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

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

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

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.