Statistics and Probability Letters 135 (2018) 76–82

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On Stein's unbiased risk estimate for reduced rank estimators

Niels Richard Hansen

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5., 2100 Copenhagen Ø, Denmark

ARTICLE INFO

Article history:
Received 14 September 2017
Received in revised form 3 November 2017
Accepted 16 November 2017
Available online 6 December 2017

MSC:
62C12
62H12

Keywords:
Degrees of freedom
Reduced-rank regression
Singular value thresholding
Stein's lemma
SURE

ABSTRACT

Stein's unbiased risk estimate (SURE) is considered for matrix valued observables with low rank means. It is shown that SURE is applicable to a class of spectral function estimators including the reduced rank estimator.

1. Introduction

Low rank approximations of a matrix are useful for dimension reduction (PCA), reduced-rank regression and matrix completion, see (Mukherjee et al., 2015) and Chapter 7 in Hastie et al. (2015), and compression and noise reduction as treated by Candès et al. (2013). The computation of a low rank approximation is typically based on the singular value decomposition of the matrix. If $\mathbf{Y}$ is a $p \times q$ matrix with $p \geq q$, the singular value decomposition can be written as

$$\mathbf{Y} = \sum_{k=1}^{q} d_k u_k v_k^T$$

with $d_k \geq 0$, $u_k \in \mathbb{R}^p$ and $v_k \in \mathbb{R}^q$. The u_k-vectors as well as the v_k-vectors are orthonormal. When all the singular values are unique and ordered as $d_1 > d_2 > \cdots > d_q \geq 0$, the Eckart–Young–Mirsky theorem states that the matrix of rank at most r that approximates $\mathbf{Y}$ best in terms of the Frobenius norm as well as the spectral norm is given by

$$\hat{\mu}(r) = \sum_{k=1}^{r} d_k u_k v_k^T$$

for $r \in \{1, \ldots, q\}$. The hard threshold approximation given by

$$\overline{\mu}(\lambda) = \sum_{k=1}^{q} d_k 1(d_k \geq \lambda) u_k v_k^T$$

E-mail address: Niels.R.Hansen@math.ku.dk.
URL: http://www.math.ku.dk/∼richard.

https://doi.org/10.1016/j.spl.2017.11.006
0167-7152/

for $\lambda \geq 0$ yields the same sequence of approximations but parametrized differently. Other approximations may be obtained by shrinkage of the singular values toward zero, one example being *soft thresholding*

$$\tilde{\mu}(\lambda) = \sum_{k=1}^{q} (d_k - \lambda)_+ u_k v_k^T,$$

which was studied in detail by Candès et al. (2013).

A natural question to ask is how the parameter r or λ above should be chosen. A statistical answer is given by providing a sampling model of $\mathbf{Y}$ and regarding the low rank approximations as estimators of an unknown mean. If it is possible to estimate the risk of those estimators, r or λ can be chosen by minimizing the risk estimate. Candès et al. (2013) demonstrated that soft thresholding is a Lipschitz continuous estimator, and they used this to show that the risk can then be estimated unbiasedly via SURE. However, other estimators like the hard thresholding estimator are discontinuous, and it is of interest to understand precisely if and when SURE can be applied beyond soft thresholding.

We will throughout work with the model where $\mathbf{Y} = (Y_{ij})_{i,j}$ has independent entries with

$$Y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$$

and $\mu = (\mu_{ij})_{i,j}$ is of low rank. The estimators $\hat{\mu}(r)$, $\overline{\mu}(\lambda)$ and $\tilde{\mu}(\lambda)$ are examples from the more general class of *spectral function* estimators

$$\hat{\mu} = \sum_{k=1}^{q} f_k(d_k) u_k v_k^T, \quad \lambda_1 > \lambda_2 > \cdots > \lambda_q \geq 0$$

for some spectral functions $f_k : [0, \infty) \to [0, \infty)$. The hard thresholding estimator has $f_k(d) = d1(d \geq \lambda)$, the soft thresholding estimator has $f_k(d) = (d - \lambda)_+$, and the estimator with $f_k(d) = d1(k \leq r)$ – which gives the best rank r approximation – will be referred to as the reduced rank estimator.

In the framework of spectral function estimators, Candès et al. (2013) derived an explicit formula (formula (9) in their paper) for the divergence of $\hat{\mu}$ as a function of $\mathbf{Y}$ when $\mathbf{Y}$ has distinct singular values and f_k is differentiable in a neighborhood of d_k. This divergence is required for the computation of SURE, and it is therefore important for applications. They also demonstrated in detail (Lemma III.3) via Stein's lemma the unbiasedness of SURE in the special case of soft thresholding, but they did not demonstrate if their divergence formula can be applied to obtain unbiased risk estimation for other estimators.

Mukherjee et al. (2015) derived similar formulas for the divergence—apparently unaware of the paper by Candès et al. (2013). One difference is that Mukherjee et al. (2015) focused on the regression setup, where the columns of the observation matrix are projected onto a fixed subspace before it is subjected to a low rank approximation.

Neither Candès et al. (2013) nor Mukherjee et al. (2015) provided conditions for general spectral function estimators that ensure that Stein's lemma applies. Mukherjee et al. (2015) indicated on page 460 that the mere existence of the partial derivatives (Lebesgue) almost everywhere is sufficient for Stein's lemma, which is not the case. Candès et al. (2013) stated a version of Stein's lemma as their Proposition III.1, which correctly assumes weak differentiability of the estimator, but they did not demonstrate weak differentiability for other estimators than soft thresholding.

The purpose of this paper is to provide conditions ensuring that a spectral function estimator is, indeed, weakly differentiable so that Stein's lemma applies. In particular, we show that the reduced rank estimator is weakly differentiable so that the SURE formula as given by Candès et al. (2013) or Mukherjee et al. (2015) results in unbiased estimation of the risk. To illustrate the relevance of such sufficient conditions, we show by a small simulation that the SURE formula for singular value hard thresholding does not give unbiased estimation of the risk.

2. Degrees of freedom and Stein's lemma

In this section we state a version of Stein's lemma, which can be applied directly to a class of spectral function estimators including the reduced rank estimator. It is formulated for n-dimensional Gaussian vectors and applies to the matrix valued observations and estimators above by taking $n = pq$.

Let $y \sim N(\mu, \sigma^2 I)$ be an n-dimensional Gaussian random variable and $\hat{\mu}$ an estimator of μ with finite second moment; $E\|\hat{\mu}\|_2^2 < \infty$. Define, in addition, the effective degrees of freedom as

$$\mathrm{df} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i), \tag{1}$$

and let $\nabla \cdot \hat{\mu} = \sum_{i=1}^{n} \partial_i \hat{\mu}_i$ denote the divergence of $\hat{\mu}$ whenever it is Lebesgue almost everywhere differentiable. The following lemma gives a sufficient condition for $\nabla \cdot \hat{\mu}$ to be an unbiased estimate of the degrees of freedom. The proof of the lemma is given in the supplementary material.

Lemma 1. *Let $E \subseteq \mathbb{R}^n$ be a closed set such that $\hat{\mu} : E^c \to \mathbb{R}^n$ is continuously differentiable. If $\mathcal{H}^{n-1}(E) = 0$ and either $\nabla \cdot \hat{\mu} \geq 0$ Lebesgue almost everywhere or*

$$\sum_{i=1}^{n} E\left|\partial_i \hat{\mu}_i\right| < \infty \tag{2}$$

then

$$\mathrm{df} = E\left(\nabla \cdot \hat{\mu}\right). \tag{3}$$

When (3) holds the risk of $\hat{\mu}$ can be estimated by SURE, which is treated in greater detail in the supplementary material. It is noteworthy that some condition like $\mathcal{H}^{n-1}(E) = 0$ is necessary for (3) to hold. Results by e.g. Tibshirani (2015) and Mikkelsen and Hansen (in press) show that $\nabla \cdot \hat{\mu}$ is generally biased as an estimate of the degrees of freedom if the estimator $\hat{\mu}$ is discontinuous as a function of the data y—even if it is Lebesgue almost everywhere differentiable.

3. Spectral function and reduced rank estimators

As derived in the supplementary material, the risk of an estimator $\hat{\mu}$ using the Frobenius norm as loss function is

$$E\|\hat{\mu} - \mu\|_F^2 = E\|\mathbf{Y} - \hat{\mu}\|_F^2 - \sigma^2 pq + 2\sigma^2 \mathrm{df}.$$

When $E\left(\nabla \cdot \hat{\mu}\right) = \mathrm{df}$, and in particular when Lemma 1 applies, formula (3) in the supplementary material gives that

$$\mathrm{SURE} = \|\mathbf{Y} - \hat{\mu}\|_F^2 - \sigma^2 pq + 2\sigma^2 \nabla \cdot \hat{\mu}$$

is an unbiased estimate of the risk.

Theorem 1. *Consider a spectral function estimator*

$$\hat{\mu} = \sum_{k=1}^{q} f_k(d_k) u_k v_k^T$$

with finite second moment and with spectral functions fulfilling that $f_1, \ldots, f_{q-1}$ are continuously differentiable on $(0, \infty)$, f_q is continuously differentiable on $[0, \infty)$ with $f_q(0) = f_q'(0) = 0$, $f_k \geq f_l$ for $k < l$ and $f_k' \geq 0$. Then SURE *is an unbiased estimate of the risk, and*

$$\nabla \cdot \hat{\mu} = (p - q) \sum_{k=1}^{q} \frac{f_k(d_k)}{d_k} + \sum_{k=1}^{q} f_k'(d_k) + 2 \sum_{\substack{k,l=1 \\ k \neq l}}^{q} \frac{d_k f_k(d_k)}{d_k^2 - d_l^2} \tag{4}$$

whenever $\mathbf{Y}$ has q distinct, positive singular values.

Proof. We show that the conditions in Lemma 1 are fulfilled. We first show that the spectral function estimator $\hat{\mu}$ is continuously differentiable on the set of matrices with no two identical singular values.

Letting

$$E = \{\mathbf{Y} \in M(p, q) \mid \mathbf{Y} \text{ has at least two identical singular values}\}$$

it was shown by Mukherjee et al. (2015) that E is a proper subvariety, and it is, in particular, a closed set.

For $\mathbf{Y} \in E^c$ the singular value decomposition can be written as

$$\mathbf{Y} = \sum_{k=1}^{q} d_k u_k v_k^T$$

with $d_1 > d_2 > \cdots > d_q \geq 0$, and we can regard $d_k = d_k(\mathbf{Y}) \in \mathbb{R}$, $v_k = v_k(\mathbf{Y}) \in \mathbf{R}^q$ and $u_k = u_k(\mathbf{Y}) \in \mathbf{R}^p$ as functions of $\mathbf{Y}$. An application of Theorem 5.3 in Serre (2010) to $\mathbf{Y}^T\mathbf{Y}$ shows that the singular values $d_1, \ldots, d_q$ as well as $v_1, \ldots, v_q$ are C^∞ in a neighborhood $\mathcal{O} \subseteq E^c$ of $\mathbf{Y}$. It can be assumed that $d_1 > d_2 > \cdots > d_q \geq 0$ in that neighborhood. This gives that

$$u_k = \frac{\mathbf{Y} v_k}{d_k}$$

is also C^∞ in $\mathcal{O}$ for $k = 1, \ldots, q-1$. If $d_q > 0$ this is true for $k = q$ as well. In this case, it follows by the assumptions on f_k that $f_k(d_k) u_k v_k^T$ for $k = 1, \ldots, q$ are continuously differentiable in $\mathcal{O}$. If $d_q = 0$, u_q is not unique, but the assumptions on f_q in 0 ensure that $f_q(d_q) u_q v_q^T$ is still continuously differentiable in $\mathcal{O}$ (with derivative $\mathbf{0}$ in $\mathbf{Y}$). These arguments show that $\hat{\mu}$ is continuously differentiable in the neighborhood $\mathcal{O}$ around $\mathbf{Y}$ for any $\mathbf{Y} \in E^c$, and $\hat{\mu}$ is thus continuously differentiable on E^c.

In the next step we establish that $\mathcal{H}^{pq-1}(E) = 0$. Let $V_r(m)$ denote the Stiefel manifold of r-tuples of m-dimensional orthonormal vectors in $\mathbb{R}^m$, and introduce $h : V_{q-2}(q) \times V_q(p) \times [0,\infty)^{q-1} \to M(p,q)$ by

$$h((v_k)_k, (u_k)_k, d) = \sum_{k=1}^{q-2} d_k u_k v_k^T + d_{q-1}(u_{q-1}v_{q-1}^T + u_q v_q^T),$$

where the unit vectors v_{q-1} and v_q are chosen (by h and depending on $(v_k)_k$) orthogonal to $v_1, \ldots, v_{q-2}$. The set E is in the image of h because when two singular values are identical the singular value decomposition is not unique, and it is possible to choose an orthogonal transformation such that v_{q-1}^T and v_q^T in the singular value decomposition are those chosen by the map h. The Stiefel manifold $V_r(m)$ has dimension

$$\dim(V_r(m)) = rm - \frac{1}{2}r(r+1)$$

as a differentiable manifold. It follows that E locally is contained in the image under a Lipschitz map of a set of dimension

$$q(q-2) - \frac{1}{2}(q-2)(q-1) + pq - \frac{1}{2}q(q+1) + q - 1 = pq - 2.$$

It follows from Theorem 2.4.1 in Evans and Gariepy (1992) that E has Hausdorff dimension at most $pq-2$, whence $\mathcal{H}^{pq-1}(E) = 0$.

In the last step we verify that $\nabla \cdot \hat{\mu} \geq 0$ on E^c. The divergence formula (4) was shown by Candès et al. (2013) and is given as (9) in their paper. It applies a priori only when $d_q > 0$, but the formula extends by continuity to $d_q = 0$ due to the assumption that $f_q(0) = f_q'(0) = 0$. As f_k is positive and f_k' is also assumed positive, the first two terms in (4) are positive. For the third term we rearrange the double sum as

$$\sum_{\substack{k,l=1 \\ k\neq l}}^{q} \frac{d_k f_k(d_k)}{d_k^2 - d_l^2} = \sum_{k=1}^{q-1}\sum_{l=k+1}^{q} \frac{d_k f_k(d_k) - d_l f_l(d_l)}{d_k^2 - d_l^2}. \tag{5}$$

Using that $d_k > d_l$ for $k < l$, and that this implies that

$$f_k(d_k) \geq f_k(d_l) \geq f_l(d_l),$$

we have that for each term in this double sum

$$\frac{d_k f_k(d_k) - d_l f_l(d_l)}{d_k^2 - d_l^2} \geq \frac{(d_k - d_l) f_l(d_l)}{d_k^2 - d_l^2} \geq 0.$$

This completes the proof. □

It is of interest to know if a spectral function estimator has finite second moment. To this end observe that

$$\|\hat{\mu}\|_F^2 = \sum_{k=1}^{q} f_k(d_k)^2.$$

Hence if $f_k(d) \leq C_k d$ for some constant C_k (e.g. f_k is shrinking the singular values as is the case for reduced rank, and hard and soft thresholding), then

$$\|\hat{\mu}\|_F^2 \leq \max_k\{C_k^2\} \sum_{k=1}^{q} d_k^2 = C\|\mathbf{Y}\|_F^2,$$

and $\hat{\mu}$ has finite second moment.

The reduced rank estimator is an important example of a spectral function estimator for which the conditions in Theorem 1 are fulfilled. We state this as the following corollary, which also gives an alternative formula for the divergence in this particular case.

Corollary 1. *For the reduced rank estimator* $\hat{\mu}(r)$, SURE *is an unbiased estimate of the risk, and*

$$\nabla \cdot \hat{\mu}(r) = pr + \sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2} \tag{6}$$

whenever $\mathbf{Y}$ *has q distinct singular values.*

Proof. The spectral functions $f_k(d) = d1(k \leq r)$ for a fixed $r < q$ fulfill all the conditions for Theorem 1. Observe also that $f_k(d_k)/d_k = 1(k \leq r)$, that $f_k'(d_k) = 1(k \leq r)$ and that $d_k f_k(d_k) = d_k^2 1(k \leq r)$. Using (5) it then follows from (4) that

$$\begin{aligned}
\nabla \cdot \hat{\mu} &= (p-q)r + r + 2\sum_{k=1}^{r}\sum_{l=k+1}^{r} \underbrace{\frac{d_k^2 - d_l^2}{d_k^2 - d_l^2}}_{=1} + 2\sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{d_k^2}{d_k^2 - d_l^2} \\
&= pr - qr + r + r^2 - r + \sum_{k=1}^{r}\sum_{l=r+1}^{q}\left(1 + \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2}\right) \\
&= pr - qr + r^2 + r(q-r) + \sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2} \\
&= pr + \sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2}.
\end{aligned}$$

Finally, $\hat{\mu}(q) = \mathbf{Y}$ with $\nabla \cdot \hat{\mu}(q) = pq$. □

It should be noted that (6) is identical to the formula in Theorem 3 in Mukherjee et al. (2015). The general formula (4) is identical to (9) from Candès et al. (2013), while Mukherjee et al. (2015) showed an equivalent general formula for the divergence of spectral function estimators, which is given in their Theorem 4.

It is possible that the monotonicity requirements on the spectral functions above can be relaxed, and that (2) can be verified instead of the positivity on the divergence. It is also possible that the requirement that $f_q'(0) = 0$ can be relaxed. Neither of these potential generalizations will be pursued in this paper.

It is finally noted that $f_k(d) = d1(d \geq \lambda)$ is not continuous, and Theorem 1 thus does not apply to the singular value hard thresholding estimator $\overline{\mu}$. In the simulation study below it is demonstrated that $\nabla \cdot \overline{\mu}$ is, in fact, also biased as an estimate of the degrees of freedom.

4. Simulation

This section presents the results from a simulation that illustrates the unbiasedness of SURE for reduced rank estimation and singular value soft thresholding, whereas (4) is shown to be a biased estimate of the degrees of freedom for singular value hard thresholding.

We made $B = 5000$ simulations with $p = q = 21$ and $Y_{ij}^b \sim \mathcal{N}(0, 1)$ for $b = 1, \ldots, B$. With

$$\mathbf{Y}^b = \sum_{k=1}^{q} d_k^b u_k^b (v_k^b)^T, \qquad d_1^b \geq d_2^b \geq \cdots \geq d_q^b \geq 0$$

the singular value decomposition of the bth matrix $\mathbf{Y}^b = (Y_{ij}^b)_{i,j}$ we computed the three estimators

$$\begin{aligned}
\hat{\mu}_b(r) &= \sum_{k=1}^{r} d_k^b u_k^b (v_k^b)^T && \text{(reduced rank)} \\
\overline{\mu}_b(\lambda) &= \sum_{k=1}^{q} d_k^b 1(d_k^b \geq \lambda) u_k^b (v_k^b)^T && \text{(hard thresholding)} \\
\tilde{\mu}_b(\lambda) &= \sum_{k=1}^{q} (d_k^b - \lambda)_+ u_k^b (v_k^b)^T. && \text{(soft thresholding)}
\end{aligned}$$

By the definition of the degrees of freedom, (1), the estimate

$$\hat{\mathrm{df}}_0(r) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{tr}(\hat{\mu}_b(r)^T(\mathbf{Y}^b - \mu))$$

is an unbiased estimate of df for the reduced rank estimator, and similar estimates of df were computed for the other two estimators. Note that such estimates based on the covariance definition are of no use in real applications as they rely on knowledge of the true mean. In this simulation the true mean was $\mu = 0$.

Estimates based on the divergence were computed for each of the three estimators as follows

$$\hat{\mathrm{df}}(r) = pr + \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{r}\sum_{l=r+1}^{q} \frac{(d_k^b)^2 + (d_l^b)^2}{(d_k^b)^2 - (d_l^b)^2}$$

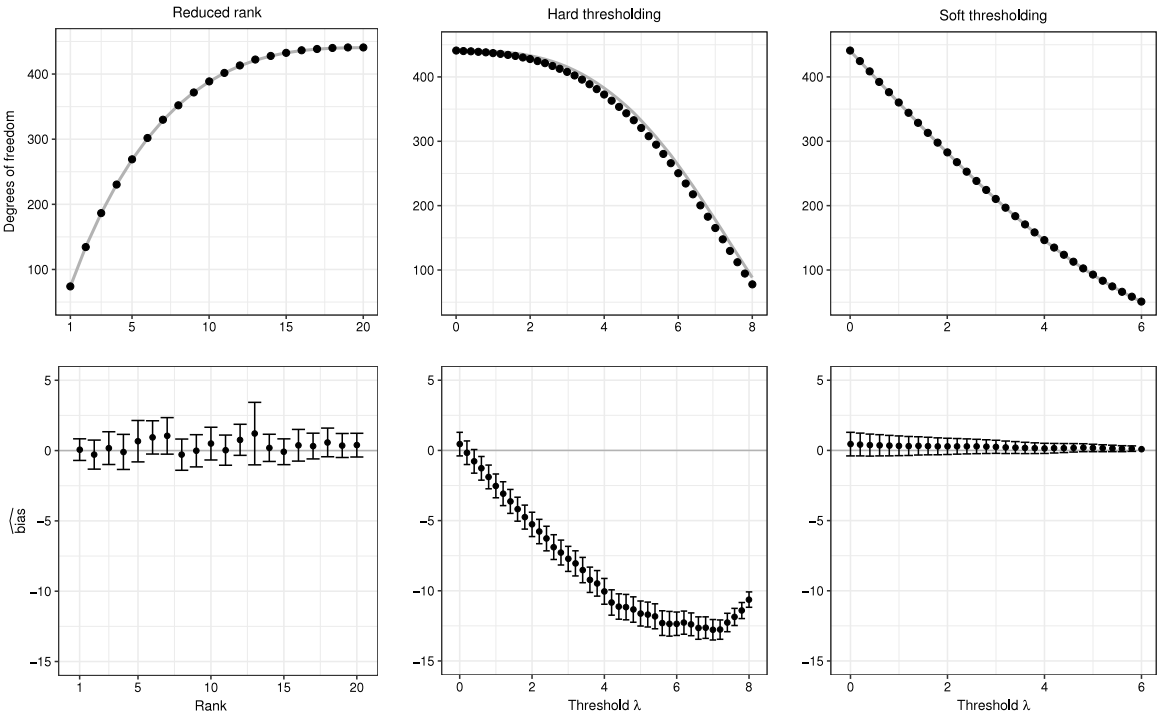

Fig. 1. The top row shows average degrees of freedom as estimated by the divergence formulas (points) and by the covariance definition (1) (gray line). The bottom row shows the bias of the divergence estimate for the three different estimators. For the bias, the 95% confidence intervals shown quantify the simulation uncertainty.

$$\overline{\mathrm{df}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\left((p-q+1)\sum_{k=1}^{q}1(d_k^b \geq \lambda) + 2\sum_{\substack{k,l=1\\k\neq l}}^{q}\frac{(d_k^b)^2\,1(d_k^b \geq \lambda)}{(d_k^b)^2-(d_l^b)^2}\right)$$

$$\tilde{\mathrm{df}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\left((p-q)\sum_{k=1}^{q}\left(1-\frac{\lambda}{d_k^b}\right)_+ + \sum_{k=1}^{q}1(d_k^b \geq \lambda) + 2\sum_{\substack{k,l=1\\k\neq l}}^{q}\frac{d_k^b(d_k^b-\lambda)_+}{(d_k^b)^2-(d_l^b)^2}\right).$$

Note that with

$$r(\lambda) = \sum_{k=1}^{q}1(d_k^b \geq \lambda)$$

it holds that $\hat{\mathrm{df}}(r(\lambda)) = \overline{\mathrm{df}}(\lambda)$.

The bias estimate is defined as

$$\widehat{\mathrm{bias}}(r) = \hat{\mathrm{df}}(r) - \hat{\mathrm{df}}_0(r)$$

for the reduced rank estimator and likewise for the other two estimators. For reduced rank and soft thresholding, the bias is zero by the theoretical results.

Fig. 1 shows the results of the simulation. It shows clearly that for singular value hard thresholding the bias is non-zero, though it is relatively small. The results for reduced rank and soft thresholding are completely in concordance with the theoretical results that the bias is 0, and with the simulation results presented by Candès et al. (2013) and Mukherjee et al. (2015).

5. Final comments

There is no claim of originality in terms of the estimators considered or the formulas presented for estimating degrees of freedom and computing SURE. These can be found in the papers by Candès et al. (2013) and by Mukherjee et al. (2015). However, neither of these two papers – nor other papers that the author is aware of – gives a complete proof of the fact that

Stein's lemma does apply to the reduced rank estimator even though this estimator is not Lipschitz continuous and have discontinuities. The purpose of this paper was to give this proof.

The proof given relies on Lemma 1 that explicitly allows for estimators with discontinuities as long as these constitute a sufficiently small exception set. It is not enough for the exception set E to have Lebesgue measure zero. In the proof of Theorem 1 the set E is easily seen to be a proper subvariety, which thus have codimension 1 and Lebesgue measure zero as argued by Mukherjee et al. (2015). However, the stronger condition that $\mathcal{H}^{pq-1}(E) = 0$ in terms of the Hausdorff measure is needed, and this was established by effectively showing that E has codimension 2. That argument is closely related to the long established fact that the set of matrices with repeated eigenvalues in the set of real symmetric matrices has codimension 2. A result credited to Neumann and Wigner, see p. 36 in Lax (2007). It appears somewhat complicated to determine the codimension of E as a subvariety, see Dana and Ikramov (2006) for the case of symmetric matrices, and the proof of Theorem 1 proceeded by counting the free parameters in the singular value decomposition instead.

We may note that the reduced rank estimator, $\hat{\mu}(r)$, and the hard thresholding estimator, $\overline{\mu}(\lambda)$, provide the exact same sequence of estimators for a given observation when viewed as functions of r and λ, respectively. Yet, for a fixed r we can by SURE obtain an unbiased risk estimate for $\hat{\mu}(r)$, while for fixed λ, the SURE formula based on the divergence estimate of degrees of freedom does not give an unbiased risk estimate for $\overline{\mu}(r)$. This is understandable as the mapping between the two sequences of estimators is data dependent, but it highlights that the parametrization matters when estimators are assessed via their frequentistic risk. When tuning parameters are selected by minimizing a risk estimate, this leads to the slightly peculiar phenomenon that different parametrizations can lead to different choices of tuning parameters. Or as is the case here, that one parametrization provides unbiased risk estimates, while another provides biased risk estimates, even though the risk estimates are identical as sets.

Acknowledgment

This work was supported by a research grant (13358) from VILLUM FONDEN.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2017.11.006.

References

Candès, E.J., Sing-Long, C.A., Trzasko, J.D., 2013. Unbiased risk estimates for singular value thresholding and spectral estimators. IEEE Trans. Signal Process. 61, 4643–4657.

Dana, M., Ikramov, K., 2006. On the codimension of the variety of symmetric matrices with multiple eigenvalues. J. Math. Sci. 137, 4780–4786.

Evans, L.C., Gariepy, R.F., 1992. Measure Theory and Fine Properties of Functions. In: Studies in Advanced Mathematics, CRC Press, Boca Raton, FL.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity. In: Monographs on Statistics and Applied Probability, vol. 143, CRC Press, Boca Raton, FL.

Lax, P.D., 2007. Linear Algebra and its Applications, second ed.. In: Pure and Applied Mathematics, Wiley-Interscience, Hoboken, NJ.

Mikkelsen, F.R., Hansen, N.R., 2018. Degrees of freedom for piecewise Lipschitz estimators. Ann. Inst. H. Poincaré Probab. Statist. (in press), arXiv:1601.03524v3.

Mukherjee, A., Chen, K., Wang, N., Zhu, J., 2015. On the degrees of freedom of reduced-rank estimators in multivariate regression. Biometrika 102, 457–477.

Serre, D., 2010. Matrices, second ed.. In: Graduate Texts in Mathematics, vol. 216, Springer, New York.

Tibshirani, R.J., 2015. Degrees of freedom and model search. Statist. Sinica 25, 1265–1296.

Supplementary Material: On Stein's unbiased risk estimate for reduced rank estimators

Niels Richard Hansen

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5. 2100 Copenhagen Ø, Denmark

Abstract

This document contains some background on risk estimation, Stein's lemma and the SURE formula, which are used in the paper, and it contains the proof of Lemma 1.

1. Squared error risk and Stein's lemma

We consider $\mathbb{R}^n$ equipped with the standard inner product and 2-norm

$$||y||^2 = y^T y = \sum_{i=1}^n y_i^2.$$

For $y, \mu \in \mathbb{R}^n$ and $\hat{\mu} = \hat{\mu}(y)$ it holds that

$$||\hat{\mu} - \mu||^2 = ||y - \hat{\mu}||^2 + 2\hat{\mu}^T(y - \mu) - 2y^T(y - \mu) + ||y - \mu||^2.$$

Now, if $y \sim \mathcal{N}(\mu, \sigma^2 I)$ and we compute the mean on both sides of the identity above we get

$$\begin{aligned} E||\hat{\mu} - \mu||^2 &= E||y - \hat{\mu}||^2 + 2\sum_{i=1}^n \operatorname{cov}(\hat{\mu}_i, y_i) - 2\sum_{i=1}^n \underbrace{\operatorname{var}(y_i, y_i)}_{=\sigma^2} + \sigma^2 n \\ &= E||y - \hat{\mu}||^2 - \sigma^2 n + 2\sigma^2 \mathrm{df} \end{aligned}$$

where the effective degrees of freedom is defined as

$$\mathrm{df} = \frac{1}{\sigma^2}\sum_{i=1}^n \operatorname{cov}(\hat{\mu}_i, y_i).$$

Thus with the squared error loss function, the risk of the estimator $\hat{\mu}$ fulfills the identity

$$\text{Risk} = E||\hat{\mu} - \mu||^2 = E||y - \hat{\mu}||^2 - \sigma^2 n + 2\sigma^2 \mathrm{df}.$$

The key result for obtaining an estimate of df, and thus an estimate of the risk, is Stein's lemma or Stein's identity (Lemma 1 or Lemma 2 in Stein (1981)). To formulate a version of this identity introduce $C_c^\infty(\mathbb{R}^n)$ as the set of smooth functions with compact support. A function $f \in \mathcal{L}^1_{\mathrm{loc}}(\mathbb{R}^n)$ is said to be weakly differentiable if there for each $i = 1, \ldots, n$ exists a function denoted $\partial_i f \in \mathcal{L}^1_{\mathrm{loc}}(\mathbb{R}^n)$ such that for all $\psi \in C_c^\infty(\mathbb{R}^n)$

$$\int \partial_i f(y)\psi(y)\mathrm{d}y = -\int f(y)\partial_i \psi(y)\mathrm{d}y. \tag{1}$$

The set of weakly differentiable functions is denoted $W^{1,1}_{\mathrm{loc}}(\mathbb{R}^n)$ and is known as a Sobolev space.

Email address: Niels.R.Hansen@math.ku.dk (Niels Richard Hansen)
URL: www.math.ku.dk/~richard (Niels Richard Hansen)

Lemma (Stein's lemma)**.** *If* $\hat{\mu} \in W^{1,1}_{\text{loc}}(\mathbb{R}^n)$ *and* $\sum_{i=1}^n E|\partial_i \hat{\mu}_i| < \infty$ *then*

$$\text{df} = E(\nabla \cdot \hat{\mu}). \tag{2}$$

Stein's lemma and its proof can be found in many variations. The version in Stein (1981) requires that $\hat{\mu}$ is almost differentiable, but as pointed out by e.g. Johnstone (1998), Fourdrinier and Wells (2012), and Candès et al. (2013), this is effectively the same as assuming that the estimator is weakly differentiable. See also Definition 2 and Lemma 5 in Tibshirani (2015). Stein's lemma basically follows directly from the defining integration-by-parts formula (1) with the Gaussian density in place of ψ. An approximation argument is required since the Gaussian density doesn't have compact support. The condition

$$\sum_{i=1}^n E|\partial_i \hat{\mu}_i| < \infty$$

allows for such an approximation via dominated convergence. In the proof of Lemma 1 below we carry out this argument under the assumption $\nabla \cdot \hat{\mu} \geq 0$, which allows for the use of monotone convergence instead.

2. SURE

When Stein's lemma applies,

$$\text{SURE} = ||y - \hat{\mu}||^2 - \sigma^2 n + 2\sigma^2 \nabla \cdot \hat{\mu}$$

is an unbiased estimate of Risk. It should be noted that weak differentiability as well as the integration condition must be checked to ensure that $\nabla \cdot \hat{\mu}$ is unbiased as an estimate of df. It's, in particular, not sufficient for $\hat{\mu}$ to be differentiable Lebesgue almost everywhere.

In the paper, SURE is applied in the setup with matrix valued observables. To see how this works, note that the Frobenius norm on the set of $p \times q$ matrices is simply the standard 2-norm if the matrices are regarded as elements in $\mathbb{R}^{pq}$. Thus it follows directly from the formula above that with the Frobenius norm as loss function,

$$\text{SURE} = ||\mathbf{Y} - \hat{\mu}||_F^2 - \sigma^2 pq + 2\sigma^2 \nabla \cdot \hat{\mu}. \tag{3}$$

We may note that $\hat{\mu}$ is then matrix valued and

$$\nabla \cdot \hat{\mu} = \sum_{k,l} \partial_{kl} \hat{\mu}_{kl}.$$

3. Proof of Lemma 1

We first show that the condition $\mathcal{H}^{n-1}(E) = 0$ implies that $\hat{\mu} \in W^{1,1}_{\text{loc}}(\mathbb{R}^n)$. This is the case if for all i and Lebesgue almost all $(y_1, \ldots, y_{n-1}) \in \mathbb{R}^{n-1}$ the function

$$t \mapsto \hat{\mu}(y_1, \ldots, y_{i-1}, t, y_i, \ldots, y_{n-1}) \tag{4}$$

is absolutely continuous on compact intervals, see Theorem 4.9.2.2 in Evans and Gariepy (1992). When $\mathcal{H}^{n-1}(E) = 0$ the projection maps

$$(y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_n) \mapsto (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$$

for $i = 1, \ldots, n$ map E onto Lebesgue null sets by Corollary 2.4.1.1 in Evans and Gariepy (1992). Thus when $\hat{\mu}$ is continuously differentiable on E^c, (4) is continuously differentiable, and thus absolutely continuous, Lebesgue almost everywhere.

In the second part of the proof we show that $\nabla \cdot \hat{\mu} \geq 0$ implies (2) for $\hat{\mu} \in W^{1,1}_{\text{loc}}(\mathbb{R}^n)$. To this end we first note that

$$\sum_{i=1}^{n} \int \hat{\mu}_i(y) \partial_i \psi(y) \mathrm{d}y = -\int \nabla \cdot \hat{\mu}(y)\ \psi(y) \mathrm{d}y$$

for all $\psi \in C_c^\infty(\mathbb{R}^n)$.

Let φ denote the density for the $\mathcal{N}(\mu, \sigma^2 I)$ distribution. Choose $\kappa \in C_c^\infty(\mathbb{R}^n)$ with $\kappa(y) \in [0,1]$ and such that $\kappa(y) = 1$ for $||y||_2 \leq 1$. Defining

$$\varphi_m(y) = \varphi(y)\kappa(m^{-1}y),$$

then $\varphi_m \in C_c^\infty(\mathbb{R}^n)$ and

$$\varphi(y) \geq \varphi_m(y) \geq \varphi(y) 1(||y||_2 \leq m) \nearrow \varphi(y)$$

for $m \to \infty$. We also observe that

$$\partial_i \varphi_m(y) = (\partial_i \varphi(y))\kappa(m^{-1}y) + \varphi(y) m^{-1} \partial_i \kappa(m^{-1}y) \to \partial_i \varphi(y)$$

for $m \to \infty$. Here we used that $\partial_i \kappa(m^{-1}y) = 0$ for $||y||_2 < m$ and $\kappa(m^{-1}y) \to 1$ for $m \to \infty$. Moreover,

$$|\hat{\mu}_i(y) \partial_i \varphi_m(y)| \leq |\hat{\mu}_i(y)(y_i - \mu_i)\varphi(y)|/\sigma^2 + C|\hat{\mu}_i(y)|\varphi(y)$$

for some constant C, and the right hand side above is integrable. Thus by dominated convergence,

$$\lim_{m\to\infty} \int \hat{\mu}_i(y) \partial_i \varphi_m(y) \mathrm{d}y = \int \hat{\mu}_i(y) \partial_i \varphi(y) \mathrm{d}y.$$

If $\nabla \cdot \hat{\mu}$ is almost everywhere positive

$$\begin{aligned}\int \nabla \cdot \hat{\mu}(y)\ \varphi(y) \mathrm{d}y &\geq \limsup_{m\to\infty} \int \nabla \cdot \hat{\mu}(y)\ \varphi_m(y) \mathrm{d}y \geq \liminf_{m\to\infty} \int \nabla \cdot \hat{\mu}(y)\ \varphi_m(y) \mathrm{d}y \\ &\geq \liminf_{m\to\infty} \int_{||y||_2 \leq m} \nabla \cdot \hat{\mu}(y)\ \varphi(y) \mathrm{d}y = \int \nabla \cdot \hat{\mu}(y)\ \varphi(y) \mathrm{d}y\end{aligned}$$

using monotone convergence for the last identity. This shows that $\int \nabla \cdot \hat{\mu}(y)\ \varphi_m(y) \mathrm{d}y \to \int \nabla \cdot \hat{\mu}(y)\ \varphi(y) \mathrm{d}y$, and combined with the dominated convergence above

$$\begin{aligned}E(\nabla \cdot \hat{\mu}) &= \int \nabla \cdot \hat{\mu}(y)\ \varphi(y) \mathrm{d}y = \lim_{m\to\infty} \int \nabla \cdot \hat{\mu}(y)\ \varphi_m(y) \mathrm{d}y \\ &= -\lim_{m\to\infty} \sum_{i=1}^{n} \int \hat{\mu}_i(y) \partial_i \varphi_m(y) \mathrm{d}y = -\sum_{i=1}^{n} \int \hat{\mu}_i(y) \partial_i \varphi(y) \mathrm{d}y \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \int \hat{\mu}_i(y)(y_i - \mu_i)\varphi(y) \mathrm{d}y = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i).\end{aligned}$$

References

Candès, E.J., Sing-Long, C.A., Trzasko, J.D., 2013. Unbiased risk estimates for singular value thresholding and spectral estimators. IEEE Transactions on Signal Processing 61, 4643–4657.

Evans, L.C., Gariepy, R.F., 1992. Measure theory and fine properties of functions. *Studies in Advanced Mathematics*, CRC Press, Boca Raton, FL.

Fourdrinier, D., Wells, M.T., 2012. On improved loss estimation for shrinkage estimators. Statist. Sci. 27, 61–81.

Johnstone, I., 1998. On inadmissibility of some unbiased estimates of loss, in: Gupta, S.S., Berger, J.O. (Eds.), Statistical decision theory and related topics. IV., Springer-Verlag, New York, pp. 361–379.

Stein, C.M., 1981. Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9, 1135–1151.

Tibshirani, R.J., 2015. Degrees of freedom and model search. Statistica Sinica 25, 1265–1296.