On the computation of symmetrized M-estimators of scatter

Jari Miettinen, Klaus Nordhausen, Sara Taskinen and David E. Tyler

Abstract This paper focuses on the computational aspects of symmetrized M-estimators of scatter, i.e. the multivariate M-estimators of scatter computed on the pairwise differences of the data. Such estimators do not require a location estimate, and more importantly, they possess the important block and joint independence properties. These properties are needed, for example, when solving the independent component analysis problem. Classical and recently developed algorithms for computing the M-estimators and the symmetrized M-estimators are discussed. The effect of parallelization is considered as well as new computational approach based on using only a subset of pairwise differences. Efficiencies and computation time comparisons are made using simulation studies under multivariate elliptically symmetric models and under independent component models.

1 Introduction

Almost all of the classical multivariate methods, including principal component analysis, multivariate regression, canonical correlation analysis, etc., are dependent on the use of the sample covariance matrix. It is well known that under the assumption of multivariate normality, the methods based on this estimator are optimal. However, if the normality assumption is not satisfied, e.g., if the data are con-

Jari Miettinen and Sara Taskinen
Department of Mathematics and Statistics, University of Jyvaskyla, 40014 Jyvaskyla, Finland e-mail: jari.p.miettinen@jyu.fi, sara.l.taskinen@jyu.fi

Klaus Nordhausen
Department of Mathematics and Statistics, University of Turku, 20014 Turku, and School of Health Sciences, University of Tampere, 30014 Tampere, Finland, e-mail: klaus.nordhausen@utu.fi

David E. Tyler
Department of Statistics, Rutgers University, NJ 08854, USA, e-mail: dtyler@rci.rutgers.edu

taminated with outlying observations or have heavier tails than that of the normal distribution, then methods based on the sample covariance matrix perform poorly.

A widely used approach for robustifying classical multivariate methods is the so-called plug-in approach. In this approach, the sample covariance matrix is replaced by a robust scatter matrix. As a consequence a vast variety of robust alternatives for the sample covariance matrix have been proposed in the literature. Some widely used robust estimators include M-estimators (Maronna, 1976; Huber, 1981), MCD-estimators (Rousseeuw, 1985) and S-estimators (Davies, 1987; Lopuhaä, 1989), among others. For an overview of robust multivariate methods, see Maronna et al (2006).

When robust plug-in methods are proposed, one important issue is often ignored, namely that a multivariate method may not be valid unless the robust scatter matrix satisfies certain crucial properties that hold for the sample covariance matrix. In Nordhausen and Tyler (2015) a thorough discussion of such properties is given. Focusing on the so-called joint and block independence properties (defined in the next section), Nordhausen and Tyler (2015) give several examples of plug-in multivariate methods, which are not valid unless the scatter matrix possesses these properties. Examples include independent component analysis, observational regression, and graphical modeling. For the role of scatter matrices in independent component analysis, see also Oja et al (2006), Nordhausen et al (2008), and Tyler et al (2009), among others.

In Oja et al (2006) it is shown that by computing any scatter matrix using pairwise differences rather than the observations themselves produces an estimator with the joint independence property. Sirkiä et al (2007) discuss general symmetrized M-estimators, and give as examples the symmetrized Huber estimators, and Dümbgen's (1998) estimator, which is a symmetrized version of Tyler's (1987) M-estimator. Croux et al (1994) and Roelant et al (2009) propose using symmetrized S-estimators in univariate and multivariate regression settings, respectively, with their main focus being on improving efficiency at the normal model.

As symmetrized estimators are defined using pairwise differences, the computations become intensive with increasing sample size. In this paper we focus on the computational aspects and consider a few practical ways to handle this problem, especially in the context of M-estimates. The paper is organized as follows. In Section 2 we recall the definitions of scatter matrix and block and joint independence, and in Section 3 the definition and main properties of symmetrized M-estimators of scatter. Section 4 provides some new approaches for computing symmetrized estimators. In Sections 5 and 6 simulation studies are given to compare efficiencies and computation times of different approaches, respectively. The paper is concluded with some discussion in Section 7.

2 Scatter matrices and block independence

Recall first the definition of a scatter matrix functional.

Definition 1. Let $\mathbf{x}$ be a p-variate random vector with cumulative distribution function $F_{\mathbf{x}}$. Then a $p \times p$ matrix valued functional $\mathbf{V} = \mathbf{V}(F_{\mathbf{x}})$ is a *scatter matrix functional* if it is symmetric, positive semi-definite and affine equivariant in the sense that

$$\mathbf{V}(F_{\mathbf{A}\mathbf{x}+\mathbf{b}}) = \mathbf{A}\mathbf{V}(F_{\mathbf{x}})\mathbf{A}^t \tag{1}$$

for any full rank $p \times p$ matrix $\mathbf{A}$ and any p-vector $\mathbf{b}$.

A scatter matrix is then naturally defined as $\hat{\mathbf{V}} = \mathbf{V}(F_n)$, where F_n is the empirical cdf. Most robust counterparts of covariance matrix satisfy (1). However, they usually do not satisfy the so-called joint and block independence properties, which are characteristic of the covariance matrix, and are defined as follows.

Definition 2. Assume that $\mathbf{x} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_k^t)^t$ is a p-vector consisting of k mutually independent subvectors with dimension p_i, $i = 1, \ldots, k$, such that $\sum_{i=1}^k p_i = p$.

i) The scatter matrix functional $\mathbf{V}(F_{\mathbf{x}})$ is said to have the *block independence property* if it is a block diagonal matrix with block sizes p_i, $i = 1, \ldots, k$.
ii) If $k = p$, which means that $\mathbf{x}$ has independent components, and $\mathbf{V}(F_{\mathbf{x}})$ is a diagonal matrix, then it is said to have the *joint independence property*.

Note that the block independence property implies the joint independence property, but not vice versa. In Nordhausen and Tyler (2015) several examples of multivariate methods are given for which it is necessary for a scatter matrix to possess the joint or block independence property.

Most scatter functionals do not posses the joint or block independence property. A common conjecture here is that only scatter matrices which can be expressed as functions of pairwise differences have this property. For example $\mathbf{COV}(\mathbf{x}) = E((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^t) = 2^{-1}E((\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^t)$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ denote independent copies of $\mathbf{x}$, can be written in such a way.

What about scatter matrices which cannot be expressed using pairwise differences? A quite simple but ingenious approach is to apply a scatter functional to the pairwise differences, which is known as symmetrization. Theorem 1 in Oja et al (2006) shows that when a scatter matrix functional is applied to the pairwise differences of the observations, then the resulting functional possesses the joint independence property. In Nordhausen and Oja (2011) and Nordhausen and Tyler (2015) it is shown that symmetrization yields to a more general block independence property.

A formal definition of symmetrization is given as follows.

Definition 3. Let $\mathbf{V}(F_{\mathbf{x}})$ be any scatter functional. Then the corresponding *symmetrized scatter functional* is defined as

$$\mathbf{V}_s(F_{\mathbf{x}}) = \mathbf{V}(F_{\mathbf{x}_1 - \mathbf{x}_2}),$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are two independent copies of $\mathbf{x}$.

In this paper we are mainly interested in computational aspects of symmetrized scatter matrices. Although the computational issues discussed herein apply to any symmetrized scatter matrix, this paper focuses on symmetrized M-estimators of

scatter (Sirkiä et al, 2007), which, as to be seen, can be made computationally feasible for fairly large sample sizes. The next section reviews the definition and basic properties of symmetrized M-estimators of scatter.

3 Symmetrized M-estimators of scatter

Write again $\mathbf{x}$ for a p-variate random vector with cumulative distribution function $F_{\mathbf{x}}$. In this paper we focus on elliptically symmetric distributions. Such a distribution family is often used in robustness studies as it includes distributions with heavy tails (e.g. elliptical Cauchy distribution) as well as distributions which can be used to generate atypical observations (e.g. contaminated normal distribution).

An elliptically symmetric distribution is obtained as an affine transformation of a spherical distribution. Recall that a p-variate random vector $\mathbf{z}$ is spherically symmetric around the origin if $\mathbf{U}\mathbf{z} \sim \mathbf{z}$ for all orthogonal $p \times p$ matrices $\mathbf{U}$. Then $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z} + \boldsymbol{\mu}$, where $\boldsymbol{\Omega}$ is a full rank $p \times p$ matrix and $\boldsymbol{\mu}$ a p-vector, has an elliptically symmetric distribution with density of the form

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, g) = |\boldsymbol{\Sigma}|^{-1/2} g(\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})),$$

where $g(\mathbf{z}) = \exp(-\rho(||\mathbf{z}||))$ represents the density of $\mathbf{z}$, with $\rho(\cdot)$ being a non-negative function, and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^t$. Without loss of generality, $\boldsymbol{\Sigma}^{1/2}$ is taken to be the symmetric positive definite square-root of $\boldsymbol{\Sigma}$. Note that the density of $\mathbf{z}$ depends only on the value of its radius $||\mathbf{z}||$, and the function $\rho(\cdot)$ does not depend on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The parameter $\boldsymbol{\mu}$ is the location center of the distribution and the scatter matrix $\boldsymbol{\Sigma}$ is proportional to the regular covariance matrix (if it exists). Examples of function $g(\cdot)$ include $g(\mathbf{z}) = (2\pi)^{-p/2} \exp(-\mathbf{z}^t\mathbf{z}/2)$, which corresponds to the p-variate normal distribution, and

$$g(\mathbf{z}) = \frac{\Gamma((p+\nu)/2)}{\Gamma(\nu/2)(\pi\nu)^{p/2}} \left(1 + \frac{\mathbf{z}^t\mathbf{z}}{\nu}\right)^{-(p+\nu)/2},$$

which corresponds to the p-variate t-distribution on ν degrees of freedom. Within the class of elliptical distribution, i.e. for unknown g, only the location $\boldsymbol{\mu}$ and the "shape" of $\boldsymbol{\Sigma}$, i.e. the value of $\boldsymbol{\Sigma}$ up to proportionality, is well defined, whereas the constant of proportionality is confounded with the function g.

Next, we recall the definition of the symmetrized M-functional as given in Sirkiä et al (2007).

Definition 4. Assume that $\mathbf{x}$ is a p-variate random vector with cdf $F_{\mathbf{x}}$, and let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two independent copies of $\mathbf{x}$. A *symmetrized M-functional* $\mathbf{V}_s = \mathbf{V}_s(F_{\mathbf{x}_1 - \mathbf{x}_2})$ is defined as a solution to

$$E[w_1(r_{12})(\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^t - w_2(r_{12})\mathbf{V}_s] = \mathbf{0},$$

where $r_{12} = [(\mathbf{x}_1 - \mathbf{x}_2)^t \mathbf{V}_s^{-1} (\mathbf{x}_1 - \mathbf{x}_2)]^{1/2}$, and w_1 and w_2 are some real-valued functions on $[0, \infty)$.

Sirkiä et al (2007) observe that the assumptions on the weight functions and on the distribution of the pairwise differences needed for the existence and uniqueness of symmetrized M-functionals follow from Huber's (1981) results for (non-symmetrized) M-functionals. When $\mathbf{x}$ has an elliptical distribution, $\mathbf{V}_s \propto \boldsymbol{\Sigma}$, with the constant of proportionality being dependent on the weight functions w_1 and w_2 and the density g, but not on the parameters $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$.

An estimator corresponding to a scatter matrix functional $\mathbf{V}_s$ is obtained when $F_{\mathbf{x}_1 - \mathbf{x}_2}$ in Definition 4 is replaced with the empirical distribution function of the pairwise differences. A symmetrized M-estimator of scatter, $\hat{\mathbf{V}}_s$, then solves

$$\binom{n}{2}^{-1} \sum_{i<j} \sum [w_1(r_{ij})(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t - w_2(r_{ij})\mathbf{V}_s] = \mathbf{0},$$

where w_1 and w_2 are real-valued functions on $[0, \infty)$. Notice that choices $w_1(r) = \rho'(r)/r$ and $w_2(r) = 2$ yield the maximum likelihood estimator under a specific elliptical distribution. The robustness properties and limiting distributions of general symmetrized M-estimators were discussed in Sirkiä et al (2007).

In this paper we consider the following symmetrized M-estimators:

- The sample covariance matrix, which corresponds to $w_1(r) = 1$ and $w_2(r) = 2$, or equivalently to $w_1(r) = 1/2$ and $w_2(r) = 1$
- The symmetrized Cauchy M-estimator, which has weight functions corresponding to those of the maximum likelihood estimator for the elliptical Cauchy distribution, i.e. $w_1(r) = (1+p)/(1+r^2)$ and $w_2(r) = 1$. It is worth noting that this is not the same as the maximum likelihood estimator based on the pairwise differences from a random sample of an elliptical Cauchy distribution.
- The symmetrized Huber estimators, which have weight functions $w_2(r) = 1$ and

$$w_1(r) = \begin{cases} 1/\sigma^2, & r^2 \leq c^2 \\ c^2/(r^2\sigma^2), & r^2 > c^2, \end{cases}$$

  where c is a tuning constant defined so that $q = Pr(\chi_p^2 \leq c^2/2)$ for a chosen q. The scaling factor σ is chosen so that $E[w_1(||\mathbf{x}_1 - \mathbf{x}_2||)] = p$, where $\mathbf{x}_1, \mathbf{x}_2 \sim N(\mathbf{0}, \mathbf{I}_p)$, which makes the estimator Fisher-consistent for $\boldsymbol{\Sigma}$ at the multivariate normal model.
- Dümbgen's (1998) estimator, which corresponds to choosing $w_1(r) = p/r^2$ and $w_2(r) = 1$.

Dümbgen's estimator is only define up to proportionality, i.e. both $\hat{\mathbf{V}}_{s,1}$ and $\hat{\mathbf{V}}_{s,2}$ satisfy the corresponding estimating equations, if and only if $\hat{\mathbf{V}}_{s,1} \propto \hat{\mathbf{V}}_{s,2}$. Furthermore, as noted previously, under sampling from an elliptical distribution, the symmetrized

Cauchy M-estimator is Fisher-consistent for the parameter $\boldsymbol{\Sigma}$ only up to proportionality. This is also true of the sample covariance matrix and the symmetrized Huber M-estimator at elliptical models other than the multivariate normal. These factors, though, are not important to the efficiency comparisons given in Section 5 since only the shape of the scatter matrices are considered in these comparisons.

4 Computation of symmetrized M-estimators

Hereafter, we consider only the case $w_2(\cdot) = 1$, which agrees with the original definition of the M-estimators given in Maronna (1976). Note that this case holds for the three M-estimators discussed in the previous section, as well as for the maximum likelihood estimators of scatter under an elliptical family of distributions, i.e. with a fixed g. A general recent overview of the M-estimators of scatter for the case $w_2(\cdot) = 1$ can be found in Dümbgen et al (2015b). They point out that the most commonly used method to compute such M-estimates is via a simple fixed point algorithm, which is known to converge under very general conditions to a unique solution, regardless of the initial value, as shown in Kent and Tyler (1991).

Assume in the following that we have a sample of n vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and the goal is to compute the symmetrized M-scatter matrix $\mathbf{V}_s$ of interest. The most naive approach would be to apply the fixed point algorithm for the unsymmetrized scatter of interest to all $n(n-1)$ pairwise differences $\mathbf{x}_i - \mathbf{x}_j$ with $i \neq j$. Notice that now the location center does not need to be estimated as for the symmetrized vectors the location center is naturally the origin. Nevertheless, even for a moderate sample size n, the computational burden can be tremendous and so new approaches are needed to to deal with this. In the following we will consider a few practical ways to reduce the computational burden and memory demand.

The number of pairwise differences can be halved since only the $n(n-1)/2$ pairwise differences $\mathbf{x}_i - \mathbf{x}_j$ with $i < j$ are needed to compute the symmetrized scatter matrix. Hence the most basic algorithm is *the fixed point algorithm* with updating step at iteration k:

$$\mathbf{V}_s^{k+1}(\mathbf{X}) = \binom{n}{2}^{-1} \sum_{i<j} \left\{ w_1(r_{ij}^k)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \right\},$$

where r_{ij}^k is based on the current scatter estimate $\mathbf{V}_s^k$. Recently, Nordhausen and Tyler (2015) suggested rewriting the above algorithm as

$$\mathbf{V}_s^{k+1}(\mathbf{X}) = 2(n(n-1))^{-1} \sum_{i=2}^{n} \mathbf{S}_i^{k+1}(\mathbf{X}),$$

where

$$\mathbf{S}_i^{k+1}(\mathbf{X}) = \sum_{j=1}^{i-1} w_1(r_{ij}^k)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t.$$

The computation of $\mathbf{S}_i^{k+1}(\mathbf{X})$ then can be naturally divided into several threads. We refer to this algorithm as the *the parallel algorithm* and study, in Section 6, how much this approach speeds up computations.

Dümbgen et al (2015a) have recently argued that using a fixed point algorithm for computing M-estimates of scatter can be less than optimal. They consider several alternative algorithms and recommended *a partial Newton (PN) algorithm*, which, in most cases, is considerably faster. The basic idea behind the PN algorithm is to first perform a few fixed point steps and to then evaluate whether shifting to a Newton-Raphson step with an approximated Hessian is better. We refer to the reader to the aforementioned paper for more details regarding the algorithm. A restriction of the PN algorithm is that the weight functions must be smooth, which excludes, for example, Huber's weight functions. Two versions of the PN algorithm were introduced in Dümbgen et al (2015a), with one version requiring all pairwise differences $\mathbf{x}_i - \mathbf{x}_j$ with $i < j$ being in the memory, and the other version being a sequential algorithm which avoids storing all pairwise differences. The sequential algorithm seems to be, in most cases, faster than the one that stores all pairwise differences.

We have, thus, several algorithms available so far for the computation of symmetrized M-estimators of scatter. However, these are all computationally intensive as they either store all pairwise differences $\mathbf{x}_i - \mathbf{x}_j$ with $i < j$ in the memory or compute sequentially all quantities of interest. This computational burden is demonstrated later in Section 6.

A possible way to ease this computational problem can be motivated by noting the resemblance of the symmetrized scatter matrix to a U-statistic of order two. Recall that for a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ a U-statistic for a parameter θ based on a symmetric kernel $h(\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(K)})$ of order K is defined as

$$U = N^{-1} \sum_{i=1}^{N} h(\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(K)}),$$

where $N = \binom{n}{K}$ and the kernel is computed for all possible subsamples of size K denoted by $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(K)}$. A simple example of a U-statistic is the sample covariance matrix, which has a kernel of order two and can be expressed as

$$h(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) = 2^{-1}(\mathbf{x}_{(1)} - \mathbf{x}_{(2)})(\mathbf{x}_{(1)} - \mathbf{x}_{(2)})^t$$

and hence

$$\mathbf{COV}(\mathbf{X}) = \binom{n}{2}^{-1} \sum_{i \neq j} 2^{-1}(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})(\mathbf{x}_{(i)} - \mathbf{x}_{(j)})^t.$$

In general, though, not all symmetrized scatter matrices can be expressed as U-statistics, since they typically have only an implicit rather than an explicit representation in terms of pairwise differences.

In the context of U-statistics, Blom (1976) noted that it is possible to use less than N terms without losing much information when estimating θ, and he called such estimates incomplete U-statistics. Such estimates have also been referred to as

weighted U-statistics, with weights 0 or 1, or as reduced U-statistics. In Blom (1976) and Brown and Kildea (1978) the statistical properties of incomplete U-statistics are derived.

Following the idea of incomplete U-statistics, many ways to choose the terms used in computations are possible. The most basic one is independent subsampling, where m sets out all N sets are chosen at random. This can, however, give different weight to different observations in the data. Another convenient choice for kernels of order $K = 2$, which gives each observation equal weight, is what we refer to as a "running average of length m". For this purpose, we treat the ordering of the data as cyclic and define an extended data matrix $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_{n+m}^*) = (\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_1, \ldots, \mathbf{x}_m)$. Our *incomplete symmetrized M-estimator of length m*, $\hat{\mathbf{V}}_I$, then solves

$$\mathbf{V}_I = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=i+1}^{i+m} w_1(r_{ij})(\mathbf{x}_i^* - \mathbf{x}_j^*)(\mathbf{x}_i^* - \mathbf{x}_j^*)^t.$$

In the following we explore the idea of computing symmetrized scatter matrices by using running averages of different lengths m, and compare the loss in efficiency to the gain in computation time. From a practical point of view, the observation order should be randomly permuted in order to avoid the effect of how the data was recorded. Using permutations in the simulations, though, are not needed since the simulated data set follows the same model as any permutation of it.

5 Efficiency comparisons

In this section, we compare the efficiencies of the incomplete (using only mn pairwise differences) symmetrized estimators to that of the corresponding complete symmetrized estimator. We include the symmetrized Huber estimator with $q = 0.90$, Dümbgen's estimator and the symmetrized Cauchy M-estimator in the comparisons. Since, as previously discussed, the estimators are only comparable up to proportionality, we standardize all estimators so that their traces are equal to p.

To compare the finite sample efficiencies, we first carried out a simulation study with samples of size n =1000, 2000 and 4000, dimensions $p = 3$ and $p = 8$, and under the normal distribution (N), the contaminated normal distribution (cN) and the t-distribution on 5 degrees of freedom (t_5). The cumulative distribution function of the contaminated normal distribution is $\Phi_{\varepsilon,c}(\mathbf{x}) = (1-\varepsilon)\Phi(\mathbf{x}) + \varepsilon\Phi(\mathbf{x}/c)$, where $\varepsilon, c > 0$ and Φ denotes the cumulative distribution function of the standard multivariate normal distribution. In our simulation settings, we used $\varepsilon = 0.1$ and $c = 3$.

The asymptotic efficiencies of the three standardized robust scatter estimators relative to the standardized sample covariance matrix are listed in Table 1. The asymptotic relative efficiencies were computed using the results in Sirkiä et al (2007), wherein they observed that the symmetrized Huber estimator and the Dümbgen's estimator are highly efficient not only at heavy tailed distributions but also at the

multivariate normal distribution. The symmetrized Cauchy M-estimator, though, suffers from some efficiency loss at the multivariate normal distribution case.

Table 1 Asymptotic efficiencies of the symmetrized Huber M-estimator, Dümbgen's estimator and the symmetrized Cauchy M-estimator relative to the sample covariance matrix. The asymptotic relative efficiencies are evaluated at the normal (N), the contaminated normal (cN) and t-distribution on 5 degrees of freedom (t_5).

	p=3			p=8		
	N	cN	t_5	N	cN	t_5
symmetrized Huber	0.99	1.67	2.12	1.00	1.65	2.12
Dümbgen	0.93	2.27	2.40	0.96	2.43	2.60
symmetrized Cauchy	0.77	2.03	1.04	0.85	2.26	1.20

To compare the finite sample efficiencies, the mean squared errors of the off-diagonal elements of the standardized scatter matrices, that is,

$$MSE(\hat{\mathbf{V}}) = \frac{2}{Np(p-1)} \sum_{k=1}^{N} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} (\hat{\mathbf{V}}_{ij}^{(k)} - \mathbf{I}_{ij})^2,$$

were computed using $N = 2000$ samples. The efficiencies were then defined by taking the ratios of the corresponding MSEs. The results are listed in Tables 2-4. For all of the estimators, there is some loss, but somewhat surprising not a large loss, in efficiency when $m = 10$, and when $m = 20$, the efficiency loss is always less than 5%. The loss in efficiency is slightly worst for the Dümbgen's estimator than for the other estimators.

Among the symmetrized scatter matrices considered in this paper, only the sample covariance matrix is a U-statistic. Nevertheless, the simulations indicate that all scatter matrices computed using running averages of length m seem to behave in a similar fashion. These empirical results suggest that theoretical results obtained for the incomplete sample covariance matrix may give us insight into the behavior of other incomplete symmetrized estimates of scatter.

In particular, results from Brown and Kildea (1978) for incomplete U-statistics, allow us to compute the asymptotic relative efficiency of the incomplete sample covariance estimator with respect to the complete sample covariance matrix. For a spherically symmetric distribution with $\mathbf{COV}(\mathbf{z}) = \mathbf{I}_p$, the efficiency of the incomplete symmetrized sample covariance matrix relative to the complete one is

$$ARE(\hat{\mathbf{V}}_s, \hat{\mathbf{V}}_I^{(m)}) = \frac{2m\kappa}{0.5 + (2m-1)\kappa}, \tag{2}$$

where $\kappa = E[z_i^2 z_j^2]/(2(E[z_i^2 z_j^2] + 1))$, and z_i and z_j are different components of $\mathbf{z}$. In Figure 1 we plot the asymptotic relative efficiency of the symmetrized incomplete sample covariance matrix as a function of m for the 3-variate normal, contaminated normal and t_5-distribution, for which $\kappa = 1/4$, 25/68, and 3/8 respectively. We also

Table 2 Finite sample relative efficiencies (MSE from 2000 samples) of the incomplete symmetrized Huber M-estimator with respect to the complete estimator.

		p=3			p=8		
	m	N	cN	t_5	N	cN	t_5
n=1000	10	0.94	0.95	0.94	0.95	0.96	0.95
	20	0.97	0.97	0.98	0.97	0.98	0.97
	50	0.98	0.98	0.98	0.98	0.98	0.98
	100	0.97	0.99	0.98	0.98	0.98	0.98
n=2000	10	0.94	0.95	0.95	0.95	0.96	0.95
	20	0.97	0.98	0.97	0.97	0.98	0.97
	50	0.99	0.99	0.98	0.98	0.99	0.99
	100	0.99	0.99	0.99	0.99	0.99	0.99
n=4000	10	0.95	0.97	0.96	0.95	0.96	0.97
	20	0.98	0.98	0.97	0.97	0.98	0.98
	50	0.99	1.00	0.99	0.99	0.99	0.99
	100	0.99	0.99	0.99	0.99	0.99	0.99

Table 3 Finite sample relative efficiencies (MSE from 2000 samples) of incomplete Dümbgen's estimators with respect to the complete estimator.

		p=3			p=8		
	m	N	cN	t_5	N	cN	t_5
n=1000	10	0.90	0.93	0.91	0.94	0.94	0.94
	20	0.95	0.95	0.96	0.96	0.97	0.97
	50	0.97	0.97	0.97	0.98	0.98	0.98
	100	0.98	0.97	0.97	0.98	0.98	0.98
n=2000	10	0.90	0.92	0.92	0.93	0.94	0.94
	20	0.94	0.96	0.96	0.97	0.97	0.97
	50	0.98	0.98	0.98	0.98	0.98	0.99
	100	0.98	0.99	0.98	0.99	0.99	0.99
n=4000	10	0.90	0.92	0.92	0.94	0.94	0.95
	20	0.95	0.96	0.96	0.97	0.97	0.97
	50	0.98	0.97	0.99	0.98	0.99	0.99
	100	0.98	0.99	0.99	0.99	0.99	0.99

simulated the finite sample efficiencies, which correspond to the dash lines in the figures, computed as ratios of MSEs using $n = 1000$ and $N = 2000$ repetitions. It can be seen that the efficiencies increase rapidly as a function of m, with a limit of one as $m \to \infty$. Interestingly, the efficiency at $m = 1$ is notably higher in the case of heavy tailed distributions than in the case of the normal distribution. The choice $m = 20$ is sufficient to produce an estimator with very high efficiency.

Table 4 Finite sample relative efficiencies (MSE from 2000 samples) of incomplete Cauchy M-estimators with respect to the complete estimator.

		p=3			p=8		
	m	N	cN	t_5	N	cN	t_5
n=1000	10	0.93	0.94	0.95	0.94	0.95	0.95
	20	0.96	0.97	0.98	0.97	0.97	0.97
	50	0.98	0.98	0.98	0.98	0.98	0.98
	100	0.98	0.99	0.98	0.98	0.98	0.98
n=2000	10	0.93	0.95	0.95	0.94	0.95	0.95
	20	0.96	0.97	0.98	0.97	0.97	0.97
	50	0.98	0.98	0.98	0.98	0.99	0.99
	100	0.98	0.99	0.99	0.99	0.99	0.99
n=4000	10	0.93	0.94	0.93	0.94	0.95	0.95
	20	0.96	0.97	0.96	0.97	0.97	0.97
	50	0.98	0.99	0.99	0.99	0.99	0.99
	100	0.98	0.99	0.99	0.99	0.99	0.99

All simulations so far have focused on data coming from an elliptical model, among which the only distribution with independent marginals is the multivariate normal distribution. However, there are many areas of applications, such as independent components analysis, for which independent marginals outside of the multivariate normal distribution are of interest. Consequently, we also simulated data for different sample sizes and dimensions from a model with mutually independent components where each component has a standard exponential distribution. Here, if the scatter functional possesses the joint independence property, then the off-diagonal values of the scatter matrix are equal to zero. For this setting, we compare

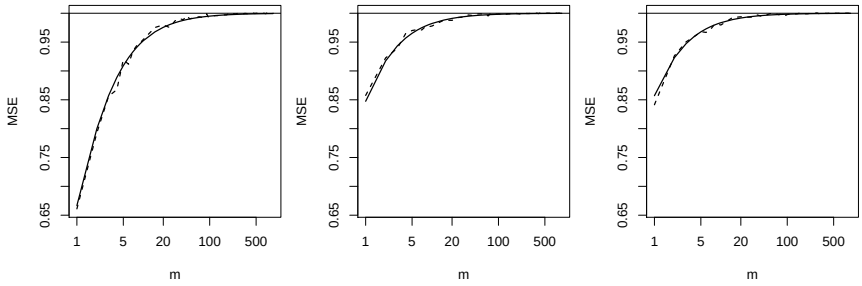

Fig. 1 Finite sample efficiencies of the incomplete symmetrized sample covariance matrix with respect to the symmetrized sample covariance matrix (dashed lines) for different distributions with $n = 1000$ and $p = 3$. The distributions from left to right are the normal distribution, the contaminated normal distribution and the t_5-distribution. The solid lines give the asymptotic relative efficiencies.

the symmetrized M-estimators (Dümbgen's estimator and the Cauchy M-estimator) based on all pairwise differences to the corresponding estimators using running averages of length 20. Figure 2 gives the mean squared errors of the off-diagonal values based on 1000 repetitions. The figure shows that in this case the incomplete estimators with $m = 20$ behave similarly to the regular symmetrized estimators.

For comparison, we also compute the corresponding non-symmetrized versions of the scatter matrices (Tyler's estimator and the Cauchy M-estimator, which corresponds to the MLE for the Cauchy distribution, respectively). These estimators do not possess the desired joint independence property, and so the corresponding functionals of these non-symmetrized versions do not have zero off-diagonal elements even though the variables are independent. Consequently, as seen in Figure 2, their MSEs do not go to zero as n increases.

As this section demonstrates, using running average sets of pairwise differences with small to moderate values of m results in only a small loss of efficiency relative to their complete version. In the next section we will see how this small loss in efficiency pays off in computation time.

6 Computation times

In comparing the computation times of the different algorithms presented in Section 4, we again chose the two dimensions $p = 3$ and $p = 8$, and use five sample sizes $n =$1000, 2000, 4000, 8000, 16000. For each combination of p and n, 50 samples from the multivariate t-distribution with 5 degrees of freedom are generated, and the computation times of the different algorithms are measured. The scatter matrices under consideration are the same as those used in previous sections. For the symmetrized Cauchy M-estimator and for the Dümbgen's estimator, both the fixed point algorithms and the partial Newton algorithms can be found in the R-packages ICSNP (Nordhausen et al, 2012) and fastM (Dümbgen et al, 2014), respectively. The symmetrized Huber estimator can also be computed using the R-package ICSNP. Currently, there are plans to implement the running average versions of the estimators in these package.

Our main interest in the following comparisons is two-fold. First, we are interested in when parallelization is beneficial, and second, in how fast the running average versions are relative to the standard implementations. In the simulations we chose $m = 20$ for the incomplete estimators as this was in all cases considered to yield highly efficient estimators. We also used the partial Newton algorithm from the fastM package that uses sequential computations as this seems to be faster than having all pairwise differences in the memory (Dümbgen et al, 2015a). All functions are mainly written in C or C++ with an R interface and should be therefore comparable (but have sometimes slightly different convergence criteria). The comparisons were done using R 3.1.1 (R Core Team, 2014) on a Intel(R) Core(TM) i7-3770 CPU with 3.40 GHz, 32 GB of memory using 64-bit Red Hat Linux.

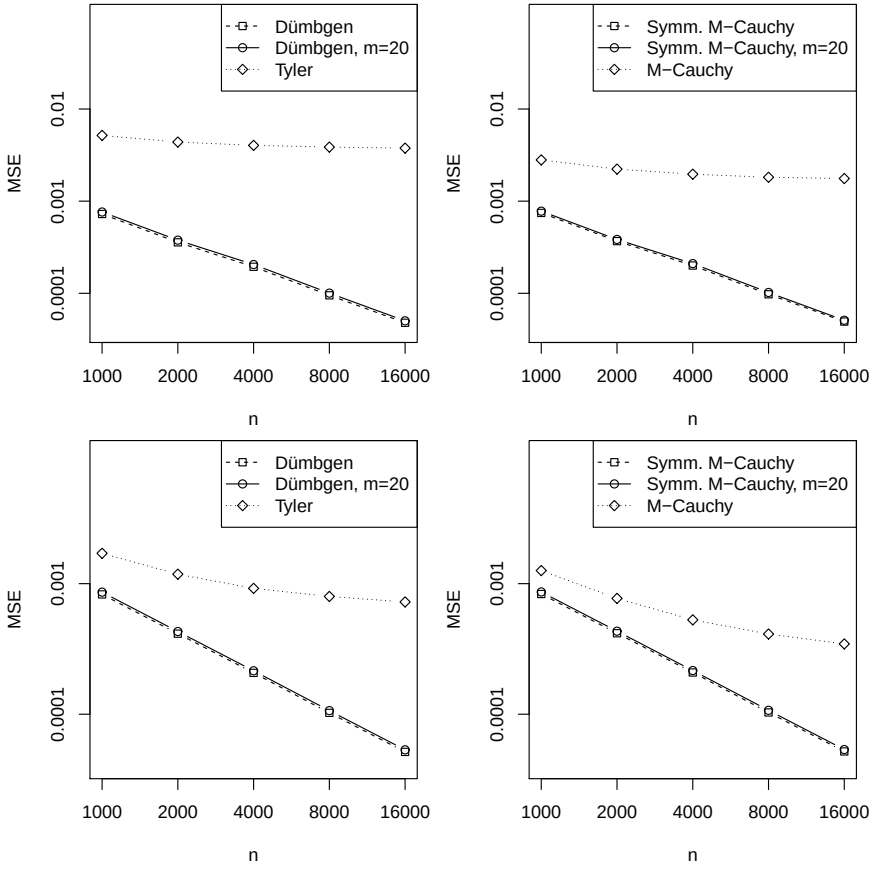

Fig. 2 Mean squared errors of the off-diagonal elements of the Dümbgen's estimate, the incomplete Dümbgen's estimate with $m = 20$, and Tyler's estimate on the left, and MSE of the symmetrized Cauchy M-estimate, the incomplete symmetrized Cauchy M-estimate with $m = 20$, and non-symmetrized Cauchy M-estimate on the right, when the three-dimensional (on the top row) and eight-dimensional (on the bottom row) data are generated from a distribution with mutually independent and exponentially (with mean 1) distributed components.

Medians of the computation times (on the logscale) of the symmetrized Cauchy M-estimator, Dümbgen's estimator and the symmetrized Huber estimator are given in Figures 3, 4 and 5, respectively.

As expected, the regular fixed point algorithm utilizing all pairwise differences is the slowest while the incomplete estimator is the fastest to compute. The ratio of their computation times is approximately the ratio of the number of pairs, which is $0.5(n-1)/m$. With large sample sizes, using two cores gains approximately 50% in computation time and using four cores approximately 75% relative to using only one core. We compared also the computation times when using six cores, but the computation times did not differ significantly different the version using four cores.

Notice that the gain percentage of the parallel computation grows with the sample size and the dimension until it reaches a limiting level. Parallelization becomes beneficial somewhere around n =2000 when $p = 3$ and around n =1000 when $p = 8$.

As already pointed out in Dümbgen et al (2015a), the partial Newton algorithm is not considerably faster than the fixed point algorithm when computing Dümbgen's estimator. However, the PN algorithm computes Dümbgen's estimator and the symmetrized Cauchy M-estimator equally fast, whereas all the other algorithms compute Dümbgen's estimator much faster than the symmetrized Cauchy M-estimator; for $p = 3$ and $p = 8$, approximately 5 and 18 times faster, respectively. Hence, the PN algorithm is superior to parallelized fixed point algorithm using four cores for the symmetrized Cauchy M-estimator, and vice versa for Dümbgen's estimator. Recall that the PN algorithm cannot be applied to the Huber estimator since the weight functions are not smooth. The computation time of the symmetrized Huber estimator is approximately the same as that of the Dümbgen's estimator when $p = 3$, but twice as long when $p = 8$.

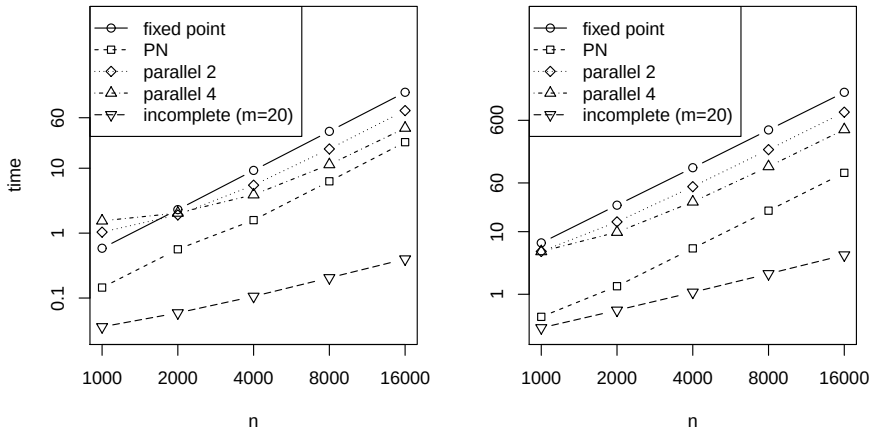

Fig. 3 Median computation times in seconds on a logscale for various algorithms used to compute the symmetrized Cauchy M-estimator. For each sample size, the median computation time is based on 50 independent random samples from the multivariate t_5-distribution. In the left panel $p = 3$ and in the right panel $p = 8$.

7 Discussion

The relevance of symmetrized scatter matrices has only recently been recognized within the statistics literature. The benefit of using such scatter matrices is twofold: (i) they do not require a location estimate, and (ii) they possess the joint and the block independence properties, which are necessary properties for many multivari-

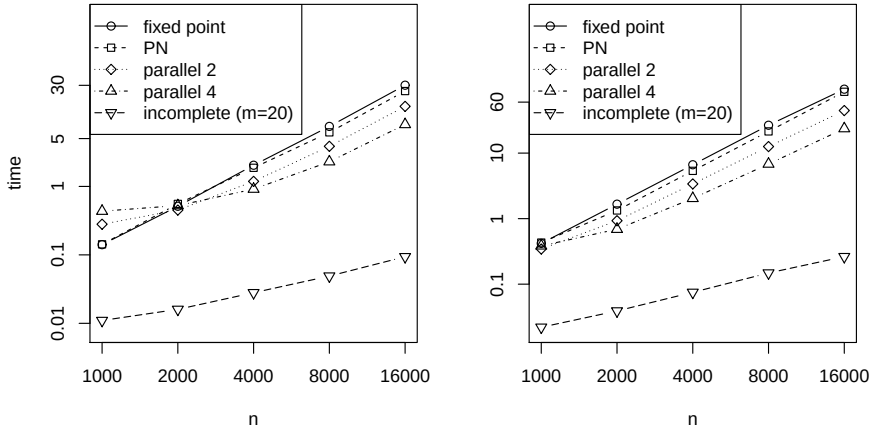

Fig. 4 Median computation times in seconds on a logscale for various algorithms used to compute Dümbgen's estimator. For each sample size, the median computation time is based on 50 independent random samples from the multivariate t_5-distribution. In the left panel $p = 3$ and in the right panel $p = 8$.

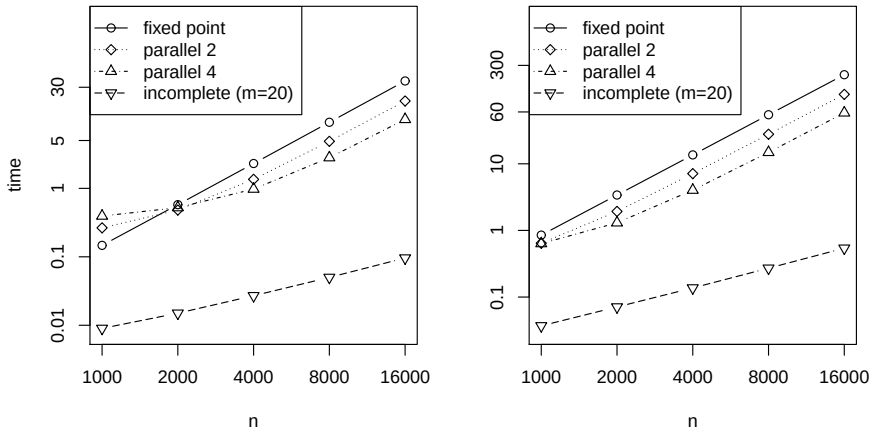

Fig. 5 Median computation times in seconds on a logscale for various algorithms used to compute the symmetrized Huber M-estimator with $q = 0.90$ For each sample size, the median computation time is based on 50 independent random samples from the multivariate t_5-distribution. In the left panel $p = 3$ and in the right panel $p = 8$.

ate methods. These benefits, however, come at a cost, namely that symmetrized scatter matrices tend to be more computationally intensive and are slightly less robust than their unsymmetrized counterparts. In this paper, the computational aspects of symmetrized M-estimators have been considered. In particular, it is shown that parallelization of the fixed-point algorithm is possible for these M-estimators and that this provides a considerable gain when, for example, four cores are used. Paralleization of the fixed point algorithm alone, though, is not as computationally efficient as

more recently proposed partial Newton algorithms. Another computational alternative, proposed within the paper, is motivated by results on incomplete U-statistics, namely to reduce the number of pairwise differences used in computations. Such an approach proves to be promising. A huge gain in computation time is achieved with only a small loss in efficiency. Finally, we note that while the parallelization approach is specific for M-estimators, the subsampling of pairwise differences can be applied to any symmetrized scatter matrix.

Acknowledgements This work was supported by the Academy of Finland under Grants 251965, 256291 and 268703. David Tyler's work for this material was supported by the National Science Foundation under Grant No. DMS-1407751. We thank Dr. Seija Sirkiä for providing us the asymptotic relative efficiencies of the symmetrized estimators.

References

Blom G (1976) Some properties of incomplete U-statistics. Biometrika 63(3):573–580

Brown BM, Kildea DG (1978) Reduced U-statistics and the hodges-lehmann estimator. Ann Statist 6(4):828–835, DOI 10.1214/aos/1176344256

Croux C, Rousseeuw PJ, Hössjer O (1994) Generalized S-estimators. Journal of the American Statistical Association 89:1271–128

Davies PL (1987) Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Annals of Statistics 15:1269–1292

Dümbgen L (1998) On Tyler's M-functional of scatter in high dimension. Annals of the Institute of Statistical Mathematics 50:471–491

Dümbgen L, Nordhausen K, Schuhmacher H (2014) fastM: Fast Computation of Multivariate M-estimators. URL http://CRAN.R-project.org/package=fastM, R package version 0.0-2

Dümbgen L, Nordhausen K, Schuhmacher H (2015a) New algorithms for M-estimation of multivariate location and scatter, arXiv:1312.6489

Dümbgen L, Pauly M, Schweizer T (2015b) M-functionals of multivariate scatter. Statistics Surveys 9:31–105

Huber PJ (1981) Robust Statistics. Wiley, New York

Kent JT, Tyler DE (1991) Redescending M-estimates of multivariate location and scatter. Ann Statist 19(4):2102–2119

Lopuhaä H (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. Annals of Statistics 17:1662–1683

Maronna RA (1976) Robust M-estimators of multivariate location and scatter. Annals of Statistics 4:51–67

Maronna RA, Martin DR, Yohai VJ (2006) Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics, Wiley

Nordhausen K, Oja H (2011) Scatter matrices with independent block property and ISA. In: Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011), pp 1738–1742

Nordhausen K, Tyler DE (2015) A cautionary note on robust covariance plug-in methods, to appear in Biometrika.

Nordhausen K, Oja H, Ollila E (2008) Robust independent component analysis based on two scatter matrices. Austrian Journal of Statistics 37:91–100

Nordhausen K, Sirkiä S, Oja H, Tyler DE (2012) ICSNP: Tools for Multivariate Nonparametrics. URL http://CRAN.R-project.org/package=ICSNP, r package version 1.0-9

Oja H, Sirkiä S, Eriksson J (2006) Scatter matrices and independent component analysis. Austrian Journal of Statistics 35:175–189

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Roelant E, Van Aelst S, Croux C (2009) Multivariate generalized S-estimators. Journal of Multivariate Analysis 100:876–887

Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G, Vincze I, Wertz W (eds) Mathematical Methods and Applications, vol B, Reidel, Dordrecht, Netherlands, pp 283–297

Sirkiä S, Taskinen S, Oja H (2007) Symmetrised M-estimators of scatter. Journal of Multivariate Analysis 98:1611–1629

Tyler DE (1987) A distribution-free M-estimator of multivariate scatter. Annals of Statistics 15:234–251

Tyler DE, Critchley F, Dümbgen L, Oja H (2009) Invariant co-ordinate selection. Journal of the Royal Statistical Society, Series B 71:549–595