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

Jérémie Bigot, Charles-Alban Deledalle, Delphine Féral

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

Jérémie Bigot, Charles Deledalle & Delphine Féral

Institut de Mathématiques de Bordeaux et CNRS (UMR 5251)
Université de Bordeaux

April 21, 2017

Abstract

We consider the problem of estimating a low-rank signal matrix from noisy measurements under the assumption that the distribution of the data matrix belongs to an exponential family. In this setting, we derive generalized Stein's unbiased risk estimation (SURE) formulas that hold for any spectral estimators which shrink or threshold the singular values of the data matrix. This leads to new data-driven spectral estimators, whose optimality is discussed using tools from random matrix theory and through numerical experiments. Under the spiked population model and in the asymptotic setting where the dimensions of the data matrix are let going to infinity, some theoretical properties of our approach are compared to recent results on asymptotically optimal shrinking rules for Gaussian noise. It also leads to new procedures for singular values shrinkage in finite-dimensional matrix denoising for Gamma-distributed and Poisson-distributed measurements.

Keywords: matrix denoising, singular value decomposition, low-rank model, Gaussian spiked population model, spectral estimator, Stein's unbiased risk estimate, random matrix theory, exponential family, optimal shrinkage rule, degrees of freedom.

AMS classifications: 62H12, 62H25.

1 Introduction

1.1 Low rank matrix denoising in an exponential family

In various applications, it is of interest to estimate a signal matrix from noisy data. Typical examples include the case of data that are produced in a matrix form, while others are concerned with observations from multiple samples that can be organized in a matrix form. In such setting, a typical inference problem involves the estimation of an unknown (non-random) signal matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ from a noisy data matrix $\boldsymbol{Y}$ satisfying the model:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}, \tag{1.1}$$

where $\boldsymbol{W}$ is an $n \times m$ noise matrix with real entries $\boldsymbol{W}_{ij}$ assumed to be independent random variables with $\mathbb{E}[\boldsymbol{W}_{ij}] = 0$ and $\mathrm{Var}(\boldsymbol{W}_{ij}) = \tau_{ij}^2$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. In this paper, we focus on the situation where the signal matrix $\boldsymbol{X}$ is assumed to have a low rank structure, and we consider the general setting where the distribution of $\boldsymbol{Y}$ belongs to a continuous exponential family parametrized by the entries of the matrix $\boldsymbol{X} = \mathbb{E}[\boldsymbol{Y}]$. For discrete observations (count data), we also consider the specific case of Poisson noise.

The low rank assumption on $\boldsymbol{X}$ is often met in practice when there exists a significant correlation between the columns of $\boldsymbol{X}$. This can be the case when the columns of $\boldsymbol{X}$ represent 2D images at different wavelength of hyperspectral data, since images at nearby wavelengths are strongly correlated [CSLT13]. Further applications, where low-rank modeling of $\boldsymbol{X}$ is relevant, can be found in genomics [WDB01, ABB00], NMR spectroscopy [NPDL11], collaborative filtering [CR09] or medical imaging [BD06, LBH+12], among many others.

Low-rank matrix estimation is classically done in the setting where the additive noise is Gaussian with homoscedastic variance. The more general case of observations sampled from an exponential family is less developed, but there exists an increasing research interest in the study of low rank matrix recovery beyond the Gaussian case. Examples of low-rank matrix recovering from Poisson distributed observations can be found in applications with count data such as network traffic analysis [BMG13] or call center data [SH05]. A theory for low-rank matrix recovery and completion in the case of Poisson observations has also been recently proposed in [CX16]. Matrix completion under a low rank assumption with additive errors having a sub-exponential distribution and belonging to an exponential family has also been considered in [Laf15]. The recent work [UHZB16] proposes a novel framework to approximate, by a low rank matrix, a tabular data set made of numerical, Boolean, categorical or ordinal observations.

1.2 The class of spectral estimators

A standard approach to estimate a low rank matrix relies on the singular value decomposition (SVD) of the data matrix

$$\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \tag{1.2}$$

where $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_{\min(n,m)} \geq 0$ denote its singular values, and $\tilde{\boldsymbol{u}}_k, \tilde{\boldsymbol{v}}_k$ denote the associated singular vectors. In this paper, we propose to consider the class of spectral estimators $\hat{\boldsymbol{X}}^f = f(\boldsymbol{Y})$, where $f : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ is a (possibly data-dependent) mapping that acts on the singular values of the data matrix $\boldsymbol{Y}$ while leaving its singular vectors unchanged. More precisely, these estimators take the form

$$\hat{\boldsymbol{X}}^f = f(\boldsymbol{Y}) = \sum_{k=1}^{\min(n,m)} f_k(\boldsymbol{Y}) \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \tag{1.3}$$

where, for each $1 \leq k \leq \min(n, m)$, $f_k(\boldsymbol{Y})$ are real positive values that may depend only on $\tilde{\sigma}_k$ (hence we write $f_k(\tilde{\sigma}_k)$) or on the whole matrix $\boldsymbol{Y}$.

1.3 Investigated spectral estimators

Typical examples of spectral estimators include the classical principal component analysis (PCA) applied to matrix denoising defined, for some $1 \leq r \leq \min(n,m)$, as

$$\hat{\boldsymbol{X}}^r = \sum_{k=1}^{r} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = f_k(\tilde{\sigma}_k) = \tilde{\sigma}_k \tag{1.4}$$

for all $1 \leq k \leq r$ and where it is implicitely understood that $f_k(\tilde{\sigma}_k) = 0$ for $k \geq r+1$. Another typical spectral estimator in matrix denoising with Gaussian measurements is the soft-thresholding [CSLT13] which corresponds to the choice

$$\hat{\boldsymbol{X}}_{\text{soft}} = \sum_{k=1}^{\min(m,n)} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = f_k(\boldsymbol{Y}) = \left(1 - \frac{\lambda(\boldsymbol{Y})}{\tilde{\sigma}_k}\right)_+ \tilde{\sigma}_k, \tag{1.5}$$

for all $1 \leq k \leq \min(n,m)$ and where $\lambda(\boldsymbol{Y}) > 0$ is a possibly data-dependent threshold parameter, and $(x)_+ = \max(x,0)$ for any $x \in \mathbb{R}$. Finaly, we will consider a more general class of shrinkage estimators, encompassing the PCA and the soft-thresholding, that perform

$$\hat{\boldsymbol{X}}_w = \sum_{k=1}^{\min(m,n)} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = f_k(\boldsymbol{Y}) = w_k(\boldsymbol{Y}) \tilde{\sigma}_k, \tag{1.6}$$

where $w_k(\boldsymbol{Y}) \in [0,1]$ is a possibly data-dependent shrinking weight.

1.4 Main contributions

Under the assumption that the distribution of $\boldsymbol{Y}$ belongs to an exponential family, the goal of this paper is to derive data-driven choices for the weights $w_k(\boldsymbol{Y})$ in (1.3).We construct estimators via a two-step procedure. First, an active set of non-zero singular values is defined. Then, in a second step, weights $w_k(\boldsymbol{Y})$ associated with non-zero singular values are optimized, and shown to reach desired asymptotical properties in the Gaussian spiked population model. The main contributions of the paper are then the following ones.

1.4.1 An AIC inspired criterion for rank and singular values locations estimation

When no *a priori* is available on the rank of the signal matrix $\boldsymbol{X}$, optimizing for the weights w_k, for all $1 \leq k \leq \min(m,n)$, can lead to estimators with large variance (*i.e.*, overfitting the noise). We propose an automatic rule to prelocalize the subset of non-zero singular values. An active set $s^\star \subseteq \mathcal{I} = \{1, 2, \ldots, \min(n,m)\}$ of singular values is defined as the minimizer of a penalized log-likelihood criterion that is inspired by the Akaike information criterion (AIC)

$$s^* \in \arg\min_{s \subseteq \mathcal{I}} \; -2 \log q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) + 2|s| p_{n,m} \quad \text{with} \quad p_{n,m} = \frac{1}{2}\left(\sqrt{m} + \sqrt{n}\right)^2, \tag{1.7}$$

where $\tilde{\boldsymbol{X}}^s = \sum_{k \in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$, $|s|$ is the cardinal of s, and $q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s)$ is the likelihood of the data in a given exponential family with estimated parameter $\tilde{\boldsymbol{X}}^s$. For the case of Gaussian measurements

with homoscedastic variance τ^2, one has that $q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) = \|\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s\|_F^2/2\tau^2$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and we show that the active set of singular values boils down to

$$s^\star = \{k \; ; \; \tilde{\sigma}_k > c_+^{n,m}\}, \tag{1.8}$$

where $c_+^{n,m} = \tau(\sqrt{m} + \sqrt{n})$. For Gamma and Poisson measurements, we resort to a greedy optimization procedure described in Section 4.

Once the active set has been determined, the subsequent shrinkage estimator is obtained by optimizing only for the weights within this subset while setting the other ones to zero.

1.4.2 Novel data-driven shrinkage rules minimizing SURE-like formulas

We use the principle of Stein's unbiased risk estimation (SURE) [Ste81] to derive unbiased estimation formulas for the mean squared error (MSE) risk and mean Kullback-Leibler (MKL) risks of spectral estimators. Minimizing such SURE-like formulas over an appropriate class of spectral estimators is shown to lead to novel data-driven shrinkage rules of the singular values of the matrix $\boldsymbol{Y}$. In particular, our approach leads to novel spectral estimators in situations where the variances τ_{ij}^2 of the entries $\boldsymbol{W}_{ij}$ of the noise matrix are not necessarily equal, and may depend on the signal matrix $\boldsymbol{X}$.

As an illustrative example, let us consider spectral estimators of the form

$$\hat{\boldsymbol{X}}_w^1 = f(\boldsymbol{Y}) = w_1(\boldsymbol{Y})\tilde{\sigma}_1\tilde{\boldsymbol{u}}_1\tilde{\boldsymbol{v}}_1^t, \tag{1.9}$$

which only act on the first singular value $\tilde{\sigma}_1$ of the data while setting all the other ones to zero. In this paper, examples of data-driven choices for the weight $w_1(\boldsymbol{Y})$ are the following ones:

- for Gaussian measurements with $n \leq m$ and known homoscedastic variance τ^2

$$w_1(\boldsymbol{Y}) = \left(1 - \frac{\tau^2}{\tilde{\sigma}_1^2}\left(1 + |m-n| + 2\sum_{\ell=2}^{n}\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2}\right)\right)_+ \mathbb{1}_{\{\tilde{\sigma}_1 > c_+^{n,m}\}}, \tag{1.10}$$

- for Gamma measurements with $\tau_{ij}^2 = \boldsymbol{X}_{ij}^2/L$ and $L > 2$ (see Section 2.1 for a precise definition),

$$w_1(\boldsymbol{Y}) = \min\left[1, \left(\frac{L-1}{Lmn}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\hat{\boldsymbol{X}}_{ij}^1}{\boldsymbol{Y}_{ij}} + \frac{1}{Lmn}\left(1 + |m-n| + 2\sum_{\ell=2}^{\min(n,m)}\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2}\right)\right)^{-1}\right]\mathbb{1}_{\{1\in s^*\}}, \tag{1.11}$$

- for Poisson measurements with $\tau_{ij}^2 = \boldsymbol{X}_{ij}$ (see Section 2.1 for a precise definition)

$$w_1(\boldsymbol{Y}) = \min\left[1, \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\boldsymbol{X}}_{ij}^1}\right]\mathbb{1}_{\{1\in s^*\}}. \tag{1.12}$$

Beyond the case of rank one, closed-form solutions for the weights cannot be obtained, except for the case of Gaussian measurements with homoscedastic variance τ^2. In this latter case, the rule for $w_1(\boldsymbol{Y})$ in (1.10) generalizes to other eigenvalues $w_k(\boldsymbol{Y})$ as

$$w_k(\boldsymbol{Y}) = \left(1 - \frac{\tau^2}{\tilde{\sigma}_k^2}\left(1 + |m-n| + 2\sum_{\ell=1;\ell\neq k}^{\min(n,m)} \frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right)\right)_+ \mathbb{1}_{\{\tilde{\sigma}_k > c_+^{n,m}\}}. \tag{1.13}$$

For Gamma or Poisson distributed measurements, we propose fast algorithms to get numerical approximations of the weights $w_k(\boldsymbol{Y})$ (see Section 5.2 for more details).

1.4.3 Asymptotic properties in the Gaussian spiked population model

Another contribution of the paper is to discuss the optimality of the shrinking weights (1.13) for Gaussian noise in the asymptotic setting where the dimensions of the matrix $\boldsymbol{Y}$ are let going to infinity. These theoretical results are obtained for the so-called *spiked population model* that has been introduced in the literature on random matrix theory and high-dimensional covariance matrix estimation (see e.g. [BS06, BN12, DS07, SN13]). All the theoretical and asymptotic results of the paper (other than derivation of proposed estimators) assume this model.

Definition 1.1. *The Gaussian spiked population model corresponds to the following setting:*

- *the $\boldsymbol{W}_{ij}$ in (1.1) are iid Gaussian random variables with zero mean and variance $\tau^2 = 1/m$,*
- *the $\boldsymbol{X}_{ij}$'s in (1.1) are the entries of an unknown $n \times m$ matrix $\boldsymbol{X}$ that has a low rank structure, meaning that it admits the SVD $\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t$, where $\boldsymbol{u}_k$ 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 $\sigma_1 > \sigma_2 > \ldots > \sigma_{r^*}$,*
- *the rank r^* of the matrix $\boldsymbol{X}$ is assumed to be fixed,*
- *the dimensions of the data matrix $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ are let going to infinity in the asymptotic framework where the sequence $m = m_n \geq n$ is such that $\lim_{n\to+\infty} \frac{n}{m} = c$ with $0 < c \leq 1$.*

In the Gaussian spiked population model, the asymptotic locations of the empirical singular values $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_{\min(n,m)}$ are well understood in the random matrix theory (further details are given in Section 3.1). Note that the setting where the rank r^* is not held fixed but allowed to grow with $\min(n,m)$ is very different, see e.g. [LW12] and references therein.

Under the Gaussian spiked population model, our contributions are then as follows:

- we prove the convergence of the SURE formula when the dimensions of $\boldsymbol{Y}$ tend to infinity,
- it is shown that minimizing the asymptotic value of SURE leads to the same estimator as the limiting value of the estimator obtained by minimizing the SURE,
- this model allows to show that the novel data-driven spectral estimators derived in this paper are asymptotically connected to existing optimal shrinkage rules [SN13, GD14a, Nad14] for low-rank matrix denoising,
- in this setting, we are also able to connect the choice of the penalty function $2|s|p_{n,m}$ in (1.7) with Stein's notion of degrees of freedom (see e.g. [Efr04]) for spectral estimators.

1.4.4 Numerical experiments and publicly available source code

As the theoretical properties of our estimators are studied in an asymptotic setting, we report the results of various numerical experiments to analyze the performances of the proposed estimators for finite-dimensional matrices. These experiments allow the comparison with existing shrinkage rules for Gaussian-distributed measurements and they are also used to shed some lights on the finite sample properties of the method for Gamma-distributed or Poisson-distributed measurements. We also exhibit the settings where the signal matrix $\boldsymbol{X}$ is either easy or more difficult to recover. From these experiments, the main findings are the following ones:

- the use of an appropriate active set s of singular values is an essential step for the quality of shrinkage estimators whose weights are data-driven by SURE-like estimators; taking $s = \{1, \ldots, \min(n, m)\}$ leads to poor results while the choice of $s = s^*$ minimizing the AIC criterion (1.7) appears to yield the best performances,
- for Gaussian noise, the performances of our approach are similar to those obtained by the asymptotically optimal spectral estimator proposed in [GD14a] when the true rank r^* of the signal matrix $\boldsymbol{X}$ is sufficiently small,
- for Gamma or Poisson distributed measurements, the spectral estimators proposed in this paper give better results than estimators based on PCA (restricted to the active set s^*) or soft-thresholding of singular values.

Beyond the case of Gaussian noise, the implementation of the estimators is not straightforward, and we thus provide publicly available source code at

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

to reproduce the figures and the numerical experiments of this paper.

1.5 Related results in the literature

Early work on singular value thresholding began with the work in [EY36] on the best approximation of fixed rank to the data matrix $\boldsymbol{Y}$. Spectral estimators with different amounts of shrinkage for each singular value of the data matrix have then been proposed in [EM72, EM76]. In the case of Gaussian measurements with homoscedastic variance, the problem of estimating $\boldsymbol{X}$ under a low-rank assumption has recently received a lot of attention in the literature on high-dimensional statistics, see e.g. [CSLT13, DG14, JS15, SN13]. Recent works [GD14a, Nad14] also consider the more general setting where the distribution of the additive noise matrix $\boldsymbol{W}$ is orthogonally invariant, and such that its entries are iid random variables with zero mean and finite fourth moment. In all these papers, the authors have focused on spectral estimators which shrink or threshold the singular values of $\boldsymbol{Y}$, while its singular vectors are left unchanged. In this setting, the main issue is to derive optimal shrinkage rules that depends on the class of spectral estimators that is considered, on the loss function used to measure the risk of an estimator of $\boldsymbol{X}$, and on appropriate assumptions for the distribution of the additive noise matrix $\boldsymbol{W}$.

1.6 Organization of the paper

Section 2 is devoted to the analysis of a data matrix whose entries are distributed according to a continuous exponential family. SURE-like formula are first given for the mean squared error risk, and then for the Kullback-Leibler risk. As an example of discrete exponential family, we also derive such risk estimators for Poisson distributed measurements. The computation of data-driven shrinkage rules is then discussed for Gaussian, Gamma and Poisson noises. In Section 3, we restrict our attention to the Gaussian spiked population model in order to derive asymptotic properties of our approach. We study the asymptotic behavior of the SURE formula proposed in [CSLT13, DG14] for spectral estimators using tools from RMT. This result allows to make a connection between data-driven spectral estimators minimizing the SURE for Gaussian noise, and the asymptotically optimal shrinkage rules proposed in [SN13, Nad14] and [GD14a]. In Section 4, we study the penalized log-likelihood criterion (1.7) used to select an active set of singular values. Its connection to the degrees of freedom of spectral estimators and rank estimation in matrix denoising is discussed. Various numerical experiments are finally proposed in Section 5 to illustrate the usefulness of the approach developed in this paper for low-rank denoising and to compare its performances with existing methods. The proofs of the main results of the paper are gathered in a technical Appendix A, and numerical implementation details are described in Appendix B.

2 SURE-like formulas in exponential families

For an introduction to exponential families, we refer to [Bro86]. The idea of unbiased risk estimation in exponential families dates back to [Hud78]. More recently, generalized SURE formulas have been proposed for the estimation of the MSE risk, for denoising under various continuous and discrete distributions in [RS07], and for inverse problems whithin the continuous exponential families in [Eld09]. In [Del15], SURE-like formula are derived for the estimation of the Kullback-Leibler risk that applies to both continuous and discrete exponential families. In what follows, we borrow some ideas and results from these works. We first treat the case of continuous exponential families, and then we focus on Poisson data in the discrete case.

2.1 Data sampled from a continuous exponential family

We recall that $\boldsymbol{Y}$ is an $n \times m$ matrix with independent and real entries $\boldsymbol{Y}_{ij}$. For each $1 \leq i \leq n$ and $1 \leq j \leq m$, we assume that the random variable $\boldsymbol{Y}_{ij}$ is sampled from a continuous exponential family, in the sense that each $\boldsymbol{Y}_{ij}$ admits a probability density function (pdf) $q(y; \boldsymbol{X}_{ij})$ with respect to the Lebesgue measure $\mathrm{d}y$ on the real line $\mathcal{Y} = \mathbb{R}$. The pdf $q(y; \boldsymbol{X}_{ij})$ of $\boldsymbol{Y}_{ij}$ can thus be written in the general form:

$$q(y; \boldsymbol{X}_{ij}) = h(y) \exp\left(\eta(\boldsymbol{X}_{ij}) y - A(\eta(\boldsymbol{X}_{ij}))\right), \ y \in \mathcal{Y}, \tag{2.1}$$

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, and $\boldsymbol{X}_{ij}$ is an unknown parameter of interest belonging to some open subset $\mathcal{X}$ of $\mathbb{R}$. Throughout the paper, we will suppose that the following assumption holds:

Assumption 2.1. *The link function η and the log-partition function A are such that*

$$A'(\eta(x)) = x \text{ for all } x \in \mathcal{X},$$

where A' denotes the first derivative of A.

Since $\mathbb{E}[\boldsymbol{Y}_{ij}] = A'(\eta(\boldsymbol{X}_{ij}))$ for exponential families in the general form (2.1), Assumption 2.1 implies that $\mathbb{E}[\boldsymbol{Y}_{ij}] = \boldsymbol{X}_{ij}$, and thus the data matrix satisfies the relation $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ where $\boldsymbol{W}$ is a centered noise matrix, which is in agreement with model (1.1). Now, if we let $\Theta = \eta(\mathcal{X})$, it will be also convenient to consider the expression of the pdf of $\boldsymbol{Y}_{ij}$ in the canonical form:

$$p(y; \boldsymbol{\theta}_{ij}) = h(y) \exp\left(\boldsymbol{\theta}_{ij} y - A(\boldsymbol{\theta}_{ij})\right), \ y \in \mathcal{Y}, \tag{2.2}$$

where $\boldsymbol{\theta}_{ij} = \eta(\boldsymbol{X}_{ij}) \in \Theta$ is usually called the canonical parameter of the exponential family. Finally, we recall the relation $\mathrm{Var}(\boldsymbol{Y}_{ij}) = A''(\boldsymbol{\theta}_{ij}) = A''(\eta(\boldsymbol{X}_{ij}))$ where A'' denotes the second derivative of A. Then, we denote by $\boldsymbol{\theta}$ the $n \times m$ matrix whose entries are the $\boldsymbol{\theta}_{ij}$'s.

Examples of data satisfying model (2.1) are the following ones:

Gaussian noise with known variance τ^2:

$$q(y; \boldsymbol{X}_{ij}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \boldsymbol{X}_{ij})^2}{2\tau^2}\right), \ \mathbb{E}[\boldsymbol{Y}_{ij}] = \boldsymbol{X}_{ij}, \ \mathrm{Var}(\boldsymbol{Y}_{ij}) = \tau^2,$$

$$\mathcal{Y} = \mathbb{R}, \ \mathcal{X} = \mathbb{R}, \ \Theta = \mathbb{R}, \ h(y) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{y^2}{2\tau^2}\right), \ \eta(x) = \frac{x}{\tau^2}, \ A(\theta) = \tau^2 \frac{\theta^2}{2}.$$

Gamma-distributed measurements with known shape parameter $L > 0$:

$$q(y; \boldsymbol{X}_{ij}) = \frac{L^L y^{L-1}}{\Gamma(L) \boldsymbol{X}_{ij}^L} \exp\left(-L \frac{y}{\boldsymbol{X}_{ij}}\right) \mathbb{1}_{]0,+\infty[}(y), \ \mathbb{E}[\boldsymbol{Y}_{ij}] = \boldsymbol{X}_{ij}, \ \mathrm{Var}(\boldsymbol{Y}_{ij}) = \frac{\boldsymbol{X}_{ij}^2}{L},$$

$$\mathcal{Y} = \mathbb{R}, \ \mathcal{X} =]0, +\infty[, \ \Theta =]-\infty, 0[, \ h(y) = \frac{L^L y^{L-1}}{\Gamma(L)} \mathbb{1}_{]0,+\infty[}(y), \ \eta(x) = -\frac{L}{x}, \ A(\theta) = -L \log\left(-\frac{\theta}{L}\right).$$

The matrix $\boldsymbol{\theta} = \eta(X)$ can then be estimated via the $n \times m$ matrix $\hat{\boldsymbol{\theta}}^f = \hat{\boldsymbol{\theta}}^f(\boldsymbol{Y})$ whose entries are given by

$$\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}) = \eta\left(\hat{\boldsymbol{X}}^f_{ij}\right), \text{ for all } 1 \le i \le n, \ 1 \le j \le m, \tag{2.3}$$

where $\hat{\boldsymbol{X}}^f_{ij}$ is a spectral estimator as defined in eq. (1.3).

In the rest of this section, we follow the arguments in [Eld09] and [Del15] to derive SURE-like formulas under the exponential family for the estimators $\hat{\boldsymbol{\theta}}^f$ and $\hat{\boldsymbol{X}}^f$, using either the mean-squared error (MSE) risk or the Kullback-Leibler (KL) risk.

2.1.1 Unbiased estimation of the MSE risk

We consider the following MSE risk which provides a measure of discrepancy in the space Θ of natural parameters, and then indirectly in the space of interest $\mathcal{X}$.

Definition 2.1. *The squared error (SE) risk of* $\hat{\boldsymbol{\theta}}^f$ *is* $\mathrm{SE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}\|_F^2$, *and the mean-squared error (MSE) risk of* $\hat{\boldsymbol{\theta}}^f$ *is defined as* $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathbb{E}\left[\mathrm{SE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta})\right] = \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}\|_F^2\right]$.

Using the above MSE risk to compare $\hat{\boldsymbol{\theta}}^f$ and $\boldsymbol{\theta}$ implies that the discrepancy between the estimator $\hat{\boldsymbol{X}}^f$ and the matrix of interest $\boldsymbol{X}$ is measured by the quantity $\mathrm{MSE}_\eta(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \mathrm{MSE}(\eta(\hat{\boldsymbol{X}}^f), \eta(\boldsymbol{X}))$ which is different from $\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$. For Gaussian noise, $\mathrm{MSE}_\eta(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \frac{1}{\tau^2}\mathbb{E}\left[\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2\right]$, while for Gamma distributed measurements with known shape parameter $L > 0$, it follows that

$$\mathrm{MSE}_\eta(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = L^2 \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\boldsymbol{X}_{ij} - \hat{\boldsymbol{X}}^f_{ij}}{\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}^f_{ij}}\right)^2.$$

The following proposition gives a SURE formula for the MSE risk introduced in Definition 2.1.

Proposition 2.1. *Suppose that the data are sampled from a continuous exponential family. Assume that the function* h, *in the definition* (2.2) *of the exponential family, is twice continuously differentiable on* $\mathcal{Y} = \mathbb{R}$. *If the following condition holds*

$$\mathbb{E}\left[\left|\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})\right|\right] < +\infty, \text{ for all } 1 \le i \le n,\ 1 \le j \le m, \tag{2.4}$$

then, the quantity

$$\mathrm{GSURE}(\hat{\boldsymbol{\theta}}^f) = \|\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y})\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(2\frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}) + \frac{h''(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\right) + 2\,\mathrm{div}\,\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y}), \tag{2.5}$$

where $\mathrm{div}\,\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\partial\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}}$, *is an unbiased estimator of* $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta})$

Note that $\mathrm{GSURE}(\hat{\boldsymbol{\theta}}^f)$ is an unbiased estimator of $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta})$ and not of $\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$. It is shown in Section 3.3 that the results of Proposition 2.4 coincide with the approach in [CSLT13] on the derivation of a SURE formula in the case of Gaussian noise for smooth spectral estimators. In the case of Gamma noise, assuming $L > 2$ implies that the conditions on the function h in Proposition 2.1 is satisfied, hence assuming that conditions (2.4) holds as well, and using that

$$\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}) = -\frac{L}{f_{ij}(\boldsymbol{Y})} \quad \text{and} \quad \frac{\partial\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}} = \frac{L}{|f_{ij}(\boldsymbol{Y})|^2}\frac{\partial f_{ij}(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}},$$

it follows that

$$\mathrm{GSURE}(\hat{\boldsymbol{\theta}}^f) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{L^2}{|f_{ij}(\boldsymbol{Y})|^2} - \frac{2L(L-1)}{\boldsymbol{Y}_{ij}f_{ij}(\boldsymbol{Y})} + \frac{2L}{|f_{ij}(\boldsymbol{Y})|^2}\frac{\partial f_{ij}(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}} - \frac{(L-1)(L-2)}{|\boldsymbol{Y}_{ij}|^2}. \tag{2.6}$$

2.1.2 Unbiased estimation of KL risks

Following the terminology in [Del15], let us now introduce two different notions of Kullback-Leibler risk, which arise from the non-symmetry of this discrepancy measure.

Definition 2.2. *Let $f : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ be a smooth spectral function. Consider the estimator $\hat{\boldsymbol{\theta}}^f$ defined by (2.3), where $\boldsymbol{Y}$ is a matrix whose entries $\boldsymbol{Y}_{ij}$ are independent random variables sampled from the exponential family (2.2) in canonical form:*

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

$$\mathrm{KLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \int_{\mathbb{R}} \log\left(\frac{p(y;\hat{\boldsymbol{\theta}}^f_{ij})}{p(y;\boldsymbol{\theta}_{ij})}\right) p(y;\hat{\boldsymbol{\theta}}^f_{ij})\,\mathrm{d}y,$$

  and the mean KLS risk of $\hat{\boldsymbol{\theta}}^f$ is defined as $\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathbb{E}\left[\mathrm{KLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta})\right]$,

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

$$\mathrm{KLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \int_{\mathbb{R}} \log\left(\frac{p(y;\boldsymbol{\theta}_{ij})}{p(y;\hat{\boldsymbol{\theta}}^f_{ij})}\right) p(y;\boldsymbol{\theta}_{ij})\,\mathrm{d}y,$$

  and the mean KLA risk of $\hat{\boldsymbol{\theta}}^f$ is defined as $\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathbb{E}\left[\mathrm{KLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta})\right]$.

A key advantage of the Kullback-Leibler risk is that it measures the discrepancy between the unknown distribution $p(y;\boldsymbol{\theta}_{ij})$ and its estimate $p(y;\hat{\boldsymbol{\theta}}^f_{ij})$. It is thus invariant with respect to the reparametrization $\hat{\boldsymbol{\theta}}^f = \eta(\hat{\boldsymbol{X}}^f)$ (unlike the MSE risk), and we may also write $\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathrm{MKLS}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$ and $\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathrm{MKLA}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$. As suggested in [Del15], the MKLA risk represents how well the distribution $p(y;\hat{\boldsymbol{\theta}}^f_{ij})$ explain a random variable $\boldsymbol{Y}_{ij}$ sampled from the pdf $p(y;\boldsymbol{\theta}_{ij})$. The MKLA risk is a natural loss function in many statistical problems since it takes as a reference measure the true distribution of the data, see e.g. [Hal87]. The MKLS risk represents how well one may generate an independent copy of $\boldsymbol{Y}_{ij}$ by sampling a random variable from the pdf $p(y;\hat{\boldsymbol{\theta}}^f_{ij})$. The MKLS risk has also been considered in various inference problems in statistics [HL06, Yan94].

By simple calculation, it follows that

$$\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}^f_{ij} - \boldsymbol{\theta}_{ij}\right) A'(\hat{\boldsymbol{\theta}}^f_{ij})\right] + A(\boldsymbol{\theta}_{ij}) - \mathbb{E}\left[A(\hat{\boldsymbol{\theta}}^f_{ij})\right], \tag{2.7}$$

and

$$\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{\theta}_{ij} - \hat{\boldsymbol{\theta}}^f_{ij}\right) A'(\boldsymbol{\theta}_{ij})\right] + \mathbb{E}\left[A(\hat{\boldsymbol{\theta}}^f_{ij})\right] - A(\boldsymbol{\theta}_{ij}). \tag{2.8}$$

Hence, in the case of Gaussian measurements with known variance τ^2, we easily retrieve that $\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \frac{\tau^2}{2}\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \frac{1}{2\tau^2}\mathbb{E}\left[\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2\right]$. In the case of Gamma distributed measurements with known shape parameter $L > 0$, it follows that

$$\begin{aligned}
\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) &= L\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\frac{\hat{\boldsymbol{X}}^f_{ij}}{\boldsymbol{X}_{ij}} - \log\left(\frac{\hat{\boldsymbol{X}}^f_{ij}}{\boldsymbol{X}_{ij}}\right) - 1\right], \\
\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) &= L\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\frac{\boldsymbol{X}_{ij}}{\hat{\boldsymbol{X}}^f_{ij}} - \log\left(\frac{\boldsymbol{X}_{ij}}{\hat{\boldsymbol{X}}^f_{ij}}\right) - 1\right].
\end{aligned}$$

Below, we use some of the results in [Del15] whose main contributions are the derivation of new unbiased estimators of the MKLS and MKLA risks. For continuous exponential family, the risk estimate derived in [Del15] is unbiased for the MKLS risk, while it is only asymptotically unbiased for the MKLA risk with respect to the signal-to-noise ratio. For data sampled from a continuous exponential family, this makes simpler the use of the MKLS risk to derive data-driven shinkage in low rank matrix denoising, and we have therefore chosen to concentrate our study on this risk in this setting. The following proposition establishes a SURE formula to estimate the MKLS risk in the continuous case.

Proposition 2.2. *Suppose that the data are sampled from a continuous exponential family. Assume that the function h, in the definition (2.2) of the exponential family, is continuously differentiable on $\mathcal{Y} = \mathbb{R}$. Suppose that the function A, in the definition (2.2) of the exponential family, is twice continuously differentiable on Θ. If the following condition holds*

$$\mathbb{E}\left[\left|A'(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}))\right|\right] < +\infty, \text{ for all } 1 \leq i \leq n,\ 1 \leq j \leq m, \tag{2.9}$$

then, the quantity

$$\mathrm{SUKLS}(\hat{\boldsymbol{\theta}}^f) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}) + \frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\right)A'(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})) - A(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}))\right) + \mathrm{div}\, f(\boldsymbol{Y}), \tag{2.10}$$

where $\mathrm{div}\, f(\boldsymbol{Y}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial f_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}}$, *is an unbiased estimator of* $\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) - \sum_{i=1}^{n}\sum_{j=1}^{m}A(\boldsymbol{\theta}_{ij})$.

A key difference in the formula of unbiased estimates for the MSE and the KL risks is the computation of the divergence term in (2.5) and (2.10), when $\hat{\boldsymbol{X}}^f = \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t$ is a smooth spectral estimator in the sense where each function $f_k : \mathbb{R}_+ \to \mathbb{R}_+$ is assumed to be (almost everywhere) differentiable for $1 \leq k \leq \min(n, m)$. In this setting, the divergence term in the expression of $\mathrm{GSURE}(\hat{\boldsymbol{\theta}}^f)$ depends upon the matrix $\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y}) = \eta(\hat{\boldsymbol{X}}^f)$. Therefore, when η is a nonlinear mapping, it is generally not possible to obtain a simpler expression for $\mathrm{div}\,\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y})$. To the contrary, for $\mathrm{SUKLS}(\hat{\boldsymbol{\theta}}^f)$, the divergence term is $\mathrm{div}\, f(\boldsymbol{Y})$ which has the following closed-form

expression for any smooth spectral estimators

$$\operatorname{div} f(\boldsymbol{Y}) = |m-n| \sum_{k=1}^{\min(n,m)} \frac{f_k(\tilde{\sigma}_k)}{\tilde{\sigma}_k} + \sum_{k=1}^{\min(n,m)} f_k'(\tilde{\sigma}_k) + 2 \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k) \sum_{\ell=1;\ell\neq k}^{\min(n,m)} \frac{\tilde{\sigma}_k}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\,, \tag{2.11}$$

thanks to the results from Theorem IV.3 in [CSLT13].

Note that $\mathrm{SUKLS}(\hat{\boldsymbol{X}}_w^r) = \frac{\tau^2}{2}\,\mathrm{SURE}(\hat{\boldsymbol{X}}_w^r)$ for Gaussian measurements, hence, the GSURE and SUKLS strategies match in this case. In the case of Gamma measurements, assuming that $L > 2$ implies that the conditions on the function h in Proposition 2.2 is satisfied, and by assuming that condition (2.9) holds as well, it follows that

$$\mathrm{SUKLS}(\hat{\boldsymbol{\theta}}^f) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left((L-1)\frac{f_{ij}(\boldsymbol{Y})}{\boldsymbol{Y}_{ij}} - L\log\left(f_{ij}(\boldsymbol{Y})\right)\right) - Lmn + \operatorname{div} f(\boldsymbol{Y}),$$

where the expression of $\operatorname{div} f(\boldsymbol{Y})$ is given by (2.11).

Note that it is implicitly understood in the definition of $\operatorname{div} f(\boldsymbol{Y})$ that each mapping $f_{ij} : \mathbb{R}^{n\times m} \to \mathbb{R}$ is differentiable. The differentiability of the spectral function f (and thus of its components f_{ij}) is a consequence of the assumption that the functions $f_1, \ldots, f_{\min(n,m)}$ (acting on the singular values) are supposed to be differentiable. For further details, on the differentiability of f and the f_{ij}'s, we refer to Section IV in [CSLT13]. From the arguments in [CSLT13], it follows that formula (2.11) for the divergence of f is also valid under the assumption that each function f_k is differentiable on $\mathbb{R}_+$ except on a set of Lebesgue measure zero.

2.2 The case of Poisson data

For Poisson data, the key result to obtain unbiased estimate of a given risk is the following lemma which dates back to the work in [Hud78].

Lemma 2.1. *Let $f : \mathbb{Z}^{n\times m} \to \mathbb{R}^{n\times m}$ be a measurable mapping. Let $1 \le i \le n$ and $1 \le j \le m$, and denote by $f_{ij} : \mathbb{Z}^{n\times m} \to \mathbb{R}$ a measurable function. Let $\boldsymbol{Y} \in \mathbb{Z}^{n\times m}$ be a matrix whose entries are independently sampled from a Poisson distribution on $\mathbb{Z}$. Then,*

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{X}_{ij}f_{ij}(\boldsymbol{Y})\right] = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}f_{ij}(\boldsymbol{Y}-\boldsymbol{e}_i\boldsymbol{e}_j^t)\right],$$

where, for each $1 \le i \le n$ and $1 \le j \le m$, $f_{ij}(\boldsymbol{Y})$ denotes the (i,j)-th entry of the matrix $f(\boldsymbol{Y})$, and $\boldsymbol{e}_i$ (resp. $\boldsymbol{e}_j$) denotes the vector of $\mathbb{Z}^n$ (resp. $\mathbb{Z}^m$) with the i-th entry (resp. j-th entry) equals to one and all others equal to zero.

Hudson's lemma provides a way to estimate (in an unbiased way) the expectation of the Frobenius inner product between the matrix $\boldsymbol{X}$ and the matrix $f(\boldsymbol{Y})$. To see the usefulness of this result, one may consider the following mean-squared error

$$\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \mathbb{E}\left[\left\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\right\|_F^2\right] = \mathbb{E}\left[\left\|\hat{\boldsymbol{X}}^f\right\|_F^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{X}_{ij}\hat{\boldsymbol{X}}_{ij}^f(\boldsymbol{Y}) + \|\boldsymbol{X}\|_F^2\right].$$

Therefore, by Lemma 2.1, one immediately obtains that

$$\mathrm{PURE}(\hat{\boldsymbol{\theta}}^f) = \left\|\hat{\boldsymbol{X}}^f\right\|_F^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}f_{ij}(\boldsymbol{Y}-\boldsymbol{e}_i\boldsymbol{e}_j^t), \tag{2.12}$$

is an unbiased estimate for the quantity $\mathrm{MSE}(\hat{\boldsymbol{X}}^f,\boldsymbol{X}) - \|\boldsymbol{X}\|_F^2$.

For Poisson data, one may also define the following KL risks

$$\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f,\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{X}_{ij} - \hat{\boldsymbol{X}}_{ij}^f - \hat{\boldsymbol{X}}_{ij}^f\log\left(\frac{\boldsymbol{X}_{ij}}{\hat{\boldsymbol{X}}_{ij}^f}\right)\right], \tag{2.13}$$

$$\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f,\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\hat{\boldsymbol{X}}_{ij}^f - \boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}\log\left(\frac{\hat{\boldsymbol{X}}_{ij}^f}{\boldsymbol{X}_{ij}}\right)\right], \tag{2.14}$$

which are in agreement with Definition 2.2 of KL risks for data sampled from a Poisson distribution. From the arguments in [Del15], there does not currently exist an approach to derive a SURE formula for the MKLS risk in the Poisson case since they are no unbiased formula for $\hat{\boldsymbol{X}}_{ij}^f\log\boldsymbol{X}_{ij}$. Nevertheless, as shown in [Del15], Hudson's Lemma 2.1 provides an unbiased estimator for $\boldsymbol{X}_{ij}\log\hat{\boldsymbol{X}}_{ij}^f$, and then it is possible to unbiasedly estimate the MKLA risk as follows.

Proposition 2.3. *For data sampled from a Poisson distribution, the quantity*

$$\mathrm{PUKLA}(\hat{\boldsymbol{\theta}}^f) = \sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\boldsymbol{X}}_{ij}^f - \boldsymbol{Y}_{ij}\log\left(f_{ij}(\boldsymbol{Y}-\boldsymbol{e}_i\boldsymbol{e}_j^t)\right), \tag{2.15}$$

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

2.3 Data-driven shrinkage in low-rank matrix denoising

For a matrix $\boldsymbol{X}$ with entries $\boldsymbol{X}_{ij}\in\mathcal{X}=\mathbb{R}$, we consider shrinkage estimators of the form

$$\hat{\boldsymbol{X}}_w^s = f(\boldsymbol{Y}) = \sum_{k\in s} w_k\tilde{\sigma}_k\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t, \tag{2.16}$$

with $s\subseteq\mathcal{I}=\{1,2,\ldots,\min(n,m)\}$ and $w_k\in[0,1]$, for all $k\in s$.

When the underlying matrix $\boldsymbol{X}$ is constrained to have positive entries, e.g. $\mathcal{X}=]0,+\infty[$ in the Gamma and Poisson cases, we consider instead estimators of the form

$$\hat{\boldsymbol{X}}_w^s = f(\boldsymbol{Y}) = \max\left[\sum_{k\in s} w_k\tilde{\sigma}_k\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t, \varepsilon\right], \tag{2.17}$$

where $\varepsilon>0$ is an *a priori* lower bound on the smallest value of $\boldsymbol{X}_{ij}$, where for any matrix $\boldsymbol{X}$, $\max[\boldsymbol{X},\varepsilon]_{ij} = \max[\boldsymbol{X}_{ij},\varepsilon]$, for all $1\leq i\leq m$ and $1\leq j\leq n$.

The construction of the subset s is postponed to Section 4, and we focus here in selecting the weights in a data-driven way for a fixed given s. In the following, we denote by s^c the complementary set of s in $\mathcal{I}$, *i.e.*, $s^c = \mathcal{I} \setminus s$, and we let $\hat{\boldsymbol{\theta}}^s_w = \eta\left(\hat{\boldsymbol{X}}^s_w\right)$. When $\mathcal{X} =]0, +\infty[$, we have found that considering estimators of the form (2.17) is more appropriate than trying to find shrinking weights $(w_k)_{k\in s}$ such that all the entries of the matrix $\sum_{k\in s} w_k \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ are positive, for a given subset s.

Gaussian noise with known homoscedastic variance τ^2

By applying the GSURE formula (2.5) for Gaussian distributed measurements and thanks to the expression (2.11) for the divergence of smooth spectral estimators, we obtain for $\hat{\boldsymbol{X}}^s_w$, as defined in (2.16), the SURE expression given by

$$\mathrm{SURE}(\hat{\boldsymbol{X}}^s_w) = -mn\tau^2 + \sum_{k\in s}(w_k - 1)^2\tilde{\sigma}_k^2 + \sum_{k\in s^c}\tilde{\sigma}_k^2 + 2\tau^2\sum_{k=1}^{s}\left(1 + |m-n| + 2\sum_{\ell=1;\ell\neq k}^{\min(n,m)}\frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right)w_k$$

which unbiasedly estimate $\mathrm{MSE}(\hat{\boldsymbol{X}}^s_w, \boldsymbol{X})$. Hence, for each $k \in s$, by differentiating the above expression with respect to w_k, it follows that a data-driven weight for the k-th empirical singular value is given by

$$w_k(\boldsymbol{Y}) = \left(1 - \frac{\tau^2}{\tilde{\sigma}_k^2}\left(1 + |m-n| + 2\sum_{\ell=1;\ell\neq k}^{\min(n,m)}\frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right)\right)_+, \tag{2.18}$$

which fullfils the requirement that $w_k(\boldsymbol{Y}) \in [0,1]$. Note that as $\mathrm{SUKLS}(\hat{\boldsymbol{X}}^s_w) = \frac{\tau^2}{2}\mathrm{SURE}(\hat{\boldsymbol{X}}^s_w)$ for Gaussian measurements, the exact same data-driven weight would be obtained by minimizing an estimate of the $\mathrm{MKLS}(\hat{\boldsymbol{X}}^s_w, \boldsymbol{X})$.

The case of estimators with rank one. Consider the case of estimators with rank 1, *i.e.*, let $s = \{1\}$. It follows that $\hat{\boldsymbol{X}}^1_w = \hat{\boldsymbol{X}}^{\{1\}}_w = w_1\hat{\boldsymbol{X}}^1$ where $w_1 \in [0,1]$ is given by

$$w_1(\boldsymbol{Y}) = \left(1 - \frac{\tau^2}{\tilde{\sigma}_1^2}\left(1 + |m-n| + 2\sum_{\ell=1;\ell\neq 1}^{\min(n,m)}\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2}\right)\right)_+.$$

Gamma and Poisson distributed measurements

In Gamma and Poisson cases, it is not possible to follow the same strategy as in the Gaussian case to derive optimal weights for (2.17) in a closed-form using the established SURE-like formulas. We shall investigate how data-driven shinkage can be approximated in Section 5 on numerical experiments using fast algorithms. Nevertheless, when the estimator is restricted to rank 1, optimizing KL risk estimators lead to closed-form expressions under the assumption that all the entries of the data matrix $\boldsymbol{Y}$ are strictly positive.

The case of estimators with rank one under Gamma noise. Consider again the case of estimators with rank 1, *i.e.*, let $s = \{1\}$, and let $\hat{\boldsymbol{X}}^1 = \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$ denote the PCA approximation of rank 1 of $\boldsymbol{X}$. If all the entries of the matrix $\boldsymbol{Y}$ are strictly positive, by the Perron-Frobenius theorem, all the entries of the first singular vectors $\tilde{\boldsymbol{u}}_1$ and $\tilde{\boldsymbol{v}}_1$ are strictly positive. Therefore, all the entries of $\hat{\boldsymbol{X}}^1$ belong to the set $\mathcal{X} =]0, +\infty[$, and we can consider $\hat{\boldsymbol{X}}^1_w = \hat{\boldsymbol{X}}^{\{1\}}_w = w_1 \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$ as defined in (2.16) instead of (2.17). Assuming $L > 2$ for the SUKLS formula to hold, it follows by simple calculations that

$$\mathrm{SUKLS}(\hat{\boldsymbol{\theta}}^1_w) = \sum_{i=1}^{n}\sum_{j=1}^{m}(L-1)w_1 \frac{\hat{\boldsymbol{X}}^1_{ij}}{\boldsymbol{Y}_{ij}} - mnL\log(w_1) - L\log\left(\frac{\hat{\boldsymbol{X}}^1_{ij}}{\boldsymbol{Y}_{ij}}\right) - Lmn + (1+|m-n|)w_1 + 2w_1 \sum_{\ell=2}^{\min(n,m)} \frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2}.$$

Hence, by differentiating the above expression with respect to w_1 and as it is monotonic on both sides of its unique minimum, the optimal value of $w_1 \in [0,1]$ minimizing $\mathrm{SUKLS}(\hat{\boldsymbol{\theta}}^1_w)$ is given by

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

which yields the shrinking rule (1.11) stated in the introduction of this paper. Note that it is not possible to obtain, in a closed-form, the optimal value of the weight w_1 that minimizes the criterion $\mathrm{GSURE}(\hat{\boldsymbol{\theta}}^1_w)$.

The case of estimators with rank one under Poisson noise. Using again that all the assumption that the entries of $\boldsymbol{Y}$ are positive, we can consider (by the Perron-Frobenius theorem) $\hat{\boldsymbol{X}}^1_w = \hat{\boldsymbol{X}}^{\{1\}}_w = w_1 \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$ as defined in (2.16) instead of (2.17). Then, the PURE formula (2.12) and Proposition 2.3 apply to the estimator $\hat{\boldsymbol{\theta}}^1_w = \log\left(\hat{\boldsymbol{X}}^1_w\right)$ which yield to

$$\mathrm{PURE}(\hat{\boldsymbol{\theta}}^1_w) = w_1^2\tilde{\sigma}_1^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij} w_1 \tilde{\sigma}_1^{(ij)} \tilde{\boldsymbol{u}}_{1,i}^{(ij)} \tilde{\boldsymbol{v}}_{1,j}^{(ij)},$$

$$\text{and} \quad \mathrm{PUKLA}(\hat{\boldsymbol{\theta}}^1_w) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_1 \hat{\boldsymbol{X}}^1_{ij} - \boldsymbol{Y}_{ij}\left(\log(w_1) + \log\left(\tilde{\sigma}_1^{(ij)} \tilde{\boldsymbol{u}}_{1,i}^{(ij)} \tilde{\boldsymbol{v}}_{1,j}^{(ij)}\right)\right),$$

where $\hat{\boldsymbol{X}}^1_{ij} = \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_{1,i} \tilde{\boldsymbol{v}}_{1,j}$, $\tilde{\sigma}_1^{(ij)}$ is the largest singular value of the matrix $\boldsymbol{Y} - \boldsymbol{e}_i \boldsymbol{e}_j^t$, and $\tilde{\boldsymbol{u}}_1^{(ij)}$ (resp. $\tilde{\boldsymbol{v}}_1^{(ij)}$) denotes its left (resp. right) singular vectors. Therefore, by differentiating the above expression with respect to w_1 and as it is monotonic on both sides of its unique minimum, an optimal value for $w_1 \in [0,1]$ which minimizes $\mathrm{PURE}(\hat{\boldsymbol{\theta}}^1_w)$ is given by

$$w_1(\boldsymbol{Y}) = \min\left[1, \frac{1}{\tilde{\sigma}_1^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}\tilde{\sigma}_1^{(ij)}\tilde{\boldsymbol{u}}_{1,i}^{(ij)}\tilde{\boldsymbol{v}}_{1,j}^{(ij)}\right].$$

However, this optimal shrinking rule cannot be used in practice since evaluating the values of $\tilde{\sigma}_1^{(ij)}, \tilde{\boldsymbol{u}}_1^{(ij)}, \tilde{\boldsymbol{v}}_1^{(ij)}$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$ is not feasible from a computational point of view for large values of n and m. Nevertheless, a fast algorithm to find a numerical approximation of the optimal value $w_1(\boldsymbol{Y})$ is proposed in Section 5.

To the contrary, using again that all the $\hat{\boldsymbol{X}}^1_{ij}$ are positive by the Perron-Frobenius theorem, the value of $w_1 \in [0,1]$ minimizing $\mathrm{PUKLA}(\hat{\boldsymbol{\theta}}^1_w)$ is

$$w_1(\boldsymbol{Y}) = \min\left[1, \frac{\sum_{i=1}^n \sum_{j=1}^m \boldsymbol{Y}_{ij}}{\sum_{i=1}^n \sum_{j=1}^m \hat{\boldsymbol{X}}^1_{ij}}\right],$$

which is straightforward to compute. This corresponds to the shrinkage rule (1.12) given in the introduction.

3 Gaussian spiked population model

In this section, we restrict our analysis to the Gaussian spiked population model and the asymptotic setting introduced in Definition 1.1.

3.1 Asymptotic location of empirical singular values

We summarize below the asymptotic behavior of the singular values of the data matrix $\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ in the Gaussian spiked population model.

In the case where $\boldsymbol{X} = 0$, it is well known [AGZ10, BS10] that the empirical distribution of the singular values of $\boldsymbol{Y} = \boldsymbol{W}$ (with $\tau = \frac{1}{\sqrt{m}}$) converges, as $n \to +\infty$, to the quarter circle distribution if $c = 1$ and to its generalized version if $c < 1$. This distribution is supported on the compact interval $[c_-, c_+]$ with

$$c_\pm = 1 \pm \sqrt{c}$$

where c_+ is the so-called bulk (right) edge.

When $\boldsymbol{X} \neq 0$ has a low rank structure, the asymptotic behavior of the singular values of $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ is also well understood [BN12, DS07, SN13], and generalizations to noise matrix $\boldsymbol{W}$ whose distribution is orthogonally invariant have also been recently considered in [BN12]. Below, we recall some of these results that will be needed in this paper. To this end, let us introduce the real-valued function ρ defined by

$$\rho(\sigma) = \sqrt{\frac{(1+\sigma^2)(c+\sigma^2)}{\sigma^2}} \text{ for any } \sigma > 0.$$

Then, the following result holds (see e.g. Theorem 2.8 in [BN12] and Proposition 9 in [SN13]).

Proposition 3.1. *Assume that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ is a random matrix sampled from the Gaussian spiked population model with $\tau = \frac{1}{\sqrt{m}}$ and $\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t$. Then, for any fixed $k \geq 1$, one*

has that, almost surely,

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

Moreover,

$$\lim_{n\to+\infty} \tilde{\sigma}_{\min(n,m)} = c_-.$$

In what follows, we shall also use the relation

$$\frac{1}{\sigma^2} = \frac{\rho^2(\sigma) - (c+1) - \sqrt{(\rho^2(\sigma) - (c+1))^2 - 4c}}{2c} \text{ that holds for any } \sigma > c^{1/4}, \tag{3.1}$$

which is a consequence of e.g. the results in Section 3.1 in [BN12].

3.2 Existing asymptotically optimal shrinkage rules

Below, we briefly summarize some results in [GD14a] and [Nad14] on the construction of asymptotically optimal spectral estimators. Let

$$\hat{\boldsymbol{X}}^f = f(\boldsymbol{Y}) = \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t \tag{3.2}$$

be a given smooth spectral estimator, and consider the standard squared error $\mathrm{SE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2$ as a measure of risk. The set of spectral functions minimizing $\mathrm{SE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$ is given by $f_k(\tilde{\sigma}_k) = \tilde{\boldsymbol{u}}_k^t \boldsymbol{X} \tilde{\boldsymbol{v}}_k$, for $1 \leq k \leq \min(n,m)$. However, it cannot be used in practice since $\boldsymbol{X}$ is obviously unknown. A first alternative suggested in [GD14a] and [Nad14] is to rather study the asymptotic risk

$$\mathrm{SE}_\infty(\hat{\boldsymbol{X}}^f) = \lim_{n\to\infty} \mathrm{SE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) \text{ (in the almost sure sense)} \tag{3.3}$$

in the Gaussian spiked population model. Then, it is proposed in [GD14a] and [Nad14] to find an asymptotically optimal choice of f by minimizing $\mathrm{SE}_\infty(\hat{\boldsymbol{X}}^f)$ among a given class of smooth spectral functions. The results in [GD14a] show that, among spectral estimators of the form $\hat{\boldsymbol{X}}^\eta = \sum_{k=1}^{\min(n,m)} \eta(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t$, where $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous shrinker such that $\eta(\sigma) = 0$ whenever $\sigma \leq c_+$, an asymptotically optimal shrinkage rule is given by the choice

$$\eta^*(\sigma) = \begin{cases} \frac{1}{\sigma}\sqrt{(\sigma^2 - (c+1))^2 - 4c} & \text{if } \sigma > c_+, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

In [Nad14], it is proposed to consider spectral estimators of the form $\hat{\boldsymbol{X}}^\delta = \sum_{k=1}^r \delta_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ where $\delta_1, \ldots, \delta_r$ are positive weights. By Theorem 2.1 in [Nad14], it follows that, if $\sigma_k > c^{1/4}$ for all $1 \leq k \leq r$ with $r \leq r^*$, then the weights which minimize $\mathrm{SE}_\infty(\hat{\boldsymbol{X}}^\delta)$ over $\mathbb{R}_+^r$ are given by

$$\delta_k^* = \delta_k(\sigma_k) = \frac{\sigma_k^4 - c}{\sigma_k\sqrt{(1+\sigma_k^2)(c+\sigma_k^2)}}, \text{ for all } 1 \leq k \leq r. \tag{3.5}$$

In what follows, the shrinkage rules (3.4) and (3.5) are shown to be equivalent, and they will serve as a reference of asymptotic optimality. It should be stressed that the estimators in [GD14a] and [Nad14] are not equivalent. Indeed, the method in [Nad14] requires an estimate of the rank, while the approach in [GD14a] applies the same shrinker to all empirical singular values. Nevertheless, the shrinkage function that is applied to significant singular values (either above the bulk edge in [GD14a] or up to a given rank in [Nad14]) is the same.

3.3 Asymptotic behavior of data-driven estimators based on SURE

Following the principle of SURE, a second alternative to choose a smooth spectral estimator of the form (3.2) is to study the problem of selecting a set of functions $(f_k)_{1\leq k\leq \min(n,m)}$ that minimize an unbiased estimate of $\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X}) = \mathbb{E}\left[\|\hat{\boldsymbol{X}}^f - \boldsymbol{X}\|_F^2\right]$. For any $1 \leq i \leq m$ and $1 \leq j \leq n$, we recall that $f_{ij}(\boldsymbol{Y})$ denotes the (i,j)-th entry of the matrix $\hat{\boldsymbol{X}}^f = f(\boldsymbol{Y})$. Under the condition that

$$\mathbb{E}\left[|\boldsymbol{Y}_{ij} f_{ij}(\boldsymbol{Y})| + \left|\frac{\partial f_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}}\right|\right] < +\infty, \text{ for all } 1 \leq i \leq n,\ 1 \leq j \leq m. \tag{3.6}$$

it follows from the results in [CSLT13] (or equivalently from Proposition 2.1 for Gaussian noise with $\tau^2 = 1/m$) that

$$\mathrm{SURE}\left(\hat{\boldsymbol{X}}^f\right) = -n + \|f(\boldsymbol{Y}) - \boldsymbol{Y}\|_F^2 + \frac{2}{m}\,\mathrm{div}\, f(\boldsymbol{Y}), \tag{3.7}$$

is an unbiased estimate of $\mathrm{MSE}(\hat{\boldsymbol{X}}^f, \boldsymbol{X})$, where the divergence $\mathrm{div}\, f(\boldsymbol{Y})$ admits the closed-form expression (2.11). The SURE formula (3.7) has been used in [CSLT13] to find a data-driven value for $\lambda = \lambda(\boldsymbol{Y})$ in the the case of singular values shrinkage by soft-thresholding which corresponds to the choice

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

We study now the asymptotic behavior of the SURE formula (3.7). To this end, we shall use Proposition 3.1, but we will also need the following result (whose proof can be found in the Appendix) to study some of the terms in expression (2.11) of the divergence of $f(\boldsymbol{Y})$.

Proposition 3.2. *Assume that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ is a random matrix sampled from the Gaussian spiked population model with $\tau = \frac{1}{\sqrt{m}}$ and $\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t$. Then, for any fixed $1 \leq k \leq r^*$ such that $\sigma_k > c^{1/4}$, one has that, almost surely,*

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

In what follows, we restrict our analysis to the following class of spectral estimators (the terminology in the definition below is borrowed from [GD14a]).

Definition 3.1. Let $\hat{\boldsymbol{X}}^f = f(\boldsymbol{Y}) = \sum_{k=1}^{\min(n,m)} f_k(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t$ be a smooth spectral estimator. For a given $1 \leq r \leq \min(n,m)$, the estimator f is said to be a spectral shrinker of order r that collapses the bulk to 0 if

$$\begin{cases} f_k(\sigma) = 0 & \text{whenever } \sigma \leq c_+ \text{ and } 1 \leq k \leq r, \\ f_k(\sigma) = 0 & \text{for all } \sigma \geq 0 \text{ and } k > r. \end{cases}$$

The reason for restricting the study to spectral estimators such that $f_k(\tilde{\sigma}_k) = 0$ whenever $\tilde{\sigma}_k < c_+$ is linked to the choice of the active set s^* (1.8) of singular values in the Gaussian case, as detailed in Section 4. Now, for a spectral shrinker $\hat{\boldsymbol{X}}^f$ of order r that collapses the bulk to 0, we study the asymptotic behavior of the terms in expression (3.7) that only depend on f, namely

$$\begin{aligned} \overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right) \quad = \quad & \sum_{k=1}^{r}(f_k(\tilde{\sigma}_k) - \tilde{\sigma}_k)^2 + 2\left(1 - \frac{n}{m}\right)\sum_{k=1}^{r}\frac{f_k(\tilde{\sigma}_k)}{\tilde{\sigma}_k} + \frac{2}{m}\sum_{k=1}^{r} f_k'(\tilde{\sigma}_k) \\ & + 4\frac{n}{m}\sum_{k=1}^{r} f_k(\tilde{\sigma}_k)\left(\frac{1}{n}\sum_{\ell=1;\ell\neq k}^{n}\frac{\tilde{\sigma}_k}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right) \end{aligned} \tag{3.8}$$

The reason for studying $\overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right)$ is that finding an optimal shrinkage rule that minimizes $\mathrm{SURE}\left(\hat{\boldsymbol{X}}^f\right)$ is equivalent to minimizing expression (3.8) over spectral shrinkers of order r that collapses the bulk to 0, since $\mathrm{SURE}\left(\hat{\boldsymbol{X}}^f\right) - \overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right) = -n + \sum_{k=r+1}^{n}\tilde{\sigma}_k^2$ for such $\hat{\boldsymbol{X}}^f$.

Then, using Proposition 3.1, Proposition 3.2, and the assumption that the f_k's are continuously differentiable functions on $\mathbb{R}_+$, we immediately obtain the following result.

Lemma 3.1. *Assume that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ is a random matrix sampled from the Gaussian spiked population model with $\tau = \frac{1}{\sqrt{m}}$ and $\boldsymbol{X} = \sum_{k=1}^{r^*}\sigma_k\boldsymbol{u}_k\boldsymbol{v}_k^t$. Let $\hat{\boldsymbol{X}}^f$ be a spectral shrinker of order $r \leq r^*$ that collapses the bulk to 0, such that each function f_k, for $1 \leq k \leq r$, is continuously differentiable on $]c_+, +\infty[$. Moreover, assume that $\sigma_k > c^{1/4}$ for all $1 \leq k \leq r$. Then, one has that, almost surely,*

$$\lim_{n\to+\infty}\overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right) = \sum_{k=1}^{r}(f_k(\rho(\sigma_k)) - \rho(\sigma_k))^2 + 2f_k(\rho(\sigma_k))\left(\frac{\sigma_k^2(1+c) + 2c}{\sigma_k^2\rho(\sigma_k)}\right) \tag{3.9}$$

Asymptotically optimal shrinkage of singular values. Thanks to Lemma 3.1, one may determine an asymptotic optimal spectral shrinker as the one minimizing $\lim_{n\to+\infty}\overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right)$. For this purpose, let us define the class of estimators

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} w_k\tilde{\sigma}_k\mathbb{1}_{\{\tilde{\sigma}_k > c_+\}}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t, \tag{3.10}$$

where $1 \leq r \leq r^*$ is a given integer, and the w_k's are positive weights. In practice, the estimator $\hat{\boldsymbol{X}}_w^r$ is computed by replacing the bulk edge c_+ by its approximation $c_+^{n,m} = 1 + \sqrt{\frac{n}{m}}$ in eq. (3.10).

For moderate to large values of n and m, the quantities c_+ and $c_+^{n,m}$ are very close, and this replacement does not change the numerical performances of $\hat{\boldsymbol{X}}_w^r$.

Then, provided that $\sigma_k > c^{1/4}$ for all $1 \leq k \leq r$, it follows from Lemma 3.1 that

$$\lim_{n\to+\infty} \overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}_w^r\right) = \sum_{k=1}^{r} \rho^2(\sigma_k)(w_k-1)^2 + 2w_k\left(\frac{\sigma_k^2(1+c)+2c}{\sigma_k^2}\right).$$

Differentiating the above expression with respect to each weight w_k leads to the following choice of asymptotically optimal weights

$$w_k^* = 1 - \frac{\sigma_k^2(1+c)+2c}{\sigma_k^2\rho^2(\sigma_k)} \text{ for all } 1 \leq k \leq r. \tag{3.11}$$

Therefore, if the singular values of the matrix $\boldsymbol{X}$ to be estimated are sufficiently large (namely $\sigma_k > c^{1/4}$ for all $1 \leq k \leq r$), by using Proposition 3.1 and eq. (3.11), one has that an asymptotically optimal spectral shrinker of order $r \leq r^*$ is given by the choice of functions

$$f_k^*(\rho(\sigma_k)) = \begin{cases} \left(1 - \frac{\sigma_k^2(1+c)+2c}{\sigma_k^2\rho^2(\sigma_k)}\right)\rho(\sigma_k) & \text{if } \rho(\sigma_k) > c_+, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for all } 1 \leq k \leq r. \tag{3.12}$$

Using, the relation (3.1) one may also express the asymptotically optimal shrinking rule (3.12) either as a function of $\rho(\sigma_k)$ only,

$$f_k^*(\rho(\sigma_k)) = \begin{cases} \frac{1}{\rho(\sigma_k)}\sqrt{(\rho^2(\sigma_k)-(c+1))^2-4c} & \text{if } \rho(\sigma_k) > c_+, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

or as function of σ_k only (using that $\rho(\sigma_k) > c_+$ is equivalent to $\sigma_k > c^{1/4}$),

$$f_k^*(\rho(\sigma_k)) = \begin{cases} \frac{\sigma_k^4-c}{\sigma_k\sqrt{(1+\sigma_k^2)(c+\sigma_k^2)}} & \text{if } \sigma_k > c^{1/4}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Therefore, for spectral shrinker of order r, we remark that the shrinkage rule (3.13) coincides with the rule (3.4) which has been obtained in [GD14a]. Similarly, when the quantity $f_k^*(\rho(\sigma_k))$ is expressed as a function of σ_k only in (3.14), then we retrieve the shrinking rule (3.5) derived in [Nad14]. Therefore, it appears that minimizing either the asymptotic behavior of the SURE, that is $\lim_{n\to+\infty} \overline{\mathrm{SURE}}\left(\hat{\boldsymbol{X}}^f\right)$, or the limit of SE risk (3.3) leads to the same choice of an asymptotically optimal spectral estimator.

Data-driven shrinkage of empirical singular values. From the results in Section 2.3, the principle of SURE minimisation leads to the following data-driven choice of spectral shrinker of order r that collapses the bulk to 0

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} f_k(\tilde{\sigma}_k)\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{v}}_k^t, \tag{3.15}$$

where $f_k(\tilde{\sigma}_k) = w_k(\boldsymbol{Y})\tilde{\sigma}_k \mathbb{1}_{\{\tilde{\sigma}_k > c_+\}}$, for all $1 \leq k \leq r$, with $w_k(\boldsymbol{Y})$ given by (2.18). From Proposition 3.1 and Proposition 3.2 it follows that, if $\sigma_k > c^{1/4}$, then, almost surely,

$$\lim_{n \to +\infty} f_k(\tilde{\sigma}_k) = \left(1 - \frac{\sigma_k^2(1+c) + 2c}{\sigma_k^2 \rho^2(\sigma_k)}\right) \rho(\sigma_k), \text{ for all } 1 \leq k \leq r \leq r^*.$$

Therefore, the data-driven spectral estimator $\hat{\boldsymbol{X}}_w^r$ (3.15) asymptotically leads to the optimal shrinking rule of singular values given by (3.12) which has been obtained by minimizing the asymptotic behavior of the SURE.

Note that when $\tau \neq 1/\sqrt{m}$, it suffices to replace the condition $\tilde{\sigma}_k > c_+$ by $\tilde{\sigma}_k > \tau(\sqrt{m}+\sqrt{n})$ in the definition of $\hat{\boldsymbol{X}}_w^r$, which yields the shrinking rule (1.13) stated in the introduction of this paper.

4 Estimating active sets of singular values in exponential families

In this section, we propose to formulate a new Akaike information criterion (AIC) to select an appropriate set of singular values over which a shrinkage procedure might be applied. To this end, we shall consider the estimator $\tilde{\boldsymbol{X}}^s = \sum_{k \in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ defined for a subset $s \subseteq \mathcal{I} = \{1, 2, \ldots, \min(n, m)\}$, and we address the problem of selecting an optimal subset $s^\star$ from the data $\boldsymbol{Y}$.

In the case of Gaussian measurements, the shrinkage estimators that we use in our numerical experiments are of the form $\hat{\boldsymbol{X}}^f = \sum_{k \in s^\star} f_k(\tilde{\sigma}_k) \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ where

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

for some (possibly data-dependent) shrinkage functions f_k. The set $s^\star$ is based on the knowledge of an approximation $c_+^{n,m}$ of the bulk edge c_+. Thanks to Proposition 3.1, the bulk edge c_+ is interpreted as the threshold which allows to distinguish the locations of significant singular values in the data from those due to the presence of additive noise. Interestingly, the following result shows that the active set $s^\star$ may be interpreted through the prism of model selection using the minimisation of a penalized log-likelihood criterion.

Proposition 4.1. *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*

$$s^* = \arg\min_{s \subseteq \mathcal{I}} m\|\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s\|_F^2 + 2|s| p_{n,m} \text{ with } p_{n,m} = \left(\frac{1}{2}\left(\sqrt{m} + \sqrt{n}\right)^2\right), \tag{4.1}$$

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.

Proof. We remark that $\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s = \sum_{k \notin s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$. It results that

$$m\|\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s\|_F^2 + 2|s| p_{n,m} = m \sum_{k \notin s} \tilde{\sigma}_k^2 + 2|s| p_{n,m} = \sum_{k=1}^{n} \begin{cases} m\tilde{\sigma}_k^2 & \text{if } k \notin s \\ 2p_{n,m} & \text{otherwise} \end{cases}. \tag{4.2}$$

Using that $\sqrt{2p_{n,m}/m} = c_+^{n,m}$, it follows that the set $s^\star = \{k\ ;\ \tilde{\sigma}_k > c_+^{n,m}\}$ is by definition such that $k \in s^\star$ if and only if $2p_{n,m} < m\tilde{\sigma}_k^2$. Therefore, by (4.2), the criterion $s \mapsto m\|\boldsymbol{Y} - \tilde{\boldsymbol{X}}^s\|_F^2 + 2|s|p_{n,m}$ is minimum at $s = s^\star$ which concludes the proof. □

In the model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$, where the entries of $\boldsymbol{W}$ are iid Gaussian variables with zero mean and variance τ^2, it is well known that the degrees of freedom (DOF) of a given estimator $\hat{\boldsymbol{X}}$ is defined as

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

The DOF is widely used in statistics to define various criteria for model selection among a collection of estimators, see e.g. [Efr04]. In low rank matrix denoising, the following proposition shows that it is possible to derive the asymptotic behavior of the DOF of spectral estimators.

Proposition 4.2. *Assume that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ is a random matrix sampled from the Gaussian spiked population model with $\tau = \frac{1}{\sqrt{m}}$ and $\boldsymbol{X} = \sum_{k=1}^{r^*}\sigma_k\boldsymbol{u}_k\boldsymbol{v}_k^t$. Let $\hat{\boldsymbol{X}}^f$ be a spectral shrinker of order $r \leq r^*$ that collapses the bulk to 0, such that each function f_k, for $1 \leq k \leq r$, is continuously differentiable on $]c_+, +\infty[$. Moreover, assume that $\sigma_k > c^{1/4}$ for all $1 \leq k \leq r$. Then, one has that, almost surely,*

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

Proof. Thanks to the derivation of the SURE in [Ste81] and formula (2.11) on the divergence of spectral estimators, one has that

$$\mathrm{DOF}(\hat{\boldsymbol{X}}^f) = \mathbb{E}\left[\mathrm{div}\,\hat{\boldsymbol{X}}^f\right] = \mathbb{E}\left[|m-n|\sum_{k=1}^{r}\frac{f_k(\tilde{\sigma}_k)}{\tilde{\sigma}_k} + \sum_{k=1}^{r}f_k'(\tilde{\sigma}_k) + 2\sum_{k=1}^{r}f_k(\tilde{\sigma}_k)\sum_{\ell=1;\ell\neq k}^{n}\frac{\tilde{\sigma}_k}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right].$$

By Proposition 3.1, Proposition 3.2, and our assumptions on the f_k's, one has that, almost surely,

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

which completes the proof. □

Hence, in the Gaussian spiked population model, by Proposition 4.2 and using that $\sigma_k^2 > \sqrt{c}$ for all $1 \leq k \leq r$, it follows that if $s \subseteq \{1, \ldots, r\}$ then

$$\lim_{n\to+\infty}\frac{1}{m}\mathrm{DOF}(\tilde{\boldsymbol{X}}^s) = |s|\left(1 + c + \frac{2c}{\sigma_k^2}\right) \leq |s|\left(1 + \sqrt{c}\right)^2 = |s|c_+^2. \tag{4.3}$$

Hence, the quantity $2|s|\left(\frac{1}{2}\left(\sqrt{m} + \sqrt{n}\right)^2\right)$ is asymptotically an upper bound of $\mathrm{DOF}(\tilde{\boldsymbol{X}}^s)$ (when normalized by $1/m$) for any given set $s \subseteq \{1, \ldots, r\}$.

Let us now consider the more general case where the entries of $\boldsymbol{Y}$ are sampled from an exponential family. To the best of our knowledge, extending the notion of the bulk edge to non-Gaussian data sampled from an exponential family has not been considered so far in the literature on random matrices and low rank perturbation model. Therefore, except in the Gaussian case, it is far from being trivial to find an appropriate threshold value $\bar{c}$ to define an active set in the form $\bar{s} = \{k \; ; \; \tilde{\sigma}_k > \bar{c}\}$.

Nevertheless, to select an appropriate active set of singular values, we introduce the following criterion that is inspired by the previous results on the DOF of the estimator $\tilde{\boldsymbol{X}}^s$ in the Gaussian case and the statistical literature on the well known AIC for model selection [Aka74].

Definition 4.1. *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*

$$\mathrm{AIC}(\tilde{\boldsymbol{X}}^s) = -2\log q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) + 2|s|p_{n,m} \quad \textit{with} \quad p_{n,m} = \frac{1}{2}\left(\sqrt{m}+\sqrt{n}\right)^2 \; . \tag{4.4}$$

where $|s|$ *is the cardinal of* s*, and* $q(\boldsymbol{Y}; \tilde{\boldsymbol{X}}^s) = \prod_{i=1}^n \prod_{j=1}^m q(\boldsymbol{Y}_{ij}; \tilde{\boldsymbol{X}}^s_{ij})$ *is the likelihood of the data in the general form* (2.1) *at the estimated parameters* $\boldsymbol{X}_{ij} = \tilde{\boldsymbol{X}}^s_{ij}$.

In the above definition of $\mathrm{AIC}(\tilde{\boldsymbol{X}}^s)$, the quantity $2|s|p_{n,m}$ is an approximation of the degree of freedom of $\tilde{\boldsymbol{X}}^s$, i.e., of the numbers of its free parameters as it is justified by Proposition 4.2 in the case of Gaussian measurements. The AIC allows us to define an optimal subset of active variables as

$$s^* = \underset{s\subseteq\mathcal{I}}{\arg\min}\ \mathrm{AIC}(\tilde{\boldsymbol{X}}^s).$$

For Gaussian measurements, Proposition 4.1 gives the value of the optimal set s^* in a closed-form.

Following the arguments in Section 2.3, for Gamma or Poisson measurements and for a given subset s, we consider the estimator

$$\tilde{\boldsymbol{X}}^s_\epsilon = \max\left[\sum_{k\in s} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right], \tag{4.5}$$

when $\epsilon > 0$ is an *a priori* value to satisfy the positivity constraint on the entries of an estimator in this setting. However, contrary to the case of Gaussian noise, the search of an optimal subset $s^\star \subset \underset{s\subseteq\mathcal{I}}{\arg\min}\ \mathrm{AIC}(\tilde{\boldsymbol{X}}^s_\epsilon)$ becomes a combinatorial problem in this context. In our numerical experiments, we thus choose to construct an approximation $\tilde{s}$ of $s^\star$ with a greedy search strategy that reads as follows

$$\tilde{s} = \mathcal{I} \setminus \left\{k \in \mathcal{I} \; ; \; \mathrm{AIC}(\tilde{\boldsymbol{X}}^{\mathcal{I}\setminus\{k\}}_\epsilon) \leq \mathrm{AIC}(\tilde{\boldsymbol{X}}^{\mathcal{I}}_\epsilon)\right\}. \tag{4.6}$$

For Gaussian measurements, $\tilde{s} = s^\star$ since the optimisation problem (4.6) becomes separable. In our numerical experiments, we have found that $\tilde{s}$ selects a relevant set of active singular values which separates well the structural content of $\boldsymbol{X}$ while removing most of the noise component. Further details are given in Section 5 below.

For Gaussian noise, the computation of the active set s^* of singular values may also be interpreted as a way to estimate the unknown rank r^* of the signal matrix $\boldsymbol{X}$. In this setting, one has that $s^\star = \{k\ ;\ \tilde{\sigma}_k > c_+^{n,m}\}$ which suggests the choice

$$\hat{r} = \max\{k\ ;\ \tilde{\sigma}_k > c_+^{n,m}\}, \tag{4.7}$$

as an estimator of r^*.

There exists an abundant literature of the problem of estimating the rank of an empirical covariance matrix for the purpose of selecting the appropriate number of significant components to be kept in PCA or factor analysis. It is much beyond the scope of this paper to give an overview of this topic. We point to the review in [Jol02] for a summary of existing methods to determine the number of components in PCA that are grouped into three categories: subjective methods, distribution-based test tools, and computational procedures. For recent contributions in the matrix denoising model (1.1) with Gaussian noise, we refer to the works [CTT14, GD14b] and references therein. For example for Gaussian data with know variance $\tau^2 = 1/m$, Eq. (11) in [GD14b] on optimal hard thresholding of singular values suggest to take

$$\hat{r} = \max\{k\ ;\ \tilde{\sigma}_k > \lambda(c)\},\ \text{with}\ \lambda(c) = \sqrt{2(c+1) + \frac{8c}{(c+1) + \sqrt{c^2+14c+1}}}, \tag{4.8}$$

as a simple method to estimate the rank. It should be remarked that the problem of estimating the true rank r^* of $\boldsymbol{X}$ in model (1.1) is somewhat ill-posed as, in the Gaussian spiked population model, Proposition 3.1 implies that one may only expect to estimate the so-called effective rank $r_{\text{eff}} = \max\{k\ ;\ \sigma_k > c^{1/4}\}$ (see e.g. Section II.D in [Nad14]).

In our numerical experiments, we shall compare different choices for the active set of singular values of the form $\hat{s} = \{1, \dots, \hat{r}\}$ where $\hat{r}$ is either given by (4.7), (4.8), or by the "oracle choices" $\hat{r} = r^*$ and $\hat{r} = r_{\text{eff}}$.

Other methods based on hypothesis testing [CTT14] could be used for rank estimation in the Gaussian model (1.1), but it is beyond the purpose of this paper to give a detailed comparison.

For Poisson or Gamma noise, it is more difficult to interpret the computation of s^* as a way to estimate the rank of $\boldsymbol{X}$ since, in our numerical experiments, we have found that the cardinality of s^* is generally not equal to $\max\{k\ ; k \in s^*\}$. Moreover, to the best of our knowledge, there is not so much work on the estimation of the true rank of a noisy matrix beyond the Gaussian case. Therefore, we have not included a numerical comparison with other methods for the choice of the active set of singular values in these two cases.

5 Numerical experiments

In this section, we assess of the performance of data-driven srhinkage rules under various numerical experiments involving Gaussian, Gamma and Poisson measurements.

5.1 The case of a signal matrix of rank one

We consider the simple setting where the rank r^* of the matrix $\boldsymbol{X}$ is known and equal to one meaning that

$$\boldsymbol{X} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^t,$$

where $\boldsymbol{u}_1 \in \mathbb{R}^n$ and $\boldsymbol{v}_1 \in \mathbb{R}^m$ are vectors with unit norm that are fixed in this numerical experiment, and σ_1 is a positive real that we will let varying. We also choose to fix $n = m = 100$, and so to take $c = \frac{n}{m} = 1$ and $c_+ = 2$. For the purpose of sampling data from Gamma and Poisson distribution, we took singular vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ with positive entries. The i-th entry (resp. j-th entry) of $\boldsymbol{u}_1$ (resp. $\boldsymbol{v}_1$) is chosen to be proportional to $1 - (i/n - 1/2)^2$ (resp. $1 - (j/m - 1/2)^2$). Let $\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ be an $n \times m$ matrix whose entries are sampled from model (2.1) and then satisfying $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}$.

Gaussian measurements

We first consider the case of Gaussian measurements, where $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ with $\mathbb{E}[\boldsymbol{W}_{ij}] = 0$, $\mathrm{Var}(\boldsymbol{W}_{ij}) = \tau^2$ with $\tau = \frac{1}{\sqrt{m}}$. In this context, we compare the following spectral shrinkage estimators:

- Rank-1 PCA shrinkage

$$\hat{\boldsymbol{X}}^1 = \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t,$$

- Rank-1 SURE-driven soft-thresholding

$$\hat{\boldsymbol{X}}^1_{\text{soft}} = \hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t \quad \text{with} \quad \hat{\sigma}_1 = (\tilde{\sigma}_1 - \lambda(\boldsymbol{Y}))_+,$$

- Rank-1 asymptotically optimal shrinkage proposed in [Nad14] and [GD14a]

$$\hat{\boldsymbol{X}}^1_* = \hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t \quad \text{with} \quad \hat{\sigma}_1 = \sqrt{\tilde{\sigma}_1^2 - 4}\, \mathbb{1}_{\{\tilde{\sigma}_1 > 2\}},$$

- Rank-1 SURE-driven weighted estimator that we have derived in Section 2.3

$$\hat{\boldsymbol{X}}^1_w = \hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t \quad \text{with} \quad \hat{\sigma}_1 = \left(1 - \frac{1}{\tilde{\sigma}_1^2}\left(\frac{1}{m} + \frac{2}{m}\sum_{\ell=2}^{n} \frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2 - \tilde{\sigma}_\ell^2}\right)\right)_+ \tilde{\sigma}_1 \mathbb{1}_{\{\tilde{\sigma}_1 > 2\}},$$

where the above formula follows from the results in Section 3.3 using that $c = 1$ and $c_+ = 2$ in these numerical experiments, and where, for the soft-thresholding, the value $\lambda(\boldsymbol{Y}) > 0$ is obtained by a numerical solver in order to minimize the SURE. As a benchmark, we will also consider the oracle estimator $\boldsymbol{X}^1_*$ that performs shrinkage by using the knowledge of the true singular-value σ_1 defined as

$$\boldsymbol{X}^1_* = \hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t \quad \text{with} \quad \hat{\sigma}_1 = \sqrt{\rho(\sigma_1)^2 - 4}\, \mathbb{1}_{\{\rho(\sigma_1) > 2\}}$$

which corresponds to the asymptotically optimal shrinking rule (3.13) as a function of $\rho(\sigma_k)$ in the setting $c = 1$ and $c_+ = 2$. Note that form the formula above $\hat{w}_1 = \hat{\sigma_1}/\tilde{\sigma_1}$ is necessary in the range $[0, 1]$ for all considered estimators.

In Figure 1, we compare the estimated singular-values $\hat{\sigma}_1$ and the estimated weights $\hat{w}_1 = \hat{\sigma}_1/\tilde{\sigma}_1$ as functions of σ_1 for the four aforementionned estimators. Because all estimators are subject to noise variance, we display, for all estimators, the median values and the 80% confidence intervals obtained from $M = 100$ noise realizations. It can be seen that the median curves for the eigenvelues and the weights of $\hat{\boldsymbol{X}}^1_w$ and $\hat{\boldsymbol{X}}^1_*$ coincide (up to variations that are slightly larger for the former) which is in agreement with the asymptotic analysis of shrinkage rules that has been carried out in Section 3.3. Spectral estimator obtained by SURE-driven soft-thresholding also leads to an optimal shrinkage rule.

In Figure 2, for each of the four spectral estimators above, we display for $M = 100$ noise realizations, as functions of σ_1, the following normalized MSE

$$\mathrm{NMSE}(\hat{\boldsymbol{X}}) = \frac{\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F^2}{\|\boldsymbol{X}\|_F^2}.$$

The normalized MSE of the estimators $\hat{\boldsymbol{X}}^1_{\mathrm{soft}}$, $\hat{\boldsymbol{X}}^1_*$ and $\hat{\boldsymbol{X}}^1_w$ are the same for values of σ_1 larger than $c^{1/4} = 1$, and they only differ for values of σ_1 close or below the threshold $c^{1/4} = 1$ (corresponding to values of $\rho(\sigma_1)$ below the bulk edge $c_+ = 2$). More remarkably, above $c^{1/4} = 1$, they offer similar NMSE values to the oracle shrinkage estimator $\boldsymbol{X}^1_*$, not only in terms of median but also in terms of variability, as assessed by the confidence intervals. The performances of the estimator $\hat{\boldsymbol{X}}^1$ (standard PCA) are clearly poorer. These numerical experiments also illustrate that, for finite-dimensional low rank matrix denoising with $r^* = 1$, data-driven spectral estimators obtained by minimizing a SURE criterion achieve performances that are similar to asymptotically optimal shrinkage rules.

Gamma and Poisson distributed measurements

Let us now consider the case where the entries of $\boldsymbol{Y}_{ij} \geq 0$ of the data matrix $\boldsymbol{Y}$ are independently sampled from a Gamma or Poisson distribution with mean $\boldsymbol{X}_{ij} > 0$. To satisfy the constraint that the estimators must be matrices with positive entries, we consider estimators of the form (2.17). In this context, we compare the following spectral shrinkage estimators, set for $\varepsilon = 10^{-6}$, as:

- Rank-1 PCA shrinkage

$$\hat{\boldsymbol{X}}^1 = \max\left[\tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t, \varepsilon\right],$$

- Rank-1 GSURE/SUKLS/PURE/PUKLA-driven soft-thresholding

$$\hat{\boldsymbol{X}}^1_{\mathrm{soft}} = \max\left[\hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_1 = (\tilde{\sigma}_1 - \lambda(\boldsymbol{Y}))_+,$$

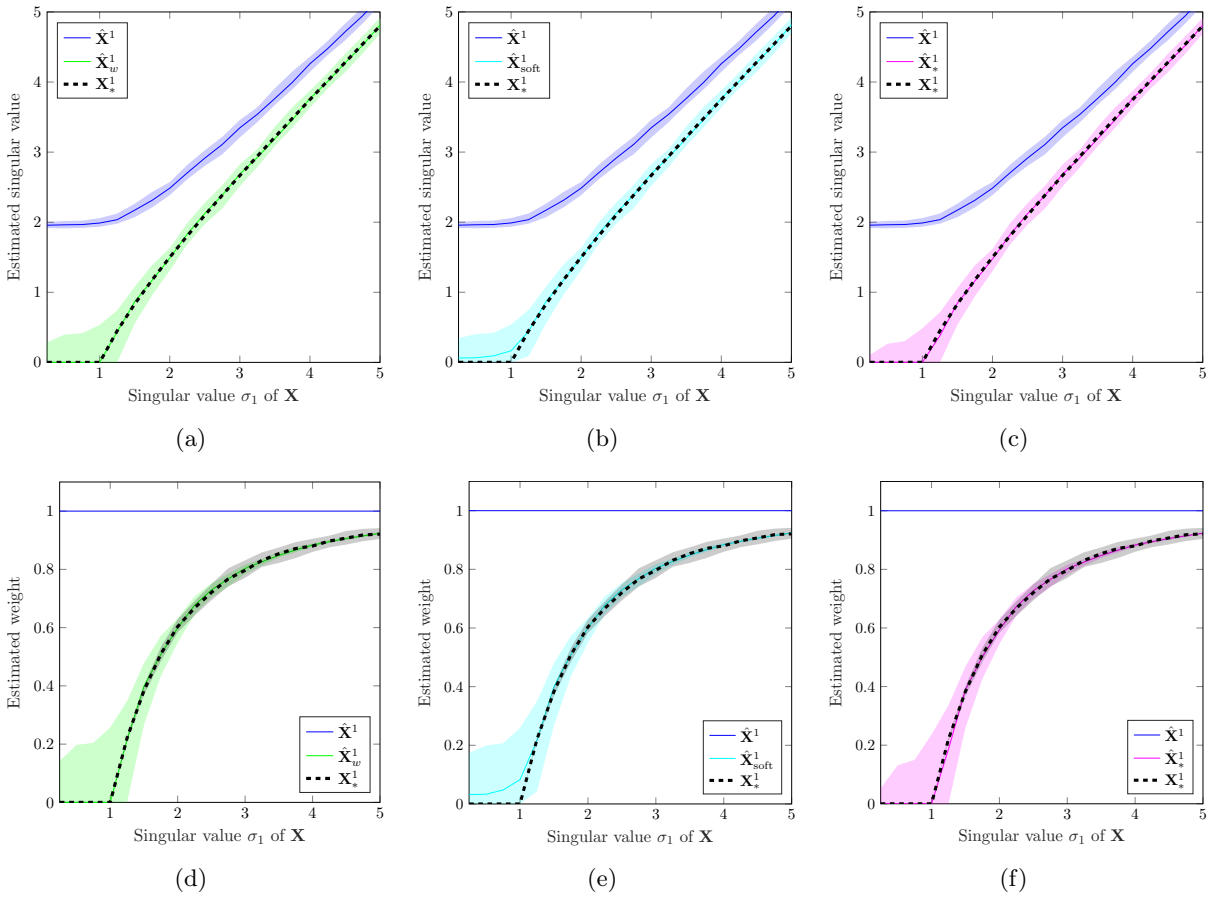
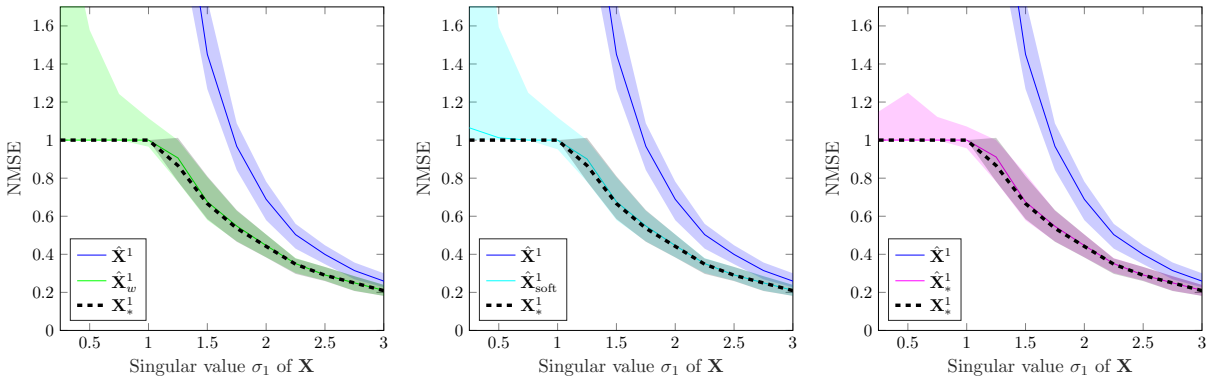


Figure 1: The case of Gaussian measurements with $m = n = 100$. Estimated first singular value $\hat{\sigma}_1$ as a function of the true underlying one σ_1, for (a) our proposed estimator $\hat{\boldsymbol{X}}^1_w$, (b) the soft-thresholding $\hat{\boldsymbol{X}}^1_{\text{soft}}$ and (c) the asymptotical one $\hat{\boldsymbol{X}}^1_*$. All of them are compared to the first singular value $\tilde{\sigma}_1$ of $\boldsymbol{Y}^1$ and the one of the oracle asymptotical estimator $\boldsymbol{X}^1_*$. (c,d,e) Same but for the corresponding weight $\hat{w}_1 = \hat{\sigma}_1/\tilde{\sigma}_1$. Curves have been computed on $M = 100$ noise realizations, only the median and an 80% confidence interval are represented respectively by a stroke and a shadded area of the same color.

- Rank-1 GSURE/SUKLS/PURE/PUKLA-driven weighted estimator

$$\hat{\boldsymbol{X}}^1_w = \max\left[\hat{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_1 = w_1(\boldsymbol{Y})\tilde{\sigma}_1 \mathbb{1}_{\{1 \in \tilde{s}\}},$$

where $\tilde{s}$ is the approximated active subset as defined in Section 4. For the soft-thresholding, the value $\lambda(\boldsymbol{Y}) > 0$ is obtained by a numerical solver in order to minimize either the GSURE or the SUKLS criterion (in the Gamma case) and either the PURE or the PUKLA criterion (in the Poisson case). The weight $w_1(\boldsymbol{Y}) \in [0, 1]$ is obtained by a numerical solver in order to minimize the GSURE and the PURE, as described in Section B. According to Section 2.3, the weight



(a) (b) (c)

Figure 2: Same as Fig. 1 but for the normalized MSE of the corresponding estimators.

$w_1(\boldsymbol{Y}) \in [0,1]$, minimizing the SUKLS criterion, has the following closed-form formula

$$w_1(\boldsymbol{Y}) = \min\left[1, \left(\frac{L-1}{Lmn}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\hat{\boldsymbol{X}}^1_{ij}}{\boldsymbol{Y}_{ij}} + \frac{1}{Lmn}\left(1+2\sum_{\ell=2}^{n}\frac{\tilde{\sigma}_1^2}{\tilde{\sigma}_1^2-\tilde{\sigma}_\ell^2}\right)\right)^{-1}\right],$$

and for the PUKLA criterion, we have

$$w_1(\boldsymbol{Y}) = \min\left[1, \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\boldsymbol{X}}^1_{ij}}\right].$$

To evaluate the performances of these estimators, we perform again a study involving $M = 100$ noise realizations.

In the Gamma case with shape parameter $L = 3$, results are reported in Figure 3 where σ_1 ranges from 0.1 to 5. In the Poisson case, results are reported in Figure 4. To generate data from a Poisson distribution with mean value $\boldsymbol{X} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^t$, we took σ_1 ranging from 25 to 400. In this context, the entries $\boldsymbol{X}_{i,j}$ are in average ranging from 0.25 to 4. When $\sigma_1 = 25$, about 78% of the entries of $\boldsymbol{Y}$ are 0 and 20% are equals to 1 which correspond to an extreme level of noise, while when $\sigma_1 = 400$, the entries of $\boldsymbol{Y}$ concentrate around 4 with a standard deviation of 2 which correspond to a simpler noisy setting.

In these experiments, it can be seen that all the data-dependent spectral estimators achieve comparable results with really small errors in terms of MSE and MKL risks. Their performances are similar to $\hat{\boldsymbol{X}}^1 = \tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$ meaning that optimizing either SURE-like criteria leads to a spectral estimator closed to correspond to matrix denoising by ordinary PCA. However, unlike the Gamma case, it might be observed in the Poisson case that when reaching a stronger noise level, *i.e*, for small value of σ_1, the NMSE of all estimator increases as the denoising problem becomes more challenging. Nevertheless, only the weight of $\hat{\boldsymbol{X}}^1_w$ driven by PUKLA does not present a drop wich allows reaching a slightly smaller MKLA. In the Gamma case, the noise level being proportional to the signal level, the NMSE/MKLS remain constant for all σ_1.

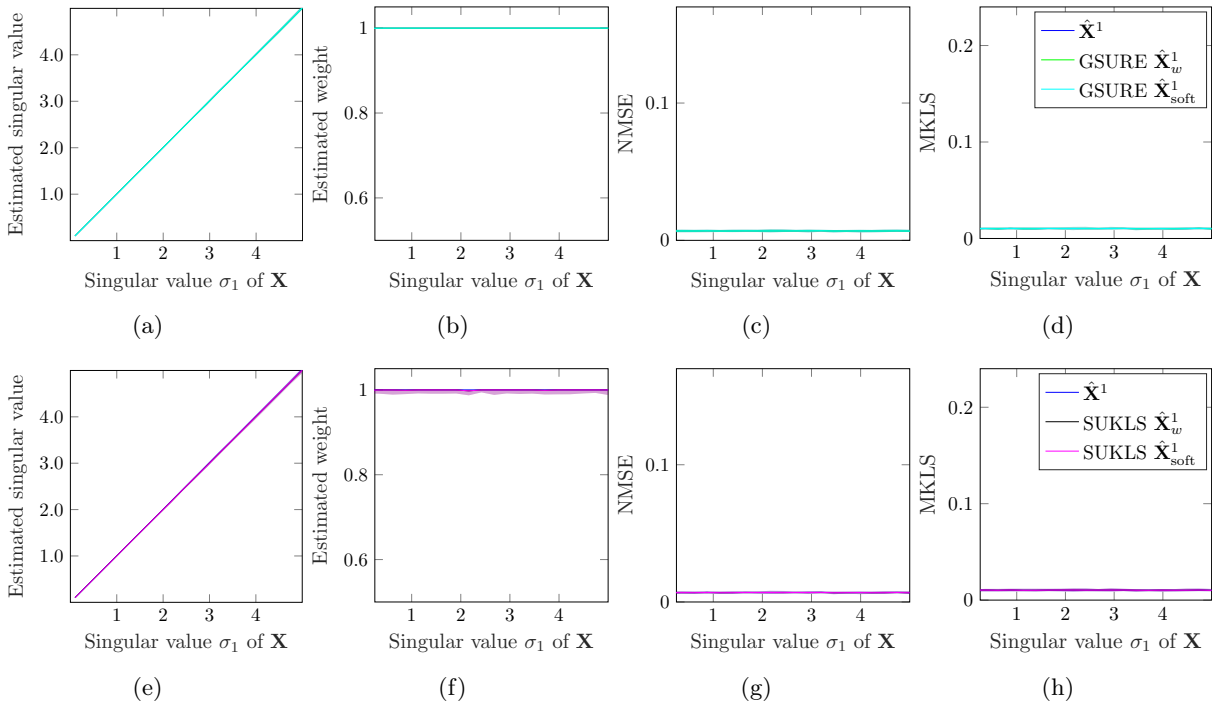


Figure 3: The case of Gamma measurements with $m = n = 100$. (a) Estimated first eigenvalue $\hat{\sigma}_1$ as a function of the true underlying one σ_1 for our proposed estimator $\hat{\boldsymbol{X}}^1_w$ and the soft-thresholding $\hat{\boldsymbol{X}}^1_{\text{soft}}$ when both are guided by the GSURE. Both of them are compared to the first singular value $\tilde{\sigma}_1$ of $\boldsymbol{Y}^1$. Same but for (b) the corresponding weights $\hat{w}_1 = \hat{\sigma}_1/\tilde{\sigma}_1$, (c) the NMSE risk and (d) the MKLS risk. (e-h) Exact same esperiments but when our proposed estimator and the soft-thresholding are both guided by SUKLS. Curves have been computed on $M = 100$ noise realizations, only the median and an 80% confidence interval are represented respectively by a stroke and a shadded area of the same color.

Finally, it should be remarked that a comparison with the asymptotically optimal shrinkage rule $\hat{\boldsymbol{X}}^1_*$ proposed in [Nad14] and [GD14a] for Gaussian noise with variance $\tau^2 = \frac{1}{m}$ is not realistic in the case of Gamma or Poisson measurements. Indeed, as remark in [GD14a], to use the estimator $\hat{\boldsymbol{X}}^1_*$ in a Gaussian model with homoscedastic variance $\tau^2 \neq \frac{1}{m}$, one may take the estimator $\hat{\boldsymbol{X}}^1_* = \sqrt{m}\tau f^*_1(\tilde{\sigma}_1/(\sqrt{m}\tau))\tilde{\boldsymbol{u}}_1\tilde{\boldsymbol{v}}^t_1$. However, this clearly means that the approach in [Nad14] and [GD14a] requires the knowledge of the variance of the entries $\boldsymbol{Y}_{ij}$ of the data matrix $\boldsymbol{Y}$ which is not possible for Gamma or Poisson measurements as either $\text{Var}(\boldsymbol{Y}_{ij}) = \boldsymbol{X}^2_{ij}/L$ or $\text{Var}(\boldsymbol{Y}_{ij}) = \boldsymbol{X}_{ij}$ in these settings.

While all estimators behave similarly in the rank 1 setting, we will see in the next section that they can significantly differ when the rank is let to be larger than 2.

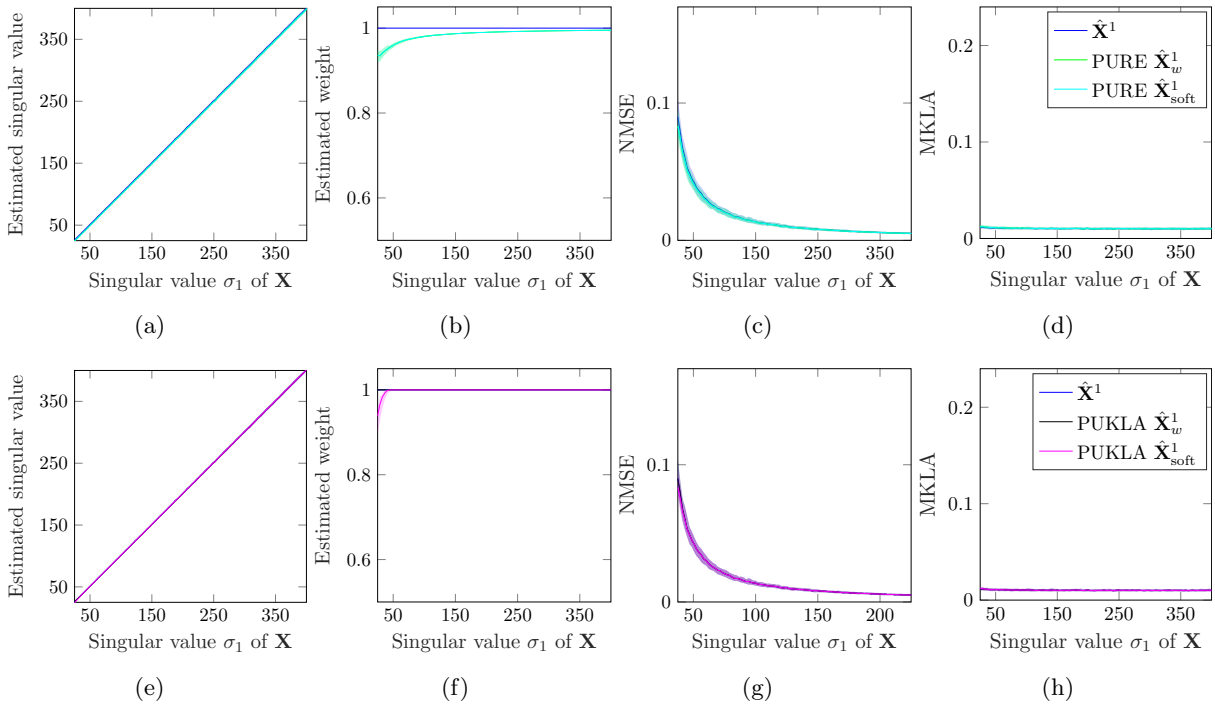


Figure 4: The case of Poisson measurements with $m = n = 100$. (a) Estimated first eigenvalue $\hat{\sigma}_1$ as a function of the true underlying one σ_1 for our proposed estimator $\hat{\boldsymbol{X}}^1_w$ and the soft-thresholding $\hat{\boldsymbol{X}}^1_{\text{soft}}$ when both are guided by the PURE. Both of them are compared to the first singular value $\tilde{\sigma}_1$ of $\boldsymbol{Y}^1$. Same but for (b) the corresponding weights $\hat{w}_1 = \hat{\sigma}_1/\tilde{\sigma}_1$, (c) the NMSE risk and (d) the MKLA risk. (e-h) Exact same esperiments but when our proposed estimator and the soft-thresholding are both guided by PUKLA. Curves have been computed on $M = 100$ noise realizations, only the median and an 80% confidence interval are represented respectively by a stroke and a shadded area of the same color.

5.2 The case of a signal matrix of rank larger than two

We now consider the more complex an realistic setting where the rank r^* of the matrix $\boldsymbol{X}$ is unknown and potentially larger than two, i.e.,

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

where $\boldsymbol{u}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^m$ are vectors with unit norm that are fixed in this numerical experiment, and σ_k are positive real values also fixed in this experiment. We also choose to fix $n = 100$ and $m = 200$, while the true rank is $r^* = 9$ as shown by the red curve in Figure 5(i). Again, let $\boldsymbol{Y} = \sum_{k=1}^{\min(n,m)} \tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t$ be an $n \times m$ matrix whose entries are sampled from model (2.1) and then satisfying $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}$.

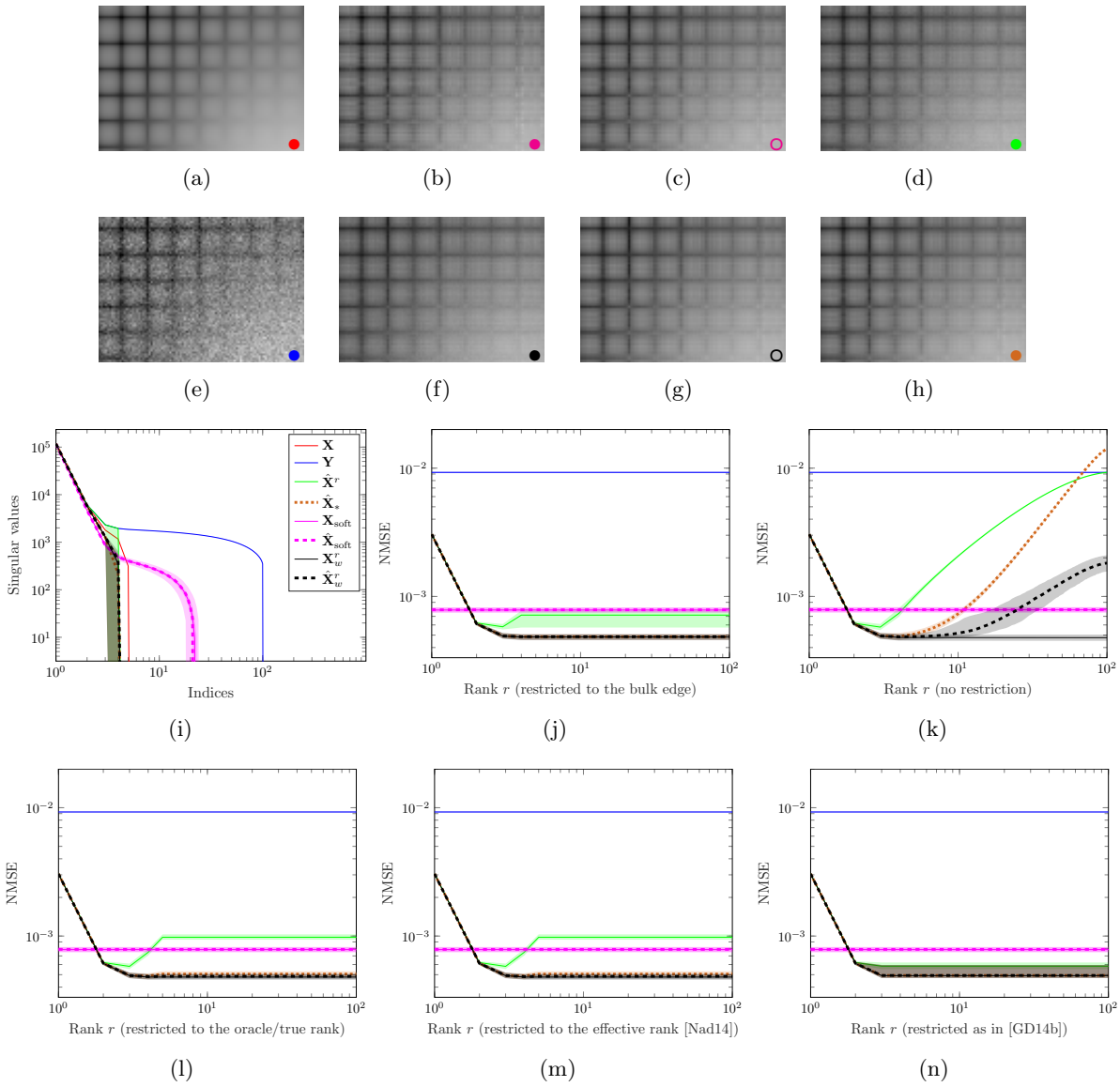


Figure 5: (a) Zoom on a 100×200 noise-free matrix and (e) a single realization of corrupted version by Gaussian noise ($\tau = 80$). (b,c) Oracle soft-thresholding $\boldsymbol{X}_{\text{soft}}$ and data-driven soft-thresholding $\hat{\boldsymbol{X}}_{\text{soft}}$. (d) PCA full rank $\hat{\boldsymbol{X}}^{r_{\max}}$, i.e., $r_{\max} = \min(n, m)$. (f,g,h) Oracle full rank approximation $\boldsymbol{X}_w^{r_{\max}}$, and data-driven full rank estimation $\hat{\boldsymbol{X}}_w^{r_{\max}}$ and $\hat{\boldsymbol{X}}_*^{r_{\max}}$. (i) Their corresponding singular values. (j) NMSE of the various approximations as a function of the rank r. (k) Same but without knowledge the bulk edge, namely $c_+ = 0$. (l,m,n) Same when the active set of singular values is of the form $\hat{s} = \{1, \dots, \hat{r}\}$ where $\hat{r}$ is either given by $\hat{r} = r^*$ (oracle/true rank), $\hat{r} = r_{\text{eff}}$ (effective rank) or by (4.8). In all the figures, the solid curves correspond to oracle estimators and the dashed curves correspond to data-driven estimators, obtained over $M = 1,000$ noise realizatrions. The grey areas represent a 80% confidence interval.

Gaussian distributed measurements

We first consider the case of Gaussian measurements, where $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ with $\mathbb{E}[\boldsymbol{W}_{ij}] = 0$, $\mathrm{Var}(\boldsymbol{W}_{ij}) = \tau^2$ with $\tau = \frac{1}{\sqrt{m}}$. In the following numerical experiments, we study the behavior of the spectral estimator:

- PCA shrinkage

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

- SURE-driven soft-thresholding

$$\hat{\boldsymbol{X}}_{\mathrm{soft}} = \sum_{k=1}^{\min(m,n)} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = (\tilde{\sigma}_k - \lambda(\boldsymbol{Y}))_+,$$

- Asymptotically optimal shrinkage proposed in [Nad14] and [GD14a]

$$\hat{\boldsymbol{X}}_*^r = \sum_{k=1}^{r} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = \frac{1}{\tilde{\sigma}_k} \sqrt{\left(\tilde{\sigma}_k^2 - (c+1)\right)^2 - 4c} \, \mathbb{1}_{\{k \leq \hat{r}\}},$$

- SURE-driven weighted estimator that we have derived in Section 2.3

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = \left(1 - \frac{1}{\tilde{\sigma}_k^2}\left(\frac{k}{m} + \frac{2}{m}\sum_{\ell=2}^{n} \frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}\right)\right)_+ \tilde{\sigma}_k \mathbb{1}_{\{k \leq \hat{r}\}},$$

where $r \in [1, \min(n,m)]$, and for the soft-thresholding, the value $\lambda(\boldsymbol{Y}) > 0$ is obtained by a numerical solver in order to minimize the SURE. Otherwise specified, we consider $\hat{r} = \max\{k \ ; \ \tilde{\sigma}_k > c_+^{n,m}\}$, *i.e.*, an estimator of the rank using knowledge of the bulk edge $c_+ \approx c_+^{n,m}$, hence, $\mathbb{1}_{\{k \leq \hat{r}\}} = \mathbb{1}_{\{\tilde{\sigma}_k > c_+^{n,m}\}}$. As discussed in Section 4, we compare, in these experiments, the influence of rank estimation by analyzing the performances of the same estimators when either $\hat{r} = r_{\max} = \min(n,m)$ (i.e. without knowledge the bulk edge, namely $c_+ = 0$), $\hat{r} = r^*$ (oracle/true rank), $\hat{r} = r_{\mathrm{eff}}$ (effective rank [Nad14]) or by (4.8) (from hard-thresholding of singular values in [GD14b]).

In order to assess the quality of SURE as an estimator of the MSE, we also compare the aforementioned approach with their oracle counterparts given by

$$\boldsymbol{X}_{\mathrm{soft}} = \sum_{k=1}^{\min(m,n)} \hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t \quad \text{with} \quad \hat{\sigma}_k = (\tilde{\sigma}_k - \lambda^{\mathrm{oracle}}(\boldsymbol{Y}))_+, \quad \text{and}$$

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

where $\lambda^{\text{oracle}}(\boldsymbol{Y})$ minimizes the squared error SE (non-expected risk) over the sets and soft-thresholding approximations respectively. Note that $\boldsymbol{X}_w^r$ and $\boldsymbol{X}_{\text{soft}}$ are ideal approximations of $\boldsymbol{X}$ that cannot be used in practice but serve as benchmarks to evaluate the performances of the data-driven estimators $\hat{\boldsymbol{X}}^r$, $\hat{\boldsymbol{X}}_{\text{soft}}$, $\hat{\boldsymbol{X}}_*^r$ and $\hat{\boldsymbol{X}}_w^r$. In order to shed some light on the variance of these estimators, and indirectly on the variance of the SURE, we perform this experiments over $M = 1000$ independent realizations of $\boldsymbol{Y}$.

The results are reported on Figure 5. For an estimator of the rank given either by $\hat{r} = \max\{k\ ;\ \tilde{\sigma}_k > c_+^{n,m}\}$ (knowledge of the bulk edege), $\hat{r} = r^*$ (oracle/true rank), $\hat{r} = r_{\text{eff}}$ (effective rank) or by (4.8), it can be observed that $\hat{\boldsymbol{X}}_w^r$, $\hat{\boldsymbol{X}}_*^r$ and $\boldsymbol{X}_w^r$ achieve comparable performances for all $r \in [1, \min(m,n)]$ even though the two first do not rely on the unknown matrix $\boldsymbol{X}$. Similarly $\hat{\boldsymbol{X}}_{\text{soft}}$ and $\boldsymbol{X}_{\text{soft}}$ achieve also comparable performances showing again that the SURE accurately estimates the MSE. In terms of error bands for the NMSE, $\hat{\boldsymbol{X}}_w^r$, $\hat{\boldsymbol{X}}_*^r$ and $\boldsymbol{X}_w^r$ outperform $\hat{\boldsymbol{X}}_{\text{soft}}$ and $\boldsymbol{X}_{\text{soft}}$ provided that r is large enough. Moreover, the performance of $\hat{\boldsymbol{X}}_w^r$ plateaus to its optimum when the rank r becomes large. This allows us to choose $r = \min(n,m)$ when we do not have *a priori* on the true or effective rank.

Interestingly, Fig. 5.(k) shows that when the above estimators are used without the knowledge of the bulk edge (i.e. by taking $c_+^{n,m} = 0$ in their computation instead of $c_+^{n,m} = 1 + \sqrt{\frac{n}{m}}$, which corresponds to the choice $\hat{r} = r_{\max} = \min(n,m)$), the performance of $\hat{\boldsymbol{X}}_w^r$ actually decreases when the rank r becomes too large. Indeed, it is clear from Fig. 5.(k), that the the error band of the NMSE of $\hat{\boldsymbol{X}}_w^r$ becomes much larger as the rank r increases. This illustrates that the SURE suffers from estimation variance in the case of over parametrization when r becomes too large, and thus it cannot be used to estimate jointly a too large number of weights. Therefore, the knowledge of an appropriate estimator $\hat{r}$ of the rank (e.g. using the bulk edge) seems to provide a relevant upper bound on the number of weights that can be jointly and robustly estimated with the SURE.

Gamma and Poisson measurements

Let us now consider the case where the entries of $\boldsymbol{Y}_{ij} > 0$ of the data matrix $\boldsymbol{Y}$ are independently sampled from a Gamma or Poisson distribution with mean $\boldsymbol{X}_{ij} > 0$. We again consider estimators of the form (2.17). In this context, we compare the following spectral shrinkage estimators, set for $\varepsilon = 10^{-6}$, as:

- PCA shrinkage

$$\hat{\boldsymbol{X}}^r = \sum_{k=1}^{r} \max\left[\tilde{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_k = \tilde{\sigma}_k \mathbb{1}_{\{k \in \tilde{s}\}},$$

- GSURE/SUKLS/PURE/SUKLA driven soft-thresholding

$$\hat{\boldsymbol{X}}_{\text{soft}} = \sum_{k=1}^{\min(m,n)} \max\left[\hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_k = (\tilde{\sigma}_k - \lambda(\boldsymbol{Y}))_+,$$

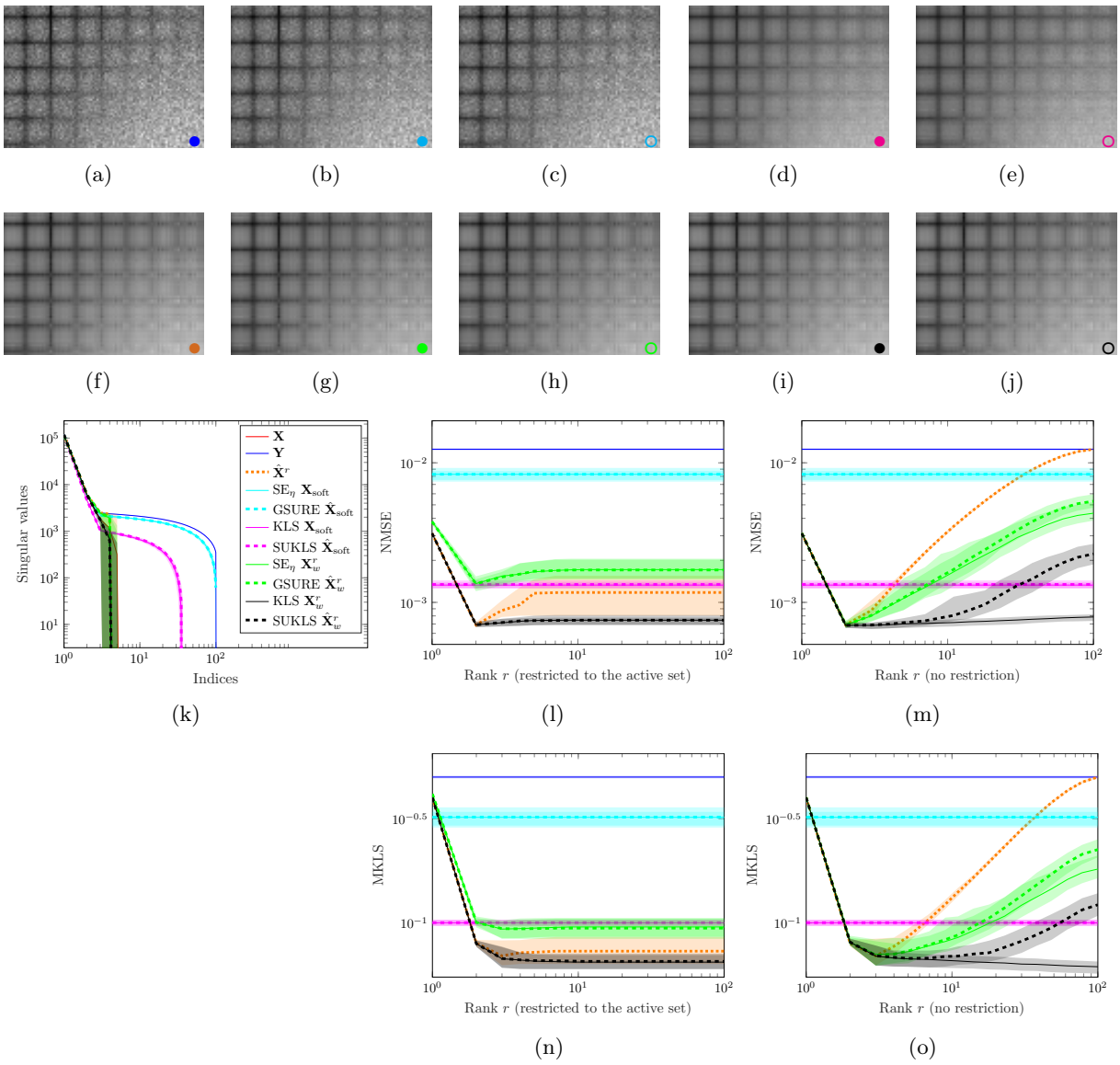


Figure 6: (a) A single realization of corrupted version by Gamma noise ($L = 80$) with zoom on a 100×200 matrix. (b,c,d,e) Oracle soft-thresholding $\boldsymbol{X}_{\mathrm{soft}}$ and data-driven soft-thresholding $\hat{\boldsymbol{X}}_{\mathrm{soft}}$ respectively for SE_η, GSURE, KLS and SUKLS. (f) PCA $\hat{\boldsymbol{X}}^{r_{\max}}$ with full rank approximation i.e. $r_{\max} = \min(n, m)$. (g,h,i,j) Oracle full rank approximation $\boldsymbol{X}_w^{r_{\max}}$, and data-driven full rank estimation $\hat{\boldsymbol{X}}_w^{r_{\max}}$ respectively for SE_η, GSURE, KLS and SUKLS. (k) Their corresponding singular values averaged over $M = 100$ noise realizations. (l,m) NMSE averaged over $M = 100$ noise realizations as a function of the rank r with and without using the active set. (n,o) Same but with respect to MKLS.

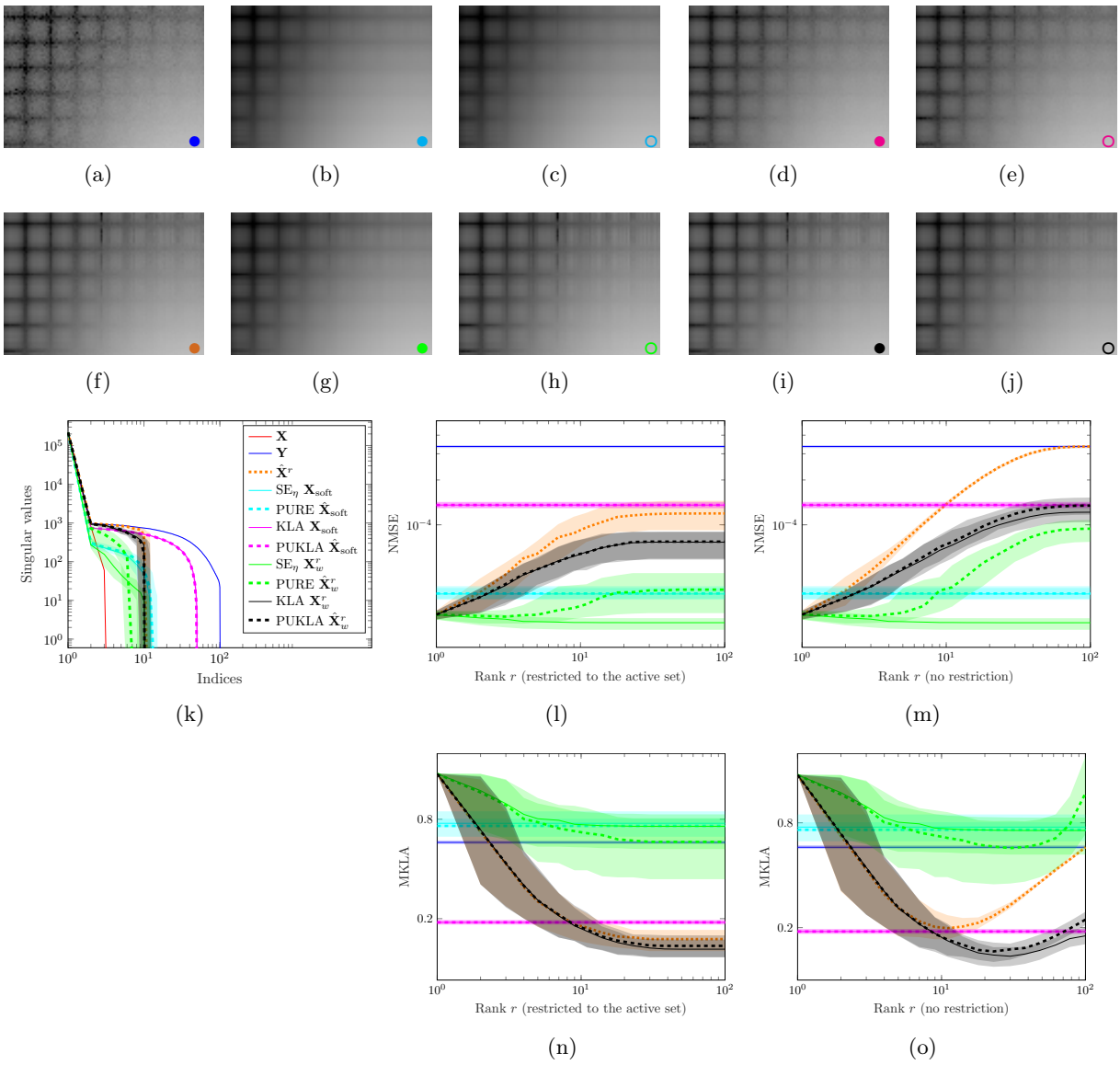


Figure 7: (a) A single realization of corrupted version by Poisson noise with zoom on a 100×200 noise-free matrix (b,c,d,e) Oracle soft-thresholding $\boldsymbol{X}_{\text{soft}}$ and data-driven soft-thresholding $\hat{\boldsymbol{X}}_{\text{soft}}$ respectively for SE, PURE, KLA and PUKLA. (f) PCA $\hat{\boldsymbol{X}}^{r_{\max}}$ with full rank approximation i.e. $r_{\max} = \min(n, m)$. (g,h,i,j) Oracle full rank approximation $\boldsymbol{X}_w^{r_{\max}}$, and data-driven full rank estimation $\hat{\boldsymbol{X}}_w^{r_{\max}}$ respectively for SE, PURE, KLA and PUKLA. (k) Their corresponding singular values averaged over 200 noise realizations. (l,m) NMSE averaged over 200 noise realizations as a function of the rank r with and without using the active set. (n,o) Same but with respect to MKLA. (Matrix entries are displayed in log-scale for better visual assessment.)

- GSURE/SUKLS/PURE/SUKLA driven weighted estimator

$$\hat{\boldsymbol{X}}_w^r = \sum_{k=1}^{r} \max\left[\hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_k = w_k(\boldsymbol{Y})\tilde{\sigma}_k \mathbb{1}_{\{k\in\tilde{s}\}},$$

where $r \in [1, \min(n,m)]$, and $\tilde{s}$ is the approximated active subset as defined in Section 4. For the soft-thresholding, the value $\lambda(\boldsymbol{Y}) > 0$ is obtained by a numerical solver in order to minimize either the GSURE or the SUKLS criterion (in the Gamma case) and either the PURE or the PUKLA criterion (in the Poisson case). As shown in Section 2.3, in the case of Gamma (resp. Poisson) measurements, the value of $w_k(\boldsymbol{Y})$ for $k \in \tilde{s}$ which minimizes the GSURE (resp. PURE) or the SUKLS (resp. PUKLA), cannot be obtained in closed form. As an alternative, we adopt a greedy one-dimensional optimization strategy starting from the matrix $\tilde{\sigma}_1 \tilde{\boldsymbol{u}}_1 \tilde{\boldsymbol{v}}_1^t$ and next updating the weights w_ℓ sequentially by starting $\ell = 1$ to $\ell = \min(n,m)$, with the constraint that, for all $\ell \notin \tilde{s}$, the weight w_ℓ is set to zero. To this end, we resort to one-dimensional optimization techniques in the interval $[0,1]$ using Matlab's command `fminbnd`. This strategy is used for GSURE, SUKLS, PURE and PUKLA by evaluating them as described in Section B. As in the Gaussian setting, we compare this spectral estimators with their oracle counterparts given by

$$\boldsymbol{X}_{\text{soft}} = \sum_{k=1}^{\min(m,n)} \max\left[\hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_k = (\tilde{\sigma}_k - \lambda^{\text{oracle}}(\boldsymbol{Y}))_+, \quad \text{and}$$

$$\boldsymbol{X}_w^r = \sum_{k=1}^{r} \max\left[\hat{\sigma}_k \tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{v}}_k^t, \varepsilon\right] \quad \text{with} \quad \hat{\sigma}_k = w_k^{\text{oracle}}(\boldsymbol{Y})\tilde{\sigma}_k \mathbb{1}_{\{k\in\tilde{s}\}}.$$

where $w_k^{\text{oracle}}(\boldsymbol{X})$ and $\lambda^{\text{oracle}}(\boldsymbol{Y})$ minimizes one of the objective SE_η, KLS, SE or KLA (non-expected risks) over the set of matrices sharing with $\boldsymbol{Y}$ the same r first left and right singular vectors, and soft-thresholding approximations respectively. Note again that $\boldsymbol{X}_w^r$ and $\boldsymbol{X}_{\text{soft}}$ are ideal approximations of $\boldsymbol{X}$ that cannot be used in practice but serve as benchmarks to evaluate the performances of the data-driven estimators $\hat{\boldsymbol{X}}_w^r$ and $\hat{\boldsymbol{X}}_{\text{soft}}$.

The results for the Gamma noise are reported on Figure 6. As in the Gaussian setting, it can be observed that $\hat{\boldsymbol{X}}_w^r$ and $\boldsymbol{X}_w^r$ achieve comparable performances, as well as $\hat{\boldsymbol{X}}_{\text{soft}}$ and $\boldsymbol{X}_{\text{soft}}$ showing that the GSURE (resp. SUKLS) accurately estimates the MSE_η (resp. KLS). Visual inspection of the restored matrices tends to show that the estimators driven by MSE_η or GSURE produce less relevant results compared to KLS or SUKLS, as confirmed by the curves of NMSE and MKLS. Performance in terms of NMSE also illustrates that minimizers of SE_η do not coincides with those of SE. As in the Gaussian setting, $\hat{\boldsymbol{X}}_w^r$ and $\boldsymbol{X}^r$ outperform $\hat{\boldsymbol{X}}_{\text{soft}}$, $\boldsymbol{X}_{\text{soft}}$ and standard PCA $\hat{\boldsymbol{X}}^r$ provided that r is large enough. Moreover, the performance of $\hat{\boldsymbol{X}}_w^r$ obtained with KL objectives plateaus to its optimum when the rank r becomes large. Again, this allows us to choose $r = \min(n,m)$ when we do not have *a priori* on the true rank $r^\star$.

The results for the Poisson noise are reported on Figure 7. The conclusions are similar to the Gaussian and Gamma cases. Obviously, the NMSE is smaller for approximations that minimizes SE (or PURE) than for those minimizing KLA (or PUKLA). However, visual inspection of the obtained matrices tends to demonstrate that minimizing such objectives might be less relevant than minimizing KL objectives. In this setting, the performance of $\hat{\boldsymbol{X}}_w^r$ is on a par with the one

of $\hat{\boldsymbol{X}}_{\text{soft}}$ based on PUKLA. In fact, for other choices of matrices $\boldsymbol{X}$, $\hat{\boldsymbol{X}}_w^r$ based on PUKLA might improve, in terms of MKLS, much more on $\hat{\boldsymbol{X}}_{\text{soft}}$, and might improve not as much on $\hat{\boldsymbol{X}}_w^r$ based on PURE. Nevertheless, whatever $\boldsymbol{X}$, we observed that $\hat{\boldsymbol{X}}_w^r$ driven by PUKLA always reaches at least as good performance in terms of MKLS as the best of $\hat{\boldsymbol{X}}_w^r$ driven by SE and $\hat{\boldsymbol{X}}_{\text{soft}}$.

Fig. 6.(m), Fig. 6.(o), Fig. 7.(m) and Fig. 7.(o) show that when the above estimators are used without the active set (i.e., by choosing $\tilde{s} = [1, \min(n, m])$, the performance of $\hat{\boldsymbol{X}}_w^r$ actually decreases when the rank r becomes too large. As in the Gaussian setting, this can be explained by the fact that the GSURE, SUKLS, PURE and PUKLA suffer from estimation variance in the case of over parametrization, hence, they cannot be used to estimate jointly a too large number of weights. The active set $\tilde{s}$ (in the same manner as the bulk edge) seems to provide a relevant selection of the weights that can be jointly and robustly estimated in a data driven way.

5.3 Signal matrix with equal singular values and increasing rank

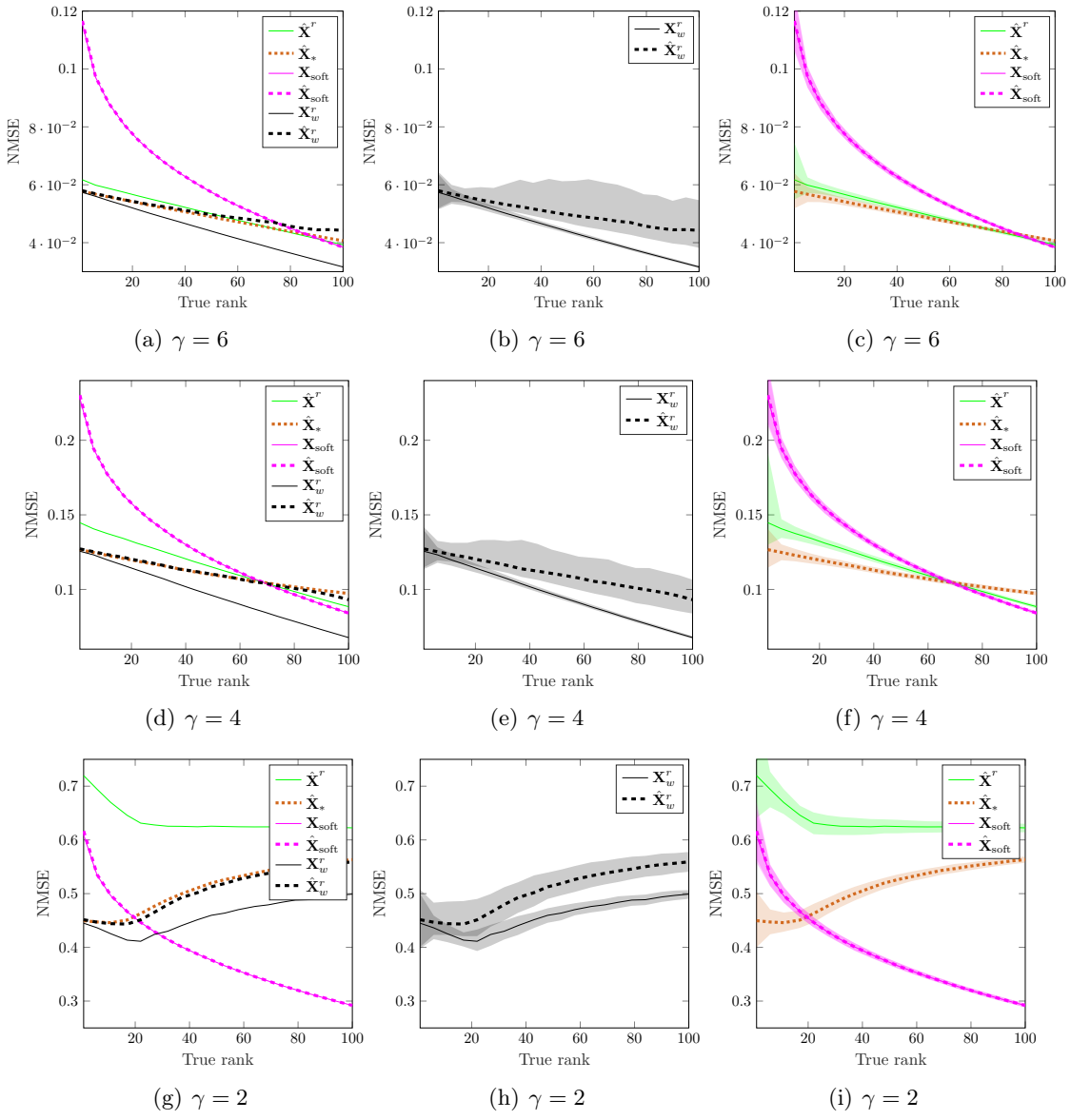
We finally propose to highlight potential limitations of our approach in the situation where the rank r^* of the matrix $\boldsymbol{X} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t$ is let growing and all positive singular values σ_k of $\boldsymbol{X}$ are equal, namely

$$\boldsymbol{Y} = \sum_{k=1}^{r^*} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^t + \boldsymbol{W} \quad \text{with} \quad \sigma_k = \gamma c_{n,m}^{1/4} \text{ for all } 1 \leq k \leq r^*, \tag{5.1}$$

where $\boldsymbol{u}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^m$ are vectors with unit norm that are fixed, $c_{n,m} = \frac{n}{m}$ and $\boldsymbol{W}$ is centered random matrix whose entries are iid Gaussian variables with variance $\tau^2 = 1/m$. We again choose to fix $n = 100$ and $m = 200$, while the true rank is r^* let growing from 1 to $\min(n, m)$ in the following numerical experiments. The constant γ is chosen to be larger than 1. Hence, eq. (5.1) corresponds to the Gaussian spiked population model in the setting where all positive singular values are equal and larger than the threshold $c_{n,m}^{1/4}$. The choice $\sigma_k = \gamma c_{n,m}^{1/4}$ with $\gamma > 1$ is motivated by the results from Proposition 3.1.

For a given value of the true rank r^*, we performed experiments involving $M = 1000$ realizations from model (5.1) to compare the NMSE of the estimators by oracle soft-thresholding $\boldsymbol{X}_{\text{soft}}$, data-driven soft-thresholding $\hat{\boldsymbol{X}}_{\text{soft}}$, PCA full rank $\hat{\boldsymbol{X}}^{r_{\max}}$ i.e. $r_{\max} = \min(n, m)$, oracle full rank approximation $\boldsymbol{X}_w^{r_{\max}}$, and data-driven full rank estimation $\hat{\boldsymbol{X}}_w^{r_{\max}}$ and $\hat{\boldsymbol{X}}_*^{r_{\max}}$. All these estimators have been introduced in Section 5.2.

In Figure 8, we report the results of numerical experiments by displaying errors bars of the NMSE of these estimators as functions of the true rank r^*. For low values of the true rank ($r^* \leq 20$), the data-driven estimators $\hat{\boldsymbol{X}}_w^{r_{\max}}$ (our approach) and $\hat{\boldsymbol{X}}_*^{r_{\max}}$ (shrinkage rule from [GD14a]) achieve the best performances that are similar in term of median value of the NMSE. However, our approach has some limitations with respect to the performances of the estimator from [GD14a] or data-driven soft-thresholding [CSLT13] in the setting where the signal matrix has equal positive singular values and when its rank is increasing. Moreover, the error bands of the NMSE for our approach becomes significantly larger than those of the other data-driven estimators when the true rank r^* increases. This illustrates that SURE minimization leads to estimators with a high variance in the case of over parametrization, that is, when there exists a large number of significant singular values in the signal matrix.



(a) $\gamma = 6$ (b) $\gamma = 6$ (c) $\gamma = 6$

(d) $\gamma = 4$ (e) $\gamma = 4$ (f) $\gamma = 4$

(g) $\gamma = 2$ (h) $\gamma = 2$ (i) $\gamma = 2$

Figure 8: Comparison of NMSE as a function of the true rank r^* in model (5.1) for different values of γ for the estimator by oracle soft-thresholding $\boldsymbol{X}_{\text{soft}}$, data-driven soft-thresholding $\hat{\boldsymbol{X}}_{\text{soft}}$, PCA full rank $\hat{\boldsymbol{X}}^{r_{\max}}$ i.e. $r_{\max} = \min(n, m)$, oracle full rank approximation $\boldsymbol{X}_w^{r_{\max}}$, and data-driven full rank estimation $\hat{\boldsymbol{X}}_w^{r_{\max}}$ and $\hat{\boldsymbol{X}}_*^{r_{\max}}$. The active set set of singular values is of the form $\hat{s} = \{1, \ldots, \hat{r}\}$ where $\hat{r} = \max\{k\ ;\ \tilde{\sigma}_k > c_+^{n,m}\}$ is an estimator of the rank using knowledge of the bulk edge $c_+ \approx c_+^{n,m}$ (a), (d), (g) Median value of the NMSE of the various estimators over $M = 1000$ Gaussian noise realizations in model (5.1) as a function of the true rank r^*. (b), (c), (e), (f), (h), (i) The grey areas represent error bands of the NMSE of data-driven and oracle estimators.

A Proof of the main results

A.1 Proof of Proposition 3.2

Let us first introduce some notation and definitions to be used in the proof. For all $1 \leq \ell \leq n$, let $\tilde{\lambda}_\ell$ be the eigenvalues of $\boldsymbol{Y}\boldsymbol{Y}^t$ namely $\tilde{\lambda}_\ell = \tilde{\sigma}_\ell^2$. For a fixed $1 \leq k \leq r^*$ such that $\sigma_k > c^{1/4}$, let us introduce the complex-valued function g_k defined by

$$g_k(z) = \frac{1}{n} \sum_{\ell=1;\ell\neq k}^{n} \frac{1}{z - \tilde{\lambda}_\ell} \quad \text{for } z \in \mathbb{C} \setminus \mathrm{supp}(\mu_\mathrm{k}),$$

where $\mathrm{supp}(\mu_\mathrm{k}) = \left\{\tilde{\lambda}_\ell; 1 \leq \ell \leq \mathrm{n},\ \ell \neq \mathrm{k}\right\}$ is the support of the random measure $\mu_k = \frac{1}{n}\sum_{\ell=1;\ell\neq k}^{n} \delta_{\tilde{\lambda}_\ell}$ on $\mathbb{R}_+$, where δ_λ denotes the Dirac measure at λ. It is clear that

$$g_k(z) = \int \frac{1}{z-\lambda} d\mu_k(\lambda).$$

The main difficulty in the proof is to show that, almost surely,

$$\lim_{n\to+\infty} g_k(\tilde{\sigma}_k^2) = \frac{1}{\rho^2(\sigma_k)} \left(1 + \frac{1}{\sigma_k^2}\right),$$

which is the purpose of what follows.

For a matrix $A \in \mathbb{R}^{n\times m}$ (with $n \leq m$), we denote its singular values by $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_n(A) \geq 0$. Hence, one has that $\tilde{\sigma}_\ell = \sigma_\ell(\boldsymbol{Y})$ for all $1 \leq \ell \leq n$. Now, we recall that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}$ where $\boldsymbol{X}$ is a fixed matrix of rank r^* and $\boldsymbol{W}$ is a random matrix with iid entries sampled from a Gaussian distribution with zero mean and variance $\frac{1}{m}$. The first step in the proof is to show that the random measure μ_k behaves asymptotically as the almost sure limit of the empirical spectral measure $\mu_{\boldsymbol{W}\boldsymbol{W}^t}$ of the Wishart matrix $\boldsymbol{W}\boldsymbol{W}^t$. By definition, the eigenvalues of $\boldsymbol{W}\boldsymbol{W}^t$ are $\lambda_\ell(\boldsymbol{W}) = \sigma_\ell^2(\boldsymbol{W})$ for all $1 \leq \ell \leq n$ and $\mu_{\boldsymbol{W}\boldsymbol{W}^t}$ is thus defined as

$$\mu_{\boldsymbol{W}\boldsymbol{W}^t} = \frac{1}{n}\sum_{\ell=1}^{n} \delta_{\lambda_\ell(\boldsymbol{W})}.$$

It is well know (see e.g. Theorem 3.6 in [BS10]) that, once $m = m_n \geq n$ and $\lim_{n\to+\infty} \frac{n}{m} = c$ with $0 < c \leq 1$, then, almost surely, the empirical spectral measure $\mu_{\boldsymbol{W}\boldsymbol{W}^t}$ converges weakly to the so-called Marchenko-Pastur distribution μ_{MP} which is deterministic and has the following density $\frac{d\mu_{MP}(\lambda)}{d\lambda} = \frac{1}{2\pi c\lambda}\sqrt{(c_+^2 - \lambda)(\lambda - c_-^2)}\ \mathbf{1}_{[c_-^2, c_+^2]}(\lambda)$. We recall that such a convergence can also be characterized through the so-called Cauchy or Stieltjes transform which is defined for any probability measure μ on $\mathbb{R}$ as

$$\forall z \in \mathbb{C} \text{ outside the support of } \mu, \quad g_\mu(z) = \int \frac{1}{z-\lambda} d\mu(\lambda).$$

By eq. (3.3.2) in [BS10], one obtains that, almost surely,

$$\lim_{n\to\infty} \int \frac{1}{z-\lambda} d\mu_{\boldsymbol{W}\boldsymbol{W}^t}(\lambda) = g_{MP}(z) \text{ for any } z \in \mathbb{C} \setminus \mathbb{R}, \tag{A.1}$$

where g_{MP} is the Cauchy transform of μ_{MP} and

$$g_{MP}(z) = \int \frac{1}{z-\lambda} d\mu_{MP}(\lambda) = \frac{z-(1-c)-\sqrt{(z-(c+1))^2-4c}}{2cz} \quad \text{for all } z \in \mathbb{C} \setminus [c_-^2, c_+^2].$$

Moreover, by Proposition 6 in [PL03], the convergence (A.1) is uniform over any compact subset of $\mathbb{C} \setminus \mathbb{R}$.

Then, it follows from the so-called Weyl's interlacing inequalities (see e.g. Theorem 3.1.2 in [HJ91]) that for all $1 \leq \ell \leq n$

$$\sigma_{\ell+r^*}(\boldsymbol{W}) \leq \sigma_\ell(\boldsymbol{Y}) \leq \sigma_{\ell-r^*}(\boldsymbol{W}), \tag{A.2}$$

with the convention that $\sigma_k(\boldsymbol{W}) = -\infty$ if $k > n$ and $\sigma_k(\boldsymbol{W}) = +\infty$ if $k \leq 0$. Thanks to the results that have been recalled above on the asymptotic properties of $\mu_{\boldsymbol{W}\boldsymbol{W}^t}$, one may use inequalities (A.2) to prove that, almost surely, the random measure μ_k converges weakly to the Marchenko-Pastur distribution μ_{MP}. Under the assumptions of Proposition 3.2 and using Proposition 3.1, it can be shown that there exists $\eta_k > 0$ such that, almost surely and for all sufficiently large n

$$\tilde{\lambda}_\ell \notin K_k := [\rho^2(\sigma_k) - \eta_k, \rho^2(\sigma_k) + \eta_k]$$

for any $1 \leq \ell \leq n$ with $\ell \neq k$. Now, recall that the support $\mathrm{supp}(\mu_\mathrm{k})$ of the random measure μ_k is $\left\{\tilde{\lambda}_\ell; 1 \leq \ell \leq n,\ \ell \neq k\right\}$, and that $\mathrm{supp}(\mu_\mathrm{MP}) = [\mathrm{c}_-^2, \mathrm{c}_+^2]$. Hence, for all sufficiently large n, one has that

$$\mathrm{supp}(\mu_\mathrm{k}) \cap \mathrm{K_k} = \emptyset \quad \text{and} \quad \mathrm{supp}(\mu_\mathrm{MP}) \cap \mathrm{K_k} = \emptyset.$$

Therefore, thanks to the weak convergence of μ_k to μ_{MP} and using Ascoli's Theorem, one may prove that

$$\lim_{n\to\infty} \sup_{z \in K_k} |g_k(z) - g_{MP}(z)| = 0 \text{ almost surely.} \tag{A.3}$$

Thanks to our assumptions, one has that, almost surely, $\lim_{n\to+\infty} \tilde{\sigma}_k^2 = \rho^2(\sigma_k)$ by Proposition 3.1. Hence, almost surely and for all sufficiently large n, one has that $\tilde{\sigma}_k^2 \in K_k$ and so

$$|g_k(\tilde{\sigma}_k^2) - g_{MP}(\rho^2(\sigma_k))| \leq \sup_{z \in K_k} |g_k(z) - g_{MP}(z)| + |g_{MP}(\tilde{\sigma}_k^2) - g_{MP}(\rho^2(\sigma_k))|.$$

Therefore, using the uniform convergence (A.3) of g_k to g_{MP} and the continuity of g_{MP} at $z = \rho^2(\sigma_k)$, one obtains that, almost surely,

$$\lim_{n\to+\infty} g_k(\tilde{\sigma}_k^2) = g_{MP}(\rho^2(\sigma_k)) = \frac{1}{\rho^2(\sigma_k)} \times \frac{\rho^2(\sigma_k) - 1 + c - \sqrt{(\rho^2(\sigma_k) - (c+1))^2 - 4c}}{2c}.$$

Since $g_k(\tilde{\sigma}_k^2) = \frac{1}{n}\sum_{\ell=1;\ell\neq k}^n \frac{1}{\tilde{\sigma}_k^2 - \tilde{\sigma}_\ell^2}$, using the above equation and relation (3.1), it follows immediately that $g_{MP}(\rho^2(\sigma_k)) = \frac{1}{\rho^2(\sigma_k)}\left(1 + \frac{1}{\sigma_k^2}\right)$ so that, almost surely,

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

which completes the proof.

A.2 A technical result to prove SURE-like formulas

We recall the key lemma needed to prove the SURE-like formulas in an exponential family in the continuous case. Similar results have already been formulated in different papers in the literature, see e.g. the review proposed in [Del15].

Lemma A.1. *Let $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ be a random matrix whose entries $\boldsymbol{Y}_{ij}$ are independently sampled from the continuous exponential family (2.2) in canonical form (that is the distribution of $\boldsymbol{Y}_{ij}$ is absolutely continuous with respect to the Lebesgue measure dy on $\mathbb{R}$). Suppose that the function h is continuously differentiable on $\mathcal{Y} = \mathbb{R}$. Let $1 \leq i \leq n$ and $1 \leq j \leq m$, and denote by $F_{ij} : \mathbb{R}^{n\times m} \to \mathbb{R}$ a continuously differentiable function such that*

$$\mathbb{E}\left[|F_{ij}(\boldsymbol{Y})|\right] < +\infty. \tag{A.4}$$

Then, the following relation holds

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij} F_{ij}(\boldsymbol{Y})\right] = -\mathbb{E}\left[\frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})} F_{ij}(\boldsymbol{Y}) + \frac{\partial F_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}}\right].$$

Proof. Using the expression (2.2) of the pdf of the random varibles $\boldsymbol{Y}_{ij}$, one has that

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij} F_{ij}(\boldsymbol{Y})\right] = \int_{\mathbb{R}^{n\times m}} F_{ij}(Y) h(y_{ij}) \boldsymbol{\theta}_{ij} \exp\left(\boldsymbol{\theta}_{ij} y_{ij} - A(\boldsymbol{\theta}_{ij})\right) \, \mathrm{d}y_{ij} \prod_{\substack{1\leq k\leq n \\ 1\leq \ell\leq m \\ (k,\ell)\neq(i,j)}}^{n} p(y_{k\ell}; \boldsymbol{\theta}_{k\ell}) \, \mathrm{d}y_{k\ell}.$$

where $Y = (y_{k\ell})_{1\leq k\leq n, 1\leq \ell\leq m}$. Thanks to condition (A.4), it follows that

$$\int_{\mathbb{R}^{n\times m}} F_{ij}(Y) h(y_{ij}) \exp\left(\boldsymbol{\theta}_{ij} y_{ij} - A(\boldsymbol{\theta}_{ij})\right) \, \mathrm{d}y_{ij} \prod_{\substack{1\leq k\leq n \\ 1\leq \ell\leq m \\ (k,\ell)\neq(i,j)}}^{n} p(y_{k\ell}; \boldsymbol{\theta}_{k\ell}) \, \mathrm{d}y_{k\ell} < +\infty. \tag{A.5}$$

Therefore, given that $\boldsymbol{\theta}_{ij} \exp\left(\boldsymbol{\theta}_{ij} y_{ij} - A(\boldsymbol{\theta}_{ij})\right) = \frac{\partial \exp(\boldsymbol{\theta}_{ij} y_{ij} - A(\boldsymbol{\theta}_{ij}))}{\partial y_{ij}}$, an integration by part and eq. (A.5) imply that

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij} F_{ij}(\boldsymbol{Y})\right] = -\int_{\mathbb{R}^{n\times m}} \frac{\partial F_{ij}(Y) h(y_{ij})}{\partial y_{ij}} \exp\left(\boldsymbol{\theta}_{ij} y_{ij} - A(\boldsymbol{\theta}_{ij})\right) \, \mathrm{d}y_{ij} \prod_{\substack{1\leq k\leq n \\ 1\leq \ell\leq m \\ (k,\ell)\neq(i,j)}}^{n} p(y_{k\ell}; \boldsymbol{\theta}_{k\ell}) \, \mathrm{d}y_{k\ell}.$$

Now, since $\frac{\partial F_{ij}(Y) h(y_{ij})}{\partial y_{ij}} = h'(y_{ij}) F_{ij}(Y) + \frac{\partial F_{ij}(Y)}{\partial y_{ij}} h(y_{ij})$, we finally obtain that

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij} F_{ij}(\boldsymbol{Y})\right] = -\mathbb{E}\left[\frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})} F_{ij}(\boldsymbol{Y}) + \frac{\partial F_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}}\right],$$

which completes the proof. □

A.3 Proof of Proposition 2.1

We remark that

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathbb{E}\left[|\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})|^2 - 2\boldsymbol{\theta}_{ij}\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})\right] + \boldsymbol{\theta}_{ij}^2\right). \tag{A.6}$$

Using Lemma A.1 with $F_{ij}(\boldsymbol{Y}) = \hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})$ and condition (2.4), it follows that

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij}\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})\right] = \mathbb{E}\left[\frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})\right] + \mathbb{E}\left[\frac{\partial\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}}\right]. \tag{A.7}$$

Then, by definition (2.2) of the exponential family, we remark that

$$\mathbb{E}\left[\frac{h''(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\right] = \int_{\mathbb{R}} h''(y_{ij})\exp\left(\boldsymbol{\theta}_{ij}y_{ij} - A(\boldsymbol{\theta}_{ij})\right)\,\mathrm{d}y_{ij}.$$

Hence, using an integration by parts twice, we arrive at

$$\mathbb{E}\left[\frac{h''(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}\right] = \boldsymbol{\theta}_{ij}^2\int_{\mathbb{R}} h(y_{ij})\exp\left(\boldsymbol{\theta}_{ij}y_{ij} - A(\boldsymbol{\theta}_{ij})\right)\,\mathrm{d}y_{ij} = \boldsymbol{\theta}_{ij}^2. \tag{A.8}$$

To complete the proof, it suffices to insert equalities (A.7) and (A.8) into (A.6).

A.4 Proof of Proposition 2.2

Thanks to eq. (2.8), one has that

$$\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left[\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})A'(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})) - \boldsymbol{\theta}_{ij}A'(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})) - A(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y}))\right] + A(\boldsymbol{\theta}_{ij}). \tag{A.9}$$

Using Lemma A.1 with $F_{ij}(\boldsymbol{Y}) = A'(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y}))$ and condition (2.9), it follows that

$$\mathbb{E}\left[\boldsymbol{\theta}_{ij}A'(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y}))\right] = -\mathbb{E}\left[\frac{h'(\boldsymbol{Y}_{ij})}{h(\boldsymbol{Y}_{ij})}A'(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y}))\right] - \mathbb{E}\left[\frac{\partial\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y})}{\partial\boldsymbol{Y}_{ij}}A''(\hat{\boldsymbol{\theta}}_{ij}^f(\boldsymbol{Y}))\right]. \tag{A.10}$$

Thus, inserting equality (A.10) into (A.9) implies that

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

is an unbiased estimator of $\mathrm{MKLS}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) - \sum_{i=1}^n \sum_{j=1}^m A(\boldsymbol{\theta}_{ij})$. Now recall that $f_{ij}(\boldsymbol{Y}) = \eta^{-1}\left(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})\right)$ and that $A'(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})) = \eta^{-1}\left(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})\right)$ by Assumption 2.1. Therefore, $\frac{\partial f_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}} = A''(\hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y}))\frac{\partial \hat{\boldsymbol{\theta}}^f_{ij}(\boldsymbol{Y})}{\partial \boldsymbol{Y}_{ij}}$, and thus

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

which completes the proof.

A.5 Proof of Proposition 2.3

Thanks to the expression (2.14) of the MKLA risk for data sampled from a Poisson distribution, it follows that

$$\mathrm{MKLA}(\hat{\boldsymbol{\theta}}^f, \boldsymbol{\theta}) + \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}\log\left(\boldsymbol{X}_{ij}\right) = \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}\left[\hat{\boldsymbol{X}}^f_{ij} - \boldsymbol{X}_{ij}\log\left(\hat{\boldsymbol{X}}^f_{ij}\right)\right]$$

In the case of Poisson data, one has that $\exp\left(\boldsymbol{\theta}_{ij}\right) = \boldsymbol{X}_{ij}$ and $\frac{h(\boldsymbol{Y}_{ij}-1)}{h(\boldsymbol{Y}_{ij})} = \boldsymbol{Y}_{ij}$. Therefore, by applying Hudson's Lemma 2.1 with $F_{ij}(\boldsymbol{Y}) = \log\left(\hat{\boldsymbol{X}}^f_{ij}\right)$, it follows that

$$\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \boldsymbol{X}_{ij}\log\left(\hat{\boldsymbol{X}}^f_{ij}\right)\right] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \boldsymbol{Y}_{ij}\log\left(f_{ij}(\boldsymbol{Y} - \boldsymbol{e}_i\boldsymbol{e}_j^t)\right)\right],$$

which completes the proof.

B Implementation details

We discuss below an algorithmic approach to find data-driven spectral estimators.

First, we discuss on how to compute data-driven spectral estimators from the expression of risk estimators. For SUKLS in continuous exponential families, and for SURE in the Gaussian case only, eq. (3.7) and (2.10) provide respectively a closed-form solution that can be evaluated in linear time $O(nm)$. On the contrary, the computations of GSURE (beyond the Gaussian case), PURE and PUKLA, given respectively in eq. (2.6), (2.12) and (2.15), cannot be evaluated in reasonable time. They rely respectively on the computation of the divergence $\operatorname{div}\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y})$, $\sum\sum \boldsymbol{Y}_{ij} f_{ij}(\boldsymbol{Y} - \boldsymbol{e}_i\boldsymbol{e}_j^t)$ and $\sum\sum \boldsymbol{Y}_{ij}\log\left(f_{ij}(\boldsymbol{Y} - \boldsymbol{e}_i\boldsymbol{e}_j^t)\right)$. Without further assumptions, such quantities requires $O(n^2m^2)$ operations in general. A standard approach for the computation of the divergence, suggested in [Gir89, RBU08], is to unbiasedly estimate it with Monte-Carlo simulations by sampling the following relation

$$\operatorname{div}\hat{\boldsymbol{\theta}}^f(\boldsymbol{Y}) = \mathbb{E}_{\boldsymbol{\delta}}\left[\operatorname{tr}\left(\boldsymbol{\delta}^t \frac{\partial \hat{\boldsymbol{\theta}}^f(\boldsymbol{Y})}{\partial \boldsymbol{Y}}\boldsymbol{\delta}\right)\right]$$

at random directions $\boldsymbol{\delta} \in \mathbb{R}^{n\times m}$ satisfying $\mathbb{E}[\boldsymbol{\delta}] = 0$, $\mathbb{E}[\boldsymbol{\delta}_i\boldsymbol{\delta}_i] = 1$ and $\mathbb{E}[\boldsymbol{\delta}_i\boldsymbol{\delta}_j] = 0$. Following [Del15], a similar first order approximation can be used for the other two quantities as

$$\sum\sum \boldsymbol{Y}_{ij} f_{ij}(\boldsymbol{Y} - \boldsymbol{e}_i\boldsymbol{e}_j^t) \approx \sum\sum \boldsymbol{Y}_{ij} \left[f_{ij}(\boldsymbol{Y}) - \boldsymbol{\delta}_{i,j} \left(\frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{Y}} \boldsymbol{\delta} \right)_{i,j} \right], \quad \text{and}$$

$$\sum\sum \boldsymbol{Y}_{ij} \log\left(f_{ij}(\boldsymbol{Y} - \boldsymbol{e}_i\boldsymbol{e}_j^t)\right) \approx \sum\sum \boldsymbol{Y}_{ij} \log \left[f_{ij}(\boldsymbol{Y}) - \boldsymbol{\delta}_{i,j} \left(\frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{Y}} \boldsymbol{\delta} \right)_{i,j} \right]$$

where the entries of $\boldsymbol{\delta}$ should be chosen Bernoulli distributed with parameter $p = 0.5$. The advantage of these three approximations is that they can be computed in linear time $O(nm)$ by making use of the results of [LS01, SS03, Ede05, CSLT13, DVP+12] that provide an expression for the directional derivative given by

$$\frac{\partial f(\boldsymbol{Y})}{\partial \boldsymbol{Y}} \boldsymbol{\delta} = \tilde{\boldsymbol{U}}(\mathbf{D} + \boldsymbol{S} + \boldsymbol{A})\tilde{\boldsymbol{V}}^t \tag{B.1}$$

where $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ are the matrices whose columns are $\tilde{\boldsymbol{u}}_k$ and $\tilde{\boldsymbol{v}}_k$, and $\mathbf{D}$, $\boldsymbol{S}$ and $\boldsymbol{A}$ are $n \times m$ matrices defined, for all $1 \le i \le n$ and $1 \le j \le m$, as

$$\mathbf{D}_{i,j} = \bar{\boldsymbol{\delta}}_{i,j} \times \begin{cases} f_i'(\tilde{\sigma}_i) & \text{if} \quad i = j \\ 0 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{S}_{i,j} = \frac{\bar{\boldsymbol{\delta}}_{i,j} + \bar{\boldsymbol{\delta}}_{j,i}}{2} \times \begin{cases} 0 & \text{if} \quad i = j \\ \frac{f_i(\tilde{\sigma}_i) - f_j(\tilde{\sigma}_j)}{\tilde{\sigma}_i - \tilde{\sigma}_j} & \text{otherwise,} \end{cases}$$

$$\boldsymbol{A}_{i,j} = \frac{\bar{\boldsymbol{\delta}}_{i,j} - \bar{\boldsymbol{\delta}}_{j,i}}{2} \times \begin{cases} 0 & \text{if} \quad i = j \\ \frac{f_i(\tilde{\sigma}_i) + f_j(\tilde{\sigma}_j)}{\tilde{\sigma}_i + \tilde{\sigma}_j} & \text{otherwise,} \end{cases}$$

where $\tilde{\sigma}_k$ and $f_k(\tilde{\sigma}_k)$ are extended to 0 for $k > \min(n, m)$ and $\bar{\boldsymbol{\delta}} = \tilde{\boldsymbol{U}}^t \boldsymbol{\delta} \tilde{\boldsymbol{V}} \in \mathbb{R}^{n\times m}$.

References

[ABB00] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences (PNAS)*, 97(18), august 2000.

[AGZ10] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[BD06] M. Bydder and J. Du. Noise reduction in multiple-echo data sets using singular value decomposition. *Magn Reson Imaging*, 24(7):849–56, 2006.

[BMG13] Juan Andres Bazerque, Gonzalo Mateos, and Georgios B. Giannakis. *Inference of Poisson count processes using low-rank tensor data*, pages 5989–5993. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 10 2013.

[BN12] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Analysis*, 111:120–135, 2012.

[Bro86] L. D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics, 1986.

[BS06] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382 – 1408, 2006.

[BS10] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[CR09] D. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[CSLT13] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.*, 61(19):4643–4657, 2013.

[CTT14] Y. Choi, J. Taylor, and R. Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Preprint arXiv:1405.7511*, 2014.

[CX16] Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Trans. Signal Processing*, 64(6):1609–1620, 2016.

[Del15] C. Deledalle. Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Preprint, arXiv:1512.08191*, 2015.

[DG14] D. Donoho and M. Gavish. Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.*, 42(6):2413–2440, 12 2014.

[DS07] R. B. Dozier and J. W. Silverstein. On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *J. Multivariate Anal.*, 98(4):678–694, 2007.

[DVP+12] C.-A. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal. Risk estimation for matrix recovery with spectral regularization. In *arXiv:1205.1482*, 2012. Presented at ICML'2012 workshop on Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing, Edinburgh, United Kingdom, 2012.

[Ede05] A. Edelman. Matrix jacobians with wedge products. *MIT Handout for 18.325*, 2005.

[Efr04] Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, pages 99–467, 2004.

[Eld09] Y. C. Eldar. Generalized sure for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

[EM72] B. Efron and C. Morris. Empirical Bayes on Vector Observations: An Extension of Stein's Method. *Biometrika*, 59(2):335–347, 1972.

[EM76] B. Efron and C. Morris. Multivariate empirical bayes and estimation of covariance matrices. *Ann. Statist.*, 4(1):22–32, 01 1976.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[GD14a] M. Gavish and D.L. Donoho. Optimal shrinkage of singular values. *Preprint arXiv:1405.7511*, 2014.

[GD14b] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is \(4/\sqrt {3}\). *IEEE Trans. Information Theory*, 60(8):5040–5053, 2014.

[Gir89] A Girard. A fast monte-carlo cross-validation procedure for large least squares problems with noisy data. *Numerische Mathematik*, 56(1):1–23, 1989.

[Hal87] Peter Hall. On Kullback-Leibler loss and density estimation. *Ann. Statist.*, 15(4):1491–1519, 1987.

[HJ91] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, New York, Melbourne, 1991. Suite de : Matrix analysis. 1985.

[HL06] Jan Hannig and Thomas C. M. Lee. On Poisson signal estimation under Kullback-Leibler discrepancy and squared risk. *J. Statist. Plann. Inference*, 136(3):882–908, 2006.

[Hud78] H. M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.*, 6(3):473–484, 05 1978.

[Jol02] I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

[JS15] J. Josse and S. Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, pages 1–10, 2015.

[Laf15] Jean Lafond. Low rank matrix completion with exponential family noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1224–1243, 2015.

[LBH+12] F. Lam, S. D. Babacan, J. P. Haldar, N Schuff, and Z.-P. Liang. Denoising diffusion-weighted MR magnitude image sequences using low rank and edge constraints. In *ISBI*, pages 1401–1404. IEEE, 2012.

[LS01] A.S. Lewis and H.S. Sendov. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis on Matrix Analysis and Applications*, 23:368–386, 2001.

[LW12] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.*, 40(2):1024–1060, 04 2012.

[Nad14] R. R. Nadakuditi. OptShrink: an algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inform. Theory*, 60(5):3002–3018, 2014.

[NPDL11] H. M. Nguyen, X. Peng, M. N. Do, and Z.-P. Liang. Spatiotemporal denoising of mr spectroscopic imaging data by low-rank approximations. In *ISBI*, pages 857–860. IEEE, 2011.

[PL03] L. Pastur and A. Lejay. Matrices aléatoires: statistique asymptotique des valeurs propres. In *Séminaire de Probabilités, XXXVI*, volume 1801 of *Lecture Notes in Math.*, pages 135–164. Springer, Berlin, 2003.

[RBU08] S. Ramani, T. Blu, and M. Unser. Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms. *IEEE Trans. on Image Processing*, 17(9):1540–1554, 2008.

[RS07] M. Raphan and E. P. Simoncelli. Learning to be Bayesian without supervision. In *Advances in Neural Inf. Process. Syst. (NIPS)*, volume 19, pages 1145–1152. MIT Press, 2007.

[SH05] Haipeng Shen and Jianhua Z. Huang. Analysis of call centre arrival data using singular value decomposition: Research articles. *Appl. Stoch. Model. Bus. Ind.*, 21(3):251–263, May 2005.

[SN13] A. A. Shabalin and A. B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.*, 118:67–76, 2013.

[SS03] D. Sun and J. Sun. Nonsmooth matrix valued functions defined by singular values. Technical report, Department of Decision Sciences, National University of Singapore, 2003.

[Ste81] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[UHZB16] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[WDB01] M.E. Wall, P.A. Dyck, and T.S. Brettin. Svdman-singular value decomposition analysis of microarray data. *Bioinformatics*, 17(6):566–568, 2001.

[Yan94] Takemi Yanagimoto. The Kullback-Leibler risk of the Stein estimator and the conditional MLE. *Ann. Inst. Statist. Math.*, 46(1):29–41, 1994.