ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), 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 $\hat{\mathbf{\Sigma}}_{TMIL}$, this function is related to the Kullback–Leibler discrimination information loss incurred when the $\mathbf{S}_j$ are summarized by $\hat{\mathbf{\Sigma}}$. In the second, leading to the 'Minimum Information Loss Determinant' estimate $\hat{\mathbf{\Sigma}}_{MILD}$, 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 $E[\mathbf{S}] = \mathbf{\Sigma}_0$. For a positive definite matrix $\mathbf{\Sigma}$ and a positive semidefinite $\mathbf{\Sigma}_0$, define a function

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

As noted by Kakizawa et al. (1998), $\Delta(\mathbf{\Sigma}_0, \mathbf{\Sigma})$ is the Kullback–Leibler discrimination information, measuring the loss when a Gaussian density with covariance $\mathbf{\Sigma}_0$ is approximated by one with covariance $\mathbf{\Sigma}$. The function is non-negative, and is zero if and only if $\mathbf{\Sigma} = \mathbf{\Sigma}_0$—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 $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_0$.

Define also

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

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

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

Minimization of $\bar{g}(\mathbf{\Sigma})$ 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 $\mathbf{\Sigma}$ from n normal samples with sample covariances $\{\mathbf{S}_j\}$, and again based on (1), is that – even without the normality assumption – $\bar{g}(\mathbf{\Sigma})$ is minimized uniquely by the average $\bar{\mathbf{S}} = n^{-1}\sum_{j=1}^{n}\mathbf{S}_j$. Indeed,

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

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

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

We consider a robust version of this estimate, which uses only a subset of the observed matrices for which $\Delta\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}\right)$ is smallest. We call such an estimate a 'Trimmed Minimum Information Loss' (*TMIL*) estimate.

Definition 1. For given h ($n/2 \le h \le n$) and positive definite, $p \times p$ Hermitian matrices $\mathbf{\Sigma}$, let $\Delta_{(k)}(\mathbf{\Sigma})$ ($k = 1, \ldots, h$) be the kth smallest of the values $\left\{\Delta\left(\mathbf{S}_j, \mathbf{\Sigma}\right)\right\}_{j=1}^{n}$. The Trimmed Minimum Information Loss estimate $\hat{\mathbf{\Sigma}}_{TMIL}$ is the minimizer of their average:

$$\hat{\mathbf{\Sigma}}_{TMIL} = \underset{\mathbf{\Sigma}>\mathbf{0}}{\operatorname{argmin}} \frac{1}{h}\sum_{k=1}^{h} \Delta_{(k)}(\mathbf{\Sigma}).$$

To facilitate comparisons with a related estimate to be introduced later in this article, we give an alternate, equivalent formulation. Let $\mathcal{H}$ be the set of all h-element subsets H of $\{1, 2, \ldots, n\}$, for which

$$\bar{\mathbf{S}}_H = \frac{1}{h}\sum_{j\in H} \mathbf{S}_j$$

is positive definite. Note that

$$\bar{\Delta}_H \overset{\text{def}}{=} \frac{1}{h}\sum_{j\in H} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_H\right) = \log\left|\bar{\mathbf{S}}_H\right| - \frac{1}{h}\sum_{j\in H}\log\left|\mathbf{S}_j\right|; \tag{3}$$

this can be interpreted as the loss in information when the subsample is summarized by $\bar{\mathbf{S}}_H$, and is infinite if the subsample contains any singular members. For positive definite, $p \times p$ Hermitian matrices $\mathbf{\Sigma}$, define a pair

$$\left(\hat{\mathbf{\Sigma}}, H_0\right) = \underset{\mathbf{\Sigma}>\mathbf{0}, H\in\mathcal{H}}{\operatorname{argmin}} \frac{1}{h}\sum_{j\in H} \Delta\left(\mathbf{S}_j, \mathbf{\Sigma}\right).$$

Then $\hat{\mathbf{\Sigma}}$ is the *TMIL* estimator and is given by

$$\hat{\mathbf{\Sigma}}_{TMIL} = \bar{\mathbf{S}}_{H_0},$$

where, by (3),

$$H_0 = \underset{H\in\mathcal{H}}{\operatorname{argmin}}\, \bar{\Delta}_H. \tag{4}$$

2.1. Computing the TMIL estimate

Consider the following algorithm:

Algorithm 1.

Initialization Select a starting set $H \in \mathcal{H}$. Define $H_{(1)} = H$. Compute the average $\hat{\mathbf{\Sigma}}_{(1)} = \bar{\mathbf{S}}_{H_{(1)}}$, and $\bar{\Delta}_{H_{(1)}}$ as at (3).
For $k = 1, 2, \ldots$ to convergence of $\bar{\Delta}_{H_{(k)}}$:

Iterative step Let $H_{(k+1)}$ be the set of indices j resulting in the h smallest values of $\Delta\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{(k)}\right)$. Set $\hat{\mathbf{\Sigma}}_{(k+1)} = \bar{\mathbf{S}}_{H_{(k+1)}}$, and compute $\bar{\Delta}_{H_{(k+1)}}$.

The following theorem implies that each sequence $\left\{\bar{\Delta}_{H_{(k)}}\right\}$, constructed as above, and for which the members $H_{(k)}$ remain in $\mathcal{H}$, decreases to a limit $\bar{\Delta}_\infty(H)$, and that $\left(\hat{\mathbf{\Sigma}}_{TMIL}, H_0\right)$ defined above are the limits $\left(\hat{\mathbf{\Sigma}}_{(\infty)}, H_{(\infty)}\right)$ corresponding to one such sequence.

Theorem 1. *Let $\left\{\mathbf{S}_j\right\}_{j=1}^{n}$ be a sample of n i.i.d. copies of $\mathbf{S}$ and let $H_{(1)}$ be any subset of $\mathcal{H}$. Let $H_{(2)}$ be the set of indices j resulting in the h smallest values of the $\Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right)$. If $\bar{\mathbf{S}}_{H_{(2)}}$ is positive definite, so that $H_{(2)} \in \mathcal{H}$, then*

$$\bar{\Delta}_{H_{(1)}} \ge \bar{\Delta}_{H_{(2)}}, \tag{5}$$

with equality if and only if $\bar{\mathbf{S}}_{H_{(1)}} = \bar{\mathbf{S}}_{H_{(2)}}$.

Of course it is not often feasible to apply Algorithm 1 and Theorem 1 to all $\binom{n}{h}$ subsequences $\left\{\bar{\Delta}_H\right\}$, each corresponding to a different starting set $H \in \mathcal{H}$. An option is to repeat the algorithm for many randomly chosen subsamples. But such subsamples will, with high probability, contain outlying (or inlying) values which may determine the final estimate. Thus we prefer the following approach, similar to an improved subsampling method given by Rousseeuw and van Driessen (1999) for the Minimum Covariance Determinant (MCD) (Rousseeuw, 1985) estimate. For each non-singular $\mathbf{S}_a$ in the sample,

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

define $H^{(a)}$ to be the set of indices of the sample values $\{\mathbf{S}_j\}$ with the h smallest values of $\Delta(\mathbf{S}_j, \mathbf{S}_a)$. Now apply Algorithm 1 repeatedly, starting with one of the sets $H^{(a)}$ each time. Each such 'run' results in a limit matrix $\hat{\mathbf{\Sigma}} = \bar{\mathbf{S}}_{H_{(\infty)}}$; the $\bar{\mathbf{S}}_{H_{(\infty)}}$ for which $\bar{\Delta}_\infty\left(H^{(a)}\right)$ is a minimum is the (approximation to) the *TMIL* estimate. In the – very unlikely – event that some $\bar{\mathbf{S}}_{H_{(k+1)}}$ is singular, our algorithm defaults to $\hat{\mathbf{\Sigma}} = \bar{\mathbf{S}}_{H_{(k)}}$.

We have investigated the use of a genetic algorithm as a means of improving on this selection of starting sets. In this algorithm, the sets $H^{(a)} \stackrel{\text{def}}{=} H_1^{(a)}$ described above are viewed merely as the first 'generation' of starting sets. Subsequent generations $\left\{H_g^{(a)}\right\}$ for $g = 2, 3, \ldots$ are formed via stochastic processes of 'crossover' and 'mutation'. Crossover is a process whereby 'fit parents' – pairs of starting sets with small values of $\bar{\Delta}_\infty\left(H^{(a)}\right)$ – are randomly chosen to produce 'children'. The children are starting sets formed from the parents in a manner very similar to that employed by Welsh and Wiens (in press), who used a genetic algorithm to construct sampling designs. For an application of genetic algorithms to frequency domain methods, see Mitra et al. (2006).

Todorov (1992) presented a simulated annealing algorithm for the computation of the MCD, and found it to be generally more effective than the then current competitor – the 'Iterative Improvement' algorithm – which uses an initial *random* subsample. In contrast, we will argue in the simulation study of Section 4 that our method of choosing the first generation $\left\{H_1^{(a)} | a = 1, \ldots, n\right\}$ is so efficient that improvements arising from the genetic algorithm are at best only very slight.

2.2. Breakdown properties

The breakdown point of an estimator is, roughly speaking, the maximum fraction of arbitrarily 'bad' members of the sample which the estimate can tolerate. See Rousseeuw and Leroy (1987) for further details. Given a 'clean' sample $\mathcal{S} = \{\mathbf{S}_j | j = 1, \ldots, n\}$, we define breakdown as the tending of the smallest or largest eigenvalue of the estimate to 0 or ∞, respectively, when $\mathcal{S}$ is replaced by 'contaminated' samples. To formalize this, let $\mathcal{SS}_m$ be the set of samples with at least $n - m$ members in common with $\mathcal{S}$:

$$\mathcal{SS}_m = \left\{ \mathcal{S}^\dagger = \left\{\mathbf{S}_j^\dagger\right\}_{j=1}^n : \sum I(\mathbf{S}_j^\dagger = \mathbf{S}_j \in \mathcal{S}) \geq n - m \right\}.$$

Denote by $ch_{\min}(\cdot)$ and $ch_{\max}(\cdot)$ the smallest and largest eigenvalues of a Hermitian matrix.

Definition 2. Given an estimate $\hat{\mathbf{\Sigma}}_n$ which assigns to each sample $\mathcal{S}$ a $p \times p$ positive definite Hermitian matrix $\hat{\mathbf{\Sigma}}_n(\mathcal{S})$, define the finite sample breakdown point at the sample $\mathcal{S}$ by

$$\varepsilon^\dagger(\hat{\mathbf{\Sigma}}_n, \mathcal{S}) = \frac{m(\mathcal{S}, \hat{\mathbf{\Sigma}}_n)}{n},$$

where

$$m(\mathcal{S}, \hat{\mathbf{\Sigma}}_n) = \min\left\{ m : \sup_{\mathcal{S}^\dagger \in \mathcal{SS}_m} \left[\frac{1}{ch_{\min}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right)} + ch_{\max}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right) \right] = \infty \right\}.$$

A natural requirement for a location estimate $\hat{\mathbf{\Sigma}}_n$ is that it be scale equivariant – for any $\lambda > 0$ we must have $\hat{\mathbf{\Sigma}}_n(\lambda\mathbf{S}_1, \ldots, \lambda\mathbf{S}_n) = \lambda\hat{\mathbf{\Sigma}}_n(\mathbf{S}_1, \ldots, \mathbf{S}_n)$. It is easy to check that both $\hat{\mathbf{\Sigma}}_{TMIL}$, being discussed here, and the companion estimator $\hat{\mathbf{\Sigma}}_{MILD}$ which is the subject of the next section, are scale equivariant. It is intuitively clear that no scale equivariant estimator can withstand breakdown if more than 50% of the sample is contaminated; the best one can hope for is to attain a breakdown point of 0.5, and then, typically, only asymptotically. This observation is made rigorous by the following result.

Lemma 1. *Let $\hat{\mathbf{\Sigma}}_n$ be as in Definition 2, and suppose as well that $\hat{\mathbf{\Sigma}}_n$ is scale equivariant. Denote by $v(\mathcal{S})$ the number of singular members of $\mathcal{S}$. Then for all n and $1 < v_0 < n$, there exists samples $\mathcal{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$, with $v(\mathcal{S}) = v_0$, such that $\varepsilon^\dagger(\hat{\Sigma}_n, \mathcal{S}) \leq (n - v_0)/2n$.*

Together with Lemma 1 the following theorem asserts that the *TMIL* estimate with $h = n/2$ has the maximum possible breakdown point among equivariant estimators in samples with $v(\mathcal{S}) = 0$, and asymptotically the maximum possible breakdown point assuming that $v(\mathcal{S}) = o(n)$.

Theorem 2. *Let $\mathcal{S} = \{\mathbf{S}_j\}_{j=1}^n$ be a set of positive semidefinite Hermitian matrices. Then for $h \geq n/2$, the finite sample breakdown point $\varepsilon_{TMIL}^\dagger(\mathcal{S}) \stackrel{\text{def}}{=} \varepsilon^\dagger(\hat{\mathbf{\Sigma}}_{TMIL}, \mathcal{S})$ of $\hat{\mathbf{\Sigma}}_{TMIL}$ satisfies*

$$\varepsilon_{TMIL}^\dagger(\mathcal{S}) \geq \frac{n - h - v(\mathcal{S})}{n}.$$

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

2.3. Bias properties

The bias of an estimator is always in reference to a given location parameter and a nominal distribution. For example, in the case of location is the sample median biased? If the goal is to estimate the population mean then the sample median is biased except for some particular nominal distributions such as those with a symmetric distribution around the centre of location. However, when the goal is to estimate the population median, the sample median is asymptotically unbiased. Note that the population mean is very unstable under small changes in the distribution and therefore in many cases it is more convenient to choose the median as the location parameter to estimate. On the other hand, in a completely non-parametric setup which includes asymmetric distributions there are no asymptotically unbiased robust estimators of the mean.

Similar considerations apply when estimating the location parameter of a population of positive semidefinite Hermitian matrices. If the goal is to estimate the mean, our proposed estimators are biased. However these estimators themselves define location parameters – their corresponding asymptotic values – and for these location parameters they are obviously asymptotically unbiased. Again these location parameters may be better choices than the population mean.

We do not know how to exactly compute the asymptotic bias of our estimators with respect to the mean in a given nominal distribution. Some estimates are obtained, by simulation, in Section 4.

3. The MILD estimate

For fixed subsets H of $\mathcal{H}$, there is an interesting, alternate formulation of the estimate $\hat{\mathbf{\Sigma}}_{TMIL}$. First define

$$\Omega\left(\mathbf{\Sigma}_0, \mathbf{\Sigma}\right) = \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_0\right),$$

and recall the definition of $g(\cdot)$ at (2).

Proposition 1. *For any $\mathbf{\Sigma}$, there exists $\mathbf{\Sigma}'$ with $\Omega\left(\mathbf{\Sigma}_0, \mathbf{\Sigma}'\right) = p$ and $g\left(\mathbf{\Sigma}'\right) \le g\left(\mathbf{\Sigma}\right)$.*

An immediate consequence of Proposition 1 is the following lemma.

Lemma 2. *The following problems are equivalent:*

P1: *find a matrix $\mathbf{\Sigma}$ minimizing $g(\mathbf{\Sigma})$;*
P2: *find a matrix $\mathbf{\Sigma}$ minimizing $|\mathbf{\Sigma}|$ subject to the constraint*

$$\Omega\left(\mathbf{\Sigma}_0, \mathbf{\Sigma}\right) = p. \tag{6}$$

For $H \in \mathcal{H}$, define

$$\bar{\Omega}_{H,\mathbf{\Sigma}} = \frac{1}{h}\sum_{j\in H} \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}_j\right).$$

Lemma 2 yields the following characterization of $\hat{\mathbf{\Sigma}}_{TMIL}$ of Section 2, for a *fixed* set $H \in \mathcal{H}$.

Theorem 3. *Let $H_0 \in \mathcal{H}$ be as at (4). Then $\hat{\mathbf{\Sigma}}_{TMIL}$ is the matrix with minimum determinant among all $p \times p$ positive definite Hermitian matrices satisfying $\bar{\Omega}_{H_0,\mathbf{\Sigma}} = p$.*

We note that, if $\mathbf{S}_k$ is a covariance matrix arising from centred data $\{\mathbf{x}_{jk}\}$, then $\Omega\left(\mathbf{S}_k, \mathbf{\Sigma}\right)$ is the average of the (squared) Mahalanobis distances $\mathbf{x}_{jk}^*\mathbf{\Sigma}^{-1}\mathbf{x}_{jk}$. Here and elsewhere we denote by '*' the conjugate transpose of a possibly complex-valued vector or matrix.

An alternate estimate, which we call the Minimum Information Loss Determinant (*MILD*) estimate, arises from minimizing the determinant $|\mathbf{\Sigma}|$ over subsets of $\mathcal{H}$. Note that $\bar{\Omega}_{H,\mathbf{\Sigma}} = \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\bar{\mathbf{S}}_H\right)$. By Corollary 1 of the Appendix, $|\mathbf{\Sigma}|$ is minimized, subject to $\bar{\Omega}_{H,\mathbf{\Sigma}} = p$, by $\mathbf{\Sigma} = \bar{\mathbf{S}}_H$.

Definition 3. Let $\mathcal{H}$ be as defined in Section 2 and let $\mathcal{S}_H$ be the set of positive definite Hermitian matrices $\mathbf{\Sigma}$ for which $\bar{\Omega}_{H,\mathbf{\Sigma}} = p$. Define

$$\mathbf{\Sigma}_H^* = \arg\min_{\mathbf{\Sigma}\in\mathcal{S}_H} |\Sigma|,$$

and

$$H_1 = \arg\min_{H\in\mathcal{H}} |\mathbf{\Sigma}_H^*|.$$

Then the Minimum Information Loss Determinant estimate $\hat{\mathbf{\Sigma}}_{MILD}$ is given by $\hat{\mathbf{\Sigma}}_{MILD} = \mathbf{\Sigma}_{H_1}^* = \bar{\mathbf{S}}_{H_1}$.

The following example shows that, in general, the set H_0 determining $\hat{\mathbf{\Sigma}}_{TMIL}$ does not coincide with the set H_1 determining $\hat{\mathbf{\Sigma}}_{MILD}$, and so the estimates need not agree.

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

Example. Take $p = 1$ and consider a sample $\{S_j\} = \{1, 100, 100, 100, 100, 100, 100, 100, 100, 100\}$. Take $h = n/2 = 5$. Then the *TMIL* estimate minimizes

$$\frac{1}{h}\sum_{j\in H}\left(\frac{S_j}{\Sigma} - \log\frac{S_j}{\Sigma}\right) - 1,$$

and so arises from $\{S_j\}_{j\in H_0} = \{100, 100, 100, 100, 100\}$, with $\hat{\Sigma}_{TMIL} = 100$ and $\bar{\Delta}_{H_0} = 0$. The *MILD* estimate arises from $\{S_j\}_{j\in H_1} = \{1, 100, 100, 100, 100\}$ and is $\hat{\Sigma}_{MILD} = 80.2$. This example also shows that when $p = 1$ the *MILD* estimate is merely the average of the smallest h of the S_j.

3.1. Computing the MILD estimate

We compute $\hat{\mathbf{\Sigma}}_{MILD}$ via the following algorithm.

Algorithm 2.

Initialization Select a starting set $H \in \mathcal{H}$. Define $H_{(1)} = H$. Compute the average $\hat{\mathbf{\Sigma}}_{(1)} = \bar{\mathbf{S}}_{H_{(1)}}$, and the determinant $\left|\hat{\mathbf{\Sigma}}_{(1)}\right|$.

For $k = 1, 2, \ldots$ to convergence of $\left|\hat{\mathbf{\Sigma}}_{(k)}\right|$:

Iterative Step Let $H_{(k+1)}$ be the set of indices j resulting in the h smallest values of $\Omega\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{(k)}\right)$. Set $\hat{\mathbf{\Sigma}}_{(k+1)} = \bar{\mathbf{S}}_{H_{(k+1)}}$, and compute $\left|\hat{\mathbf{\Sigma}}_{(k+1)}\right|$.

Theorem 4 implies that each sequence $\left\{\left|\hat{\mathbf{\Sigma}}_{(k)}\right|\right\}$ constructed as above, and for which the sets $H_{(k)}$ remain in $\mathcal{H}$ (note that by construction they are then also in $\mathcal{H}_1$) decreases to a limit, and that $\hat{\mathbf{\Sigma}}_{MILD}$ arises from one such sequence. It is analogous to Theorem 1 of Rousseeuw and van Driessen (1999).

Theorem 4. *Let $\{\mathbf{S}_j\}_{j=1}^n$ be a sample of n i.i.d. copies of $\mathbf{S}$ and let $H_{(1)}$ be any subset of $\mathcal{H}$. If $\left|\bar{\mathbf{S}}_{H_{(1)}}\right| > 0$, compute the $\Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right)$. Let $H_{(2)}$ be the set of indices of the h smallest of the $\Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right)$. Then $H_{(1)}, H_{(2)} \in \mathcal{H}_1$ and*

$$\left|\bar{\mathbf{S}}_{H_{(1)}}\right| \geq \left|\bar{\mathbf{S}}_{H_{(2)}}\right|,$$

with equality if and only if $\bar{\mathbf{S}}_{H_{(1)}} = \bar{\mathbf{S}}_{H_{(2)}}$.

We initialize Algorithm 2 in much the same manner as Algorithm 1. For each non-singular $\mathbf{S}_a$ in the sample, define $\tilde{H}^{(a)}$ to be the indices of the sample values $\{\mathbf{S}_j\}$ with the h smallest values of $\Omega\left(\mathbf{S}_j, \mathbf{S}_a\right)$. Then apply Algorithm 2 repeatedly, starting with one of the sets $\tilde{H}^{(a)}$ each time. Again this can be followed by a genetic search for improved starting sets, as described in Section 2.1, the only difference being in the specification of the loss function.

3.2. Breakdown and bias

The following theorem asserts that the *MILD* estimate has, in common with $\hat{\mathbf{\Sigma}}_{TMIL}$, the maximum possible breakdown point.

Theorem 5. *Let $\mathcal{S} = \{\mathbf{S}_j\}_{j=1}^n$ be a set of positive definite Hermitian matrices and let $v(\mathcal{S})$ be as in Lemma 1. For $h \geq n/2$, the finite sample breakdown point $\varepsilon^{\dagger}_{MILD}(\mathcal{S}) \stackrel{\text{def}}{=} \varepsilon^{\dagger}(\hat{\mathbf{\Sigma}}_{MILD}, \mathcal{S})$ of $\hat{\mathbf{\Sigma}}_{MILD}$ satisfies*

$$\varepsilon^{\dagger}_{MILD}(\mathcal{S}) \geq \frac{n - h - v(\mathcal{S})}{n}.$$

We note that the same considerations as in Section 2.3 apply to the bias of $\hat{\mathbf{\Sigma}}_{MILD}$.

4. Simulations

In this section, we report the results of a small simulation study to compare the two estimation methods—*TMIL* and *MILD*. Samples of $np \times p$ Hermitian matrices $\mathbf{S} = \mathbf{X}\mathbf{X}^*$, with $\mathbf{X} = \mathbf{U}/\sqrt{2L} + i\mathbf{V}/\sqrt{2L}$, were constructed from independent $p \times L$ matrices $\mathbf{U}$ and $\mathbf{V}$. A fraction ε_1 of these $2n$ matrices $(\mathbf{U}, \mathbf{V})$ were contaminated; the remainder had i.i.d. $N(0, 1)$ elements.

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

Table 1
Performance of the sample average $\bar{\mathbf{S}} = \hat{\mathbf{\Sigma}}_{MEAN}$: values of per-element root mean squared error $(\text{rmse}_{MEAN})^{a}$ and its standard error $[\cdot]^{a}$, followed by the bias $\langle \text{bias}_{MEAN} \rangle^{a}$ and average values of $\left\{ \min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p} \right\}$; $p = 30$, $n = 100$ and $N = 100$.

	$f_{small} = 0$	$f_{small} = 0.5$	$f_{small} = 1$
$\varepsilon_1 = 0.3$, $\varepsilon_2 = 0.4$:	(177.94) [0.15] ⟨175.04⟩ {16.44, 1.95}	(80.43) [0.10] ⟨77.71⟩ {12.01, 1.42}	(23.17)[0.03] ⟨20.04⟩ {9.79, 0.89}
$\varepsilon_1 = 0.5$, $\varepsilon_2 = 0.2$:	(148.35) [0.14] ⟨146.02⟩ {12.69, 1.79}	(67.41) [0.09] ⟨64.78⟩ {10.55, 1.35}	(20.37) [0.03] ⟨16.63⟩ {9.31, 0.91}
$\varepsilon_1 = 0$:	$(\text{rmse}_{MEAN})^{a}$= (12.87) [0.03], $\langle \text{bias}_{MEAN} \rangle^{a}$= ⟨1.30⟩, average $\left\{ \min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p} \right\}$ = {9.12, 1.00}.		

[a] Values have been multiplied by 1000.

Table 2
Without a genetic search for improved starting sets: values of per-element root mean squared errors $(\text{rmse}_{TMIL}, \text{rmse}_{MILD})^{a}$ and their pooled standard errors $[\cdot]^{a}$, followed by the biases $\langle \text{bias}_{TMIL}, \text{bias}_{MILD} \rangle^{a}$ and average values of $\left\{ \min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p} \right\}$; $p = 30$, $n = 100$ and $N = 100$.

	$f_{small} = 0$	$f_{small} = 0.5$	$f_{small} = 1$
$\varepsilon_1 = 0.3$, $\varepsilon_2 = 0.4$:	(18.34, 18.27) [0.04] ⟨1.86, 2.69⟩ {8.85, 0.98}	(18.41, 18.45) [0.05] ⟨1.90, 2.80⟩ {8.86, 0.98}	(18.37, 44.35) [0.04] ⟨1.90, 41.88⟩ {8.85, 0.77}
$\varepsilon_1 = 0.5$, $\varepsilon_2 = 0.2$:	(18.28, 18.28) [0.04] ⟨1.85, 1.85⟩ {9.05, 0.99}	(18.27, 18.27) [0.04] ⟨1.83, 1.83⟩ {9.07, 1.00}	(27.53, 36.40) [0.16] ⟨21.99, 33.22⟩ {8.92, 0.81}
$\varepsilon_1 = 0$:	$(\text{rmse}_{TMIL}, \text{rmse}_{MILD})^{a}$= (18.49, 18.61) [0.04], $\langle \text{bias}_{TMIL}, \text{bias}_{MILD} \rangle^{a}$= ⟨1.84, 3.85⟩, average $\left\{ \min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p} \right\}$ = {8.71, 0.98}.		

[a] Values have been multiplied by 1000.

The division by $\sqrt{2L}$ ensures that, without contamination, $\mathbf{\Sigma}_0 = E[\mathbf{S}] = \mathbf{I}_p$. Within the contaminated matrices, the elements of $\mathbf{U}$ and $\mathbf{V}$ were simulated from the distribution

$$(1 - \varepsilon_2)\, N(0, \sigma = 1) + \varepsilon_2 \cdot f_{small} N(0, \sigma = 0.3) + \varepsilon_2 \cdot f_{large} N(0, \sigma = 3),$$

the interpretation being that, of the contaminated elements, a fraction f_{small} had standard deviations of 0.3 while the remaining proportion $f_{large} = 1 - f_{small}$ had standard deviations of 3. Estimates $\hat{\mathbf{\Sigma}}_{TMIL}$ and $\hat{\mathbf{\Sigma}}_{MILD}$ were then computed from this sample. This process was carried out N times. We report the averages $\bar{\mathbf{\Sigma}}_{TMIL}$ and $\bar{\mathbf{\Sigma}}_{MILD}$ of the N estimates $\left\{ \hat{\mathbf{\Sigma}}_j \right\}_{j=1}^{N}$ so obtained, and the 'per-element' bias and root mean squared errors

$$\text{bias} = \frac{1}{p} \left\| \frac{\sum_{j=1}^{N} \hat{\mathbf{\Sigma}}_j}{N} - \mathbf{I}_p \right\|, \qquad \text{rmse} = \sqrt{\frac{\sum_{j=1}^{N} \left\| \hat{\mathbf{\Sigma}}_j - \mathbf{I}_p \right\|^2}{Np^2}};$$

here we use the Euclidean norm $\|\mathbf{A}\| = \sqrt{\text{tr}\mathbf{A}\mathbf{A}^*}$. Except where otherwise noted, in all cases reported in this section we have used $p = 30, L = 2p, n = 100, N = 100$, and $h = [(n+1)/2]$. The value of h is the integer giving the largest possible breakdown point, according to Theorems 2 and 5. Other values of these inputs gave qualitatively similar output.

We first ran the simulations using the sample average $\bar{\mathbf{S}} \overset{\text{def}}{=} \hat{\mathbf{\Sigma}}_{MEAN}$, i.e. $h = n$, obtaining the values in Table 1. For the robust estimators, when the algorithms were stopped after the first generation, i.e. no genetic searching for improved starting sets was done, we obtained the performance measures in Table 2. When the genetic algorithm was also applied, until there had been no change for 5 consecutive generations, we obtained the values in Table 3. For all estimators, as a benchmark we first took $\varepsilon_1 = 0$—no contamination. Combinations of various values of $(\varepsilon_1, \varepsilon_2, f_{small})$ yielded the remaining values in Tables 1–3.

Some observations to be made from these simulations are the following.

(i) For the sample average $\hat{\mathbf{\Sigma}}_{MEAN}$, and except in clean samples, the bias was the dominant contributor to the mean squared error (mse). The bias and mse of the proposed estimators $\hat{\mathbf{\Sigma}}_{TMIL}$ and $\hat{\mathbf{\Sigma}}_{MILD}$ were very stable under contamination at the levels $(\varepsilon_1, \varepsilon_2) = (0.3, 0.4)$ and $(\varepsilon_1, \varepsilon_2) = (0.5, 0.2)$, with bias making only a minor contribution, except for those of $\hat{\mathbf{\Sigma}}_{TMIL}$ at $(\varepsilon_1, \varepsilon_2) = (0.5, 0.2)$ with all contamination 'small' ($f_{small} = 1$), and those of $\hat{\mathbf{\Sigma}}_{MILD}$ at both levels with all contamination small. With $f_{small} = 1$ the performance of $\hat{\mathbf{\Sigma}}_{MEAN}$ was surprisingly good—especially in view of its disastrous performance in the other cases.
(ii) As could perhaps have been anticipated from the Example of Section 3, *TMIL* outperforms *MILD* in terms of mse when there is a preponderance of 'small' contamination. In other cases the differences were at most only slight.

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

Table 3
Following a genetic search for improved starting sets: values of per-element root mean squared errors (rmse_{TMIL}, rmse_{MILD})[a] and their pooled standard errors [·][a], followed by the biases ⟨bias_{TMIL}, bias_{MILD}⟩[a] and average values of $\left\{\min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p}\right\}$; $p = 30$, $n = 100$ and $N = 100$.

	$f_{small} = 0$	$f_{small} = 0.5$	$f_{small} = 1$
$\varepsilon_1 = 0.3$, $\varepsilon_2 = 0.4$:	(18.35, 18.28) [0.04] ⟨1.83, 2.69⟩ {8.84, 0.98}	(18.43, 18.45) [0.05] ⟨1.87, 2.81⟩ {8.86, 0.98}	(18.39, 44.35) [0.04] ⟨1.90, 41.89⟩ {8.85, 0.77}
$\varepsilon_1 = 0.5$, $\varepsilon_2 = 0.2$:	(18.28, 18.28) [0.04] ⟨1.85, 1.85⟩ {9.05, 0.99}	(18.27, 18.27) [0.04] ⟨1.83, 1.83⟩ {9.07, 1.00}	(22.91, 36.40) [0.17] ⟨14.86, 33.22⟩ {8.88, 0.81}
$\varepsilon_1 = 0$:	(rmse_{TMIL}, rmse_{MILD})[a] = (18.55, 18.61) [0.04], ⟨bias_{TMIL}, bias_{MILD}⟩[a] = ⟨1.80, 3.86⟩, average $\left\{\min \bar{\Delta}_H, \min |\hat{\mathbf{\Sigma}}|_H^{1/p}\right\}$ = {8.71, 0.98}		

[a] Values have been multiplied by 1000.

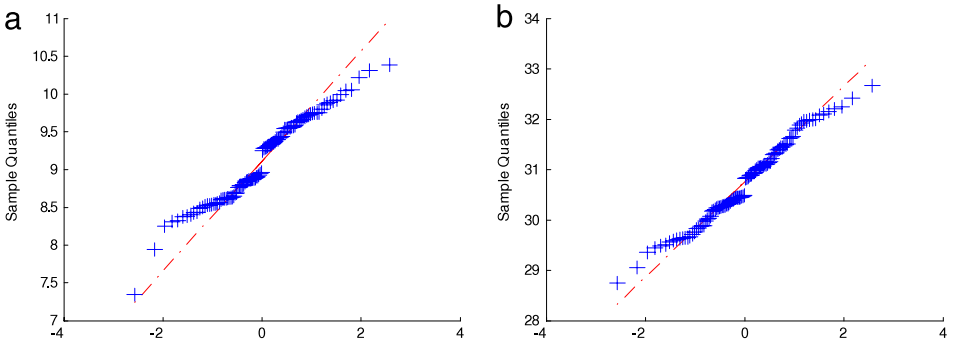

Fig. 1. Representative qq-plots of (a) $\{t_j\}_{j=1}^n$, and (b) $\{d_j\}_{j=1}^n$ for 'clean' samples.

(iii) The time required when the genetic steps were included averaged about 4.5 times that without these steps. This is perhaps a heavy price to pay, especially given that in almost all cases there was no significant reduction in the loss realized beyond the first generation of subsamples. We view this as testament to the efficiency of our method of choosing this first generation.

As diagnostic tools we computed as well, for $j = 1, \ldots, n$,

$$t_j = \Delta\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{TMIL}\right), \tag{7a}$$

$$d_j = \Omega\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{MILD}\right), \tag{7b}$$

in each sample and prepared qq-plots of the results. Some representative plots – each from one of the N samples – are shown in Figs. 1–4; we have found them to be useful discriminators between those h matrices $\mathbf{S}_j$ near the centre of the sample as defined by $\hat{\mathbf{\Sigma}}$ – these have smaller values of these measures t_j and d_j – and those $n - h$ that are farther away. Note that $t_j \geq 0$, with equality if and only if $\mathbf{S}_j = \hat{\mathbf{\Sigma}}_{TMIL}$, yielding the interpretation of t_j as a measure of the distance between $\mathbf{S}_j$ and the centre of the subsample. In view of a remark made earlier – that d_j is the *average* of the Mahalanobis distances associated with the vectors from which $\mathbf{S}_j$ is computed – we would not expect this measure to identify the contaminated vectors themselves. We would however expect the d_j, and by extension the t_j, to be approximately normally distributed, and so we plot against quantiles of the Normal distribution. See also Hardin and Rocke (2005), who investigate outlier detection and removal methods after plotting the individual Mahalanobis distances against chi-square quantiles.

5. Case study revisited

In the study described in Section 1, a portion of which is detailed here, 43-channel recordings were obtained from each of 68 healthy, male, dextral volunteers with a mean age of 26.8 ($s.d. = 7.9$) years. Recordings for each individual were obtained for a period of 3 min, during which time participants were required to look at a spot projected onto a screen at a fixed location. Each channel was digitized at a rate of 256 samples/s. using a 12-bit analog to digital converter. The 43 sensor locations represent standard sensor positions described in the American Electroencephalography Society Guidelines (Sharbrough et al., 1990). The sensor locations were AF3, AF4, FC5, FC6, FC1, FC2, C3, C4, C1, C2, TP7, TP8, CP5, CP6, CP1, CP2, P3, P4, P1, P2, PO3, PO4, AFz, Fz, FCz, Cz, CPz, Pz, POz, FP1, FP2, F7, F8, F3, F4, FT7, FT8, T7, T8, P7, P8, O1 and O2. All recordings were made relative to a left-ear reference. The digitized data were visually edited off-line in an attempt to select 20, 1-s artefact-free data segments. Data segments containing voltage spikes, large eye movements or muscle activity were not selected for analysis. The fast Fourier transform (FFT) was applied to each channel following the subtraction of the mean and the application of

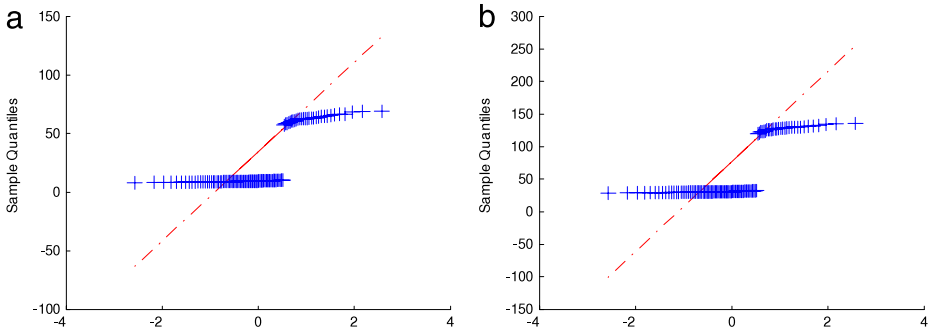

Fig. 2. Representative qq-plots of (a) $\{t_j\}_{j=1}^n$, and (b) $\{d_j\}_{j=1}^n$; $(\varepsilon_1, \varepsilon_2) = (0.3, 0.4)$ and $f_{small} = 0$.

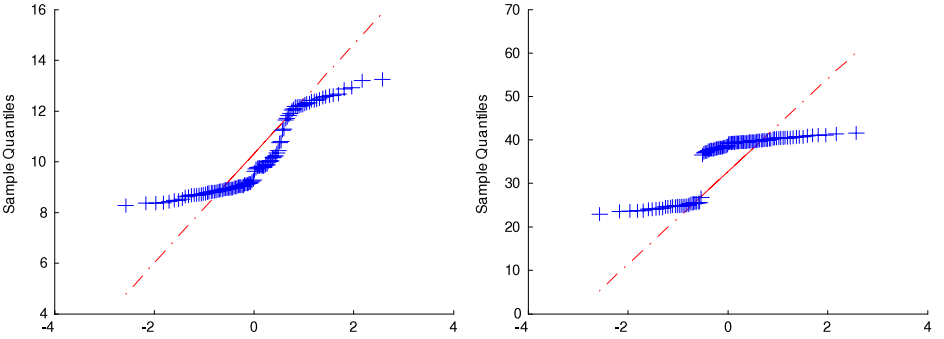

Fig. 3. Representative qq-plots of (a) $\{t_j\}_{j=1}^n$, and (b) $\{d_j\}_{j=1}^n$; $(\varepsilon_1, \varepsilon_2) = (0.3, 0.4)$ and $f_{small} = 1$.

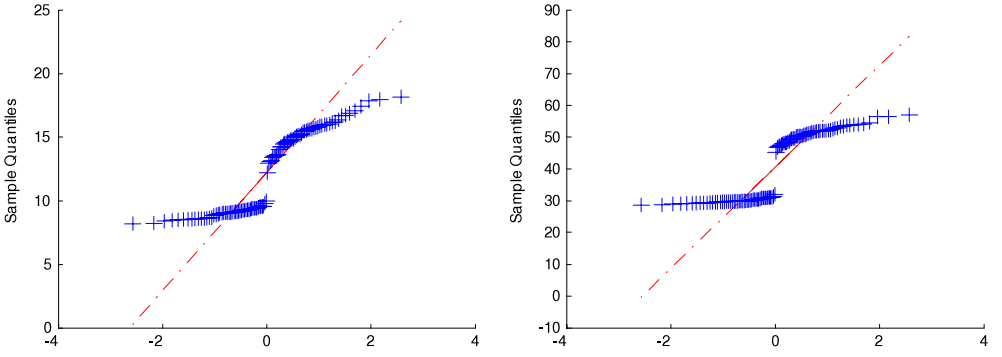

Fig. 4. Representative qq-plots of (a) $\{t_j\}_{j=1}^n$, and (b) $\{d_j\}_{j=1}^n$; $(\varepsilon_1, \varepsilon_2) = (0.5, 0.2)$ and $f_{small} = 0.5$.

a 50% Hamming taper $\{h_t\}$ (Brillinger, 1981, p. 55). For each individual these transforms were then averaged over the 20 segments and over the 6 frequencies in the band from 8 to 13 Hz. Specifically, let $\left\{\mathbf{x}_t^{(j,s)} | t = 1, \ldots, 256, s = 1, \ldots, 20\right\}$ be the centred, artefact-free segments for the jth participant ($j = 1, \ldots, 68$). The FFT (of the tapered data) at frequency ω is

$$\mathbf{X}^{(j,s)}(\omega) = \frac{1}{\sqrt{256}} \sum_{t=1}^{256} h_t \mathbf{x}_t^{(j,s)} e^{-2\pi i \omega t}.$$

We form 43×120 matrices $\mathbf{X}^{(j)}$, whose columns are the vectors $\mathbf{X}^{(j,s)}(\omega)$ for $s = 1, \ldots, 20$ and $\omega = 8/256, 9/256, \ldots, 13/256$. We then compute $\mathbf{S}_j = \mathbf{X}^{(j)}\mathbf{X}^{(j)*}/120$. The $p \times p$ ($p = 43$) Hermitian matrices $\mathbf{S}_j$ summarize the 'alpha band' for each of the $n = 68$ subjects.

Both the TMIL and MILD estimates were computed using a value of h determined by comparing qq-plots for $h = n - 1, n - 2, \ldots$, until additional trimming failed to show an improvement in the linear fit of the estimate. After an examination of quantile plots obtained from the discrimination measures (7), 27 of the 68 matrices were identified as outliers with respect to the TMIL measure and removed, yielding a trimmed sample of 41 matrices. Similarly, 17 matrices were removed after computing the MILD estimate. Quantile plots for the untrimmed and trimmed samples appear in Fig. 5.

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

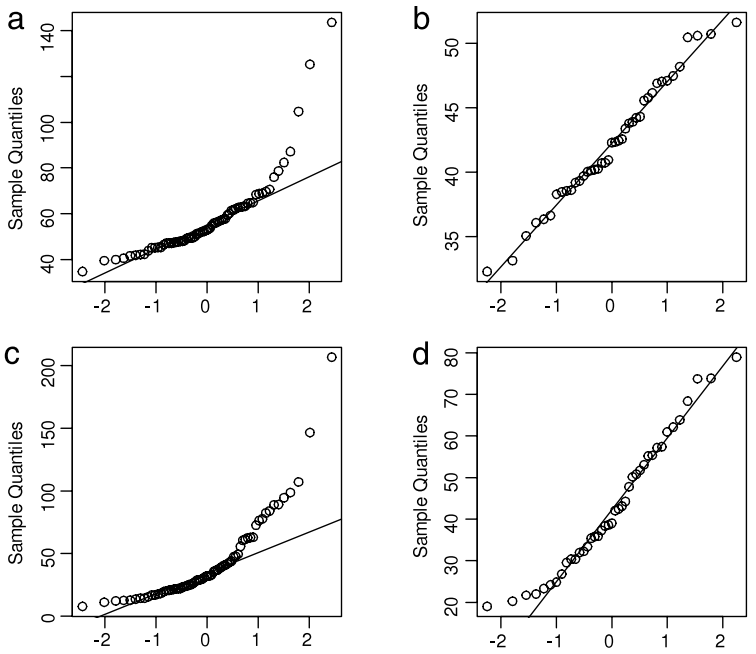

Fig. 5. Top: qq-plots of $\left\{t_j = \Delta\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{TMIL}\right)\right\}$ for the (a) untrimmed and (b) trimmed samples of EEG spectral matrices. Bottom: qq-plots of $\left\{d_j = \Delta\left(\mathbf{S}_j, \hat{\mathbf{\Sigma}}_{MILD}\right)\right\}$ for the (c) untrimmed and (d) trimmed samples.

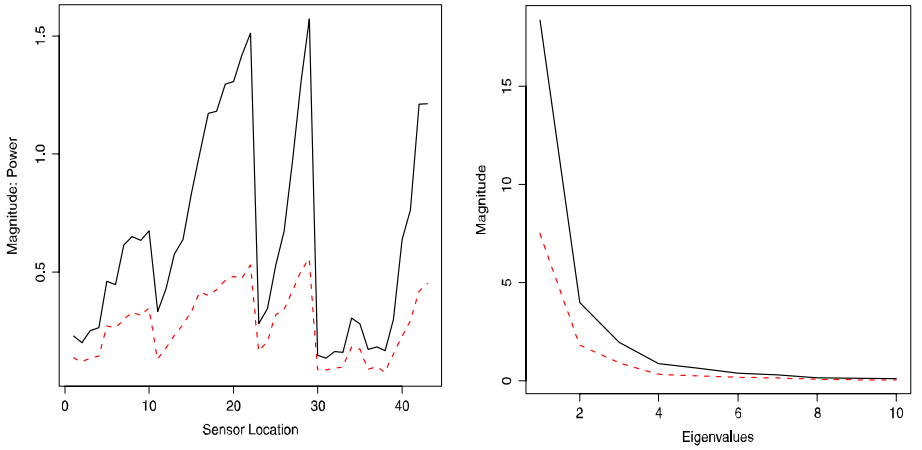

Fig. 6. Left: Diagonal elements of mean cross-spectrum matrices obtained from the untrimmed (solid line) and trimmed (broken line) samples of EEG spectral matrices. Right: First 10 eigenvalues of the mean cross-spectrum matrices obtained from the untrimmed (solid line) and trimmed (broken line) samples of EEG spectral matrices.

Heaviness in the lower tail shown in Fig. 5(d) suggests the presence of 'inliers' in the data that were undetected by the MILD estimate. The TMIL estimate therefore seems preferable, and so the rest of the output is for this estimate only.

Fig. 6 (left) contains the diagonal elements, or spectral power values at each sensor location, of the mean of the cross-spectral matrices for the trimmed and untrimmed matrix samples. A reduction in spectral power of the trimmed matrix estimate is observed across all sensor locations with a large reduction in spectral power occurring at some sites. The reduction in overall spectral power is reflected in the corresponding reduction in the magnitude of the eigenvalues of the mean trimmed matrix sample shown in Fig. 6 (right).

In order to examine the sensitivity of the eigenvectors to the presence of outliers, plots of the real and imaginary components for the first 3 eigenvectors are shown in Fig. 7. Using the standard method for normalizing principal components, the eigenvectors were normalized for the ith variable and jth eigenvalue λ_j, according to $a'_{ij} = a_{ij}\sqrt{\lambda_j}/s_i$ which yields the coherence of a_{ij} with the jth principal component. The first eigenvector appears to be the most sensitive to the effects of trimming as shown in Fig. 7(a), (d) where differences between the trimmed and untrimmed samples are present

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

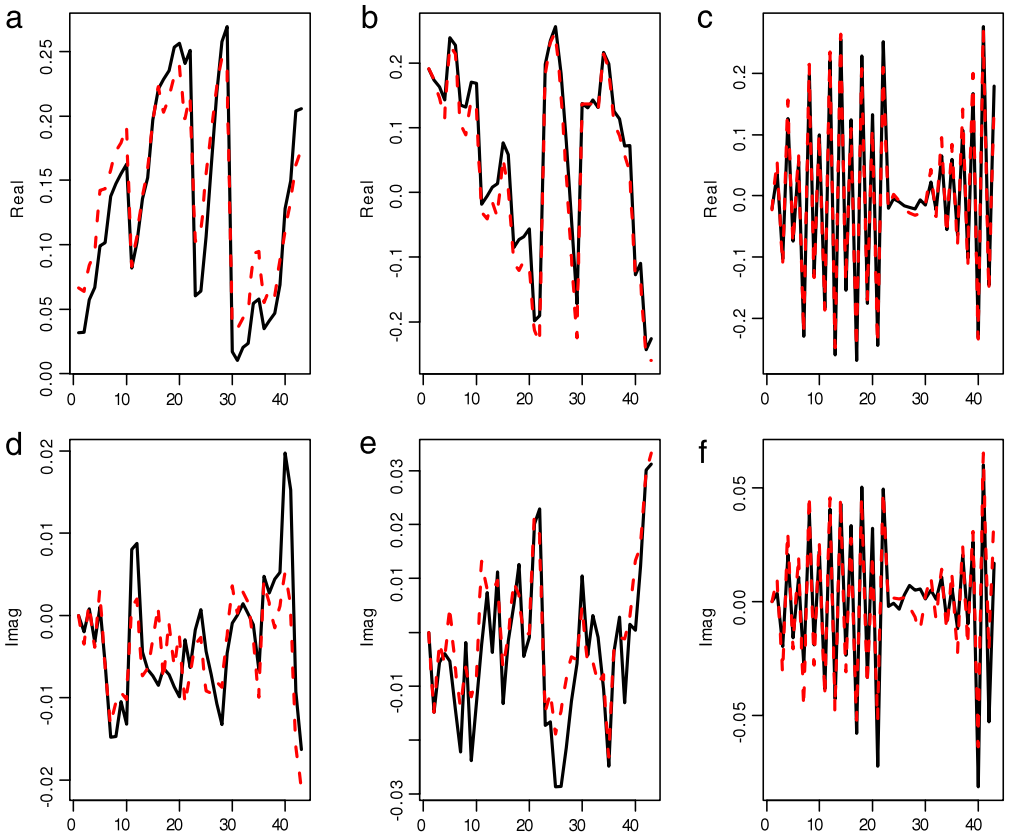

Fig. 7. (a), (b), (c): Real parts of the elements of the first three eigenvectors plotted against the sensor locations {1, 2, . . . , 43}. (d), (e), (f): Imaginary parts of the elements of these eigenvectors plotted against the sensor locations. Solid lines correspond to untrimmed samples of EEG spectral matrices, and broken lines to trimmed samples.

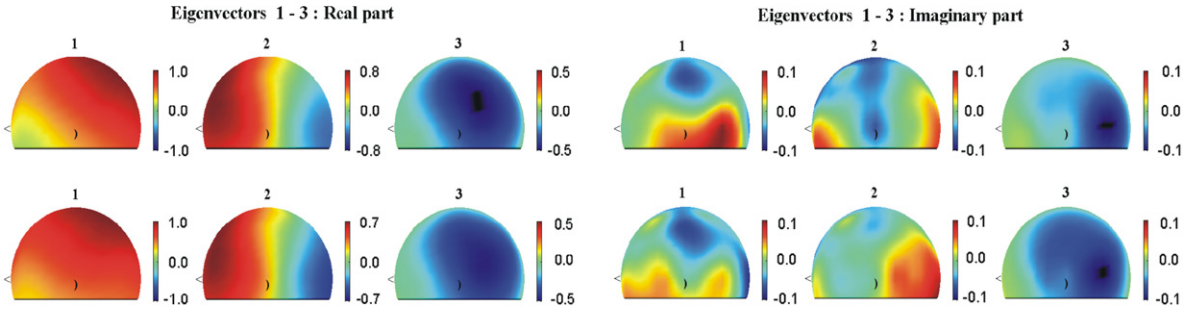

Fig. 8. Left side view of spherical spline interpolated eigenvector coefficients from mean cross-spectrum matrices. Top row: untrimmed matrix sample. Bottom row: trimmed matrix sample.

across all sensor locations in both the real and imaginary components. For all 3 vectors the imaginary parts appear to be the most sensitive to the effects of trimming. The oscillations observed in Fig. 7(c), (f) are the result of positive eigenvector coefficients for sensor locations on the right side of the scalp, while the corresponding locations on the left side have negative values.

The eigenvector differences between the trimmed and untrimmed mean spectral matrices are summarized in the form of topographical displays shown in Figs. 8 and 9. These displays represent a spherical approximation to the surface of the scalp and were constructed using spherical spline interpolation (Wahba, 1981) through the 43 sensor locations. An examination of these displays shows that differences in the real part of the first eigenvector appear to be associated with the frontal regions of the head, while the largest differences in the imaginary components are in the temporal regions associated with the first two eigenvectors.

Overall, the results from this example indicate that sets of cross-spectral matrices are sensitive to the presence of outlying or unusual matrices in the data. The results further show that an important property of the TMIL estimate is that it is effective

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

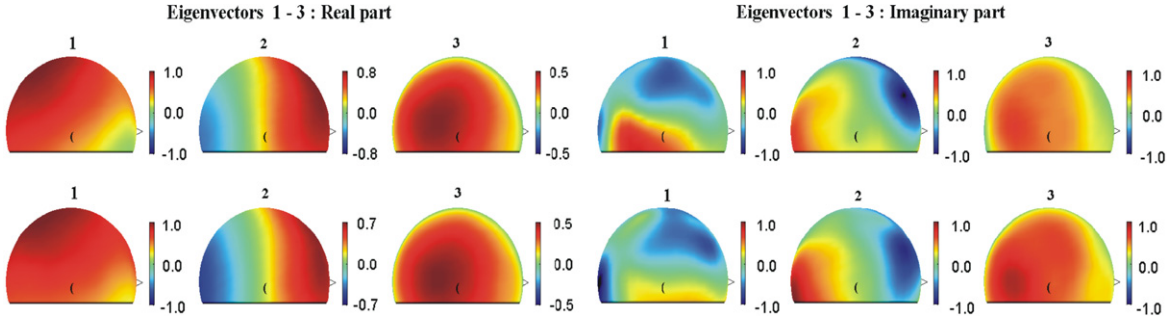

Fig. 9. Right side view of spherical spline interpolated eigenvector coefficients from mean cross-spectrum matrices. Top row: untrimmed matrix sample. Bottom row: trimmed matrix sample.

for the identification of matrices that can be considered inliers. The value of using these techniques in practice is supported by the fact that these methods were able to identify subsets of matrix outliers even though the data had been visually edited for gross artefacts prior to analysis. This has implications for subsequent analyses such as PCA, or methods based on PCA, as well as EEG or MEG imaging methods that rely on the analysis of eigenvector subsets. The use of robust estimates of the spectrum and cross-spectrum therefore provides a means of improving the reliability of experimental findings with large array recordings, and this will in turn aid in improving the understanding of brain pathology and function.

Acknowledgements

The research of D. Wiens is supported by the Natural Sciences and Engineering Research Council of Canada, and was largely carried out while enjoying the hospitality of the Departmento de Matemática, Facultad de Ciencias Exactas y Naturales at the Universidad de Buenos Aires. Victor Yohai was partially supported by Grants X-018 from the University of Buenos Aires, PIP 5505 from CONICET, Argentina and PICT 00899 from ANPCYT, Argentina. We are grateful to Zoltan Koles, of the Department of Electrical and Computer Engineering at the University of Alberta, for helpful suggestions regarding the EEG recordings in the case study. We are also grateful to Pierre Flor-Henry, of the Centre for Psychiatric Assessment and Therapeutics, Alberta Hospital Edmonton, for providing the data used in the case study. The research has benefited from the incisive comments of several anonymous reviewers.

Appendix. Derivations

Proof of Theorem 1. Inequality (5) follows from

$$\bar{\Delta}_{H_{(1)}} = \frac{1}{h}\sum_{j\in H_{(1)}} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) \geq \frac{1}{h}\sum_{j\in H_{(2)}} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) \geq \frac{1}{h}\sum_{j\in H_{(2)}} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(2)}}\right) = \bar{\Delta}_{H_{(2)}}. \tag{A.1}$$

The first inequality in (A.1) follows from the definition of $H_{(2)}$ as the set of minimizing indices, and the second is the assertion that for any H, $h^{-1}\sum_{j\in H}\Delta\left(\mathbf{S}_j, \boldsymbol{\Sigma}\right)$ is minimized by $\boldsymbol{\Sigma} = \bar{\mathbf{S}}_H$. If equality holds in (5) then the inequalities in (A.1) must be equalities, and then $\bar{\mathbf{S}}_{H_{(1)}} = \bar{\mathbf{S}}_{H_{(2)}}$ follows from the fact that $h^{-1}\sum_{j\in H_{(2)}}\Delta\left(\mathbf{S}_j, \hat{\boldsymbol{\Sigma}}\right)$ is *uniquely* minimized by $\bar{\mathbf{S}}_{H_{(2)}}$. □

Proof of Lemma 1. Take a sample $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$ such that the members $\mathbf{S}_1, \ldots, \mathbf{S}_{v_0}$ have all elements equal to 0 and the remaining members are arbitrary, but positive definite, Hermitian matrices. It suffices to show that for any $m \geq (n - v_0)/2$ there is a sequence of contaminated samples $\mathcal{S}_k \in \mathcal{SS}_m$ culminating in breakdown, i.e. with

$$\frac{1}{ch_{\min}\left(\hat{\Sigma}_n\left(\mathcal{S}_k\right)\right)} + ch_{\max}\left(\hat{\Sigma}_n\left(\mathcal{S}_k\right)\right) \to \infty. \tag{A.2}$$

Thus let $m \geq (n - v_0)/2$ and put

$$\mathcal{S}_k^{(1)} = \left\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{v_0}, k\mathbf{S}_{v_0+1}, \ldots, k\mathbf{S}_{v_0+m}, \mathbf{S}_{v_0+m+1}, \ldots, \mathbf{S}_n\right\}, \quad 1 \leq k < \infty.$$

Define $L = \sup_{1<k<\infty} ch_{\max}\left(\hat{\Sigma}_n\left(\mathcal{S}_k^{(1)}\right)\right)$. The number of members of $\mathcal{S}_k^{(1)}$ not in $\mathcal{S}$ is equal to m, so that $\mathcal{S}_k^{(1)} \in \mathcal{SS}_m$; therefore if $L = \infty$ we are through. Suppose on the contrary that $L < \infty$. In that case we consider the sequence of samples $\mathcal{S}_k^{(2)} = \mathcal{S}_k^{(1)}/k$, i.e.

$$\mathcal{S}_k^{(2)} = \left\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{v_0}, \mathbf{S}_{v_0+1}, \ldots, \mathbf{S}_{v_0+m}, \frac{\mathbf{S}_{v_0+m+1}}{k}, \ldots, \frac{\mathbf{S}_n}{k}\right\} \in \mathcal{SS}_m.$$

The number of members of $\mathcal{S}_k^{(2)}$ not in $\mathcal{S}$ is equal to $n - v_0 - m \leq m$ and so $\mathcal{S}_k^{(2)} \in \mathcal{S}\mathcal{S}_m$. Moreover, by scale equivariance,

$$ch_{\min}\left(\hat{\Sigma}_n\left(\mathcal{S}_k^{(2)}\right)\right) \leq ch_{\max}\left(\hat{\Sigma}_n\left(\mathcal{S}_k^{(2)}\right)\right) = ch_{\max}\left(\frac{\hat{\Sigma}_n\left(\mathcal{S}_k^{(1)}\right)}{k}\right) \leq L/k,$$

and therefore $ch_{\min}\left(\hat{\Sigma}_n\left(\mathcal{S}_k^{(2)}\right)\right) \to 0$. This establishes (A.2). □

Proof of Theorem 2. Let $\lambda_{01}(\mathcal{S}) = \min\{ch_{\min}(S_j) : 1 \leq j \leq n, \lambda_1(S_j) > 0\}$ be the minimum of the smallest eigenvalues of the positive definite members of $\mathcal{S}$. It is enough to show that if $m < n - h - v(\mathcal{S})$ then there exist constants $\gamma_- = \gamma_-(\mathcal{S})$ and $\gamma_+ = \gamma_+(\mathcal{S})$ for which

$$\inf_{\mathcal{S}^\dagger \in \mathcal{S}\mathcal{S}_m} ch_{\min}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right) \geq \gamma_- > 0, \tag{A.3}$$

and

$$\sup_{\mathcal{S}^\dagger \in \mathcal{S}\mathcal{S}_m} ch_{\max}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right) \leq \gamma_+ < \infty, \tag{A.4}$$

since then $m(\mathcal{S}, \hat{\mathbf{\Sigma}}_n) \geq n - h - v(\mathcal{S})$.

Take $m < n - h - v(\mathcal{S})$ and let $\mathcal{S}^\dagger \in \mathcal{S}\mathcal{S}_m$. Since $m + v(\mathcal{S}) < n - h \leq h$, for any $H \in \mathcal{H}$ there exists $j \in H$ such that $\mathbf{S}_j^\dagger = \mathbf{S}_j \in \mathcal{S}$ with $\mathbf{S}_j$ positive definite. Then $ch_{\min}(\mathbf{S}_j^\dagger) \geq \lambda_{01}$ and so $ch_{\min}\left(\bar{S}_H\right) \geq \lambda_{01}/h$. Since this lower bound does not depend on $\mathcal{S}^\dagger$ we have that $\inf_{\mathcal{S}^\dagger \in \mathcal{S}\mathcal{S}_m} ch_{\min}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right) \geq \lambda_{01}/h$, and so (A.3) holds with $\gamma_-(\mathcal{S}) = \lambda_{01}(\mathcal{S})/n$.

For (A.4), recall that the estimate is obtained by running Algorithm 1 repeatedly, initialized each time by one of finitely many sets $H^{(a)}$. Since $n - m - v(\mathcal{S}) > h$, we can find a set $H_0 \in \mathcal{H}$ such that for all $j \in H_0$ we have $\mathbf{S}_j^\dagger = \mathbf{S}_j \in \mathcal{S}$ with $\mathbf{S}_j$ positive definite. Thus there is at least one set $H^{(a)}$ for which $\bar{\mathbf{S}}_{H^{(a)}}$ is positive definite. For each such set, and in the notation of Theorem 1, we set $H_{(1)} = H^{(a)}$ and construct a sequence $\left\{H_{(k)}\right\}$ for which the sequence $\left\{\frac{1}{h}\sum_{j \in H_{(k+1)}} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(k)}}\right)\right\}$ is bounded. In particular, there is a positive constant $\alpha = \alpha\left(H^{(a)}\right)$ for which $\alpha \geq \max_{j \in H_{(k+1)}} \Delta\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(k)}}\right) = \max_{j \in H_{(k+1)}} \sum_{l=1}^p f\left(\lambda_{l,j}\right)$, where $f(\cdot)$ is as at (1) and $\left\{\lambda_{l,j}\right\}_{l=1}^p$ is the set of eigenvalues of $\bar{\mathbf{S}}_{H_{(k)}}^{-1}\mathbf{S}_j$. Since f is non-negative, $f\left(\lambda_{l,j}\right) \leq \alpha$ uniformly in j, l. Then if $\lambda_+ = \lambda_+\left(H^{(a)}\right) > 1$ is the largest root of the equation $f(\lambda) = \alpha$, we have, for all $j \in H_{(k+1)}$, that $ch_{\max}\left(\bar{\mathbf{S}}_{H_{(k)}}^{-1}\mathbf{S}_j\right) \leq \lambda_+$. Thus

$$ch_{\max}\left(\bar{\mathbf{S}}_{H_{(k)}}^{-1}\bar{\mathbf{S}}_{H_{(k+1)}}\right) \leq \frac{1}{h}\sum_{j \in H_{(k+1)}} ch_{\max}\left(\bar{\mathbf{S}}_{H_{(k)}}^{-1}\mathbf{S}_j\right) \leq \lambda_+,$$

and so $\left|\bar{\mathbf{S}}_{H_{(k)}}^{-1}\bar{\mathbf{S}}_{H_{(k+1)}}\right| \leq \lambda_+^p$, i.e. $\left|\bar{\mathbf{S}}_{H_{(k+1)}}\right| \leq \lambda_+^p \left|\bar{\mathbf{S}}_{H_{(k)}}\right|$. Iterating this inequality gives

$$\left|\bar{\mathbf{S}}_{H_{(k+1)}}\right| \leq \lambda_+^{kp}\left|\bar{\mathbf{S}}_{H_{(1)}}\right|.$$

Now let k_a be the number of iterations required, in Algorithm 1, before convergence is declared. This in the last inequality gives

$$\left|\bar{\mathbf{S}}_{H_{(k+1)}}\right| \leq \lambda_+^{k_a p}\left|\bar{\mathbf{S}}_{H_{(1)}}\right| = \left[\lambda_+\left(H^{(a)}\right)\right]^{k_a p}\left|\bar{\mathbf{S}}_{H^{(a)}}\right|.$$

Since the estimate is the limit of one of the sequences $\left\{\bar{\mathbf{S}}_{H_{(k)}}\right\}$, we have

$$\left|\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right| \leq \beta(\mathcal{S}), \tag{A.5}$$

for

$$\beta = \max_a\left\{\left[\lambda_+\left(H^{(a)}\right)\right]^{k_a p}\left|\bar{\mathbf{S}}_{H^{(a)}}\right|\right\}.$$

By this and (A.3),

$$\beta \geq \left|\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right| = \prod_{i=1}^p \lambda_i\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right) \geq \gamma_-^{p-1} ch_{\max}\left(\hat{\mathbf{\Sigma}}_n(\mathcal{S}^\dagger)\right),$$

and then (A.4) holds with $\gamma_+(\mathcal{S}) = \beta(\mathcal{S})/[\gamma_-(\mathcal{S})]^{p-1}$. □

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

Proof of Proposition 1. For $\mu > 0$ define

$$k(\mu) = g(\mu\boldsymbol{\Sigma}) = \frac{1}{\mu}\Omega(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}) + p\log\mu - E\left[\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{S}\right|\right] - p.$$

It suffices to show that if μ_0 minimizes $k(\cdot)$, then $\boldsymbol{\Sigma}' \overset{\text{def}}{=} \mu_0\boldsymbol{\Sigma}$ satisfies $\Omega\left(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}'\right) = p$. For this we note that the only critical point of $k(\cdot)$ is $\mu_0 = \Omega(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma})/p$, with $\Omega(\boldsymbol{\Sigma}_0, \mu_0\boldsymbol{\Sigma}) = \Omega(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma})/\mu_0 = p$, as required. Since $k(\mu) \to \infty$ as $\mu \to 0, \infty$, μ_0 furnishes a minimum of $k(\cdot)$. □

Before proving Theorem 4 we must establish two preliminary results. The first characterizes $\boldsymbol{\Sigma}_0$ as the solution to a minimization problem; it is analogous to a result of Grübel (1988) which gives a motivation for the MCD estimator.

Lemma 3. *Among all $\boldsymbol{\Sigma} > \mathbf{0}$ satisfying (6), $|\boldsymbol{\Sigma}|$ is uniquely minimized by $\boldsymbol{\Sigma}_0$.*

Proof of Lemma 3. First note that $\Omega(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_0) = \mathrm{tr}\mathbf{I}_p = p$, so that $\boldsymbol{\Sigma}_0$ satisfies (6). Now suppose that $\boldsymbol{\Sigma} > \mathbf{0}$ satisfies (6) and let $\{\lambda_j\}_{j=1}^p$ be the eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0$. By the arithmetic–geometric mean inequality,

$$1 = \frac{1}{p}\Omega(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}) = \frac{1}{p}\sum_{j=1}^p \lambda_j \geq \left(\prod_{j=1}^p \lambda_j\right)^{1/p} = \left|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right|^{1/p},$$

so that $|\boldsymbol{\Sigma}| \geq |\boldsymbol{\Sigma}_0|$. The inequality is an equality iff all λ_j are equal, necessarily to 1, so that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. □

Taking expectations with respect to the empirical distribution gives the following corollary.

Corollary 1. *For $H \in \mathcal{H}$, among all $\boldsymbol{\Sigma} > \mathbf{0}$ such that $\bar{\Omega}_{H,\boldsymbol{\Sigma}} = p$, $|\boldsymbol{\Sigma}|$ is uniquely minimized by $\bar{\mathbf{S}}_H$.*

Proof of Theorem 4. Assume that $\left|\bar{\mathbf{S}}_{H_{(2)}}\right| > 0$, else there is nothing to prove. That $H_{(m)} \in \mathcal{H}_1$ for both $m = 1$ and $m = 2$ is true by construction, since in each case $\boldsymbol{\Sigma}_{(m)} = \bar{\mathbf{S}}_{H_{(m)}}$ is the average of the members of $H_{(m)}$. By the definition of $H_{(2)}$ as the set of *minimizing* indices,

$$\lambda \overset{\text{def}}{=} \frac{1}{hp}\sum_{j\in H_{(2)}} \Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) \leq \frac{1}{hp}\sum_{j\in H_{(1)}} \Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) = 1. \tag{A.6}$$

Note also that $\lambda > 0$, else $\Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) = 0$ for all $j \in H_{(2)}$ and hence $\left|\bar{\mathbf{S}}_{H_{(2)}}\right| = 0$. By the definition of λ,

$$\frac{1}{hp}\sum_{j\in H_{(2)}} \Omega\left(\mathbf{S}_j, \lambda\bar{\mathbf{S}}_{H_{(1)}}\right) = \frac{1}{\lambda hp}\sum_{j\in H_{(2)}} \Omega\left(\mathbf{S}_j, \bar{\mathbf{S}}_{H_{(1)}}\right) = \frac{\lambda}{\lambda} = 1.$$

Thus

$$\frac{1}{h}\sum_{j\in H_{(2)}} \Omega\left(\mathbf{S}_j, \lambda\bar{\mathbf{S}}_{H_{(1)}}\right) = p,$$

so that by the inequality established in (A.6), followed by Corollary 1,

$$\left|\bar{\mathbf{S}}_{H_{(1)}}\right| \geq \left|\lambda\bar{\mathbf{S}}_{H_{(1)}}\right| \geq \left|\bar{\mathbf{S}}_{H_{(2)}}\right|,$$

with equality iff $\bar{\mathbf{S}}_{H_{(2)}} = \lambda\bar{\mathbf{S}}_{H_{(1)}}$ and $\lambda = 1$, i.e. iff $\bar{\mathbf{S}}_{H_{(1)}} = \bar{\mathbf{S}}_{H_{(2)}}$. □

Proof of Theorem 5. It suffices to establish (A.3) and (A.4). That (A.3) holds is shown exactly as in the proof of Theorem 2. For (A.4), first let $H_0 \in \mathcal{H}$ be as in the proof of that theorem. Then (A.4) will follow, exactly as before, if we establish a bound of the form (A.5). For this, let $\lambda_{0p} = \lambda_{0p}(\mathcal{S}) = \max_{H\in\mathcal{H}}\{ch_{\max}(\bar{S}_H)\}$ be the maximum of the largest eigenvalues of the h-element averages of the members of $\mathcal{S}$. Then $ch_{\max}\left(\bar{\mathbf{S}}_{H_0}\right) \leq \lambda_{0p}$, and therefore $\left|\bar{\mathbf{S}}_{H_0}\right| \leq \lambda_{0p}^p$. Since $\hat{\boldsymbol{\Sigma}}_n(\mathcal{S}^\dagger)$ minimizes $\left|\bar{\mathbf{S}}_H\right|$ we have $\left|\hat{\boldsymbol{\Sigma}}_n(\mathcal{S}^\dagger)\right| \leq \lambda_{0p}^p$; thus (A.5) holds with $\beta(\mathcal{S}) = \lambda_{0p}^p$. □

References

Borgiotti, G.V., Kaplan, L.J., 1979. Superresolution of uncorrelated interference sources using adaptive array techniques. IEEE Transactions on Antennas and Propagation 27, 842–845.
Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden-Day.
Grübel, R., 1988. A minimal characterization of the covariance matrix. Metrika 35, 49–52.
Hardin, J., Rocke, D.M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14, 928–946.
Kakizawa, Y., Shumway, R.H., Taniguchi, M., 1998. Discrimination and clustering for multivariate time series. Journal of the American Statistical Association 93, 328–340.

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011

Kleiner, B., Martin, R.D., Thomson, D.J., 1979. Robust estimation of power spectra. Journal of the Royal Statistical Society. Series B 41, 313–351.
Mitra, A., Kundu, D., Agrawal, G., 2006. Frequency estimation of undamped exponential signals using genetic algorithms. Computational Statistics & Data Analysis 51, 1965–1985.
Mosher, J.C., Lewis, P.S., Leahy, R.M., 1992. Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Transactions on Biomedical Engineering 39, 541–557.
Nunez, P.L., Srinivasan, R., 2006. Electric Fields of the Brain: The Neurophysics of EEG, second ed. Oxford University Press.
Pascual-Marqui, R.D., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity of the brain. International Journal of Psychophysiology 18, 49–65.
Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications. Reidel, Dordrecht, pp. 283–297.
Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley.
Rousseeuw, P.J., van Driessen, K., 1999. A Fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Sharbrough, F., Chatrian, G., Lesser, R.P., Luders, H., Nuwer, M., Picton, T.W., 1990. Guidelines for Standard Electrode Position Nomenclature. Electroencephalographic Society, Bloomfield, CT.
Shumway, R.H., Stoffer, D.S., 2006. Time Series Analysis and its Applications: With R Examples. Springer.
Srivastava, M.S., Khatri, C.G., 1979. An Introduction to Multivariate Statistics. North Holland.
Stoffer, D.S., 1999. Detecting common signals in multiple time series using the spectral envelope. Journal of the American Statistical Association 94, 1341–1356.
Todorov, V., 1992. Computing the minimum covariance determinant estimator (MCD) by simulated annealing. Computational Statistics & Data Analysis 14, 515–525.
Wahba, G., 1981. Spline interpolation and smoothing on a sphere. SIAM Journal on Scientific and Statistical Computing 2, 5–16.
Welsh, A.H., Wiens, D.P., 2011. Robust model-based sampling designs. Statistics and Computing, in press (http://dx.doi.org/10.1007/s11222-012-9339-3).

Please cite this article in press as: Lind, J.C., et al., Robust minimum information loss estimation. Computational Statistics and Data Analysis (2012), doi:10.1016/j.csda.2012.06.011