

Adaptive shrinkage of singular values

Julie Josse
Agrocampus Ouest
 and
 Sylvain Sardy
Université de Genève

November 25, 2014

Abstract

To recover a low rank structure from a noisy matrix, truncated singular value decomposition has been extensively used and studied. Recent studies suggested that the signal can be better estimated by shrinking the singular values. We pursue this line of research and propose a new estimator offering a continuum of thresholding and shrinking functions. To avoid an unstable and costly cross-validation search, we propose new rules to select two thresholding and shrinking parameters from the data. In particular we propose a generalized Stein unbiased risk estimation criterion that does not require knowledge of the variance of the noise and that is computationally fast. A Monte Carlo simulation reveals that our estimator outperforms the tested methods in terms of mean squared error on both low-rank and general signal matrices across different signal to noise ratio regimes. In addition, it accurately estimates the rank of the signal when it is detectable.

Keywords: denoising, singular values shrinking and thresholding, Stein's unbiased risk estimate, adaptive trace norm, rank estimation

1 Introduction

In many applications such as image denoising, signal processing, collaborative filtering, it is common to model the data \mathbf{X} , an $N \times P$ matrix, as

$$\mathbf{X} = \mathbf{W} + \mathbf{E}, \quad (1)$$

where the unknown matrix \mathbf{W} is measured with i.i.d. $N(0, \sigma^2)$ errors \mathbf{E} . The matrix \mathbf{W} is assumed to have low rank $R < \min(N, P)$, which means that its singular value decomposition (SVD) $\mathbf{W} = \mathbf{PDQ}^T$ has R non-zero singular values $d_1 \geq \dots \geq d_R$. Note that model (1) is also known as bilinear model [Mandel, 1969] in analysis of variance, as fixed factor score model [de Leeuw et al., 1985] or fixed effect models [Cauissinus, 1986] in principal component analysis. Such models describe well data in many sciences, such as genotype-environment data in agronomy, or relational data in social science and in biological networks, where the variation between rows and columns is of equal interest [Hoff, 2007].

To denoise the data, an old approach consists in performing the SVD of the matrix $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ and defining $\hat{\mathbf{W}} = \hat{\mathbf{P}}\hat{\mathbf{D}}\hat{\mathbf{Q}}^T$ with $\hat{\mathbf{P}} = \mathbf{U}$, $\hat{\mathbf{Q}} = \mathbf{V}$ and keeping the first R singular values while setting the others to zero. In other words, the so-called truncated SVD keeps the empirical directions \mathbf{U} and \mathbf{V} , and estimates the singular values by

$$\hat{d}_i = \lambda_i \cdot 1(i \leq R) = \lambda_i \cdot 1(\lambda_i \geq \tau), \quad (2)$$

which can be parametrized either in the rank R or the threshold τ (here $1(\cdot)$ is the indicator function). This estimate is also solution [Eckart and Young, 1936] to

$$\min_{\mathbf{W} \in \mathbb{R}^{N \times P}} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{W}) \leq R, \quad (3)$$

where $\|\mathbf{M}\|_F$ is the Frobenius norm of the matrix \mathbf{M} . The truncation (2) and its penalty formulation (3) are reminiscent of the hard thresholding [Donoho and Johnstone, 1994] solution to best subset variable selection in regression. The truncated SVD requires R as tuning parameter, which can be selected using cross-validation [Owen and Perry, 2009, Josse and Husson, 2012] or Bayesian considerations [Hoff, 2007]. While this approach is still extensively used, recent studies [Chatterjee, 2013, Donoho and Gavish, 2014a] suggested an optimal hard threshold for singular values with better asymptotic mean squared error than thresholding at R or at the bulk edge (the limit of detection). More precisely, Donoho and Gavish [2014a] considered the asymptotic framework, in which the matrix size is much larger than the rank of the signal matrix to be recovered, and the signal-to-noise ratio of the low-rank piece stays constant while the matrix grows, and showed that the optimal threshold is $(4/\sqrt{3}\sqrt{p}\sigma)$ in the case of a square ($p \times p$) matrix and σ known. The other thresholds for the cases of rectangular matrices and unknown σ are also detailed in their paper.

Another popular and recent estimation strategy consists in applying a soft thresholding rule to the singular values

$$\hat{d}_i = \lambda_i \max(1 - \tau/\lambda_i, 0), \quad (4)$$

where any singular value smaller than the threshold τ is set to zero. The estimate $\hat{\mathbf{W}} = \mathbf{U}\hat{\mathbf{D}}\mathbf{V}^T$ with \hat{d}_i in (4) is also the closed form solution [Mazumder et al., 2010, Cai et al., 2010] to

$$\min_{\mathbf{W} \in \mathbb{R}^{N \times P}} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\|_F^2 + \tau \|\mathbf{W}\|_*, \quad (5)$$

where $\|\mathbf{W}\|_* = \sum_{i=1}^{\min(N,P)} d_i$ is the trace norm of the matrix \mathbf{W} . The regularization (4) and its penalty formulation (5) are inspired by soft thresholding [Donoho and Johnstone, 1994] and lasso [Tibshirani, 1996]. The tuning parameter τ is often selected by cross-validation. Recently, Candès et al. [2013] defined a Stein unbiased risk estimate (SURE) [Stein, 1981] to select τ more efficiently considering the noise variance σ^2 as known.

Finally, other reconstruction schemes involving nonlinear shrinkage of the singular values have been proposed in the literature [Verbanck et al., 2013, Raj Rao, 2013, Shabalin and Nobel, 2013]. More precisely, using the same asymptotic framework as previously and asymptotic results on the distribution of the singular values and singular vectors [Johnstone, 2001, Baik and Silverstein, 2006, Paul, 2007], Shabalin and Nobel [2013] and Gavish and Donoho

[2014] showed that the shrinkage estimator $\hat{\mathbf{W}}$ closest to \mathbf{W} in term of mean squared error has the form $\hat{\mathbf{W}} = \mathbf{U}\hat{\mathbf{D}}\mathbf{V}^T$ with, when $\sigma = 1$,

$$\hat{d}_i = \frac{1}{\lambda_i} \sqrt{(\lambda_i^2 - \beta - 1)^2 - 4\beta} \cdot 1(i \geq (1 + \sqrt{\beta})), \quad (6)$$

and $N/P \rightarrow \beta \in (0, 1]$. The case with unknown σ is also covered in their paper.

With different asymptotics, considering that the noise variance tends to zero and N and P are fixed, Verbanck et al. [2013] also reached similar estimates and suggested the following heterogeneous shrinkage estimate

$$\hat{d}_i = \lambda_i \left(1 - \frac{\sigma^2}{\lambda_i^2}\right) \cdot 1(i \leq R). \quad (7)$$

Unlike soft thresholding, the smallest singular values are more shrunk than the largest ones. It is a two-step procedure: first select R , then shrink the R largest singular values. It is a compromise between hard and soft thresholding.

In regression, Zou [2006] also bridged the gap between soft and hard thresholdings by defining the adaptive lasso estimator governed by two parameters chosen by cross-validation to control thresholding and shrinkage. To avoid expensive resampling, Sardy [2012] selected them by minimizing a Stein unbiased estimate of the risk. Adaptive lasso has oracle properties and has shown good results in terms of prediction accuracy, especially using Stein unbiased risk.

In this paper, we propose in Section 2.1 an estimator inspired by adaptive lasso to recover \mathbf{W} . It thresholds and shrinks the singular values in a single step using two parameters that parametrize a continuum of thresholding and shrinking functions. We propose in Section 2.2 simple though efficient strategies to select the two tuning parameters from the data, without relying on the unstable and costly cross-validation. One approach consists in estimating the ℓ_2 -loss, the other in selecting the threshold at the detection limit, estimated empirically given the data matrix \mathbf{X} , and the last one can be applied when the variance σ^2 is unknown. Finally, we assess the method on simulated data in Section 3 and show that it outperforms the state of the art methods in terms of mean squared error and rank estimation.

2 Adaptive trace norm

2.1 Definition

The past evolution of regularization for reduced rank matrix estimation reveals that the empirical singular values should not simply be thresholded (with hard thresholding), but should also be shrunk (with soft thresholding) or more heavily shrunk, as in (6) and (7). Inspired by adaptive lasso [Zou, 2006], we propose a continuum of functions indexed by two parameters (τ, γ) . The following thresholding and shrinkage function,

$$\hat{d}_i = \lambda_i \max\left(1 - \frac{\tau^\gamma}{\lambda_i^\gamma}, 0\right), \quad (8)$$

is defined for a positive threshold τ , and encompasses soft thresholding (4) for $\gamma = 1$ and hard thresholding (2) when $\gamma \rightarrow \infty$. We call the associated estimator

$$\hat{\mathbf{W}}_{\tau,\gamma} = \sum_{i=1}^{\min(N,P)} U_i \lambda_i \max\left(1 - \frac{\tau^\gamma}{\lambda_i^\gamma}, 0\right) V_i' \quad (9)$$

the adaptive trace norm estimator (ATN) with $\tau \geq 0$ and $\gamma \geq 1$. Note that as a byproduct, the rank of the matrix is also estimated as $\hat{R} = \sum_i^{\min(N,P)} 1(\hat{d}_i \geq 0)$.

In comparison to the hard and soft thresholding rules, the advantage of using the single and more flexible thresholding and shrinkage function (8) is twofold. First (8) parametrizes a rich family of functions that can more closely approach an ideal thresholding and shrinking function to recover well the structure of the underlying matrix \mathbf{W} , given the noise level σ^2 . Second, the specific multiplicative factors $(1 - \tau^\gamma/\lambda_i^\gamma)$ fit the rationale that the largest singular values correspond to stable directions and should be shrunk mildly. In comparison to other non linear thresholding rules, (8) does not rely on asymptotic derivations. Instead it selects its parameters $(\hat{\tau}, \hat{\gamma})$ from the data, which leads to smaller MSE in many scenarii as illustrated in Section 3.

Our estimator is related to penalized Frobenius norm regularization. In a matrix completion and regression context, Mazumder et al. [2010], Gaiffas and Lecue [2011] and Chen et al. [2013] proved that, for a weakly increasing weight sequence ω , the optimization problem

$$\min_{\mathbf{W} \in \mathbb{R}^{N \times P}} \frac{1}{2} \|\mathbf{X} - \mathbf{W}\|_F^2 + \alpha \|\mathbf{W}\|_{*,\omega} \quad \text{with} \quad \|\mathbf{W}\|_{*,\omega} = \sum_{i=1}^{\min(N,P)} \omega_i d_i \quad (10)$$

has the closed form solution $\hat{\mathbf{W}} = \mathbf{U}\hat{\mathbf{D}}\mathbf{V}^T$ with $\hat{d}_i = \max(\lambda_i - \alpha\omega_i, 0)$. So the adaptive trace norm estimator (9) has weights inversely proportional to the empirical singular values, and corresponds to $\alpha = \tau^\gamma$ and $\omega_i = 1/\lambda_i^{\gamma-1}$.

2.2 Selection of τ and γ

The parameters τ and γ can be estimated using cross-validation. Leave-one-out cross-validation, first consists in removing one cell $(i; j)$ of the data matrix \mathbf{X} . Then, for one pair (τ, γ) , it consists in predicting its value using the estimator obtained from the dataset that excludes this cell. The value of the predicted cell is denoted \hat{X}_{ij}^{-ij} . Finally, the prediction error is computed $(X_{ij} - \hat{X}_{ij}^{-ij})^2$ and the operation is repeated for all the cells in \mathbf{X} and for each (τ, γ) . The pair that minimizes the error of prediction is selected. Such a procedure requires a method which provides an estimator despite the missing values. Such methods estimating the singular vectors and singular values from incomplete data exist but use computationally intensive iterative algorithms [Ilin and Raiko, 2010, Mazumder et al., 2010, Josse and Husson, 2012]. This makes the cross-validation procedure difficult to use in practice even with a K -fold strategy.

As an alternative, we suggest three methods which strength is to select τ and γ adapting to the signal \mathbf{W} and the noise level σ^2 .

2.2.1 When σ is known

The first method seeks good ℓ_2 -risk. The mean squared error $\text{MSE} = \mathbb{E}\|\hat{\mathbf{W}} - \mathbf{W}\|^2$, or risk, of the estimator $\hat{\mathbf{W}} = \mathbf{U}\hat{\mathbf{D}}\mathbf{V}^T$ depends on the unknown \mathbf{W} and

cannot be computed explicitly. However, it can be estimated unbiasedly using Stein unbiased risk estimate ($\mathbb{E}(\text{SURE}) = \text{MSE}$). For estimators that threshold and shrink singular values, SURE still has a classical form with the residual sum of squares (RSS) penalized by the divergence of the operator:

$$\text{SURE} = -NP\sigma^2 + \text{RSS} + 2\sigma^2 \text{div}(\hat{\mathbf{W}}). \quad (11)$$

However, the form of the divergence is not straightforward and Candes et al. [2013, Theorem 4.3] showed that it has the following closed form expression:

$$\text{div}(\hat{\mathbf{W}}) = \sum_{s=1}^{\min(N,P)} \left(f'_i(\lambda_i) + |N - P| \frac{f'_i(\lambda_i)}{\lambda_i} \right) + 2 \sum_{t \neq s, t=1}^{\min(N,P)} \frac{\lambda_i f'_i(\lambda_i)}{\lambda_s^2 - \lambda_t^2},$$

where $f_i(\lambda_i)$ is the thresholding and shrinking function. The derivation of SURE requires this function to be weakly differentiable in Stein sense, which means differentiable except on a set of measure zero. This is for instance the case for the soft-thresholding function used by Candes et al. [2013]. Likewise the adaptive thresholding function (8) is differentiable on \mathbb{R} except at $\lambda_i = \tau$ for the range of interest $\gamma \in [1, \infty)$. Hence SURE for adaptive trace norm is

$$\text{SURE}(\tau, \gamma) = -NP\sigma^2 + \sum_{s=1}^{\min(N,P)} \lambda_s^2 \min\left(\frac{\tau^{2\gamma}}{\lambda_s^{2\gamma}}, 1\right) + 2\sigma^2 \text{div}(\hat{\mathbf{W}}_{\tau, \gamma}), \quad (12)$$

where

$$\begin{aligned} \text{div}(\hat{\mathbf{W}}_{\tau, \gamma}) &= \sum_{s=1}^{\min(N,P)} \left(1 + (\gamma - 1) \frac{\tau^\gamma}{\lambda_s^\gamma} \right) \cdot 1(\lambda_s \geq \tau) + |N - P| \max\left(1 - \frac{\tau^\gamma}{\lambda_s^\gamma}, 0\right) \\ &+ 2 \sum_{t \neq s, t=1}^{\min(N,P)} \frac{\lambda_s^2 \max\left(1 - \frac{\tau^\gamma}{\lambda_s^\gamma}, 0\right)}{\lambda_s^2 - \lambda_t^2}. \end{aligned}$$

A selection rule for $\tau \geq 0$ and $\gamma \geq 1$ finds the pair (τ, γ) that minimizes the bivariate function $\text{SURE}(\tau, \gamma)$ in (12). It is not computationally costly, unlike cross-validation, but supposes the variance of the noise σ^2 as known.

The second selection method is primarily driven by a good estimation of the rank of the matrix \mathbf{W} . The parameter that determines the estimated rank is the threshold τ since any empirical singular value $\lambda_i \leq \tau$ is set to zero by (8). Inspired by the universal rule of Donoho and Johnstone [1994] and thresholding tests [Sardy, 2013], we propose to use the $(1 - \alpha)$ -quantile of the distribution of the largest empirical singular value λ_1 of \mathbf{X} under the null hypothesis that \mathbf{W} has rank zero to determine the selected threshold. With α tending to zero with the sample size, null rank estimation is guaranteed with probability tending to one under the null hypothesis. Donoho and Johnstone [1994] implicitly used level of order $\alpha = O(1/\sqrt{\log N})$ when $N = P$, so we choose a similar level tending to zero with the maximum of N and P . This leads to the definition of universal threshold for reduced rank mean matrix estimation:

$$\tau_{\max(N,P)} = \sigma F_{\Lambda_1}^{-1} \left(1 - \frac{1}{\sqrt{\log(\max(N, P))}} \right), \quad (13)$$

where F_{Λ_1} is the cumulative distribution function of the largest singular value under Gaussian white noise with unit variance. We then select the shrinkage parameter γ by minimizing SURE (12) in γ for $\tau = \tau_{\max(N,P)}$. In practice the finite sample distribution of Λ_1 of an $N \times P$ matrix of independent and identically distributed Gaussian random variables $N(0, 1)$ is known [Zanella et al., 2009] but difficult to use. Thus, we simulate random variables from that distribution and take the appropriate quantile to estimate $\tau_{\max(N,P)}$ in (13). Alternatively, we could use results of Shabalin and Nobel [2013], who derived the asymptotic distribution of the singular values of model (1) based on results from random matrix theory [Johnstone, 2001, Baik and Silverstein, 2006, Paul, 2007].

2.2.2 When σ is unknown

Both previous methods need as an input the noise variance σ^2 . In some applications such as image denoising [Talebi and Milanfar, 2013, Candes et al., 2013], it is known or a good estimation is available. However, very often, this is not the case and the formula (12) cannot be used as such. Inspired by generalized cross validation [Craven and Wahba, 1979], we propose generalized SURE:

$$\text{GSURE}(\tau, \gamma) = \frac{\sum_{s=1}^{\min(N,P)} \lambda_s^2 \min\left(\frac{\tau^{2\gamma}}{\lambda_s^{2\gamma}}, 1\right)}{(1 - \text{div}(\hat{\mathbf{W}}_{\tau,\gamma})/(NP))^2}. \quad (14)$$

Using a first order Taylor expansion $1/(1 - \epsilon)^2$ of (14), we get that $\text{GSURE} \approx \text{RSS} \left(1 + 2 \text{div}(\hat{\mathbf{W}}_{\tau,\gamma})/(NP)\right)$; then considering the estimate of variance $\hat{\sigma}^2 = \text{RSS}/(NP)$, one sees how GSURE approximates SURE (11).

The GSURE criterion has the great advantage of not requiring any input value and can be applied straightforwardly to select both tuning parameters.

3 Simulations

3.1 Gaussian setting with large N and P

We compare the adaptive trace norm estimator to existing ones by reproducing the simulation of Candes et al. [2013]. Here, matrices of size 200×500 are generated according to model (1) with four signal-to-noise ratios $\text{SNR} \in \{0.5, 1, 2, 4\}$ (calculated as one over $\sigma\sqrt{NP}$) and two values for the rank $R \in \{10, 100\}$. For each combination, 50 datasets are generated. We consider five estimators:

- Truncated SVD (TSVD). We use the common one with the true rank R for (2) as well as the ones proposed by Donoho and Gavish [2014a] with asymptotic MSE optimal choices of hard threshold $\tau = \lambda_*(\frac{N}{P})\sqrt{P}\sigma$ when σ is known, and $\tau = w(\frac{N}{P})\text{median}(\lambda_i)$ when σ is unknown. The values for the coefficients $\lambda_*(\frac{N}{P})$ and $w(\frac{N}{P})$ are given in their Tables 1 and 4.
- Optimal shrinkage (OS) of Shabalin and Nobel [2013] and Gavish and Donoho [2014] when σ is known as well as when σ is unknown as defined in Section 7 of Gavish and Donoho [2014].
- Singular value soft thresholding (SVST) with τ selected to minimize SURE [Candes et al., 2013] for (4), knowing σ .

R	SNR	ATN GSURE	TSVD τ	OS	Lower bound
MSE					
10	4	0.004	0.004	0.004	0.0012
100	4	0.037	0.409	0.335	0.0106
10	2	0.017	0.017	0.017	0.0024
100	2	0.142	0.755	0.606	0.0212
10	1	0.067	0.072	0.067	0.0048
100	1	0.454	1.000	0.892	0.0424
10	0.5	0.254	0.321	0.250	0.0097
100	0.5	0.978	1.000	0.994	0.0845
Rank					
10	4	11 (1.8)	10 (0.0)	10 (0.0)	
100	4	102 (1.7)	49 (1.2)	78 (0.7)	
10	2	11 (1.4)	10 (0.0)	10 (0.0)	
100	2	112 (2.8)	20 (1.6)	48 (1.3)	
10	1	11 (1.3)	10 (0.0)	10 (0.0)	
100	1	140 (4.3)	0 (0.0)	16 (1.6)	
10	0.5	15 (1.6)	10 (0.0)	10 (0.0)	
100	0.5	14 (7.6)	0 (0.0)	2 (1.2)	

Table 1: Monte Carlo results in terms of mean squared errors (top) and rank estimation with its standard deviation (bottom). R is the true rank (10 or 100) and SNR is the signal-to-noise ratio. Three fully automatically estimators are considered: adaptive trace norm (ATN) based on GSURE (14), truncated SVD (TSVD) using $\tau = \omega(0.4)\text{median}(\lambda_i)$ [Donoho and Gavish, 2014a] and optimal shrinkage (OS) using the estimation for the noise variance [Gavish and Donoho, 2014]. The lower bound of Donoho and Gavish [2014b] is indicated. Sample size is $N = 200$ individuals and number of variables is $P = 500$. Results correspond to the mean over the 50 simulations. Best results linewise are indicated in **bold**.

- the 2-step estimator with the true rank R [Verbanck et al., 2013] for (7).
- Adaptive trace norm (ATN) with three selections of the two parameters indexing a family of shrinkers. With σ known: SURE (12), and SURE as a function of γ only (τ is set to the universal threshold (13)). With σ unknown: GSURE (14).

We report in Table 1 and 2 the estimated mean squared error between the fitted matrix $\hat{\mathbf{W}}$ and the true signal \mathbf{W} , and the estimated rank (the number of singular values that are not set to zero). We also include the lower bound on worst-case MSE for any matrix denoiser as given in Donoho and Gavish [2014b] which can be used as a baseline. The standard deviations of the MSEs are very small for all the estimators and vary from the order of 10^{-5} for high SNR to 10^{-3} for small SNR. Thus, the MSEs can be directly analysed to compare the estimators. We indicate the standard deviations for the rank.

Table 1 reports the performance with no oracle information, while Table 2 is when some parameters are known, either the true rank or the true noise variance. Comparing the two Tables allows to assess the performance loss by having to estimate all parameters, like in most real life applications.

Looking at Table 1, we see that the proposed adaptive trace norm estimator performs remarkably well, owing to its flexibility (with two parameters) and to a good selection of the appropriate model with GSURE. It can even outperform oracle estimators (in Table 2) that are governed by a single parameter. Here GSURE results are very similar to its corresponding SURE results, showing that

R	SNR	ATN		TSVD		OS	SVST	2-steps
		SURE	universal	R	τ			
MSE								
10	4	0.004	0.004	0.004	0.004	0.004	0.008	0.004
100	4	0.037	0.037	0.038	0.038	0.037	0.045	0.037
10	2	0.017	0.017	0.017	0.017	0.017	0.033	0.017
100	2	0.142	0.147	0.152	0.158	0.146	0.156	0.141
10	1	0.067	0.067	0.072	0.072	0.067	0.116	0.067
100	1	0.448	0.623	0.733	0.856	0.600	0.448	0.491
10	0.5	0.253	0.251	0.321	0.321	0.250	0.353	0.257
100	0.5	0.852	0.957	3.164	1.000	0.961	0.852	1.477
Rank								
10	4	11 (1.6)	10 (0.0)		10	10	65 (2.3)	
100	4	103 (1.9)	100 (0.0)		100	100	193 (0.8)	
10	2	11 (1.0)	10 (0.1)		10	10	63 (2.2)	
100	2	114 (2.2)	100 (0.0)		100	100	181 (1.1)	
10	1	11 (1.2)	10 (0.0)		10	10	59 (1.7)	
100	1	154 (1.8)	65 (0.8)		38 (0.6)	64 (0.7)	154 (1.2)	
10	0.5	15 (1.6)	10 (0.0)		10	10	51 (2.7)	
100	0.5	87 (3.3)	16 (0.8)		0	15 (0.8)	86 (2.6)	

Table 2: Same setting as for the Monte Carlo simulations of Table 1. The same estimators are considered, except that oracle quantities, either the true rank R or the true variance σ^2 , are used. Considered estimators are: ATN with two selection rules, SURE and universal, for the two parameters, assuming σ is known; TSVD knowing true rank R and $\tau = \lambda_*(0.4)\sqrt{500}\sigma$ [Donoho and Gavish, 2014a]; Optimal shrinkage (OS) with σ known [Gavish and Donoho, 2014]; singular value soft thresholding (SVST) with σ is known [Candes et al., 2013]; two-steps knowing true rank R [Verbanck et al., 2013].

(14) is a good approximation to (11) thanks to a large value for NP . Figure 1 illustrates the striking ability of GSURE to approximate the true loss in different regimes. On the top, the estimated risk of ATN GSURE is represented as a function of (τ, γ) for $R=10$ and $\text{SNR}=1$ on the left (row 6 of Table 1) and for $R=100$ and $\text{SNR}=0.5$ (last row of Table 1) on the right. On the bottom, the true loss function is plotted as a function of τ and γ . The estimated values (top) and the optimal choice (bottom) located by a cross are close.

Looking at Table 2, we see that when the SNR is high (equal to 4 and 2), the ATN, 2-step, TSVD (with R and τ) and OS estimators give results in term of MSE of the same order of magnitude and clearly outperform the SVST approach. When the SNR decreases, the TSVD, and the 2-step and OS methods to a lower extend, collapse; this is when the SVST provides better results especially for the difficult setting when the rank $R = 100$. The good behavior of the 2-step approach in many situations highlights the fact that it is often a good strategy to apply a different amount of shrinkage to each singular value. These simulations provide good insights into the regimes for which each estimator is well suited: low noise regime for the TSVD, moderate noise regime for the 2-step, and high noise for the SVST. But, if one is interested in a single estimator, regardless of the unknown underlying structure, then ATN becomes the method of choice. Figure 2 illustrates the adaptation of ATN to various SNR and rank by representing a typical shrinking and thresholding function selecting (τ, γ) with SURE. As expected when SNR is 0.5, ATN is close to soft thresholding, whereas when SNR is 4, it is close to hard thresholding.

As far as rank estimation is concerned, even if it is not the primary objective, ATN gives a very good estimation of the rank and SVST considerably over-

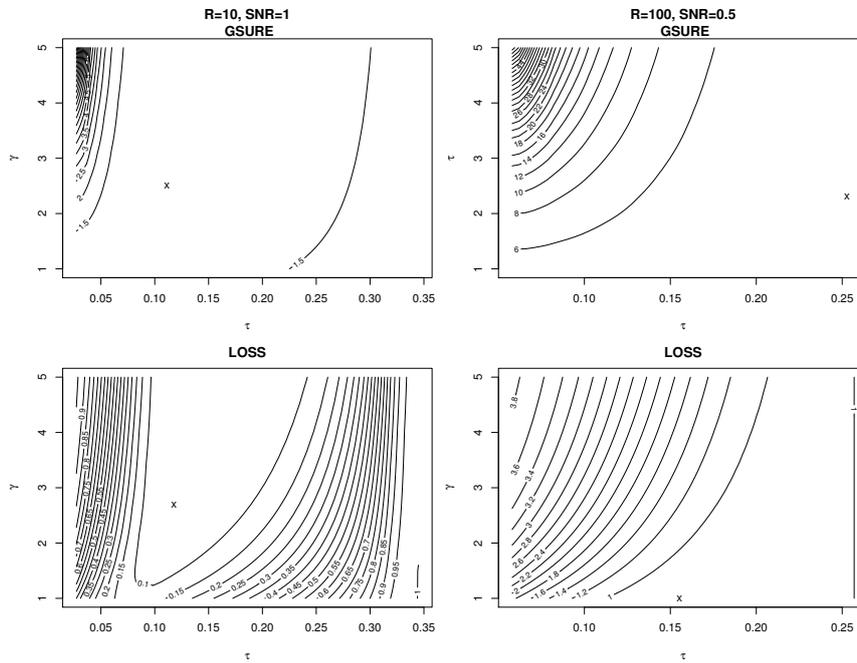


Figure 1: Top: Generalized Stein unbiased risk estimate (GSURE); Bottom: corresponding true ℓ_2 -loss. Both are plotted as a function of (τ, γ) for data generated as in Table 1. Left: true rank $R = 10$ and signal to noise ratio $\text{SNR} = 1$; Right: $R = 100$ and $\text{SNR} = 0.5$. The cross 'x' points to the minimum of the bivariate curve. Comparing columnwise, we see good fit between the location of the minima, especially on the left.

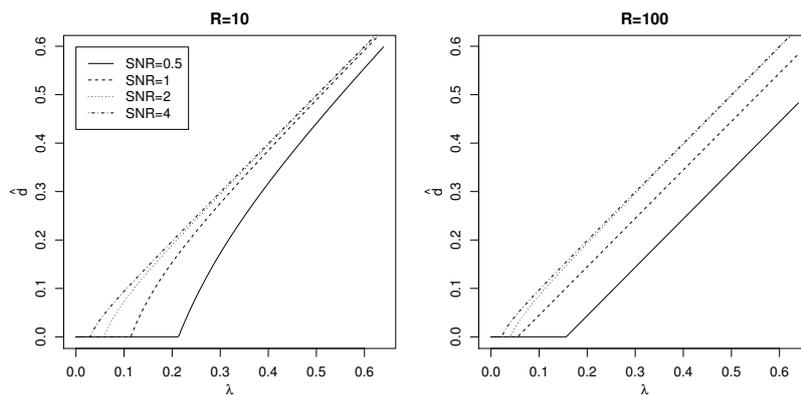


Figure 2: Typical thresholding function selected by ATN with SURE for ranks $R = 10$ (left) and $R = 100$ (right), and for four values of SNR.

estimates. This phenomenon is also known in the setup of regression where the lasso tends to select too many variables [Zou, 2006, Zhang and Huang, 2008]. Note that ATN using the universal threshold (column 2 of Table 2) was designed for estimating the rank and indeed provided a very good estimate.

Finally, the results of TSVD and OS (oracle or not) when $R = 100$ are not as competitive as when $R = 10$, because these methods are based on asymptotics and assume low rank compared to matrix size. In addition, the MSEs are very similar across both Tables for $R = 10$ whereas there are differences for $R = 100$ which indicates that the estimation of σ encounters difficulties.

The case SNR=0.5 and $R=100$ in Table 2 is also worth a comment. Here, the data are so noisy that part of the signal is indistinguishable from the noise: only 16 singular values are greater than the ones that would be obtained under the null hypothesis that the rank of the $N \times P$ matrix \mathbf{W} is zero. Nevertheless, ATN with SURE estimates on average a matrix of rank 87, which is quite remarkable. In this situation, the same amount of shrinkage is applied to all the singular values with soft thresholding (see that the selected $\hat{\gamma} = 1$ on Figure 2 in this situation) leading to the smallest MSE.

3.2 Non-Gaussian noise

The methods considered above are all based on the assumption of Gaussian noise. We now assess their sensitivity to Student noise with 5 degrees of freedom, based on the same simulations as for Table 1. We also considered a more difficult situation with a matrix size divided by 10, that is $N = 20$ and $P = 50$. Figure 3 points to two noticeable consequences: the boxplots are more variable with Student, yet centered around the same median, except for ATN with GSURE that sees its efficiency drop when both N and P are small (bottom right).

3.3 Simulations based on a real small data set

We consider here a realistic simulation based on a wine dataset with $N = 21$ wines described by $P = 30$ sensory descriptors (the data are available in the R package FactoMineR [Lê et al., 2008]). We used the fitted rank- R matrix as the true signal matrix, and then added Gaussian noise to perform a Monte Carlo simulation. Note that in practice, it makes sense to center the data before using the estimators since the values are shrunk toward zero. On this small sample case, we found the same trends as observed previously. As illustrated in Figure 4 for a case with two levels of noise and $R=8$, the estimators often manage to improve on the usual truncated SVD, the SVST fits well for high noise regime and poor otherwise, and ATN remains very powerful. Note that ATN with universal τ has results similar to those of the optimal shrinker (OS), which provides an empirical interpretation of the OS estimator and highlights the capability of ATN to find the optimal way to shrink the singular values. Finally, GSURE, although still the best method among the blind estimators (the last 3 estimators on the graphics) in term of median MSE, is more variable.

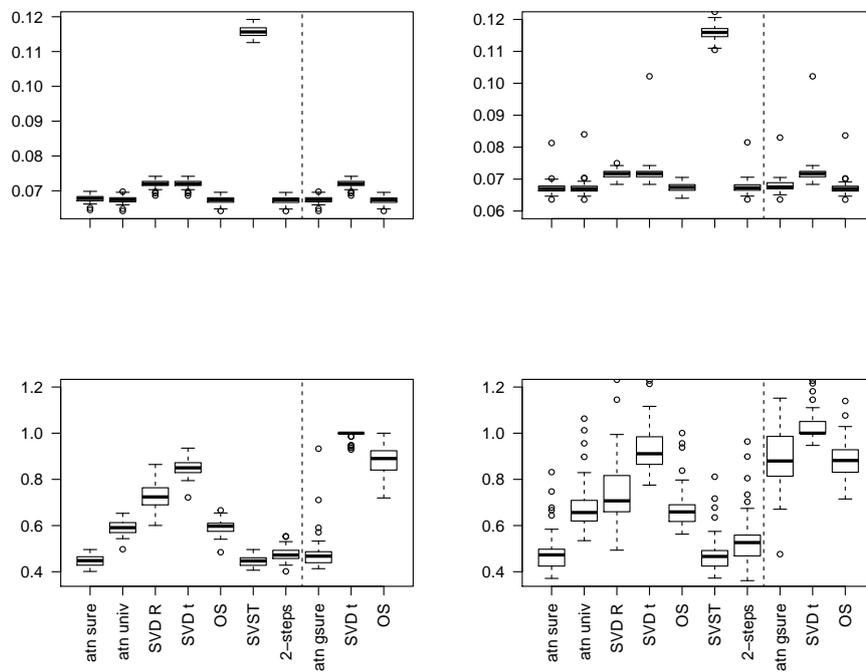
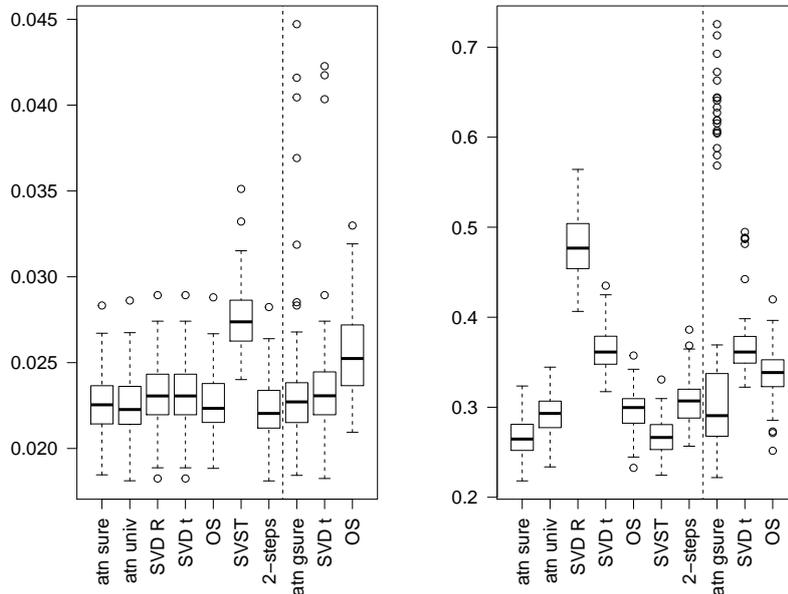


Figure 3: MSE boxplots for $R = 10$ and $\text{SNR}=1$. Top: $N = 200, P = 500$; bottom: $N = 20, P = 50$. Left: Gaussian; right: Student.

Figure 4: Distribution of the MSE for the wine dataset simulations for $R = 8$.

4 Conclusions

Recovering a reduced rank matrix from noisy data is a hot topic that has aroused the scientific community for a few years, as testified by the abundant recent literature on the subject. The adaptive trace norm estimator combines the strength of the hard, soft and the two-step procedures by means of a shrinking and a thresholding parameter indexing a family of shrinkers. The method adapts to the data which ensures good estimation of both low rank and general signal matrices, whatever the regime encountered in practice. The tuning parameters are estimated without using computationally intensive resampling methods thanks to the SURE and GSURE formulae. The latter version has the great advantage of not requiring knowledge of the noise variance. In addition, the rank is also estimated accurately, especially with the universal version designed for it. Our method outperforms the competitors on simulations and thus can be recommended to users.

We showed that Student noise affected the results, but other corruptions such as strong outliers alter the performances of the estimators to a greater extent. To tackle this issue, a natural extension could be to derive robust estimators using for instance the robust Huber loss function ρ instead of the Frobenius norm in (10). It leads to an estimator solution to

$$\min_{\mathbf{W} \in \mathbb{R}^{N \times P}} \|\mathbf{X} - \mathbf{W}\|_{\rho} + \alpha \|\mathbf{W}\|_{*,\omega} \quad \text{with} \quad \|\mathbf{W}\|_{*,\omega} = \sum_{i=1}^{\min(N,P)} \omega_i d_i, \quad (15)$$

where $\|H\|_\rho = \sum_{i=1}^N \sum_{j=1}^P \rho(h_{ij})$ with

$$\rho(h) = \begin{cases} h^2/2, & |h| \leq \tilde{\alpha} \\ \tilde{\alpha}|h| - \tilde{\alpha}^2/2, & |h| > \tilde{\alpha} \end{cases}$$

is the Huber loss with cutpoint $\tilde{\alpha}$ [Huber, 1981]. Extending the results of Sardy et al. [2001], we could rewrite (15) as

$$\min_{\mathbf{W}, \mathbf{R} \in \mathbb{R}^{N \times P}} \|\mathbf{X} - \mathbf{W} - \mathbf{R}\|_F^2 + \alpha \|\mathbf{W}\|_{*,\omega} + \tilde{\alpha} \|\mathbf{R}\|_1 \quad \text{with} \quad \|\mathbf{R}\|_1 = \sum_{i=1}^N \sum_{j=1}^P |R_{ij}|.$$

It would allow to solve the problem by block coordinate relaxation and could be used as an alternative to the robust estimator of Candès et al. [2009].

Two others extensions should be considered. First, assessing our estimator in a missing data framework as an alternative to iterative soft thresholding algorithms [Mazumder et al., 2010]. Second, assessing our estimator to denoise inner product matrices or covariance matrices as in [Ledoit and Wolf, 2012]. Finally, we can mention the work of Hoff [2013] who suggested a Bayesian treatment of Tucker decomposition methods to analyze arrays datasets. To better fit data, he pointed to hierarchical priors learning the values of the hyper-parameters from the data with an empirical Bayesian approach, which follows essentially the same goal as our method.

The results are reproducible with the R code provided by the first author on her webpage.

Acknowledgment

The authors are grateful for the helpful comments of the reviewers and editors. J.J. is supported by an AgreeSkills fellowship of the European Union Marie-Curie FP7 COFUND People Programme. S.S. is supported by the Swiss National Science Foundation. This work started while both authors were visiting Stanford University and the authors would like to thank the Department of Statistics for hosting them and for its stimulating seminars.

References

- J. Baik and J. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.
- J. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principle Component Analysis? *Journal of ACM*, 58:1–37, 2009.
- E. J. Candès, C.A. Sing-Long, and Trzasko J.D. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657, 2013.

- H. Caussinus. *Models and uses of principal component analysis (with discussion)*, pages 149–178. DSWO Press, 1986.
- S. Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv:1212.1247*, 2013.
- K. Chen, H. Dong, and K.-S. Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.
- P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.
- J. D. de Leeuw, A. Mooijaart, and R. van der Leeden. *Fixed Factor Score Models with Linear Restrictions*. University of Leiden, 1985.
- D. L. Donoho and M. Gavish. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on information theory*, 60 (8):5040–5053, 2014a.
- D. L. Donoho and M. Gavish. Minimax risk of matrix denoising by singular value thresholding. *Annals of Statistics, To Appear*, 2014b.
- D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- S. Gaïffas and G. Lecue. Weighted algorithms for compressed sensing and matrix completion. *arXiv:1107.1638*, 2011.
- M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *arXiv:1405.7511v2*, 2014.
- P. D. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685, 2007.
- P. D. Hoff. Equivariant and scale-free tucker decomposition models. *arXiv:1312.6397*, 2013.
- P. J. Huber. *Robust Statistics*. John Wiley, New York, 1981.
- A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.
- I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.
- J. Josse and F. Husson. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, 56 (6):1869–1879, 2012.

- J. Josse and F. Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):1–21, 2012.
- S. Lê, J. Josse, and F. Husson. Factominer: An r package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 3 2008.
- O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40:1024–1060, 2012.
- J. Mandel. The partitioning of interaction in analysis of variance. *Journal of the research of the national bureau of standards, Series B*, 73:309–328, 1969.
- R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 99:2287–2322, 2010.
- A. B. Owen and P. O. Perry. Bi-cross-validation of the svd and the nonnegative matrix factorization. *Annals of Applied Statistics*, 3(2):564–594, 2009.
- D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.
- N. Raj Rao. Optshrink - low-rank signal matrix denoising via optimal, data-driven singular value shrinkage. *arXiv:1306.6042*, 2013.
- S. Sardy. Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood’s block gradient. *Journal of the American Statistical Association*, 107:800–813, 2012.
- S. Sardy. Blockwise and coordinatewise thresholding to combine tests of different natures in modern anova. *arXiv:1302.6073*, 2013.
- S. Sardy, P. Tseng, and A. G. Bruce. Robust wavelet denoising. *IEEE Transactions on Signal Processing*, 49:1146–1152, 2001.
- A. A. Shabalin and B. Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118(0):67 – 76, 2013.
- C. M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9:1135–1151, 1981.
- H. Talebi and P. Milanfar. Global image denoising. *IEEE Transactions on Image Processing*, 2013.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological*, 58:267–288, 1996.
- M. Verbanck, F. Husson, and J. Josse. Regularized PCA to denoise and visualize data. *Statistics and Computing*, forthcoming, 2013.
- A. Zanella, M. Chiani, and M. Z. Win. On the marginal distribution of the eigenvalues of Wishart matrices. *IEEE Transactions on Communications*, 57(4):1050–1060, 2009.

- C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.
- H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.