Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising

Charles Deledalle

Joint work with **Jérémie Bigot** and **Delphine Féral**

Institut de Mathématiques de Bordeaux
CNRS – University of Bordeaux

ISNPS, Salerno, Italy
Session High-dimensional matrix estimation
June 2018

Motivations

Problem of matrix denoising

- Estimate an unknown signal matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ from a noisy data matrix $\boldsymbol{Y}$ satisfying the model:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W},$$

where $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ is a noise matrix.

- $\boldsymbol{W}_{ij}$ are assumed to be **independent random variables** with

$$\mathbb{E}(\boldsymbol{W}_{ij}) = 0 \text{ and } \mathrm{Var}(\boldsymbol{W}_{ij}) = \tau_{ij}^2$$

for $1 \leq i \leq n$ and $1 \leq j \leq m$.

- Homoscedastic: $\tau_{ij} = \tau$ constant,
- Heteroscedastic: τ_{ij} vary (but dependent on $\boldsymbol{X}$).

Motivations

Assumption (Low-rank signal matrix)

- *The signal matrix $\boldsymbol{X}$ is assumed to have a* **low rank structure**, *with singular value decomposition (SVD)*

$$\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t .$$

- $\boldsymbol{u}_k$ *and* $\boldsymbol{v}_k$ *are the left and right singular vectors associated to the singular value* $\sigma_k > 0$, *for each* $1 \leq k \leq r^*$, *with* $\sigma_1 > \sigma_2 > \ldots > \sigma_r$.
- $0 \leq r^* \leq \min(m, n)$ *is the rank of* $\boldsymbol{X}$.

Unlike $\boldsymbol{X}$, the **noisy** data matrix $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ has almost surely full rank

$$\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

where $\tilde{\sigma}_k, \tilde{\boldsymbol{u}}_k, \tilde{\boldsymbol{v}}_k$ denotes its SVD (empirical SVD).

Motivations

Definition (Spectral estimators)

- *Given the SVD of* $$\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$
- *A* **spectral estimator:** $$\hat{\boldsymbol{X}}^f = \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k) \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

where $0 \leq f_k(\tilde{\sigma}_k) \leq \tilde{\sigma}_k$ *depends only on the singular value* $\tilde{\sigma}_k$.

Examples:

- PCA: $f_k(\tilde{\sigma}_k) = \tilde{\sigma}_k$ if $k \leq r$, 0 otherwise,
- Soft-thresholding: $f_k(\tilde{\sigma}_k) = (\tilde{\sigma}_k - \lambda)_+$,
- This talk: $f_k(\tilde{\sigma}_k) = w_k \tilde{\sigma}_k$

Goal

Ideally, one would like to select a set of functions $(f_k)_{1 \leq k \leq \min(n,m)}$ that minimize the mean-squared error (with respect to the noise $\boldsymbol{W}$)

$$\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \mathbb{E}\left(\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2 \right).$$

Not feasible since X is unknown!

Two main alternatives in the literature:

- **asymptotic optimal shrinkage** rules (setting $\min(n,m) \to \infty$) with a noise matrix $\boldsymbol{W}$ whose distribution is assumed to be **orthogonally invariant** (e.g., in the Gaussian spiked population model). (Gavish & Donoho, 2014), (Nadakuditi, 2014)
- non-asymptotic **soft-thresholding** rules which minimize an **unbiased estimate of the MSE** in the **Gaussian case**. (Candès, Sing-Long & Trzasko, 2013), (Donoho & Gavish, 2014)

Outline

1. Asymptotic optimal shrinkage
2. Non-asymptotic rules in the case of Gaussian noise
3. Generalization to noises in the exponential family
4. Active set of singular values
5. Evaluation and discussion

Asymptotic optimal shrinkage

Definition (Gaussian spiked population model)

$$\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t + \boldsymbol{W},$$

where $1 \leq r^* \leq \min(n,m)$ *is fixed and* $\boldsymbol{W}_{ij} \underset{iid}{\sim} \mathcal{N}(0, \frac{1}{m})$.

Asymptotic setting:

the sequence $m = m_n \geq n$ is such that $\lim_{n \to +\infty} \frac{n}{m} = c > 0$.

Asymptotic behavior of singular values

Proposition (Bai and Silverstein (2010))

Assume that $\boldsymbol{Y}$ is sampled from the Gaussian spiked population model. Then, for any fixed $k \geq 1$, one has that, almost surely,

$$\lim_{n \to +\infty} \tilde{\sigma}_k = \begin{cases} \rho\left(\sigma_k\right) & \text{if } \sigma_k > c^{1/4}, \\ c_+ & \text{otherwise.} \end{cases}$$

where $\rho\left(\theta\right) = \sqrt{\frac{(1+\theta^2)(c+\theta^2)}{\theta^2}}$ for any $\theta > 0$

and $c_+ = 1 + \sqrt{c}$ is the so-called **bulk edge**.

Asymptotic optimal shrinkage

Asymptotic optimal shrinkage (Gavish & Donoho, 2014)

As a consequence, the spectral estimator

$$\hat{\boldsymbol{X}}^f = \sum_{k=1}^{\min(n,m)} f(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t,$$

where

$$f(\tilde{\sigma}_k) = \begin{cases} \frac{1}{\tilde{\sigma}_k}\sqrt{(\tilde{\sigma}_k^2 - (c+1))^2 - 4c} & \text{if } \tilde{\sigma}_k > c_+, \\ 0 & \text{otherwise} \end{cases}$$

is **asymptotically optimal** in the sense that it minimizes $\lim_{n\to\infty} \|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2$ almost surely among the class of continuous spectral shrinkers that collapses the bulk to 0 (*i.e.*, $f(\tilde{\sigma}) = 0$ if $\tilde{\sigma} \leq c_+$).

Remark: equivalent expression in Nadakuditi (2014) but where the bulk edge constraint $\tilde{\sigma}_k > c_+$ is replaced by a rank assumption $k \leq r \leq r^\star$.

Non-asymptotic rules in the case of Gaussian noise

Non-asymptotic rules in the case of Gaussian noise

Alternatively, use the principle of **Stein's Unbiased Risk Estimate** *i.e.* find a data-based quantity $\mathrm{SURE}(\hat{\boldsymbol{X}}^f)$ satisfying

$$\mathbb{E}(\mathrm{SURE}(\hat{\boldsymbol{X}}^f)) = \mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \mathbb{E}\left(\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2\right).$$

Proposition (SURE, Stein 1981)

Assume f is differentiable (or at least weakly) and $\boldsymbol{W}_{ij} \underset{iid}{\sim} \mathcal{N}(0, \tau^2)$. If

$$\mathbb{E}\left(\left|\hat{\boldsymbol{X}}^f_{ij}\right|\right) < +\infty, \text{ for all } 1 \leq i \leq n,\ 1 \leq j \leq m,$$

then, the quantity

$$\mathrm{SURE}(\hat{\boldsymbol{X}}^f) = \|\hat{\boldsymbol{X}}^f - \boldsymbol{Y}\|_F^2 - mn\tau^2 + 2\tau^2 \,\mathrm{div}\left(\hat{\boldsymbol{X}}^f\right)$$

is an **unbiased estimator** *of $\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$, where*

$$\mathrm{div}\left(\hat{\boldsymbol{X}}^f\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\partial \hat{\boldsymbol{X}}^f_{ij}}{\partial \boldsymbol{Y}_{ij}}$$

Non-asymptotic rules in the case of Gaussian noise

Proposition (Candès, Sing-Long & Trzasko (2013))

If the functions $f_1, \ldots, f_{\min(n,m)}$ (acting on the singular values) are differentiable, then

$$\begin{aligned} \operatorname{div}\left(\hat{\boldsymbol{X}}^f\right) &= |m-n| \sum_{k=1}^{\min(n,m)} \frac{f_k(\tilde{\sigma}_k)}{\tilde{\sigma}_k} + \sum_{k=1}^{\min(n,m)} f_k'(\tilde{\sigma}_k) \\ &\quad +2 \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k) \sum_{\ell=1;\ell\neq k}^{\min(n,m)} \frac{\tilde{\sigma}_k}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}. \end{aligned}$$

In CST (2013), this formula leads to **data-dependent soft-thresholding**

$$f_k(\tilde{\sigma}_k) = (\tilde{\sigma}_k - \lambda)_+, \text{ for all } 1 \leq k \leq \min(n,m),$$

relevant for Gaussian noise and where $\lambda > 0$ is a parameter chosen to minimize $\mathrm{SURE}(\hat{\boldsymbol{X}}^f)$.

Non-asymptotic rules in the case of Gaussian noise

We consider the class of spectral estimators of the form

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

with w_k non-negative weights and $1 \leq r \leq \min(n,m)$ a targeted rank.

Default setting: Choose r as the largest integer such that $\tilde{\sigma}_k > c_+$.

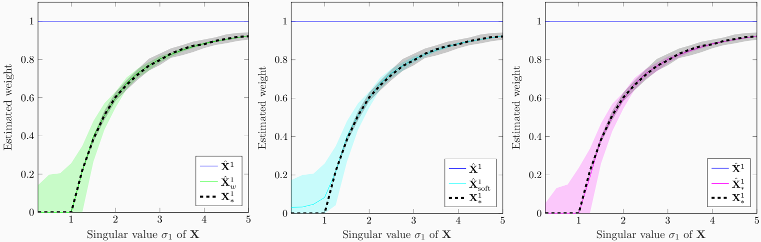
Proposition (Bigot, D. and Féral, 2017)

Assume $\boldsymbol{W}_{ij} \underset{iid}{\sim} \mathcal{N}(0,\tau^2)$. *Computing the weights minimizing* $\mathrm{SURE}(\hat{\boldsymbol{X}}_w)$ *leads to the choice*

$$w_k = \left(1 - \frac{\tau^2}{\tilde{\sigma}_k^2}\left(1 + |m-n| + 2\sum_{\ell=1;\ell\neq k}^{\min(n,m)} \frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right)\right)_+$$

for all $1 \leq k \leq r$.

Numerical experiments – $m = n = 100$ **with** $r = r^* = 1$
100 Gaussian noises $\tau^2 = 1/m$



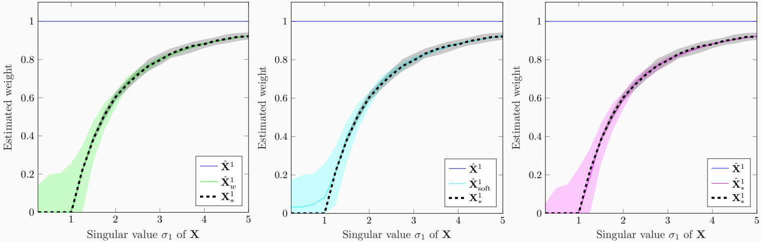
SURE / Soft-Thresholding / Asymptotic rule (Gavish & Donoho, 2014)

The black curve is an oracle rule (minimizing the true MSE)

Non-asymptotic rules in the case of Gaussian noise

Numerical experiments – $m = n = 100$ **with** $r = r^* = 1$

100 Gaussian noises $\tau^2 = 1/m$



SURE / Soft-Thresholding / Asymptotic rule (Gavish & Donoho, 2014)

The black curve is an oracle rule (minimizing the true MSE)

Is our non-asymptotic rule, asymptotically optimal?

Non-asymptotic rules in the case of Gaussian noise

Proposition (Bigot, D. and Féral (2017))

Assume that $\boldsymbol{Y}$ is sampled from the Gaussian spiked population model. Then, for any fixed $1 \leq k \leq r^*$ such that $\sigma_k > c^{1/4}$, one has that, almost surely,

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{\ell=1; \ell \neq k}^{n} \frac{\tilde{\sigma}_k}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2} = \frac{1}{\rho(\sigma_k)} \left(1 + \frac{1}{\sigma_k^2}\right).$$

A direct consequence is that our **spectral estimator is asymptotically optimal** (same limit as in by Gavish & Donoho (2014), Nadakuditi (2014)).

$$\underset{\hat{\boldsymbol{X}}_w^r}{\operatorname{arginf}} \lim_{n \to \infty} \|\hat{\boldsymbol{X}}_w^r - \boldsymbol{X}\|_F^2 = \lim_{n \to \infty} \underset{\hat{\boldsymbol{X}}_w^r}{\operatorname{arginf}} \operatorname{SURE}(\hat{\boldsymbol{X}}_w^r) = \underset{\hat{\boldsymbol{X}}_w^r}{\operatorname{arginf}} \lim_{n \to \infty} \operatorname{SURE}(\hat{\boldsymbol{X}}_w^r)$$

No optimal rules for non-Gaussian noise.
Is there instead a non-asymptotic rule in this case?

Generalization to noises in the exponential family

Generalization to noises in the exponential family

Assumption (Noise in the exponential family)

We assume the noise $\boldsymbol{W}$ is such that the distribution of $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ belongs to **the exponential family** *(with independent entries) parameterized by $\boldsymbol{X}$ and such that $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}$.*

The random variable $\boldsymbol{Y}_{ij}$ is sampled from a continuous or discrete exponential family of distributions on $\mathbb{R}$ with pdf

$$q(y; \boldsymbol{X}_{ij}) = h(y) \exp\left(\eta(\boldsymbol{X}_{ij})y - A(\eta(\boldsymbol{X}_{ij}))\right), \; y \in \mathbb{R},$$

where

- η (the link function) is a one-to-one and smooth function,
- A (the log-partition function) is a twice differentiable mapping,
- h is a known function,
- $\boldsymbol{X}_{ij} \in \mathbb{R}$ is an unknown real parameter of interest.

Remark: $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X} \Rightarrow A'(\eta(x)) = x \qquad (A'$ should be one-to-one).

Examples of noise models in the exponential family

- Homoscedastic and **known** variance in the **Gaussian case**

$$q(y; \boldsymbol{X}_{ij}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \boldsymbol{X}_{ij})^2}{2\tau^2}\right), \text{ and } \operatorname{Var}(\boldsymbol{Y}_{ij}) = \tau$$

- Heteroscedastic and **unknown** variance (but function of $\boldsymbol{X}$) in the **Gamma case** (with **known** shape parameter $L > 0$)

$$q(y; \boldsymbol{X}_{ij}) = \frac{L^L y^{L-1}}{\Gamma(L) \boldsymbol{X}_{ij}^L} \exp\left(-L \frac{y}{\boldsymbol{X}_{ij}}\right) \mathbb{1}_{\mathbb{R}+}(y), \text{ and } \operatorname{Var}(\boldsymbol{Y}_{ij}) = \frac{\boldsymbol{X}_{ij}^2}{L}$$

Spectral estimator in the natural parameter space

Consider the pdf of $\boldsymbol{Y}_{ij}$ in the **canonical form**:

$$p(y; \boldsymbol{\theta}_{ij}) = h(y) \exp\left(\boldsymbol{\theta}_{ij} y - A(\boldsymbol{\theta}_{ij})\right) \quad \text{where } \boldsymbol{\theta}_{ij} = \eta(\boldsymbol{X}_{ij}) \in \Theta$$

Generalized SURE formula are available for $\hat{\boldsymbol{\theta}}^f \in \mathbb{R}^{n \times m}$ whose entries are

$$\hat{\boldsymbol{\theta}}^f_{ij} = \eta(\hat{\boldsymbol{X}}^f_{ij}), \text{ for all } 1 \leq i \leq n, \ 1 \leq j \leq m,$$

where $f_{ij}(\boldsymbol{Y})$ is the (i,j)-th entry of the matrix $\hat{\boldsymbol{X}}^f$.

(Hudson, 1978), (Stein, 1981), (Raphan and Simoncelli, 2007), (Eldar, 2009)

Alternative to measuring the risk in the natural parameter space?

Definition

- *The mean-squared error (MSE) risk of* $\hat{\boldsymbol{\theta}}^f$ *is defined as*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathbb{E}\left(\|\hat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}\|_F^2\right) = \mathbb{E}\left(\|\eta(\hat{\boldsymbol{X}}^f) - \eta(\boldsymbol{X})\|_F^2\right) \neq \mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$$

- *The Kullback-Leibler (KL) risk of* $\hat{\boldsymbol{\theta}}^f$ *is defined as*

$$\begin{aligned}\mathrm{KL}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left(\int_{\mathbb{R}}\log\left(\frac{p(y;\hat{\boldsymbol{\theta}}^f_{ij})}{p(y;\boldsymbol{\theta}_{ij})}\right)p(y;\hat{\boldsymbol{\theta}}^f_{ij})dy\right)\\ &= \mathrm{KL}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})\end{aligned}$$

Remark: KL is invariant to the reparameterization $\hat{\boldsymbol{\theta}}^f = \eta(\hat{\boldsymbol{X}}^f)$ since it is a discrepancy measure between distributions!

Stein Unbiased estimator for Kullback Leibler risk

Proposition (Bigot, D. and Féral (2017))

Assume that the function h is C^1 on $\mathbb{R}$. Suppose that the function A is C^2 on Θ. If the following condition holds

$$\mathbb{E}\left(\left|A'(\hat{\boldsymbol{\theta}}^f_{ij})\right|\right) < +\infty, \text{ for all } 1 \leq i \leq n,\ 1 \leq j \leq m,$$

then, if f is differentiable, the quantity

$$\mathrm{SUKL}(\hat{\boldsymbol{\theta}}^f) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left(\hat{\boldsymbol{\theta}}^f_{ij} + \frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\right)A'(\hat{\boldsymbol{\theta}}^f_{ij}) - A(\hat{\boldsymbol{\theta}}^f_{ij})\right) + \mathrm{div}\left(\hat{\boldsymbol{X}}^f\right),$$

where

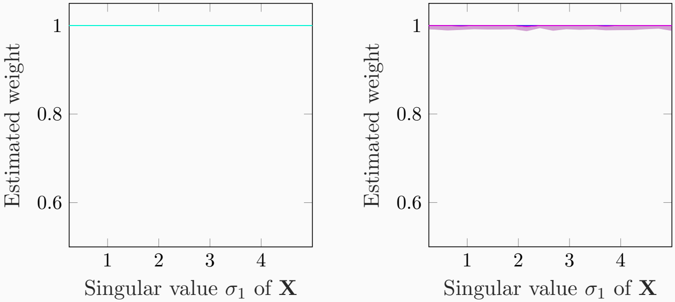
$$\mathrm{div}\left(\hat{\boldsymbol{X}}^f\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial \hat{\boldsymbol{X}}^f_{ij}}{\partial \boldsymbol{Y}_{ij}}.$$

is an **unbiased estimator** *of* $\mathrm{KL}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) - \sum_{i=1}^{n}\sum_{j=1}^{m} A(\boldsymbol{\theta}_{ij})$.

Generalization to noises in the exponential family

Gamma distributed measurements: $m = n = 100$, $r = r^* = 1$

SUKL (MKL risk) / GSURE (MSE risk)



Optimal data-driven weights

What if $r^\star > 1$?

Active set of singular values

Active set of singular values

A problem of model selection

- **Gaussian case:** choose an estimator collapsing the bulk to 0 of the form

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

  where r is the largest integer such that $\tilde{\sigma}_k > c_+$.

- **Non-Gaussian cases:** no notions of bulk edge. We will consider

$$\tilde{\boldsymbol{X}}_w^s = \sum_{k \in s} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$$

  for a subset $s \subseteq \mathcal{I} = \{1, 2, \ldots, \min(n, m)\}$.

Question: how to select a relevant subset $s^\star$?

Active set of singular values

The case of Gaussian noise

In the Gaussian case, the bulk edge constraint leads us to consider:

$$s^\star = \{k \; ; \; \tilde{\sigma}_k > c_+^{n,m}\} \text{ with } c_+^{n,m} = 1 + \sqrt{\frac{n}{m}}.$$

Proposition

Assume that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ where the entries of $\boldsymbol{W}$ are iid Gaussian variables with zero mean and standard deviation $\tau = 1/\sqrt{m}$. Then, we have

$$s^* \in \underset{s \subseteq \mathcal{I}}{\arg\min} \; m\|\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s\|_F^2 + |s| \left(\sqrt{m} + \sqrt{n}\right)^2,$$

where $\tilde{\boldsymbol{X}}^s = \sum_{k \in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ for $s \in \mathcal{I} = \{1, 2, \ldots, \min(n, m)\}$, and $|s|$ is the cardinal of s.

Remark: we have shown that $|s| \left(\sqrt{m} + \sqrt{n}\right)^2$ is an upper bound of the degree of freedom (in the sense of Efron (2004)) such that the above rule can be seen as Akaike Information Criterion (AIC) (Akaike, 1974).

The general case of an exponential family

This allows us to introduce a rule for non-Gaussian noise.

Definition

The AIC associated to $\tilde{\boldsymbol{X}}^s = \sum_{k \in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ *is*

$$\mathrm{AIC}(\tilde{\boldsymbol{X}}^s) = -2 \log q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) + |s| \left(\sqrt{m} + \sqrt{n}\right)^2,$$

where $|s|$ *is the cardinal of* s*, and*

$$q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) = \prod_{i=1}^{n} \prod_{j=1}^{m} q(\boldsymbol{Y}_{ij}; \tilde{\boldsymbol{X}}_{ij}^s)$$

is the likelihood given the data $\boldsymbol{Y}$ *are sampled from the exponential family with estimated parameters* $\boldsymbol{X}_{ij} = \tilde{\boldsymbol{X}}_{ij}^s$*.*

Evaluation and discussion

Algorithmic approach and numerical optimization

Given an active set $s^\star$ of singular values, we compute a spectral estimator of the form

$$\hat{\boldsymbol{X}}_w = \sum_{k \in s^\star} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

where optimal weights w_k for $k \in s^\star$ are obtained by (exact or numerical) minimization of an **unbiased risk formula**.

Remark: for Gamma noise, numerical optimization has to be used to find the optimal weights with the constraint that the entries of $\hat{\boldsymbol{X}}_w$ remain positive.

Matlab codes available at:

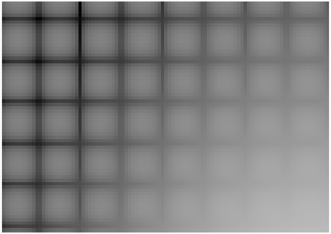
https://www.math.u-bordeaux.fr/~cdeledal/gsure_low_rank.php

Setting of numerical experiments

Consider the setting where $r^* \geq 2$ is unknown and

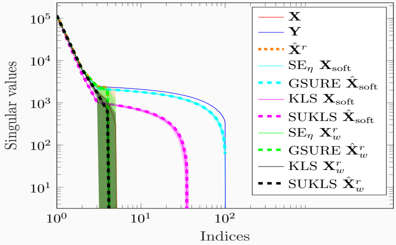
$$\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t,$$

where $\boldsymbol{u}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^m$ are fixed unit vectors, and σ_k are fixed positive real values (with $n = 100$ and $m = 200$) such that $\boldsymbol{X}_{ij} \geq 0$.



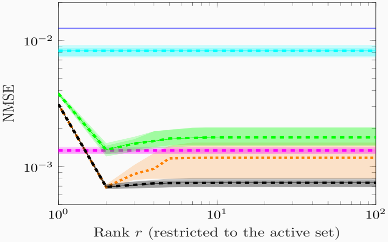
Monte-Carlo simulations with $M = 200$ repetitions

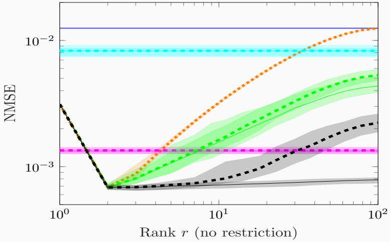
The case of Gamma noise - with confidence bands



$\boldsymbol{X}$ oracle (based on true risk),
$\hat{\boldsymbol{X}}$ based on estimated risk.

$\hat{\boldsymbol{X}}^r$ PCA of rank r,
$\hat{\boldsymbol{X}}_{\text{soft}}$ soft-thresholding,
$\hat{\boldsymbol{X}}^r_w$ our estimator.





s^* based on AIC versus $s^* = \{1 \leq k \leq r\}$

Summary in one slide: a two step procedure

- estimation of an **active set** $s^\star \subseteq \mathcal{I} = \{1, 2, \ldots, \min(n, m)\}$ **of singular values** using a criterion inspired by AIC's model selection

$$s^* \in \underset{s \subseteq \mathcal{I}}{\arg\min} \; -2 \log q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) + |s| \left(\sqrt{m} + \sqrt{n}\right)^2,$$

- given the knowledge of $s^\star$, compute a spectral estimator of the form

$$\hat{\boldsymbol{X}}_w = \sum_{k \in s^\star} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t,$$

  where optimal weights w_k for $k \in s^\star$ are obtained by minimizing an **unbiased estimation formula** of the mean Kullback-Leibler (**MKL**) risk.

Open questions: How to extend the asymptotic analysis to the spiked population model for non-Gaussian noise, and to derive asymptotically optimal shrinkage rules? Beyond the exponential family?

Thanks for your attention!

- **Further reading:**

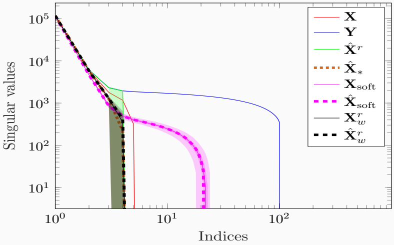
Bigot, J., Deledalle, C. and Féral, D. (2017). *Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising,* Journal of Machine Learning Research, 18(1), 4991-5040.

Deledalle, C. A. (2017). Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. Electronic journal of statistics, 11(2), 3141-3164.

- **Online code:**

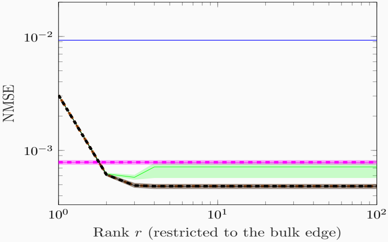
https://www.math.u-bordeaux.fr/~cdeledal/gsure_low_rank.php

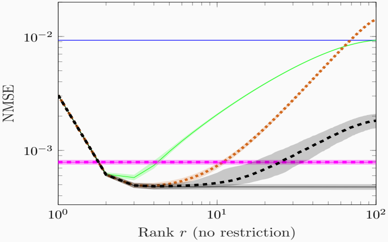
The case of Gaussian noise - with confidence bands



$\boldsymbol{X}$ oracle (based on true risk),
$\hat{\boldsymbol{X}}$ based on estimated risk.

$\hat{\boldsymbol{X}}^r$ PCA of rank r,
$\hat{\boldsymbol{X}}_*$ optimal asymptotic rule,
$\hat{\boldsymbol{X}}_{\mathrm{soft}}$ soft-thresholding,
$\hat{\boldsymbol{X}}_w^r$ our estimator.





$s^* = \{1 \leq k \leq r$ such that $\tilde{\sigma}_k > c_+^{n,m}\}$ versus $s^* = \{1 \leq k \leq r\}$

Definition (Efron (2004))

The degrees of freedom (DOF) of a given estimator $\hat{\boldsymbol{X}}$ *is defined as*

$$\mathrm{DOF}(\hat{\boldsymbol{X}}) = \frac{1}{\tau^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{Cov}(\hat{\boldsymbol{X}}_{ij}, \boldsymbol{Y}_{ij}) = \frac{1}{\tau^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}(\hat{\boldsymbol{X}}_{ij}\boldsymbol{W}_{ij}).$$

Proposition (Bigot, D. and Féral (2017))

Assume that $\boldsymbol{Y}$ *is sampled from the Gaussian spiked population model. Suppose that* $\hat{\boldsymbol{X}}^f$ *is a spectral estimator such that each function* f_k *is smooth, and that* $\sigma_k > c^{1/4}$ *for all* $1 \leq k \leq r \leq r^*$*. Then, one has that a.s.*

$$\lim_{n\to+\infty}\frac{1}{m}\mathrm{DOF}(\hat{\boldsymbol{X}}^f) = \sum_{k=1}^{r}\frac{f_k(\rho(\sigma_k))}{\rho(\sigma_k)}\left(1 + c + \frac{2c}{\sigma_k^2}\right).$$

Hence, if $\sigma_k^2 > \sqrt{c}$ for all $1 \leq k \leq r \leq r^*$, it follows that if $s \subseteq \{1, \ldots, r\}$ then

$$\lim_{n \to +\infty} \frac{1}{m} \mathrm{DOF}(\tilde{\boldsymbol{X}}^s) = |s| \left(1 + c + \frac{2c}{\sigma_k^2}\right) \leq |s| \left(1 + \sqrt{c}\right)^2 = |s| c_+^2,$$

where

$$\tilde{\boldsymbol{X}}^s = \sum_{k \in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t.$$

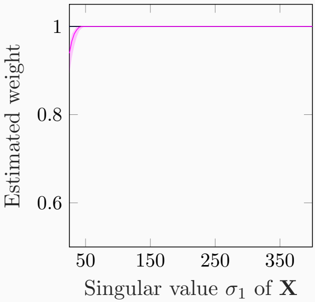
Hence, the quantity

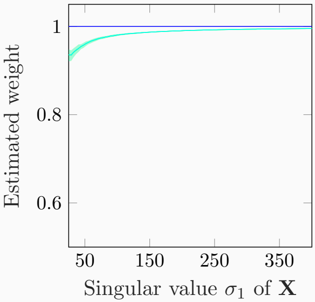
$$2|s| p_{n,m} = |s| \left(\sqrt{m} + \sqrt{n}\right)^2$$

is asymptotically an upper bound of $\mathrm{DOF}(\tilde{\boldsymbol{X}}^s)$ (when normalized by $1/m$) for any given set $s \subseteq \{1, \ldots, r\}$.

SURE formula also available for the case of Poisson noise

PUKL (MKL risk) / PURE (MSE risk)





Optimal data-driven weights

Extra slides

Example

Gamma noise with shape parameter $L > 0$: $\tau_{ij}^2 = \mathrm{Var}(\boldsymbol{Y}_{ij}) = \frac{\boldsymbol{X}_{ij}^2}{L}$

Consider rank-one approximation $r = 1$ with the spectral estimator

$$\hat{\boldsymbol{X}}_w = \eta(\hat{\boldsymbol{\theta}}_w) \quad \text{where} \quad \hat{\boldsymbol{X}}_w = w_1 \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t, \text{ for some } w_1 \geq 0.$$

Computing the weights minimizing $\mathrm{SUKL}(\hat{\boldsymbol{\theta}}_w)$ leads to the choice

$$w_1(\boldsymbol{Y}) = \frac{L/mn}{L-1} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\tilde{\sigma}_1 \boldsymbol{\alpha}_{ij}}{\boldsymbol{Y}_{ij}} + \frac{1}{(L-1)} \left(1 + |m-n| + 2 \sum_{\ell=2}^{\min(n,m)} \frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2} \right) \right)^{-1}$$

where $\boldsymbol{\alpha}_{ij}$ denotes the (i, j)-th entry of the $n \times m$ matrix $\boldsymbol{\alpha} = \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$.

Remark: no closed-form expressions for the weights minimizing $\mathrm{SUKL}(\hat{\boldsymbol{\theta}}_w)$ (neither for $\mathrm{GSURE}(\hat{\boldsymbol{\theta}}_w)$) beyond the case $r = 1$!