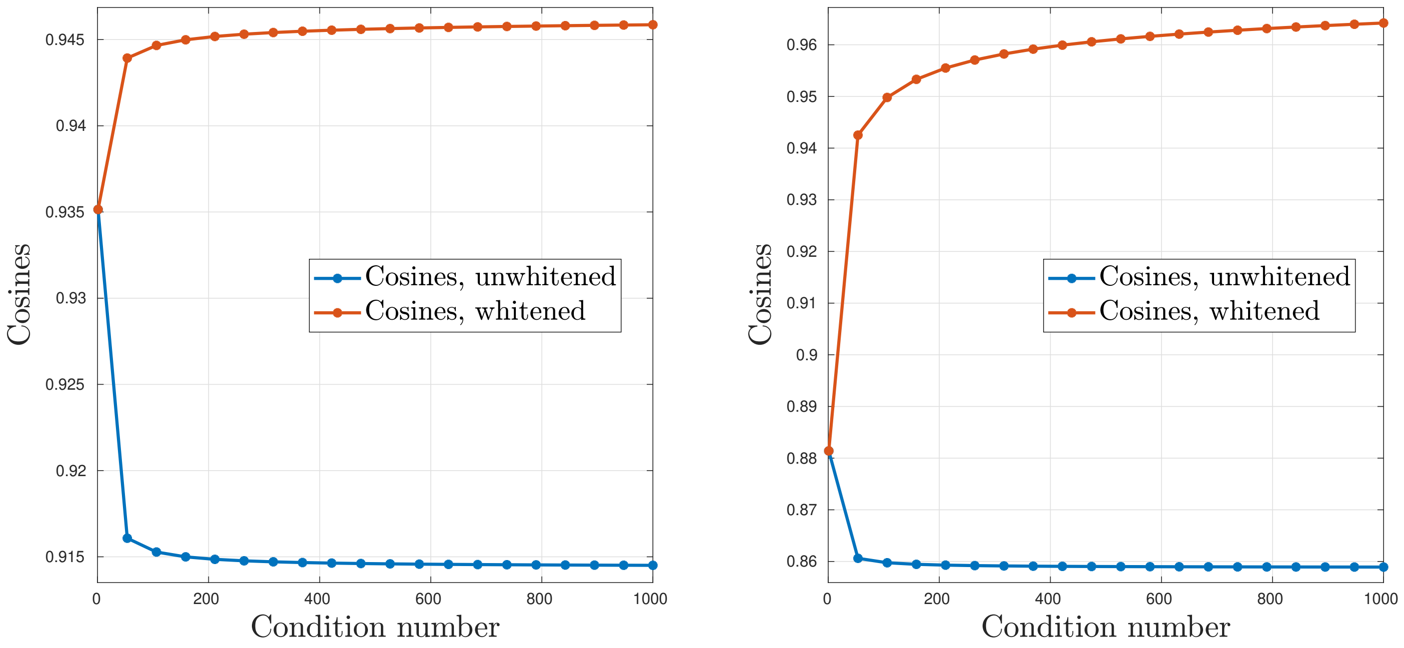


- The MSE is plotted as a function of the condition number of $\mathbf{A}^{1/2}$ and $\mathbf{B}^{1/2}$, the noise covariance matrices.

- The total energy in the noise is constant.

Whitening improves subspace estimation:

- Suppose $\mathbf{Y}' = \mathbf{X}' + \Sigma^{1/2}\mathbf{G}$.
- Compare singular vectors of $\mathbf{Y}'$ with the vectors from whitening, SVD'ing, unwhitening.



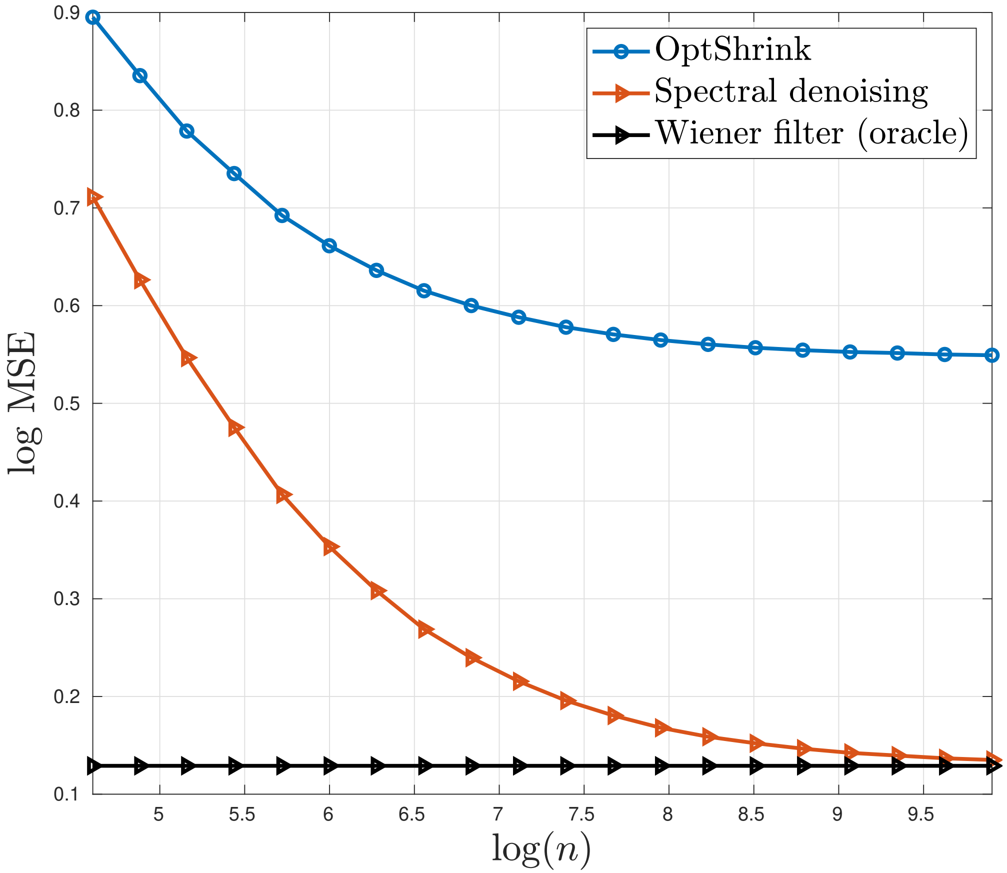
- In Leeb and Romanov (2019), we prove that

$$\frac{|\text{unwhitened cosine}|}{|\text{whitened cosine}|} \leq f(\kappa)$$

where $\kappa = \frac{1}{p}\mathrm{tr}(\Sigma_\varepsilon) \cdot \frac{1}{p}\mathrm{tr}(\Sigma_\varepsilon^{-1})$, and $f(\kappa) < 1$ for $\kappa > 1$ and is decreasing.

Relation with linear prediction:

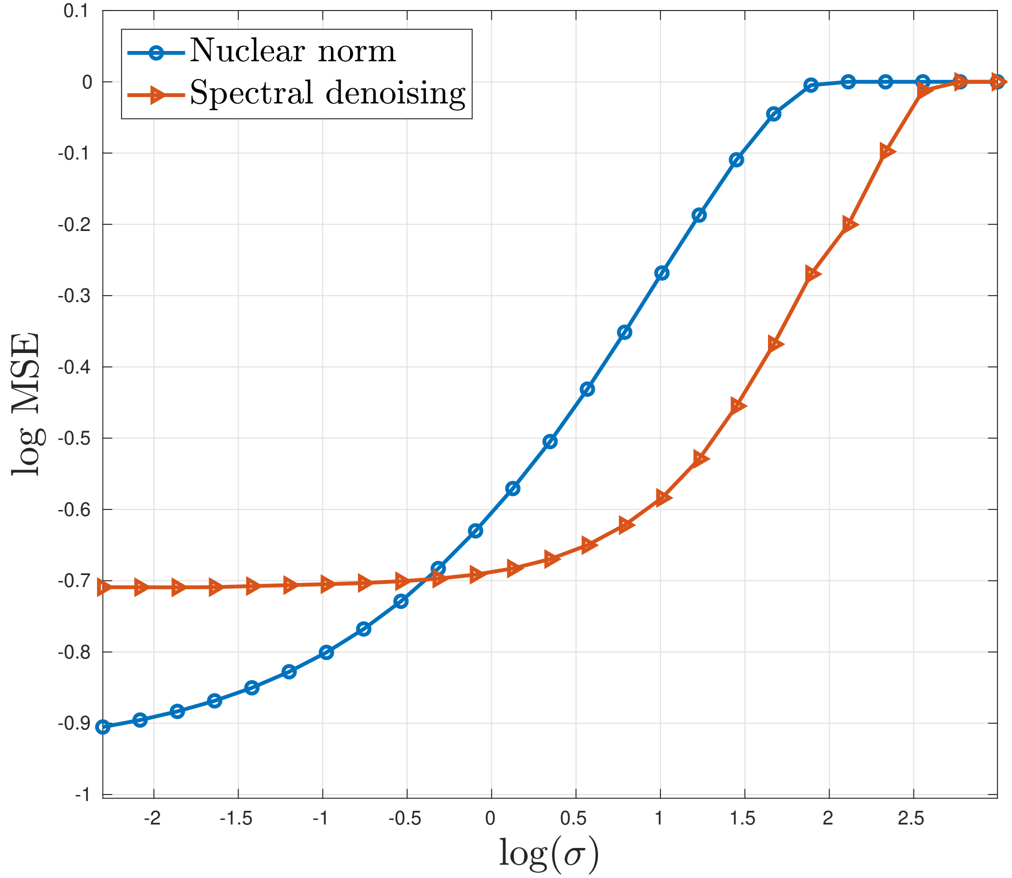
- Optimal spectral denoising with whitening converges to the Wiener filter as $p/n \to 0$.



- Optimal spectral shrinkage converges to a suboptimal linear filter.

Missing data:

- We observe $\mathcal{F}(\mathbf{Y}')$, M random entries of $\mathbf{Y}' = \mathbf{X}' + \mathbf{G}$.
- Rank 1 sampling structure, with row and column sampling probabilities $\mathbf{P}$ and $\mathbf{Q}$.



- Estimate $\mathbf{X} = \mathbf{P}^{-1/2}\widetilde{\mathbf{X}}\mathbf{Q}^{-1/2}$ with optimal spectral denoiser with respect to loss

$$\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\mathbf{P}^{1/2}(\widehat{\mathbf{X}} - \mathbf{X})\mathbf{Q}^{1/2}\|_F^2,$$

  and define $\widehat{\widetilde{\mathbf{X}}} = \mathbf{P}^{1/2}\widehat{\mathbf{X}}\mathbf{Q}^{1/2}$.
- Compare to nuclear-norm regularized least squares of Candès and Plan (2010).

Summary:

- We study the problem of estimating low-rank $\mathbf{X}$ from $\mathbf{Y} = \mathbf{X} + \mathbf{G}$.
- We use weighted loss of the form $\mathcal{L}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\Omega(\widehat{\mathbf{X}} - \mathbf{X})\Pi^T\|_F^2$.
- We have introduced spectral denoisers of the form $\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{B}}\widehat{V}^T$.
- Using new asymptotic results for the spiked model, we derived the optimal $\widehat{\mathbf{B}}$.
- Applications include submatrix estimation; heteroscedastic noise; and missing data.

References

- W. Leeb and E. Romanov. Optimal spectral shrinkage and PCA with heteroscedastic noise. arXiv 1811.02201 (2019)
- W. Leeb. Matrix denoising for weighted loss functions and heterogeneous signals. arXiv:1902.09474 (2020)
- E. Dobriban, W. Leeb, and A. Singer. Optimal prediction in the linearly transformed spiked model. Annals of Statistics 48(1), 491-513 (2019)

Additional references

- D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica, 1617-1642 (2007)
- R. R. Nadakuditi. OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. IEEE Transactions on Information Theory 60(5), 3002-3018 (2014)
- D. Donoho and M. Gavish. Minimax risk of matrix denoising by singular value thresholding. The Annals of Statistics 42(6), 2413-2440 (2014).
- A. A. Shabalin and A. B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. Journal of Multivariate Analysis 118, 67-76 (2013).
- E. J. Candès and Y. Plan. Matrix completion with noise. Proceedings of the IEEE 98(6), 925-936 (2010)

Thank you

I acknowledge support from the NSF BIGDATA program and the BSF.