Matrix Denoising with Weighted Loss

William Leeb
University of Minnesota, Twin Cities

IMA Data Science Seminar
September 22\textsuperscript{nd}, 2020
The observation model:

- We observe a “signal plus noise” matrix $Y$ of the form

$$Y = X + G$$

- $Y$ is of size $p$-by-$n$, where $p$ and $n$ are large.

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

- $G$ is a matrix of additive Gaussian noise.

**The goal:** Estimate $X$ from $Y$. 
A more detailed look:

- Write the SVD of \( X \):
  \[
  X = \sum_{k=1}^{r} t_k u_k v_k^T, 
  \]
  where \( t_k > 0 \), and the \( u_k, v_k \) are orthonormal vectors.

- Write the SVD of \( Y \):
  \[
  Y = \sum_{k=1}^{\min\{p,n\}} \lambda_k \hat{u}_k \hat{v}_k^T, 
  \]
  where \( \lambda_k > 0 \), and the \( \hat{u}_k, \hat{v}_k \) are orthonormal vectors.

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

- We study the problem as \( p, n \) grow to infinity, and \( r \) stays fixed:
  \[
  \lim_{n \to \infty} \frac{p}{n} = \gamma < \infty 
  \]

- This is a version of the spiked model of Johnstone (2001).
The loss function:

- Since $\|X\|_F \ll \|G\|_F$, consistent estimation of $X$ is not possible. Gavish and Donoho (2017) show that the optimal denoiser depends on the loss function.

- We measure the error between $X$ and $\hat{X}$ with a *weighted loss*, of the form:
  \[
  \mathcal{L}(\hat{X}, X) = \|\Omega(\hat{X} - X)\Pi^T\|_F^2
  \]

- Here, $\Omega$ and $\Pi$ 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 $X_0$ of $X$.

• We use information from the entire matrix $X$, but only penalize errors on $X_0$.

• Let $\Omega$ and $\Pi$ project onto the rows and columns of $X_0$; then the natural loss is

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

• It can be shown (Leeb, 2020) that denoising the full $X$ and projecting onto $X_0$ is typically better than denoising $X_0$ directly.
Heteroscedastic noise:

- Observe \( Y' = X' + N \), where \( N \) has rank 1 variance structure:

\[
N = S^{1/2}GT^{1/2}.
\]

Our goal is to estimate \( X' \).

- **Whiten** the noise:

\[
Y = S^{-1/2}Y'T^{-1/2} = S^{-1/2}X'T^{-1/2} + G \equiv X + G
\]

- Estimate \( X = S^{-1/2}X'T^{-1/2} \) with a method tailored for white noise, and then unwhiten:

\[
\hat{X}' = S^{1/2}\hat{X}T^{1/2}
\]

- The mean squared error is then

\[
\|\hat{X}' - X'\|^2_F = \|S^{1/2}\hat{X}T^{1/2} - S^{1/2}XT^{1/2}\|^2_F = \|S^{1/2}(\hat{X} - X)T^{1/2}\|^2_F,
\]

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

- Assume rank 1 sampling, with row and column sampling probabilities $P$ and $Q$.

- It can be shown (Dobriban, Leeb and Singer, 2020; and Leeb, 2020) that

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

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

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

- The mean squared error is then:

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

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 $\lambda_j$ with new singular values $t_k$, leaving the observed singular vectors fixed:

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

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

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

  1. The angles between the population singular vectors $u_k$ and $v_k$ of $X$ and the empirical singular vectors $\hat{u}_k$ and $\hat{v}_k$ of $Y$.

  2. Estimates of the population singular values $t_k$ of $X$ from the empirical singular values $\lambda_k$ of $Y$.

• 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:

\[
\begin{align*}
\langle u_j, \hat{u}_k \rangle^2 &\rightarrow c_{j,k}^2 = \begin{cases} 
\frac{1-\gamma/t_k^4}{1+\gamma/t_k^2}, & \text{if } j = k \text{ and } t_k^2 > \sqrt{\gamma} \\
0, & \text{otherwise}
\end{cases} \\
\langle v_j, \hat{v}_k \rangle^2 &\rightarrow \tilde{c}_{j,k}^2 = \begin{cases} 
\frac{1-\gamma/t_k^4}{1+1/t_k^2}, & \text{if } j = k \text{ and } t_k^2 > \sqrt{\gamma} \\
0, & \text{otherwise}
\end{cases}
\end{align*}
\]
• The asymptotic mean squared error (AMSE) is:

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

• 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, $\mathcal{L}(\hat{X}, X) = \|\Omega(\hat{X} - X)\Pi^T\|_F^2$?

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

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

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

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

$$\hat{B} = \arg\min_{\hat{B}'} \mathcal{L}(\hat{U} \hat{B}' \hat{V}^T, X)$$
Optimal spectral denoising:

• Solving for the optimal $\hat{B}$ is easy in principle:

$$\hat{B} = D^{-1}C \text{diag}(t)\tilde{C}^T\tilde{D}^{-1},$$

where $t = (t_1, \ldots, t_r)$, and

$$D = \hat{U}^T\Omega^T\Omega\hat{U}$$

$$\tilde{D} = \hat{V}^T\Pi^T\Pi\hat{V}$$

$$C = \hat{U}^T\Omega^T\Omega U$$

and

$$\tilde{C} = \hat{V}^T\Pi^T\Pi V$$

• These are the matrices of weighted inner products between singular vectors of $X$ and $Y$. 
Estimating $\hat{B} = D^{-1}C\text{diag}(t)\tilde{C}^T\tilde{D}^{-1}$:

- The singular values $t_1, \ldots, t_r$ are estimable, as we’ve seen.

- The matrices $D = \hat{U}^T \Omega^T \Omega \hat{U}$ and $\tilde{D} = \hat{V}^T \Pi^T \Pi \hat{V}$ are observed.

- We must estimate $C$ and $\tilde{C}$, or all inner products of the form

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

and

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

for $1 \leq k, l \leq r$.

- We will show the formulas on the next slide.
Estimating the weighted inner products:

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

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

and

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

where

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

and

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

- Note that $c_k$ and $\tilde{c}_k$ are estimable, as we’ve seen already.
Sketch of the derivation:

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

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

- The unit vector $\tilde{u}_k$ is orthogonal to $u_1, \ldots, u_r$, and uniformly random.

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

$$\tilde{u}_k^T A \tilde{u}_k \sim \frac{1}{p} \text{tr}(A).$$
Sketch of the derivation:

• Applying $\Omega$ gives:

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

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

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

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

from which we can solve for $\|\Omega u_k\|^2$. 
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.

From these weighted inner products, we can estimate $\hat{\mathbf{B}} = D^{-1} C \text{diag}(t) \tilde{C}^T \tilde{D}^{-1}$, and the optimal spectral denoiser $\hat{\mathbf{X}} = \hat{\mathbf{U}} \hat{\mathbf{B}} \hat{\mathbf{V}}^T$. 
Submatrix estimation:

- We estimate a submatrix \( X_0 = \Omega X \Pi^T \) of \( X \) by estimating \( X \) using spectral denoising with loss \( \mathcal{L}(\hat{X}, X) = \|\Omega(\hat{X} - X)\Pi^T\|_F^2 \), and taking \( \hat{X}_0 = \Omega \hat{X} \Pi^T \).

- We compare this approach with optimal singular value shrinkage applied to \( Y_0 = \Omega Y \Pi^T \).

Errors are plotted against the fraction of \( X \)'s energy contained in \( X_0 \).

- We prove that unless \( X_0 \) contains an overwhelming fraction of \( X \)'s energy, using the full matrix outperforms denoising \( Y_0 \) directly.
Heteroscedastic noise:

- Observe \( Y' = X' + N \), where \( N \) has rank 1 variance structure:
  \[
  N = S^{1/2}GT^{1/2}
  \]

- **Whiten** the noise:
  \[
  Y = S^{-1/2}Y'T^{-1/2} = S^{-1/2}X'T^{-1/2} + G \equiv X + G
  \]

- Estimate \( X = S^{-1/2}X'T^{-1/2} \) with optimal spectral denoising with weighted loss
  \[
  \mathcal{L}(\hat{X}, X) = \|S^{1/2}(\hat{X} - X)T^{1/2}\|_F^2
  \]

- Finally, unwhiten: \( \hat{X}' = S^{1/2}\hat{X}T^{1/2} \).
Comparison with OptShrink:

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

![Graph showing MSE plotted as a function of log condition number](image)

• The MSE is plotted as a function of the condition number of $S^{1/2}$ and $T^{1/2}$, 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 \( \kappa = \frac{1}{p} \text{tr}(\Sigma) \cdot \frac{1}{p} \text{tr}(\Sigma^{-1}) \), 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.
• We observe $F(Y')$, $M$ random entries of $Y' = X' + G$.
• Rank 1 sampling structure, with row and column sampling probabilities $P$ and $Q$.

![Graph showing log MSE vs log(σ)](image)

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

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

and define $\tilde{X} = P^{-1/2}\hat{X}Q^{-1/2}$.
• Compare to nuclear-norm regularized least squares of Candès and Plan (2010).
Localized denoising:

- We tile the matrix into submatrices and applying the submatrix denoising method to each one.
- The resulting method, called *localized denoising*, is never worse than singular value shrinkage but can outperform it when the matrix is *heterogeneous*.

![Images showing MIT logo denoising](image)

- Above, we denoise the MIT logo (rank 5) using singular value shrinkage (bottom left) and localized denoising (bottom right).
- The relative error of localized denoising is $7.41 \times 10^{-2}$; the relative error of singular value shrinkage is $1.25 \times 10^{-1}$. 
Summary:

- We study the problem of estimating low-rank $X$ from $Y = X + G$.

- We use weighted loss of the form $\mathcal{L}(\hat{X}, X) = \|\Omega(\hat{X} - X)\Pi^T\|_F^2$.

- We have introduced spectral denoisers of the form $\hat{X} = \hat{U}\hat{B}\hat{V}^T$.

- Using new asymptotic results for the spiked model, we derived the optimal $\hat{B}$.

- Applications include submatrix estimation; heteroscedastic noise; and missing data.
References


Additional references


Thank you

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