New highly efficient high-breakdown estimator of multivariate scatter and location for elliptical distributions

Justin FISHBONE* and Lamine MILI

Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Falls Church, Virginia, U.S.A.

Key words and phrases: Covariance matrix estimation; robust estimation; \( S \)-estimator; shape matrix estimation; \( S_q \)-estimator.

MSC 2020: Primary 62F35; secondary 62F12.

Abstract: High-breakdown-point estimators of multivariate location and shape matrices, such as the \( MM \)-estimator with smoothed hard rejection and the Rocke \( S \)-estimator, are generally designed to have high efficiency for Gaussian data. However, many phenomena are non-Gaussian, and these estimators can therefore have poor efficiency. This article proposes a new tunable \( S \)-estimator, termed the \( S_q \)-estimator, for the general class of symmetric elliptical distributions, a class containing many common families such as the multivariate Gaussian, \( t \)-, Cauchy, Laplace, hyperbolic, and normal inverse Gaussian distributions. Across this class, the \( S_q \)-estimator is shown to generally provide higher maximum efficiency than other leading high-breakdown estimators while maintaining the maximum breakdown point. Furthermore, the \( S_q \)-estimator is demonstrated to be distributionally robust, and its robustness to outliers is demonstrated to be on par with these leading estimators while also being more stable with respect to initial conditions. From a practical viewpoint, these properties make the \( S_q \)-estimator broadly applicable for practitioners. These advantages are demonstrated with an example application—the minimum-variance optimal allocation of financial portfolio investments.

Résumé: Les estimateurs à haut point de rupture des matrices de localisation et de dispersion multivariées, tels que l’estimateur \( MM \) avec rejet dur lissé et l’estimateur \( S \) de Rocke, sont généralement conçus pour atteindre une efficacité maximale lorsqu’ils sont utilisés sur des données gaussiennes. Or, comme de nombreux phénomènes ne suivent pas la loi gaussienne, ces estimateurs sont susceptibles d’avoir une faible efficacité. Pour parer à cette limitation, le présent article propose un nouvel estimateur ajustable, appelé l’estimateur \( S_q \), pour la classe générale de lois elliptiques symétriques qui comprend de nombreuses familles classiques telles que les lois gaussienne, de Student, de Cauchy, de Laplace, hyperbolique et gaussienne inverse multivariées. C’est dans ce cadre que les auteurs de cet article montrent que, tout en conservant le plus haut point de rupture possible, l’estimateur \( S_q \) offre une efficacité maximale plus élevée que celle d’autres estimateurs à haut point de rupture, qu’il présente une stabilité accrue vis-à-vis des conditions initiales, et qu’il offre une robustesse aux valeurs aberrantes comparable à celles des meilleurs estimateurs à haut point de rupture. En termes pratiques, ces caractéristiques font de l’estimateur \( S_q \) une méthode bien utile aux professionnels. Cela est illustré par un exemple concret, à savoir, l’allocation optimale d’un portefeuille financier à variance minimale.

*Corresponding author: Justin Fishbone fishbone@vt.edu

[Correction added on 16 May 2023, after first online publication: “MM-SHR” has been changed to “MM-estimator” in the Abstract.]

© 2023 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique published by Wiley Periodicals LLC on behalf of Statistical Society of Canada / Société statistique du Canada. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
1. INTRODUCTION

Huber (1964) defined what is the most common class of robust estimators, $M$-estimators. Although Huber’s original focus concerned the location case, Maronna (1976) expanded the definition to include multivariate location and scatter. After the sample median, perhaps the most common robust $M$-estimators are those using general rho functions such as the Huber or Tukey bisquare functions.

However, the drawback of using general rho functions is that they have limited efficiency when applied to parameter estimation for many probability distributions. To remedy this deficiency, various $M$-estimator approaches have been taken to iteratively reweight maximum likelihood estimator (MLE) weights based on the estimated probability density function (PDF); see, e.g., Windham (1995), Basu et al. (1998), Choi & Hall (2000), and Ferrari & Yang (2010).

Even with these improvements, multivariate $M$-estimators inherently have limited robustness to outliers. For example, Maronna (1976) showed that the upper bound on the breakdown point for $p$-dimensional $M$-estimators is $1/(p + 1)$, which converges to zero with large $p$. To combat this lack of robustness, Davies (1987) introduced a high-breakdown-point class of multivariate estimators of location and scatter called multivariate $S$-estimators. Davies showed that the asymptotic breakdown point of multivariate $S$-estimators can be set to $1/2$, which is the theoretical maximum of any equivariant estimator.

In practical scenarios, however, estimators may have large biases at considerably lower contamination levels than the breakdown point. For many years, the Tukey bisquare was the standard rho function for $S$-estimators; see, e.g., Lopuhaä (1989) and Rocke (1996). However, in the context of multivariate $S$-estimators, the bisquare is not tunable, so its robustness decreases with increasing $p$. For this reason, Rocke (1996) introduced the tunable biflat and translated biweight rho functions. Maronna et al. (2006, Sec. 6.4.4) slightly modified the biflat, proposing the Rocke rho function. The Rocke $S$-estimator (shortened here to $S$-Rocke) is currently the recommended high-breakdown estimator for moderate and large dimensions ($p \geq 15$); see Maronna & Yohai (2017) and Maronna et al. (2019, Sec. 6.10). The recommended estimator for lower dimensions is the $MM$-estimator with the smoothed hard rejection function ($MM$-SHR).

The $S$-Rocke estimator has two major shortcomings, which will be discussed in this article. First, it has low efficiency for small dimensions, $p$. Although this is an inherent disadvantage of all $S$-estimators, it is exceptionally acute for the $S$-Rocke. Second, the $S$-Rocke has poor efficiency for most common non-Gaussian distributions. This is a common weakness for general-purpose estimators such as the Rocke and bisquare $S$-estimators, the SHR $MM$-estimator, and the Huber and bisquare $M$-estimators. Examples of common phenomena that are frequently modelled by non-Gaussian distributions include stock returns, radar sea clutter, and speech signals, which approximately follow generalized hyperbolic (Konlack Socgnia & Wilcox, 2014), $K$- (Ward, Baker & Watts, 1990), and Laplace distributions (Gazor & Zhang, 2003), respectively.

In what follows, we propose and explore a new subclass of tunable, maximum-breakdown-point $S$-estimators that is applicable across common continuous elliptical distributions. This estimator, named the $S_q$-estimator, uses a density-based reweighting to attain generally higher maximum efficiencies across the elliptical class as compared to the $S$-Rocke and $MM$-SHR estimators. We compare these three estimators from the viewpoints of statistical and computational efficiency, robustness, and stability.

Although the focus on elliptical distributions sounds limiting, as we outline in the next section, many common continuous multivariate distributions—such as the Gaussian, $t$-, Laplace, and hyperbolic distributions—fall into this class. As Frahm (2009) discussed, this assumption of an elliptical distribution is “fundamental in multivariate analysis,” and it allows researchers to derive theoretical results.

The rest of the article is organized as follows. Section 2 defines the new estimator and indicates its mathematical formulation for the most common elliptical distributions. Section 3
outlines the asymptotic distribution of the \( S_q \)-estimator and compares the maximum achievable efficiencies of the \( S_q \), \( S \)-Rocke, and \( MM \)-SHR estimators; we also investigate the efficiency of the \( S_q \)-estimator under model mismatch. In Section 4, the finite-sample breakdown point of the \( S_q \)-estimator is discussed, the theoretical influence functions of the estimators are compared, and the empirical finite-sample robustness to outliers of the estimators is briefly explored. Section 5 assesses two computational aspects of the estimators: computational efficiency, and stability with respect to initial estimates. A practical example in Section 6 demonstrates the application of the estimators for the minimum-variance optimal allocation of financial portfolio investments. Some conclusions are summarized in Section 7. Proofs are provided in the Appendix, which also provides tables helping to define the \( S_q \)-estimator.

2. DEFINING THE \( S_q \)-ESTIMATOR

This section builds the definition of the proposed \( S_q \)-estimator. First, we review the elliptical class of distributions. Next, we review the definition of the multivariate \( S \)-estimator, and then we define our \( S_q \)-estimator.

2.1. Elliptical Distributions

The term elliptical distribution refers to a general class of multivariate probability distributions encompassing many familiar subclasses such as the symmetric Gaussian, \( t \), Cauchy, Laplace, hyperbolic, variance gamma, and normal inverse Gaussian distributions. Table A1 (see the Appendix) summarizes the most common elliptical distributions (Fang, Kotz & Ng, 1990, p. 69; Deng & Yao, 2018; Mériaux et al., 2019).

Symmetric elliptical distributions are defined as being a function of the squared Mahalanobis distance: 
\[
d(x, \mu, \Sigma) = (x - \mu)^\top \Sigma^{-1} (x - \mu),
\]
where \( x \in \mathbb{R}^p \), the location \( \mu \in \mathbb{R}^p \), and the \( p \times p \) positive definite symmetric \( PDS(p) \) scatter matrix \( \Sigma \in PDS(p) \). When the PDF is defined, which we assume in this article, it has the form 
\[
f_X(x) = \alpha_p |\Sigma|^{-1/2} \phi(d(x, \mu, \Sigma))
\]
for some generating function \( \phi \), where the operator \(| \cdot |\) denotes matrix determinant, and \( \alpha_p \) is a constant which ensures that \( f_X \) integrates to 1. Common generating functions are listed in Table A1. When the covariance matrix exists, it is proportional to the scatter matrix, \( \Sigma \). The corresponding shape matrix is commonly defined as
\[
\Omega = \Sigma / |\Sigma|^{1/p}.
\]
The PDF of \( d(x, \mu, \Sigma) \) is given by Kelker (1970)
\[
f_D(d) = \beta_p \ d^{p/2-1} \phi(d),
\]
where \( \beta_p = \alpha_p \pi^{p/2} / \Gamma(p/2) \). Hereafter, all densities, \( f \), refer to the density of \( d(x, \mu, \Sigma) \) identified in Equation (2), so the subscript \( D \) will be omitted.

2.2. \( S \)-Estimators

Given a set \( \{x_1, \ldots, x_n\} \) of \( n \) \( p \)-dimensional observations, \( S \)-estimates of location and shape are defined as (Maronna, Martin & Yohai, 2006, Sec. 6.4.2)
\[
\left( \hat{\mu}, \hat{\Omega} \right) = \arg \min \hat{\sigma} \begin{array}{l}
\text{subject to } |\hat{\Omega}| = 1, \\
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{d(x_i, \mu, \Omega)}{\hat{\sigma}} \right) = b
\end{array}
\]

DOI: 10.1002/cjs.11770
for some scalar rho function, \( \rho \). A proper \( S \)-estimator rho function should be a continuously differentiable, nondecreasing function in \( t \geq 0 \) with \( \rho(0) = 0 \), and where there exists a point \( c \) such that \( \rho(t) = \rho(\infty) \) for \( t \geq c \). These restrictions enable the high efficiency and high breakdown point of proper \( S \)-estimators. For simplicity, and without loss of generality, the rho functions in this article will be normalized so \( \rho(\infty) = 1 \). The parameter \( b \) is a scalar that affects the efficiency (see Section 3) and robustness of the estimator. The purpose of \( S \)-estimators is to achieve high robustness, so they are usually configured with \( b = 1/2 - (p + 1)/(2n) \), which achieves the maximum theoretical breakdown point that any affine equivariant estimator may have (see Section 4.1). To understand the derivation of the proposed estimator in the next section, note that \( \hat{\sigma} \) in the constraint is an \( M \)-estimate of the scale of the squared Mahalanobis distances \( d(x, \mu, \Omega) \).

Local solutions of Equation (3) can be found iteratively using the weighted sums

\[
\sum_{i=1}^{n} w\left(\frac{d_i}{\hat{\sigma}}\right) (x_i - \hat{\mu}) = 0,
\]

\[
\sum_{i=1}^{n} w\left(\frac{d_i}{\hat{\sigma}}\right) (x_i - \hat{\mu})(x_i - \hat{\mu})^\top \propto \hat{\Omega},
\]

where the weight function \( w(t) = \rho'(t) \) and where, for each iteration \( j \), \( \hat{\Omega}^{(j)} \) is first re-normalized, \( d_i^{(j)} = d(x_i, \hat{\mu}^{(j)}, \hat{\Omega}^{(j)}) \) is then calculated, and then \( \hat{\sigma}^{(j)} \) is evaluated using the constraint indicated in Equation (3). For the empirical results reported in this article, the estimators will all be evaluated using this weighted-sum algorithm.

To estimate a scatter matrix that is consistent at the assumed family, a separate estimator of \(|\Sigma|^{1/p}\) can then be used to scale \( \hat{\Omega} \) using Equation (1). Maronna et al. (2006, p. 186) discussed a simple estimator to scale \( \hat{\Omega} \) to \( \hat{\Sigma} \). When \( x \) is normally distributed, \( d \) has a chi-squared distribution with \( p \) degrees of freedom. Therefore, they suggested using

\[
\hat{\Sigma} = \frac{\text{Median}\left\{d\left(x_1, \hat{\mu}, \hat{\Omega}\right), \ldots, d\left(x_n, \hat{\mu}, \hat{\Omega}\right)\right\}}{\chi_p^2(0.5)} \hat{\Omega},
\]

where \( \chi_p^2(0.5) \) is the 50th percentile of the chi-squared distribution with \( p \) degrees of freedom. For the general case of elliptical distributions, we propose extending this to

\[
\hat{\Sigma} = \frac{\text{Median}\left\{d\left(x_1, \hat{\mu}, \hat{\Omega}\right), \ldots, d\left(x_n, \hat{\mu}, \hat{\Omega}\right)\right\}}{F^{-1}(0.5)} \hat{\Omega},
\]

where \( F \) is the distribution function corresponding to Equation (2), and therefore \( F^{-1}(0.5) \) is the 50th percentile of the same distribution.

For the location and shape matrices, the \( S \)-estimator formulation identified in Equation (3) is equivalent to the alternative one given by

\[
\left(\hat{\mu}, \hat{\Sigma}\right) = \arg\min_{\mu} |\Sigma| \text{ subject to } \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{d(x_i, \mu, \Sigma)}{\sigma}\right) = b,
\]

(4)
which requires that $\sigma$ be defined such that $E[\rho(d(x; \mu, \Sigma)/\sigma)] = b$ for a consistent estimator of $\Sigma$ at an assumed elliptical distribution (Rocke, 1996). While the first formulation is better for understanding the derivation of the proposed $S_q$-estimator, this second formulation is better for defining and understanding its properties; see, e.g., Lopuhaä (1989). The scale parameters in the two formulations are related asymptotically by $\sigma = |\Sigma|^{1/p}E[\hat{\sigma}]$, at the assumed distribution.

The two most common multivariate $S$-estimators are the bisquare and Rocke estimators (Maronna et al., 2019, Sec. 6.4.2, 6.4.4). The $S$-bisquare is given by $\rho_{\text{bisq}}(t) = \min\{1, 1 - (1 - t)^2\}$ and $w_{\text{bisq}}(t) = 3(1 - t^2)I(t \leq 1)$; notice that it does not have a tuning parameter to control efficiency and robustness. The $S$-Rocke estimator is defined with

$$
\rho_R(t) = \begin{cases} 
0 & \text{if } 0 \leq t \leq 1 - \gamma \\
\frac{t^{-1}}{4\gamma} \left[3 - \left(\frac{t-1}{\gamma}\right)^2\right] + \frac{1}{2} & \text{if } 1 - \gamma < t < 1 + \gamma, \\
1 & \text{if } 1 + \gamma \leq t 
\end{cases}
$$

and

$$
w_R(t) = \frac{3}{4\gamma} \left[1 - \left(\frac{t-1}{\gamma}\right)^2\right] I(1 - \gamma \leq t \leq 1 + \gamma),
$$

where the parameter $\gamma \in (0, 1]$ tunes the estimator’s efficiency and robustness. The $S$-Rocke estimator’s maximum efficiency is generally limited at $\gamma = 1$, where all points $t \geq 2$ are rejected, and which is extremely restrictive for small $p$. Both $\rho_{\text{bisq}}$ and $\rho_R$ are generic functions that do not depend on the underlying distribution. In the following section, we define an alternative $S$-estimator that accounts for the underlying distribution and generally exhibits better performance across the most common elliptical distributions. It also does not have the same inherent restrictions for small $p$ as $\rho_R$.

2.3. The Elliptical Density-Based $S_q$-Estimator

The rho function corresponding to the maximum likelihood estimate, $\hat{\sigma}_{\text{mle}}$, of the scale of $d(\mu, \Omega)$, or equivalently $d(\mu, \Sigma)$, is $\rho_{\text{mle}}(t) = -f'(t)/f(t)$. We propose weighting this rho function by the power transform of the density, that is $\tilde{\rho}_q(t) = f(t)^{1-q} \rho_{\text{mle}}(t)$, where the scalar $q \leq 1$ controls the estimator robustness, with $q = 1$ corresponding to the maximum likelihood function, and with decreasing $q$ corresponding to increased estimator robustness. In most cases, this rho function is not monotone, as required by $S$-estimators, so it is marked by a tilde. This rho function is equivalent to the ML$q$ and other $M$-estimators proposed, for example, by Windham (1995), Basu et al. (1998), Choi & Hall (2000), and Ferrari & Yang (2010). However, in this particular case of estimating the scale of the squared Mahalanobis distance of an elliptically distributed random vector, the density and rho function do not need to be recomputed with each numerical iteration $j$ based on the estimates $\hat{\mu}^{(j)}$ and $\hat{\Sigma}^{(j)}$. Substituting the PDF identified in Equation (2), we get

$$
\tilde{\rho}_q(t) = -\left(\beta_\rho \phi(t)\right)^{s_q t^q s_q} \left(t \frac{\phi'(t)}{\phi(t)} + s_p\right),
$$

where $s_p = p/2 - 1$ and $s_q = 1 - q$. Taking the derivative of $\tilde{\rho}_q$, the corresponding weight function is given by

$$
\tilde{w}_q(t) = -\left(\beta_\rho \phi(t)\right)^{s_q t^q s_q} \left(\frac{s_q s_t^2}{t} + (2s_q s_p + 1) \frac{\phi'(t)}{\phi(t)} - qt \left(\frac{\phi'(t)}{\phi(t)}\right)^2 + t \frac{\phi''(t)}{\phi(t)}\right).
$$
For simplicity, when $q < 1$, the scalar $\beta_p$ can be dropped from the calculation of $\tilde{\rho}_q$ and $\tilde{w}_q$ in Equations (5) and (6), respectively. When $\phi$ is positive only over a finite domain (e.g., a Pearson type II distribution), then we define $\tilde{\rho}_q$ and $\tilde{w}_q$ to be zero outside this domain.

For the common elliptical distributions listed in Table A1, $\tilde{\rho}_q$ is monotone in its central region between its global extrema when using appropriate values for $q$; these values are defined below. The first extremum is the minimum, which we label point $a$, and the second is the maximum, labelled $c$; this value $c$ is the same as that discussed in Section 2.2 for proper $S$-estimators. The distance between $a$ and $c$ varies monotonically with respect to $q$. We use this property to define a tunable, double-hard-rejection $S$-estimator rho function. The value of $\tilde{\rho}_q$ is held constant between zero and $a$ at the value $\tilde{\rho}_q(a)$, which hard-rejects inliers, and the value of $\tilde{\rho}_q$ is held constant above $c$ at the value $\tilde{\rho}_q(c)$, which hard-rejects outliers. The resulting monotonic function is then scaled and shifted so it ranges from 0 to 1. Thus, the definition of our $S_q$-estimator is the following:

**Definition 1.** Assuming that $\phi$ is twice continuously differentiable over its region of support and

$$
\frac{s_q s_p^2}{t} + (2 s_q s_p + 1) \frac{\phi'(t)}{\phi(t)} - qt \left( \frac{\phi'(t)}{\phi(t)} \right)^2 + t \frac{\phi''(t)}{\phi(t)}
$$

has one or two zeros in $t \in (0, \infty)$ for $q < 1$, the $S_q$-estimator is the $S$-estimator with the rho function given by

$$
\rho_q(t) = \begin{cases} 
0 & \text{if } q < 1 \text{ and } t \leq a \\
\frac{s_q}{s_1} (\tilde{\rho}_q(t) - \tilde{\rho}_q(a)) & \text{if } q < 1 \text{ and } a < t < c \\
1 & \text{if } q < 1 \text{ and } t \geq c \\
\tilde{\rho}_q(t) & \text{if } q = 1
\end{cases}
$$

(7)

where $s_1 = (\tilde{\rho}_q(b) - \tilde{\rho}_q(a))^{-1}$. The $S_q$-estimator of Type I is the case with one zero (i.e., $a = 0$), and the Type II $S_q$-estimator is the case with two zeros.

For most distributions, $\lim_{q \to 1} c = \infty$, or at $q = 1$, $\tilde{\rho}_q$ is not bounded. Therefore, we do not scale or shift $\tilde{\rho}_q$ in this case, and $\rho_q$ is not a proper $S$-estimator rho function. However, when $q = 1$ and $b = 1$, the MLE of the scale of $d(x; \mu, \Omega)$ is obtained. The $S_q$-weight function is the derivative of $\rho_q$ and is given by

$$
\tilde{w}_q(t) = \begin{cases} 
0 & \text{if } q < 1 \text{ and } t \leq a \\
\frac{1}{s_1} \tilde{w}_q(t) & \text{if } q < 1 \text{ and } a < t < c \\
0 & \text{if } q < 1 \text{ and } t \geq c \\
\tilde{w}_q(t) & \text{if } q = 1
\end{cases}
$$

(8)

Table A2 (see the Appendix) lists expressions for the inlier rejection point, $a$, and the outlier rejection point, $c$, for the common elliptical distributions listed in Table A1. For most of these distributions, the equation $\tilde{w}_q(t) = 0$ is quadratic, which provides a closed-form solution for the values of $a$ and $c$.

The asymptotic rejection probability (ARP) is defined as $Pr(d/\hat{\sigma} \geq c)$ (Rocke, 1996). Table A2 can be used to determine $q$ from a desired ARP using $F^{-1}(ARP)$. However, since $w_q$ is very tapered (i.e., applying little weight to values just below $c$), practitioners may choose...
alternative approaches to tuning that allow for higher estimator efficiencies. For example, the approach used in this article as well as in Maronna & Yohai (2017) is to tune the estimators to a desired expected efficiency, which we define in the next section.

The general definition identified in Equation (7) specifies that \( q \leq 1 \). In a few particular cases, however, there are some minor restrictions on \( q \) (when \( q < 1 \)) in order to ensure that \( a \) and \( c \) belong to the support of \( f \). Table A3 (see the Appendix) lists these restrictions.

Although the \( S_q \)-estimator is parameterized by \( \phi \) from an assumed elliptical distribution model, as we will illustrate in the following sections, the efficiency of the \( S_q \)-estimator is highly robust to errors in this assumed model. Therefore, \( S_q \)-estimators can serve as general estimators that do not require precise knowledge of the underlying data distribution.

Figure 1 illustrates examples of the \( S_q \) functions \( \tilde{\rho}_q \), \( \rho_q \), and \( w_q \) for the five-dimensional Gaussian (\( S_q \) Type II) and Laplace (\( S_q \) Type I) distributions and for various values of \( q \). As \( q \) decreases, the region of positive weights (the area between points \( a \) and \( c \)) narrows, corresponding to increased robustness. The plot also includes the PDF, illustrating how \( w_q \) roughly follows \( f \) in the central region.

Figure 2 compares the \( S_q \) asymptotic weights with those of the \( MM \)-SHR, \( S \)-Rocke, \( S \)-bisquare, and maximum likelihood estimators, and it also includes the corresponding PDF. The underlying model is a 10-dimensional standard Gaussian distribution. The \( MM \)-SHR and \( S_q \)-estimators have been tuned to 80% asymptotic efficiency relative to the MLE. The \( S \)-Rocke estimator is tuned to its maximum efficiency, which is only 77% in this case. The estimators have been set to the maximum breakdown point, with \( b = 1/2 \), which results in the shifts of the peaks of the weight curves relative to the PDF.

From Figure 2, it is clear that the Gaussian MLEs (i.e., the sample mean and covariance matrices) give uniform weight to all observations, no matter how improbable. The \( S \)-Rocke estimator has a quadratic weight function, which is a hard cutoff that cannot capture the tails of \( f \), and which has a maximum width (in this example, hard-rejecting all observations with \( d \geq 19.5 \)). The SHR weight function is cubic, and its shape is more aligned with the shape of the right half of the PDF. However, the SHR function is designed to approximate a step function, which is poorly suited for many distributions; for example, compare \( w_{\text{SHR}} \) in Figure 2 with the Laplace \( f \) plotted in Figure 1. Only the \( S_q \) weight function follows the general shape of the PDF—assigning less weight to less probable observations.
FIGURE 2: Example comparison of normalized weight functions for various estimators. For the 10-dimensional Gaussian distribution, the plot depicts the asymptotic weights for the $S_q$ ($w_q$) and MM-SHR ($w_{\text{SHR}}$) estimators tuned to 80% asymptotic relative efficiency, the $S$-Rocke ($w_r$) estimator tuned to its maximum efficiency (77% for this case), and the non-tunable MLE ($w_{\text{mle}}$) and $S$-bisquare ($w_{\text{bisq}}$) estimators. The estimators are set to their maximum breakdown points.

3. THE ASYMPTOTIC DISTRIBUTION AND RELATIVE EFFICIENCIES

In this section, we first derive the asymptotic distribution of the $S_q$-estimator, and then define various measures of efficiency. Finally, we compare the efficiency of the $S_q$-estimator with that of leading high-breakdown-point estimators, and also explore its efficiency under model mismatch.

3.1. The Asymptotic Distribution

For the asymptotic distribution of the $S_q$-estimate, we use the alternative $S$-estimator formulation identified in Equation (4). Lopuhaä (1997) derived the distribution of $S$-estimators under assumptions (A1) and (A2), where

\[(A1) \quad b = E[\rho_q \left( d \left( x_i, \mu, \Sigma \right) / \sigma \right)],\]

\[x \sim f_X (x, \mu, \Sigma, \phi(d)), \text{ where the } x_i \text{ are i.i.d.,}\]

\[\phi(d) \text{ is non-increasing,}\]

\[\phi(d) \text{ and } -\rho_q(d/\sigma) \text{ have common point(s) of decrease,}\]

and

\[(A2) \quad \phi'_p(d) \text{ is decreasing with } \phi'_p(d) < 0.\]

These assumptions may appear to be stringent, but many common symmetric elliptical distributions satisfy (A1) and (A2), including those listed in Table A1. Assumptions such as these or the ones used by Davies (1987) are commonly employed in order to derive theoretical estimator properties.

Here we use the following notation. The matrix $I_{p^2}$ is the $p^2 \times p^2$ identity matrix, $K_{p^2}$ is the $p^2 \times p^2$ commutation matrix, $\otimes$ is the Kronecker product operator, and the operator vec($\Sigma$) stacks the columns of $\Sigma$ into a column vector.
Theorem 1. Given (A1) and (A2), the asymptotic distribution of the $S_q$-estimate of $(\widehat{\mu}_n, \widehat{\Sigma}_n)$ is given by $\sqrt{n}(\widehat{\mu}_n - \mu, \widehat{\Sigma}_n - \Sigma) \rightarrow (a, B)$, with $a \perp B$. The vector $a \sim \mathcal{N}(0, \Gamma_\mu)$ where

$$\Gamma_\mu = \omega_1 \omega_2^{-2} \Sigma,$$

with

$$\omega_1 = p^{-1} E \left[ dw_q^2(d/\sigma) \right],$$

$$\omega_2 = -2\beta \int_0^\infty p^{-1} d^{p/2} w_q(d/\sigma) \phi'(d) \, dd.$$

The vector $\text{vec}(B) \sim \mathcal{N}(0, \Gamma_\Sigma)$ where

$$\Gamma_\Sigma = \xi_1 \left( I_p^2 + K_p^2 \right) (\Sigma \otimes \Sigma) + \xi_2 \text{vec}(\Sigma) \text{vec}(\Sigma)^T,$$

with

$$\xi_1 = p(p + 2) \lambda_1^{-2} E \left[ (d/\sigma)^2 w_q^2(d/\sigma) \right],$$

$$\xi_2 = \lambda_2^{-2} E \left[ (\rho_q(d/\sigma) - b)^2 \right] - 2p^{-1} \xi_1,$$

$$\lambda_1 = -2\beta \int_0^\infty \sigma^{-1} d^{p/2+1} w_q(d/\sigma) \phi'(d) \, dd,$$

$$\lambda_2 = -\beta \int_0^\infty d^{p/2} \left( \rho_q(d/\sigma) - b \right) \phi'(d) \, dd.$$

For a proof of Theorem 1, see Corollary 2 in Lopuhaä (1997).

Frahm (2009) derived the asymptotic distribution of shape matrix estimates for affine equivariant estimators. This result enables us to state the asymptotic distribution of the shape $S_q$-estimate, which is applicable using either $S$-estimator formulation; see Equations (3) and (4).

Theorem 2. Given (A1) and (A2), the asymptotic distribution of the $S_q$-estimate of shape is given by $\sqrt{n} \text{vec}(\widehat{\mathbf{Q}}_n - \mathbf{Q}) \sim \mathcal{N}(0, \Gamma_\mathbf{Q})$, where

$$\Gamma_\mathbf{Q} = \xi_1 \left( I_p^2 + K_p^2 \right) (\mathbf{Q} \otimes \mathbf{Q}) - \frac{2\xi_1}{p} \text{vec}(\mathbf{Q}) \text{vec}(\mathbf{Q})^T$$

with $\xi_1$ defined as in Theorem 1.

For a proof of Theorem 2, see Corollary 1 in Frahm (2009).

3.2. Measures of Efficiency

The asymptotic efficiency of an estimator, at an assumed distribution, is defined as the ratio of the asymptotic variance of the MLE to the variance of the estimator of interest. For multivariate estimation, this definition of efficiency is of large dimension—$p \times p$ for location and $p^2 \times p^2$ for shape and scatter. However, for affine equivariant estimation of location and scatter of elliptical distributions, the variance of the estimate depends only on a scalar. Specifically, Equations (9)–(11) are general, with only the scalars $\omega_1/\omega_2^2$ (Bilodeau & Brenner, 1999), and $\xi_1$ and
\( \zeta_2 \) (Tyler, 1982) depending on the estimator and the generating function \( \phi \). Therefore, the asymptotic efficiency of the estimate \( \hat{\mu} \) can also be defined as

\[
\text{eff}_\infty (\hat{\mu}) = \frac{\omega_{1, \text{mle}}^2}{\omega_{2, \text{mle}}^2} \cdot \frac{\omega_{1, \hat{\mu}}}{\omega_{2, \hat{\mu}}^2}
\]

and the asymptotic efficiency of the estimate \( \hat{\Omega} \) can also be defined as

\[
\text{eff}_\infty (\hat{\Omega}) = \frac{\zeta_{1, \text{mle}}}{\zeta_{1, \hat{\Omega}}}. \tag{12}
\]

It is common to define asymptotic efficiency in this way; see, e.g., Tyler (1983) and Frahm (2009).

Comparing the shape \( S_q \)-estimator’s efficiency with that of another estimator can likewise be achieved analytically using, for example, \( \zeta_{1, r} / \zeta_{1, q} \) for the \( S \)-Rocke estimator. When the resulting quotient is greater than 1, it follows that the \( S_q \)-estimator has higher asymptotic efficiency than the \( S \)-Rocke estimator. For other \( S \)-estimators, the asymptotic distribution parameters \( \omega_1, \omega_2, \) and \( \zeta_1 \) are calculated as we have indicated in Theorems 1 and 2, but using their respective weight functions. \( MM \)-estimators have the same asymptotic variance and influence function as \( S \)-estimators (Rousseeuw & Hubert, 2013). For \( MM \)-estimators, however, \( \sigma \) is effectively the tuning parameter, and it can be set accordingly.

In general, finite-sample performance measures are difficult to derive analytically. Instead, it is common to characterize finite-sample performance by empirically evaluating the behaviour of metrics derived from the Kullback–Leibler divergence between the estimated and true distributions; see, e.g., Huang et al. (2006) and Ferrari & Yang (2010). For \( t \)-distributions, which includes the Gaussian distribution, the Kullback–Leibler divergence between \( t_\nu (\mu, \Sigma) \) and \( t_\nu (\hat{\mu}, \hat{\Sigma}) \) is given by Abusev (2015)

\[
D (\mu, \Sigma; \hat{\mu}, \hat{\Sigma}) = \frac{1}{2} \left( \text{Tr} \left( \Sigma^{-1} \hat{\Sigma} \right) + (\mu - \hat{\mu})^T \Sigma^{-1} (\mu - \hat{\mu}) - p - \ln \left( \frac{|\hat{\Sigma}|}{|\Sigma|} \right) \right).
\]

Following Maronna & Yohai (2017), we then define the joint location and scatter finite-sample relative efficiency as

\[
\text{eff}_n \left( \hat{\mu}, \hat{\Sigma}; \hat{\mu}_{\text{mle}}, \hat{\Sigma}_{\text{mle}} \right) = \frac{\text{E} \left[ D \left( \mu, \Sigma; \hat{\mu}_{\text{mle}}, \hat{\Sigma}_{\text{mle}} \right) \right]}{\text{E} \left[ D \left( \mu, \Sigma; \hat{\mu}, \hat{\Sigma} \right) \right]},
\]

where \( \hat{\mu}_{\text{mle}} \) and \( \hat{\Sigma}_{\text{mle}} \) are the maximum likelihood estimates of the location and scatter matrices, respectively, and where the expectation is calculated empirically using the sample mean over \( m \) Monte Carlo trials. The location and the scatter finite-sample relative efficiencies are then respectively defined as \( \text{eff}_n (\mu, \Sigma; \hat{\mu}_{\text{mle}}, \Sigma) \) and \( \text{eff}_n (\mu, \hat{\Sigma}; \mu, \hat{\Sigma}_{\text{mle}}) \). Likewise, we define the shape matrix finite-sample relative efficiency as

\[
\text{eff}_n \left( \hat{\Omega}; \hat{\Omega}_{\text{mle}} \right) = \frac{\text{E} \left[ D \left( \mu, \Omega; \mu, \hat{\Omega}_{\text{mle}} \right) \right]}{\text{E} \left[ D \left( \mu, \Omega; \mu, \hat{\Omega} \right) \right]}. \tag{13}
\]
3.3. Comparing Estimator Efficiencies

Any estimator must provide a good estimate in the absence of contamination and when tuned to its maximum efficiency. In this section, we compare the maximum achievable efficiencies of the $S_q$, $S$-Rocke, and $MM$-SHR estimators when each is set to its maximum breakdown point. To evaluate the $S_q$-estimator’s distributional robustness, we also examine the efficiency of the $S_q$-estimator under perturbations to the assumed $S_q$ model. The results below cover large swaths of the most common elliptical families listed in Table A1 for a moderate dimension of $p = 20$. These swaths were specifically chosen to cover everyday distributions: Gaussian, Cauchy, Laplace, hyperbolic, and normal inverse Gaussian.

Robust scatter matrix estimation is generally “more difficult” than the estimation of location (Maronna et al., 2019), and as Maronna & Yohai (2017) demonstrated, divergence and efficiency metrics for scatter matrix estimators are generally much worse than for the corresponding estimators of location. Likewise, the underlying shape matrix is the most difficult part of estimating the scatter matrix. Additionally, many practical applications such as multivariate regression, principal components analysis, linear discriminant analysis, and canonical correlation analysis require only the shape matrix and not the full scatter or covariance matrices (Frahm, 2009). Therefore, unless otherwise noted, the performance results reported here are for the shape matrix, with metrics given by Equations (12) and (13), and $D(\mu, \Omega; \hat{\mu}, \hat{\Omega})$.

The maximum efficiencies of the $S_q$ and $S$-Rocke estimators generally occur when their parameters $q$ and $\gamma$ are set to 1—although the maximum breakdown point of the $S_q$-estimator is only achieved when $q < 1$. However, the maximum efficiency of the $MM$-SHR estimator must be determined by a search, as depicted in Figure 3, which plots, as an example, asymptotic efficiency versus tuning parameter for the estimators for the 20-dimensional Cauchy distribution. In the limit, as the $MM$-SHR parameter approaches infinity, all observations receive equal weight, which is the MLE for the Gaussian distribution but not for distributions such as the Cauchy. In general, for each tunable estimator, its efficiency decreases while its robustness increases as its parameter decreases. At the lower limit of its parameter space, its weight function is a delta function that may reject all the observations and may result in zero efficiency. At this point, the robustness is high, but the weighted-sum solution depends entirely on the initial estimates $\hat{\mu}^{(0)}$ and $\hat{\Omega}^{(0)}$.

![Figure 3: Estimator asymptotic relative efficiency versus tuning parameter. The relative efficiency of each estimator is plotted as a function of the tuning parameter for the Cauchy distribution with $p = 20$. All estimators are set to the maximum breakdown point. The $S$-Rocke parameter is in $[0, 1]$, the $MM$-SHR parameter is in $(0, \infty)$, and the $S_q$ parameter is in $(-\infty, 1)$ for the maximum breakdown point.](image-url)

DOI: 10.1002/cjs.11770
It should be noted that although the $S_q$-estimator generally exhibits high efficiency, at its limit with $q = 1$, it is not necessarily the maximum likelihood estimate for location and scatter. The MLE weight function for location and scatter is given by Tyler (1982)

$$w_{\text{mle}}(t) = -2 \frac{\phi'(t)}{\phi(t)},$$

(14)

whereas at $q = 1$, Equation (8) becomes

$$w_{q=1}(t) = - \frac{\phi'(t)}{\phi(t)} + t \left( \frac{\phi'(t)}{\phi(t)} \right)^2 - t \frac{\phi''(t)}{\phi(t)}.$$

(15)

**Theorem 3.** Assuming $b = E[\rho_q(d(x; \mu, \Sigma))]$, the asymptotic $S_q$-estimator with $q = 1$ is the maximum likelihood estimator for the location and scatter matrices for distributions, where

$$t \frac{\phi''(t)}{\phi(t)} - t \left( \frac{\phi'(t)}{\phi(t)} \right)^2 = y \frac{\phi'(t)}{\phi(t)}$$

for some value $y$. Therefore, asymptotically, the $S_q$-estimator can achieve the Cramér–Rao lower bound for these distributions.

A proof of Theorem 3 may be found in the Appendix.

**Remark.** Although Theorem 3 inherently assumes the alternative $S$-estimator formulation given by Equation (4), the result still holds true for location and shape matrices using the primary $S$-estimator formulation in Equation (3).

**Remark.** When this theorem holds and

$$\lim_{q \to 1} \frac{E[\rho_q(d(x; \mu, \Sigma))]}{\tilde{\rho}_q(c)} \leq \frac{1}{2},$$

then, in the limit as $q \to 1$, the Cramér–Rao lower bound can be achieved with a breakdown point of $\lim_{q \to 1} \frac{E[\rho_q(d(x; \mu, \Sigma))]}{\tilde{\rho}_q(c)}$. This means that if $\lim_{q \to 1} \tilde{\rho}_q(c)$ is finite, then, in the limit, the maximum theoretical efficiency can be obtained with a positive breakdown point.

An example family that satisfies this theorem is the Kotz type with parameter $N = 0$. Note, however, that $\lim_{q \to 1} \tilde{\rho}_q(c) = \infty$, so a high breakdown point cannot be achieved simultaneously. This example is illustrated in Figure 4, which provides the estimators’ maximum achievable asymptotic shape efficiencies for the Kotz-type distribution with parameters $N = 0$ and $r = 1/2$ as a function of parameter $s$. In this example, the $S_q$-estimator efficiency is plotted for its maximum absolute efficiency with $q = 1$ and for its approximate maximum efficiency with $q = 0.99$ at its maximum breakdown point. As seen in Figure 4, the high cost of the high breakdown point is particularly acute for large $s$.

The remainder of this article will focus on maximum efficiency at the maximum breakdown point. Figure 4 also provides the maximum efficiencies of the $S$-Rocke and MM-SHR estimators at their respective maximum breakdown points. The MM-SHR estimator efficiency peaks at $s = 1$, which is expected given that this is the Gaussian distribution, and the $S$-Rocke efficiency peaks just above this point. Their efficiencies fall off precipitously for larger and smaller values of $s$. Conversely, the efficiency of the $S_q$-estimator increases towards unity for smaller $s$. 

---

*The Canadian Journal of Statistics / La revue canadienne de statistique* DOI: 10.1002/cjs.11770
FIGURE 4: Estimator maximum achievable asymptotic efficiency for Kotz-type distribution versus parameter $s$. Maximum achievable asymptotic shape efficiencies for the maximum breakdown point are plotted for parameters $N = 0$ and $r = 1/2$. The maximum absolute asymptotic shape efficiency of the $S_q$-estimator for $q = 1$ is also shown.

FIGURE 5: Estimator maximum achievable efficiency for $t$-distribution versus the distribution parameter $\nu$. Maximum achievable asymptotic (left) and small-sample ($n = 3p$; right) efficiencies for the maximum breakdown point are plotted.

To examine the distributional robustness of the $S_q$-estimator, Figure 4 also includes the $S_q$-estimator fixed at the Gaussian model. Across the depicted range of $s$, the efficiency of the Gaussian $S_q$-estimator degrades from that of the matched $S_q$-estimator by at most 0.027 (from 99.1% to 96.3% at $s = 0.3$), which indicates high robustness to model mismatch.

The estimators’ maximum achievable asymptotic efficiencies for the $t$-distribution as a function of the distribution parameter $\nu$ are plotted in the left panel of Figure 5. When $\nu = 1$, the $t$-distribution corresponds to a Cauchy distribution, and when $\nu \to \infty$, it corresponds to the Gaussian distribution. The $S_q$-estimator exhibits the highest efficiency of the three estimators for thicker tails. The $S_q$-estimator fixed at the Cauchy model is also included in the figure and compared with the matched $S_q$-estimator. Its efficiency degrades by at most 0.0046 (from 95.79% to 95.32% at $\nu = 10$) within the depicted range of $\nu$, again indicating high robustness to model mismatch.

The maximum achievable small-sample relative efficiencies using $n = 3p$ are plotted in the right panel of Figure 5. The initial estimates were obtained using the Peña & Prieto (2007) kurtosis plus specific directions (KSD) estimator as recommended and provided by Maronna & Yohai (2017). Comparing these finite-sample results with the corresponding asymptotic values
Maximum achievable asymptotic efficiencies for the variance gamma distribution with parameter $\psi = 2$ (left) and for the generalized hyperbolic distribution with parameters $\psi = 2$ and $\chi = 1$ (right). Displayed in the left panel, one can observe that the relative results are similar. This general similarity implies that the relative performance of the asymptotic efficiencies can often be a good surrogate for the relative performance of the finite-sample efficiencies when there is no closed-form expression for the divergence appearing in Equation (13).

The estimators’ maximum achievable asymptotic efficiencies for the variance gamma distribution with $\psi = 2$ are plotted as a function of parameter $\lambda$ in the left panel of Figure 6. The plot highlights the Laplace ($\lambda = 1$) and multivariate hyperbolic ($\lambda = (p + 1)/2$) distributions. The $S_q$-estimator exhibits good performance for the hyperbolic distribution and remarkably good performance for the Laplace distribution. Fixing the $S_q$-estimator at the Laplace model results in a negligible degradation because of model mismatch of at most 0.0012 (from 98.47% to 98.35% at $\lambda = 10$) within the depicted range of $\lambda$.

The maximum achievable asymptotic efficiencies of the estimators for the generalized hyperbolic distribution with $\psi = 2$ and $\chi = 1$ are plotted as a function of parameter $\lambda$ in the right panel of Figure 6. The plot highlights the normal inverse Gaussian ($\lambda = -1/2$) and hyperbolic ($\lambda = (p + 1)/2$) distributions. Again, the $S_q$-estimator exhibits good performance for the hyperbolic distribution and remarkably good performance for the normal inverse Gaussian distribution. The $S_q$-estimator’s high distributional robustness is further illustrated by fixing the $S_q$-estimator at the normal inverse Gaussian model and observing that this model mismatch results in efficiency degradation of at most 0.0091 (from 99.11% to 98.2% at $\lambda = -15$) within the depicted range of $\lambda$.

4. ROBUSTNESS ANALYSIS

In this section, we explore the robustness of the $S_q$-estimator to outliers. First, we identify the breakdown point, and then we investigate the influence function. Finally, finite-sample simulation results are provided to further illustrate the robustness of the high-breakdown estimators.

4.1. Breakdown Point

The finite-sample breakdown point of a multivariate estimator of location or scatter is defined as the proportion of the sample that can be set to either drive $||\hat{\mu}|| = \infty$ or drive an eigenvalue of $\hat{\Sigma}$ to either zero or infinity. Unlike multivariate $M$-estimators, which have a maximum possible breakdown point of $1/(p + 1)$ (Maronna, 1976), $S$-estimators are able to achieve the maximum possible finite-sample breakdown point that any affine equivariant estimator may
have (Davies, 1987, Theorem 6). For the following theorem, \([x]\) denotes the integer part of \(x\), and the term sample in general position means that no more than \(p\) observations are contained in any hyperplane of dimension less than \(p\).

**Theorem 4.** Assuming (A1) and \(q < 1\), when a sample of size \(n\) is in general position and \(n(1 - 2b) \geq p + 1\), the breakdown point of the \(S_q\)-estimator is \((\lfloor nb \rfloor + 1)/n\).

**Corollary.** The maximum breakdown point is \((n - p + 1)/n\), which is achieved when \(b = 1/2 - (p + 1)/(2n)\). Asymptotically, this value is \(1/2\) when \(b = 1/2\).

### 4.2. Influence Function

The influence function (IF) of an estimator characterizes its sensitivity to an infinitesimal point contamination at \(z \in \mathbb{R}^p\), standardized by the mass of the contamination \(e\). The influence function for estimator \(T\), at the nominal distribution \(F\), is defined as

\[
\text{IF}(z; T, F) = \lim_{e \to 0^+} \frac{T((1-e)F + e\Delta_z) - T(F)}{e},
\]

where \(e\) is the proportion of the sample that is a point-mass, \(\Delta_z\), located at \(z\).

**Theorem 5.** Assuming (A1) and (A2), the influence functions of the \(S_q\)-estimators of \(\hat{\mu}\) and \(\hat{\Sigma}\) are

\[
\text{IF}(z; \mu, F) = \frac{\sqrt{d_z}w_q(d_z/\sigma)}{\omega_2} \frac{z_c}{\sqrt{d_z}},
\]

\[
\text{IF}(z; \Sigma, F) = \frac{\rho_q(d_z/\sigma) - b}{\lambda_2} \Sigma + \frac{p(p + 2)(d_z/\sigma)w_q(d_z/\sigma)}{\lambda_1} \left( \frac{z_c\Sigma^\top z_c}{d_z} - \frac{1}{\sigma} \Sigma \right),
\]

where \(z_c = z - \mu\) and \(d_z = z^\top \Sigma^{-1} z_c\), and where the scalars \(\omega_2, \lambda_1,\) and \(\lambda_2\) were defined in Theorem 1.

For a proof of Theorem 5, see Corollary 5.2 in Lopuhaä (1989) and Remark 2 in Lopuhaä (1997).

By the definition of \(S\)-estimators with normalized rho function, the magnitude of the first term in Equation (16) is clearly bounded above by \(\lambda_2^{-1}\Sigma\). Therefore, to compare the influence functions of the \(S_q\), \(S\)-Rocke, and MM-SHR estimators, we focus on the second term. From this term, define \(a_{z}(d_z) = \lambda_1^{-1} p(p + 2)(d_z/\sigma)w(d_z/\sigma)\) for each estimator. In Figure 7, we plot \(a_{z}(d_z)\) at the 10-dimensional Gaussian distribution for the estimators depicted in Figure 2.

By definition, all highly robust estimators have bounded IFs, and for the three estimators considered here, their IFs are continuous. This means that small amounts of contamination have only limited effects on the corresponding estimated values. The gross-error sensitivity of an estimator is the maximum of \(\text{IF}(z)\), and in this example, the \(S_q\)-estimator exhibits a lower gross-error sensitivity than both the \(S\)-Rocke and MM-SHR estimators. By its definition, the MM-SHR has an inlier rejection point of zero, meaning inliers can negatively influence the corresponding estimated values. However, proper Type II \(S_q\) functions have positive inlier rejection points, which provide robustness against inliers.
Relative to the \( S \)-Rocke and \( MM \)-SHR estimators, the similarly tuned \( S_q \)-estimator often has larger outlier rejection points. A higher rejection point is the cost of its generally higher efficiency and its ability to reject inliers. However, owing to its continuity, the influence near this point is still greatly attenuated.

4.3. Finite-Sample Robustness

To compare the finite-sample robustness of the estimators empirically, we employed the simulation method used by Maronna & Yohai (2017) and plot the shape matrix divergence\( D(\mu, \Omega; \hat{\mu}, \hat{\Omega}) \) versus shift contamination value \( k \). For a contamination proportion \( \epsilon \), the first element of each of the \( \lfloor \epsilon n \rfloor \) contaminated observations was replaced with the value \( k \), that is, \( x_1 = k \). The initial estimates of the weighted algorithm were determined with the KSD estimator. Figure 8 provides divergence plots for normally distributed data with \( \epsilon = 10\% \) contamination, for dimensions \( p = 5 \) and \( p = 20 \), and for sample sizes \( n = 5p \) and \( n = 100p \). For the cases where \( p = 20 \), the estimators were tuned to 90\% uncontaminated relative efficiency. When \( p = 5 \), the \( S \)-Rocke estimator has poor maximum efficiency, so the estimators were tuned to match the observed maximum efficiency of the \( S \)-Rocke estimator.

These plots show that the robustness of the \( S_q \)-estimator is similar to that of the other two estimators. Consistent with the results from Maronna & Yohai (2017), the relative worst-case performance of the estimators varies according to factors such as dimension, sample size, and contamination percentage. For example, the \( S_q \)-estimator exhibits its best performance here for \( p = 5 \), \( n = 100p \), but the \( MM \)-SHR estimator is the best for \( p = 5 \), \( n = 5p \).

5. COMPUTATIONAL ASPECTS

This section explores computational aspects of the \( S_q \)-estimator and other high-breakdown estimators when using the weighted-sum algorithm. We first assess the stability of the estimators by comparing their sensitivities to the initial estimates \( \hat{\mu}^{(0)} \) and \( \hat{\Omega}^{(0)} \). We then evaluate their computational efficiencies by comparing the computational convergence rates of the estimators.

5.1. Stability

The primary criticism of high-breakdown estimators is that their solutions are highly sensitive to the initial estimates \( \hat{\mu}^{(0)} \) and \( \hat{\Omega}^{(0)} \) because of the non-convexity of their objective functions. The \( S_q \)-estimator helps mitigate this drawback with a generally wider weight function; see, e.g., Figure 2. To demonstrate that the \( S_q \)-estimator is more stable with respect to the initial
estimates, 1000 Gaussian Monte Carlo simulation trials were run, where for each trial, $\Omega$ was estimated twice using different initializations. For the first estimate, $\hat{\Omega}_1^{(0)}$ was set to be the true uncontaminated sample shape matrix. For the second estimate, $\hat{\Omega}_2^{(0)}$ was determined using the sample estimate from just 25% of the same uncontaminated observations, meaning that $\hat{\Omega}_2^{(0)}$ was expected to vary considerably from the reference $\hat{\Omega}_1^{(0)}$. For each trial, the estimators were all initialized with the same $\hat{\Omega}_1^{(0)}$ and $\hat{\Omega}_2^{(0)}$, and the divergence between the two final estimates $D(\hat{\Omega}_1^{(0)}, \hat{\Omega}_2^{(0)})$ was then calculated. The same values of $p$, $n$, and $\epsilon$ were used as in the simulations that we reported in Section 4.3. The contamination method was also the same, and the value of $k$ was set to the worst-case value for that estimator and for the values of $p$, $n$, and $\epsilon$ (see Figure 8).

The results are displayed in the centre of Table 1. We observe that the $S_q$-estimator was consistently the most stable of the three estimators, and the $MM$-SHR estimator was generally the most sensitive. For the near-asymptotic cases, the $S_q$-estimator exhibited no measurable differences between the two computed estimates, unlike the $MM$-SHR estimator. Like the $S_q$-estimator, the $S$-Rocke estimator exhibited no measurable difference between the two estimates for the uncontaminated near-asymptotic cases, but under contamination, its mean divergence was roughly comparable to that exhibited by the $MM$-SHR estimator.

5.2. Computational Efficiency

To compare the relative computational efficiencies of the high-breakdown estimators, we calculated the median number of iterations required for the estimators to converge for normally

---

**FIGURE 8:** Estimator mean divergences versus contamination value $k$ at the Gaussian distribution. Gaussian shape matrix divergences are plotted for $p = 5$ (left) and $p = 20$ (right), and for small sample ($n = 5p$; top) and large sample ($n = 100p$; bottom) sizes.
Table 1: Estimator stability and computational efficiency for normally distributed data.

| Dim. | Size | Contam. | Mean divergence | Median no. of iterations |
|------|------|---------|-----------------|-------------------------|
|      |      |         | $S_q$           | $S$-Rocke               | MM-SHR |
| $p$  | $n$  | $\epsilon$ |                |                         |        |
| 5    | 100p | 0%      | 0               | 0                       | 8e-5   | 13     | 14     | 14     |
| 5    | 100p | 10%     | 0               | 7e-4                    | 5e-4   | 14     | 15     | 15     |
| 20   | 100p | 0%      | 0               | 0                       | 5e-5   | 7      | 7      | 7      |
| 20   | 100p | 10%     | 0               | 1e-3                    | 3e-2   | 18     | 8      | 16     |
| 5    | 5p   | 0%      | 2e-1            | 4e-1                    | 2e-0   | 32     | 14     | 15     |
| 5    | 5p   | 10%     | 3e-1            | 5e-1                    | 2e-0   | 31     | 15     | 15     |
| 20   | 5p   | 0%      | 6e-5            | 4e-2                    | 1e-0   | 28     | 18     | 37     |
| 20   | 5p   | 10%     | 1e-0            | 2e-0                    | 3e-0   | 30     | 18     | 33     |

Note: Mean divergence values listed as “0” have simulated average divergences less than the numerical convergence criterion, $D(\hat{\Omega}^{(j)}, \hat{\Omega}^{(j-1)}) < 10^{-10}$.

distributed data for various values of $p$, $n$, and $\epsilon$. All three estimators were set to use the same tight convergence criterion $D(\hat{\Omega}^{(j)}, \hat{\Omega}^{(j-1)}) < 10^{-10}$. The initial estimates were determined using the KSD estimator, and the estimators were tuned as in the simulations reported in Section 4.3. The contamination method was also the same, and the value of $k$ was set to the worst-case value for that estimator.

The observed results may be found on the right of Table 1. For the large-sample ($n = 100p$) simulations, the $S_q$-estimator converges nearly as fast as the other two estimators (except for the one case where $p = 20$, $\epsilon = 10\%$, where the $S$-Rocke estimator performs notably better). The $S$-Rocke estimator consistently converges fastest for all of the small-sample ($n = 5p$) cases, and the small-sample convergence of the $S_q$-estimator is relatively consistent—albeit at the upper end of the spectrum. The small-sample convergence of the MM-SHR estimator is on par with that of the $S$-Rocke estimator for small $p$, but worse than the others for large $p$.

6. APPLICATION TO FINANCIAL PORTFOLIO OPTIMIZATION

A common financial application of mean and covariance matrices may be found in modern portfolio theory for the optimal allocation of portfolio investments. Under modern portfolio theory’s mean-variance framework, a minimum-variance portfolio aims to minimize the risk (i.e., variance) of the portfolio return subject to a desired expected return (Markowitz, 1952). Mathematically, this is expressed as

$$\min \alpha^T \Omega_r \alpha$$
subject to $\alpha^T \mu_r = \mu_p$, $\alpha^T 1 = 1$,

where $\alpha$ is a normalized vector of portfolio allocation for each asset, $\Omega_r$ is the shape (or covariance) matrix for the asset returns, $\mu_r$ is the expected return of each asset, $\mu_p$ is the desired expected portfolio return, and $1$ is a vector of 1’s. The solution is given by Roy (1952), Merton (1972)

$$\alpha = s_r (\mu_r^T \Omega_r^{-1} \mu_r) \Omega_r^{-1} 1 - s_r (1^T \Omega_r^{-1} \mu_r) \Omega_r^{-1} \mu_r$$
$$+ s_r \mu_p (1^T \Omega_r^{-1} 1) \Omega_r^{-1} \mu_r - s_r \mu_p (\mu_r^T \Omega_r^{-1} 1) \Omega_r^{-1} 1,$$

$$\text{(17)}$$
where $s_r$ is a scalar that ensures the elements of $\alpha$ sum to 1.

In this section, we compare the performance of the MM-SHR, $S$-Rocke, and $S_q$-estimators for the optimal allocation of investments in the component stocks of the Dow Jones Industrial Average. For each estimator, the parameters $\Omega_r$ and $\mu_r$ were estimated for the daily returns from the component stocks. Then, using a desired portfolio daily return of $\mu_p = 0.038\%$ (corresponding to a 10\% annual return), the optimal allocations $\alpha$ were calculated using Equation (17). Using $\alpha$ for each estimator, the portfolio return was then calculated for each business day of the verification period, assuming a daily rebalance of investments. Finally, each estimator’s performance was characterized by the variance of these daily returns. The variance is a measure of the volatility of the portfolio, so for this reason, the sample variance estimator provides the most comprehensive characterization of volatility. However, for completeness, we also calculated a robust estimate of variance using an $M$-estimator with Huber’s weight function.

For the $S_q$-estimator, Konlack Socgnea & Wilcox (2014) have shown that the generalized hyperbolic distribution is a good model for stock returns and, specifically, the variance gamma subclass has good parameter stability over time. Although their analysis is for log returns, daily log returns are generally close to 1, so the variance gamma model should also fit well for gross (i.e., linear) returns. For the variance gamma $S_q$-estimator, a density-weighted $M$-estimator was used to estimate the model parameters $\lambda$ and $\psi$.

To demonstrate the robustness of the $S_q$-estimator, we begin by noting that the first quarter of 2020 contained a once-in-a-generation period of extremely high volatility due to the COVID-19 pandemic; see Figure 9. This volatility started on approximately February 21. Each estimator’s performance was assessed by estimating the parameters $\Omega_r$ and $\mu_r$ using all the returns from the first quarter, and then comparing the variances of the daily portfolio returns for only the pre-pandemic (i.e., prior to February 21) period. Each estimator was set to its maximum breakdown point. Each estimator was then tuned to its maximum asymptotic efficiency with respect to the variance gamma distribution with parameters estimated using a maximum likelihood approach and using the daily returns for the years 2016–2019.

Table 2 summarizes our observed results, listing the variances of the daily returns. The $S_q$-estimator performed best with the lowest variance, which indicates high robustness. The MM-SHR estimator exhibited the second-best performance, followed by the $S$-Rocke estimator.
Table 2: Variances of achieved daily returns for January 1, 2020 to February 20, 2020.

| Estimator   | Sample variance | Robust variance |
|-------------|-----------------|-----------------|
| $S_q$       | 75.6            | 73.0            |
| MM-SHR     | 118.6           | 110.4           |
| $S$-Rocke   | 146.6           | 148.9           |
| Sample      | 176.3           | 162.8           |

Table 3: Variances of achieved daily returns by year.

| Year | $S_q$ | MM-SHR | $S$-Rocke | $S_q$ | MM-SHR | $S$-Rocke |
|------|-------|--------|-----------|-------|--------|-----------|
| 2016 | 35.1  | 35.6   | 41.0      | 32.8  | 33.4   | 35.9      |
| 2017 | 11.8  | 12.4   | 13.4      | 11.6  | 12.0   | 12.7      |
| 2018 | 67.6  | 69.1   | 73.1      | 51.0  | 54.7   | 61.6      |
| 2019 | 36.0  | 35.0   | 44.9      | 32.2  | 31.3   | 37.4      |
| Sample mean | 37.7 | 38.0   | 43.1      | 31.9  | 32.8   | 36.9      |

The sample estimator of mean and covariance was also included to demonstrate its poor robustness.

Next, to demonstrate estimator efficiency, daily return variances were compared for a non-volatile period: 2016 through 2019. Using the same methodology and configuration as before, for each year and each estimator, $\Omega$ and $\mu_r$ were estimated. Then $\alpha$ was calculated and applied to each day of that year. The variances of each year’s daily portfolio returns are listed in Table 3, with the lowest variance highlighted with bold font. The $S_q$-estimator resulted in the lowest portfolio variance for 3 of the 4 years and the lowest variance on average, indicating high estimator efficiency. On average, the performance of the MM-SHR estimator was worse than that of the $S_q$-estimator, and the $S$-Rocke estimator gave substantially worse performance. Even with the errors in the estimates of $\lambda$ and $\psi$, together with uncertainty concerning the assumed variance gamma model, the $S_q$-estimator was able to provide efficient results because of its superior distributional robustness.

7. CONCLUSION

We have introduced the $S_q$-estimator, a new tunable multivariate estimator of location, scatter, and shape matrices for elliptical distributions. This new estimator is a subclass of $S$-estimators, which achieve the maximum theoretical breakdown point. We also compared its performance with that of the leading high-breakdown-point estimators. Across elliptical distributions, the $S_q$-estimator has generally higher efficiency and stability, it is robust to model mismatch, and its robustness to outliers is on par with that of the leading estimators. Additionally, the $S_q$-estimator provides a monotonic and upper-bounded efficiency tuning parameter, which provides simpler tuning than the MM-SHR estimator. The $S_q$-estimator is therefore broadly applicable, providing practitioners with a good general high-breakdown multivariate estimator that can be used across a broad range of practical applications, such as the optimal portfolio problem explored in Section 6.
Code for the $S_q$-estimator has been implemented in MATLAB and is available along with the simulations presented here at https://github.com/JAFishbone/.

REFERENCES

Abusev, R. A. (2015). On the distances between certain distributions in multivariate statistical analysis. *Journal of Mathematical Sciences*, 205, 2–6.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85, 549–559.

Bilodeau, M. & Brenner, D. (1999). *Theory of Multivariate Statistics*, Springer-Verlag, New York.

Choi, B. Y. E. & Hall, P. (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, 87, 453–465.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15, 1269–1292.

Deng, X. & Yao, J. (2018). On the property of multivariate generalized hyperbolic distribution and the Stein-type inequality. *Communications in Statistics - Theory and Methods*, 47, 5346–5356.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London.

Ferrari, D. & Yang, Y. (2010). Maximum L$q$-likelihood estimation. *The Annals of Statistics*, 38, 753–783.

Frahm, G. (2009). Asymptotic distributions of robust shape matrices and scales. *Journal of Multivariate Analysis*, 100, 1329–1337.

Gazor, S. & Zhang, W. (2003). Speech probability distribution. *IEEE Signal Processing Letters*, 10, 204–207.

Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93, 85–98.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A* (1961–2002), 32, 419–430.

Konlack Socgnia, V. & Wilcox, D. (2014). A comparison of generalized hyperbolic distribution models for equity returns. *Journal of Applied Mathematics*, 2014, 263465. https://doi.org/10.1155/2014/263465.

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

Lopuhaä, H. P. (1997). Asymptotic expansion of $S$-estimators of location and covariance. *Statistica Neerlandica*, 51, 220–237.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7, 77–91.

Maronna, R. A. (1976). Robust $M$-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51–67.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*, 1st ed., John Wiley & Sons, West Sussex, England.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust Statistics Theory and Methods (with R)*, 2nd ed., John Wiley & Sons, UK.

Maronna, R. A. & Yohai, V. J. (2017). Robust and efficient estimation of multivariate scatter and location. *Computational Statistics and Data Analysis*, 109, 64–75. https://doi.org/10.1016/j.csda.2016.11.006

Mériaux, B., Ren, C., El Korso, M. N., Arnaud, B., & Forster, P. (2019). Robust estimation of structured scatter matrices in (mis)matched models. *Signal Processing*, 165, 163–174.

Merton, R. C. (1972). An analytic derivation of the efficient portfolio frontier. *The Journal of Financial and Quantitative Analysis*, 7, 1851–1872.

Peña, D. & Prieto, F. J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16, 228–254.

Rocke, D. M. (1996). Robustness properties of $S$-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24, 1327–1345.

DOI: 10.1002/cjs.11770
Rousseeuw, P. & Hubert, M. (2013). High-breakdown estimators of multivariate location and scatter. In Becker, C., Fried, R., & Kuhnt, S. (Eds.), Robustness and Complex Data Structures, Springer, New York, 49–66.
Roy, A. D. (1952). Safety first and the holding of assets. Econometrica, 20, 431–449.
Tyler, D. E. (1982). Radial estimates and the test for sphericity. Biometrika, 69, 429–436.
Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. Biometrika, 70, 411–420.
Ward, K. D., Baker, C. J., & Watts, S. (1990). Maritime surveillance radar. Part 1. Radar scattering from the ocean surface. IEEE Proceedings (Radar and Signal Processing), 137, 51–62.
Windham, M. P. (1995). Robustifying model fitting. Journal of the Royal Statistical Society Series B (Methodological), 57, 599–609.

APPENDIX

Proof of Theorem 3. The $S$-estimator scaling of $b = E[\rho_q(d(x; \mu, \Sigma))]$ results in an estimator that is invariant to scaling of the weight function. Therefore, this theorem follows directly from proportionally equating Equations (14) and (15).

Proof of Theorem 4. As we discussed in Section 2.3, $q < 1$ ensures a proper $S$-estimator with finite value $c$ and bounded rho function. For the rest of the proof, see Theorem 5 in Davies (1987).
| Distribution name                                  | Generating function, \( \phi(d) \)                                                   | Range of parameters          |
|---------------------------------------------------|--------------------------------------------------------------------------------------|-----------------------------|
| Kotz type                                         | \( d^N \exp(-rd^s) \)                                                               | \( \{ r > 0, s > 0, N > p/2 \} \) |
| \( W \)                                          | \( d^{r-1} \exp(-d^s/(2^r b)) \)                                                   | \( \{ b > 0, s > 0 \} \)    |
| Gaussian                                          | \( \exp(-d/2) \)                                                                     |                             |
| Pearson type II                                   | \( (1 - d)^m, \ d \in [0, 1] \)                                                     | \( \{ m > 0 \} \)          |
| Pearson type VII                                  | \( (1 + d/s)^{-N} \)                                                                 | \( \{ N > p/2, s > 0 \} \) |
| \( t \)                                           | \( (1 + d/\nu)^{-(\nu+p)/2} \)                                                     | \( \{ \nu > 0 \} \)        |
| Cauchy                                            | \( (1 + d)^{-(1+p)/2} \)                                                            |                             |
| Generalized hyperbolic                            | \( \left( \sqrt{\psi(\chi + d)} \right)^{\lambda-p/2} K_{\lambda-p/2} \left( \sqrt{\psi(\chi + d)} \right) \) | \( \{ \psi > 0, [\chi > 0, \lambda \in \mathbb{R} \text{ or } \chi = 0, \lambda > 0] \} \) |
| Variance gamma                                    | \( \left( \sqrt{\psi \cdot d} \right)^{\lambda-p/2} K_{\lambda-p/2} \left( \sqrt{\psi \cdot d} \right) \) | \( \{ \psi > 0, \lambda > 0 \} \) |
| \( K \)                                          | \( \left( \sqrt{2\nu \cdot d} \right)^{\lambda-p/2} K_{\lambda-p/2} \left( \sqrt{2\nu \cdot d} \right) \) | \( \{ \nu > 0 \} \)        |
| Laplace                                           | \( \left( \sqrt{2d} \right)^{1-p/2} K_{1-p/2} \left( \sqrt{2d} \right) \)         |                             |
| Multivariate hyperbolic                           | \( \exp(-\sqrt{\psi(\chi + d)}) \)                                                 | \( \{ \psi > 0, \chi > 0 \} \) |
| Hyperbolic with univariate marginals               | \( \left( \sqrt{\psi(\chi + d)} \right)^{1-p/2} K_{1-p/2} \left( \sqrt{\psi(\chi + d)} \right) \) | \( \{ \psi > 0, \chi > 0 \} \) |
| Normal inverse Gaussian                           | \( \left( \sqrt{\psi(\chi + d)} \right)^{-(1+p)/2} K_{-(1+p)/2} \left( \sqrt{\psi(\chi + d)} \right) \) | \( \{ \psi > 0, \chi > 0 \} \) |

DOI: 10.1002/cjs.11770

The Canadian Journal of Statistics / La revue canadienne de statistique
**Table A2:** $s_q$ inlier and outlier rejection points for common elliptical distributions.

| Distribution          | Inlier rejection point $a$ and outlier rejection point $c$ |
|-----------------------|-----------------------------------------------------------|
| **Kotz-type**         | $a, c = \left( \frac{s+2s_pN+2s_p+\sqrt{s^2+4s_pN+4s_p^2}}{2s_p} \right)^{1/s}$ |
| **Gaussian**          | $a, c = \frac{s}{1+2s_p+\sqrt{1+4s_p}}$ |
| **Pearson type II**   | $a, c = \frac{2s_p^2+m(2s_p+1)}{2(s_p^2+m(2s_p+1))} \sqrt{m^2(4s_p^2+1)+4ms_p^2}$ |
| **Pearson type VII**  | $a, c = \frac{2N^2s_p+2N-2s_p^2(N+2N^2)}{2s_p(s^2-2Ns_p+N^2)}$ |
| **Generalized hyperbolic** | $a = \begin{cases} 0 & \text{when } \chi = 0 \text{ and } \lambda = 1 \\ \{ t|\tilde{w}_q(t) = 0 \text{ and } t \in (0, c) \} & \text{otherwise} \end{cases}$ |
|                       | $c = \{ t|\tilde{w}_q(t) = 0 \text{ and } t \in (a, \infty) \}$ |

**Table A3:** Restrictions on parameter $q$ for common elliptical distributions.

| Distribution          | Valid range of $q$                                  |
|-----------------------|-----------------------------------------------------|
| **Kotz-type**         | $q \leq 1$ unless $-1-s_p < N < -s_p$, then $1+2^{-2s}s_p+N^{-1} < q \leq 1$ |
| **Gaussian**          | $q \leq 1$                                         |
| **Pearson type II**   | $q = 1$ or $q < 1-m^{-1}$                           |
| **Pearson type VII**  | $q \leq 1$                                         |
| **Generalized hyperbolic** | $q \leq 1$ unless $\chi = 0$ and $\lambda < 1$, then unknown $< q \leq 1$ |

*aEmpirically inferred. Computational precision restricts $q \not\in (0.998, 1)$, approximately.

Received 12 February 2022
Accepted 2 November 2022