Computational Statistics and Data Analysis 65 (2013) 98–112

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Robust minimum information loss estimation

John C. Lind [a], Douglas P. Wiens [b,*], Victor J. Yohai [c]

[a] Centre for Psychiatric Assessment and Therapeutics, Alberta Hospital Edmonton, Alberta, Canada T5J 2J7

[b] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1

[c] Departmento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 23 September 2011
Received in revised form 28 May 2012
Accepted 13 June 2012
Available online 18 June 2012

Keywords:
Breakdown
Covariance
Cross-spectrum matrix
Electroencephalogram recording
Genetic algorithm
Minimum covariance determinant
Minimum information loss determinant estimate
Spectrum
Trimmed minimum information loss estimate

ABSTRACT

Two robust estimators of a matrix-valued location parameter are introduced and discussed. Each is the average of the members of a subsample – typically of covariance or cross-spectrum matrices – with the subsample chosen to minimize a function of its average. In one case this function is the Kullback–Leibler discrimination information loss incurred when the subsample is summarized by its average; in the other it is the determinant, subject to a certain side condition. For each, the authors give an efficient computing algorithm, and show that the estimator has, asymptotically, the maximum possible breakdown point. The main motivation is the need for efficient and robust estimation of cross-spectrum matrices, and they present a case study in which the data points originate as multichannel electroencephalogram recordings but are then summarized by the corresponding sample cross-spectrum matrices.

1. Introduction and Summary

A frequently encountered problem in the analysis of electroencephalogram (EEG) and magnetoencephalogram (MEG) recordings is the presence of artefacts in the data. Common sources of artefacts are muscle movement, equipment malfunction, errors in experimental procedures, unusual participant responses or the presence of misclassified individuals that do not represent the population of interest. A further complication arises from the non-stationarity of the EEG recordings, which can result in frequency spectra that differ between intervals within the same recording. Visual inspection of the data is the most common approach used to identify gross artefacts; however, in light of increasing numbers of sensors in modern recording systems, in addition to experimental designs in which recordings are often obtained across several time periods and treatment conditions, artefacts or patterns of unusual activity become increasingly difficult to detect. Because atypical recordings may go undetected and therefore introduce bias into subsequent results, there is a need for robust methods that can be applied to large channel arrays. Robust estimates of the spectrum and cross-spectrum, for example, are of particular interest because frequency domain analysis is often the preferred method for the analysis of time series in applied research. In addition, the spectrum and cross-spectrum often form the basis for other analysis techniques such as principal component analysis (PCA) and discriminant analysis—see for instance the treatment of such techniques in Stoffer (1999) and Shumway and Stoffer (2006). In the area of neuro-imaging, where large array recordings

* Corresponding author.
E-mail addresses: JohnC.Lind@albertahealthservices.ca (J.C. Lind), doug.wiens@ualberta.ca (D.P. Wiens), vyohai@dm.uba.ar (V.J. Yohai).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.06.011

are common, the estimation of electrical current flow over the surface of the scalp often depends on the calculation of the surface Laplacian (Nunez and Srinivasan, 2006, pp. 334–337) which is derived from the cross-spectrum. The EEG and MEG cross-spectrum matrices also form the basis of brain imaging methods such as those based on multiple signal classification (MUSIC) algorithms (Mosher et al., 1992), low-resolution electromagnetic tomography (LORETA) (Pascual-Marqui et al., 1994) and Borgiotti and Kaplan (1979) Beamformer methods.

In this article, we introduce and discuss two robust estimators of a matrix-valued location parameter. They have been derived by us for use in problems in which the data consist of positive semidefinite Hermitian matrices $\{\mathbf{S}_j\}$. A case in point is that in which the $\mathbf{S}_j$ are cross-spectrum matrices, whose elements are cross-products of the Fourier transforms, at various frequencies, of the original data vectors. These vectors are typically not retained, in the interest of economizing data storage. As is the case in the applications described above, the need for robustness arises when cross-spectrum matrices are obtained for each individual in a group and it is necessary to identify and remove matrices corresponding to those individuals whose recordings contain outliers or atypical patterns of activity.

The 'classical' location estimate is of course the average of the $\{\mathbf{S}_j\}$; this suffers from a well-known lack of robustness due to possible outliers. Robust estimates of the frequency spectrum in time series data based on autoregressive models for single channel recordings have been proposed by Kleiner et al. (1979); however, robust methods for multichannel recordings are less readily available. An appealing property of the estimators presented here is that they can be applied to cross-spectrum matrices obtained from high dimensional arrays. In Section 5, we apply our methods to the problem of identifying differences between the two sets of cross-spectrum matrices obtained from 43-channel EEG recordings, in order to compare results obtained before and after those matrices identified as outliers have been removed.

Each of the proposed estimates is the average in a particular 'trimmed' subsample of the $\{\mathbf{S}_j\}$. The trimming selects a subsample minimizing a certain function of its average. In the first case, leading to the 'Trimmed Minimum Information Loss' estimate $\hat{\boldsymbol{\Sigma}}_{TMIL}$, this function is related to the Kullback–Leibler discrimination information loss incurred when the $\mathbf{S}_j$ are summarized by $\hat{\boldsymbol{\Sigma}}$. In the second, leading to the 'Minimum Information Loss Determinant' estimate $\hat{\boldsymbol{\Sigma}}_{MILD}$, the function is the determinant, with the subsample restricted by a certain side condition. In each case the intent is to select subsamples whose members are close to the 'centre' of the sample. Since in each case the centre of the sample is defined by the estimate itself, the computations are iterative in nature, and we propose and assess various algorithms. We also discuss the breakdown properties and show that the best possible breakdown point is attainable, asymptotically, in each case. We include, in Section 4, a simulation study in which the two estimation methods are compared; as well the use of quantile plots to identify the outlying members of a data set is described. These theoretical and simulated results, together with what is learned from the EEG example, show these estimators to be valuable additions to the arsenal of robust methods of data analysis.

Code to duplicate all computations presented here has been written in MATLAB and in R and is available from us. All derivations are in the Appendix.

2. The TMIL estimate

Throughout this article, $\mathbf{S}$ will represent a random, $p \times p$ positive semidefinite Hermitian matrix with positive definite expectation $E[\mathbf{S}] = \boldsymbol{\Sigma}_0$. For a positive definite matrix $\boldsymbol{\Sigma}$ and a positive semidefinite $\boldsymbol{\Sigma}_0$, define a function

$$\Delta(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right) - \log\left|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right| - p.$$

As noted by Kakizawa et al. (1998), $\Delta(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma})$ is the Kullback–Leibler discrimination information, measuring the loss when a Gaussian density with covariance $\boldsymbol{\Sigma}_0$ is approximated by one with covariance $\boldsymbol{\Sigma}$. The function is non-negative, and is zero if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$—this is a consequence of the inequality

$$f(\lambda) = \lambda - \log\lambda - 1 \geq 0 = f(1), \quad (\lambda > 0), \tag{1}$$

applied to the eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0$.

Define also

$$g(\boldsymbol{\Sigma}) = E[\Delta(\mathbf{S}, \boldsymbol{\Sigma})] = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right) - E\left[\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{S}\right|\right] - p. \tag{2}$$

Let $\{\mathbf{S}_j\}_{j=1}^n$ be a sample of n i.i.d. copies of $\mathbf{S}$. The empirical version of $g(\boldsymbol{\Sigma})$ is

$$\bar{g}(\boldsymbol{\Sigma}) = \frac{1}{n}\sum_{j=1}^{n}\Delta\left(\mathbf{S}_j, \boldsymbol{\Sigma}\right) = \frac{1}{n}\sum_{j=1}^{n}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_j\right) - \frac{1}{n}\sum_{j=1}^{n}\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{S}_j\right| - p.$$

Minimization of $\bar{g}(\boldsymbol{\Sigma})$ corresponds to minimum information loss estimation in Gaussian populations. A standard result of multivariate analysis, used for instance to obtain the maximum likelihood estimate of a common covariance matrix $\boldsymbol{\Sigma}$ from n normal samples with sample covariances $\{\mathbf{S}_j\}$, and again based on (1), is that – even without the normality assumption – $\bar{g}(\boldsymbol{\Sigma})$ is minimized uniquely by the average $\bar{\mathbf{S}} = n^{-1}\sum_{j=1}^{n}\mathbf{S}_j$. Indeed,

$$\bar{g}(\boldsymbol{\Sigma}) = \Delta\left(\bar{\mathbf{S}}, \boldsymbol{\Sigma}\right) - \frac{1}{n}\sum_{j=1}^{n}\log\left|\bar{\mathbf{S}}^{-1}\mathbf{S}_j\right|,$$

and so is minimized by the minimizer $\bar{\mathbf{S}}$ of $\Delta\left(\bar{\mathbf{S}}, \boldsymbol{\Sigma}\right)$. See for instance Srivastava and Khatri (1979, Section 7.6).