New Algorithms for \textit{M}-Estimation of Multivariate Location and Scatter

Lutz Dümbgen\textsuperscript{1*}, Klaus Nordhausen\textsuperscript{2} and Heike Schuhmacher\textsuperscript{1}
(\textsuperscript{1}University of Bern and \textsuperscript{2}University of Turku)

December 2013, revised January 2014

Abstract

We present new algorithms for \textit{M}-estimators of multivariate location and scatter and for symmetrized \textit{M}-estimators of multivariate scatter. The new algorithms are considerably faster than currently used fixed-point and other algorithms. The main idea is to utilize a Taylor expansion of second order of the target functional and devise a partial Newton-Raphson procedure. In connection with the symmetrized \textit{M}-estimators we work with incomplete \textit{U}-statistics to accelerate our procedures initially.

\textsuperscript{*}Work supported by Swiss National Science Foundation.

\textbf{AMS subject classifications.} 62H12, 65C60.

\textbf{Key words.} Fixed-point algorithm, matrix exponential function, Newton-Raphson algorithm, Taylor expansion.
1 Introduction

Robust estimation of multivariate location and scatter for a distribution $P$ on $\mathbb{R}^q$ is a recurring topic in statistics. For instance, different estimators of multivariate scatter are an important ingredient for independent component analysis (ICA) or invariant coordinate selection (ICS), see Nordhausen et al. (2008) and Tyler et al. (2009) and the references therein. Of particular interest are $M$-estimators and their symmetrized versions, because they offer a good compromise between robustness and computational feasibility. The most popular algorithm to compute $M$-estimators of multivariate scatter is to iterate a fixed-point equation, see Tyler (1987) and Kent and Tyler (1991). This algorithm has nice properties such as guaranteed convergence for any starting point. However, as discussed later, it can be rather slow for high dimensions and large data sets. Computation time becomes an even bigger issue in connection with symmetrized $M$-estimators. These estimators are important because of a desirable “block independence property” as explained later; see also Dümbgen (1998) and Sirkiä et al. (2007). If applied to a sample of $n$ observations $X_1, X_2, \ldots, X_n \in \mathbb{R}^q$, symmetrized $M$-estimators utilize the empirical distribution of all $\binom{n}{2}$ differences $X_i - X_j$, $1 \leq i < j \leq n$.

The remainder of this paper is organized as follows: In Section 2 we describe briefly the various $M$-estimators we are interested in and the general target functional on the space of symmetric and positive definite matrices in $\mathbb{R}^{q \times q}$ which has to be minimized. Section 3 presents some analytical properties of the target functional which are essential to understand existing and devise new algorithms. These parts follow closely a recent survey of multivariate $M$-functionals by Dümbgen et al. (2013), referred to as DPS (2013). In Section 4 we discuss the aforementioned fixed-point algorithm of Kent and Tyler (1991) and explain rigorously why it is suboptimal. Then we introduce alternative methods. Numerical experiments show that the latter algorithms are substantially faster than the fixed-point algorithms or the algorithms by Arslan et al. (1995). Proofs are deferred to Section 5.

Some Notation. The space of symmetric matrices in $\mathbb{R}^{q \times q}$ is denoted by $\mathbb{R}_{\text{sym}}^{q \times q}$, and $\mathbb{R}_{\text{sym},>0}^{q \times q}$ stands for its subset of positive definite matrices. The identity matrix in $\mathbb{R}^{q \times q}$ is written as $I_q$. The Euclidean norm of a vector $v \in \mathbb{R}^p$ is denoted by $\|v\| = \sqrt{v^\top v}$. For matrices $M, N$ with identical dimensions we write

$$\langle M, N \rangle := \text{tr}(M^\top N) \quad \text{and} \quad \|M\| := \sqrt{\langle M, M \rangle},$$
so \( \|M\| \) is the Frobenius norm of \( M \).

## 2 The \( M \)-estimators and the target functional

Let \( X_1, X_2, \ldots, X_n \) be independent random vectors with unknown distribution \( P \) on \( \mathbb{R}^q \). Our task is to estimate a certain center \( \mu(P) \in \mathbb{R}^q \) and scatter matrix \( \Sigma(P) \in \mathbb{R}^{q \times q}_{\text{sym},>0} \).

### 2.1 The scatter-only problem

Let us start with the assumption that \( \mu(P) = 0 \). We consider the following elliptically symmetric probability densities \( f_\Sigma \) on \( \mathbb{R}^q \) depending on a parameter \( \Sigma \in \mathbb{R}^{q \times q}_{\text{sym},>0} \):

\[
  f_\Sigma(x) = \det(\Sigma)^{-1/2} \exp\left(c - \rho(x^\top \Sigma^{-1} x)/2\right),
\]

where \( \rho : [0, \infty) \to \mathbb{R} \) is a given function and \( c \) a norming constant such that \( \int \exp(c - \rho(\|x\|^2)/2) \, dx = 1 \). The corresponding log-likelihood function for this model is given by

\[
  \Sigma \mapsto nc - \frac{1}{2} \left( \sum_{i=1}^{n} \rho(X_i^\top \Sigma^{-1} X_i) + \log \det(\Sigma) \right).
\]

With the empirical distribution \( \widehat{P} = n^{-1} \sum_{i=1}^{n} \delta_{X_i} \) of the data \( X_1, X_2, \ldots, X_n \), the log-likelihood at \( \Sigma \) may be written as \( n \int \log f_\Sigma \, d\widehat{P} \). Thus maximization of the log-likelihood function over \( \mathbb{R}^{q \times q}_{\text{sym},>0} \) is equivalent to minimization of \( \Sigma \mapsto L(\Sigma, \widehat{P}) \), where

\[
  L(\Sigma, Q) := 2 \int \log(f_{I_q}/f_\Sigma) \, dQ
\]

for a generic distribution \( Q \) on \( \mathbb{R}^q \). We include \( f_{I_q} \) and \( \rho(x^\top x) \), respectively, because often this increases the range of distributions \( Q \) such that \( L(\Sigma, Q) \) is well-defined in \( \mathbb{R} \). If \( L(\cdot, Q) \) has a unique maximizer over \( \mathbb{R}^{q \times q}_{\text{sym},>0} \), we denote it with \( \Sigma(Q) \). Thus we try to estimate \( \Sigma(P) \) by \( \Sigma(P) \), assuming that both exist. If \( P \) happens to have a density \( f_{\Sigma_o} \) in our working model, then \( \Sigma(P) = \Sigma_o \). If \( P \) is merely elliptically symmetric with center 0 and scatter matrix \( \Sigma_o \), then at least \( \Sigma(P) = \gamma \Sigma_o \) for some \( \gamma > 0 \).

An important example are multivariate \( t \) distributions with \( \nu > 0 \) degrees of freedom. Here \( \rho = \rho_{\nu,q} \) with

\[
  \rho_{\nu,q}(s) = (\nu + q) \log(\nu + s) \quad \text{for} \quad s \geq 0.
\]

(1)

Note that \( \rho(x^\top \Sigma^{-1} x) - \rho(x^\top x) \) equals \( (q + \nu) \log((\nu + x^\top \Sigma^{-1} x)/(\nu + x^\top x)) \), a bounded and smooth function of \( x \in \mathbb{R}^q \).
2.2 The location-scatter problem

Now we consider probability densities $f_{\mu, \Sigma}$ on $\mathbb{R}^q$ with parameters $\mu \in \mathbb{R}^q$ and $\Sigma \in \mathbb{R}^{q \times q}_{\text{sym}, > 0}$, namely,

$$f_{\mu, \Sigma}(x) = \det(\Sigma)^{-1/2} \exp\left(c - \rho((x - \mu)^\top \Sigma^{-1}(x - \mu)) / 2\right).$$

Here $(\mu(P'), \Sigma(P'))$ is defined as the minimizer of $2 \int \log(f_{0, I_q} / f_{\mu, \Sigma}) \, dP'$, where $P'$ stands for $P$ or $\hat{P}$. But now we utilize a trick of Kent and Tyler (1991) to get back to a scatter-only problem:

With

$$y = y(x) := \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{and} \quad \Gamma = \Gamma(\mu, \Sigma) := \begin{bmatrix} \Sigma + \mu \mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix},$$

we may write

$$-2 \log f_{\mu, \Sigma}(x) = -2c + \rho(y^\top \Gamma^{-1}y - 1),$$

and $\log \det(\Sigma) = \log \det(\Gamma)$. Hence $2 \int \log(f_{0, I_q} / f_{\mu, \Sigma}) \, dP'$ equals

$$L(\Gamma, Q) = \int \left[\rho(y^\top \Gamma^{-1}y - 1) - \rho(y^\top y - 1)\right] Q(dy) + \log \det(\Gamma)$$

with $Q := L(y(X'))$, where $X' \sim P'$. Consequently, if $\Gamma \in \mathbb{R}^{(q+1) \times (q+1)}_{\text{sym}, > 0}$ minimizes $L(\cdot, Q)$, and if

$$\Gamma_{q+1,q+1} = 1,$$

then we may write

$$\Gamma = \begin{bmatrix} \Sigma(P') + \mu(P')\mu(P')^\top & \mu(P') \\ \mu(P')^\top & 1 \end{bmatrix},$$

and $(\mu(P'), \Sigma(P'))$ solves the original minimization problem.

In the special case of $\rho = \rho_{\nu,q}$ with $\nu \geq 1$ we have the identity

$$\rho_{\nu,q}(s - 1) = \rho_{\nu-1,q+1}(s) \quad \text{for } s > 0,$$

where we define

$$\rho_{0,q}(s) := q \log(s) \quad \text{for } s > 0. \quad (2)$$

In case of $\nu > 1$ one can show that any minimizer $\Gamma$ of $L(\cdot, Q)$ does satisfy the equation $\Gamma_{q+1,q+1} = 1$. In case of $\nu = 1$, which corresponds to multivariate Cauchy distributions, any minimizer $\Gamma$ of $L(\cdot, Q)$ may be rescaled such that $\Gamma_{q+1,q+1} = 1$. Hence the location-scatter problem can be reduced to a scatter-only problem, indeed, at least in case of multivariate $t$ distributions with $\nu \geq 1$ degrees of freedom.
If \( P \) has a density \( f_{\mu_o,\Sigma_o} \) in our working model, then \( (\mu(P), \Sigma(P)) = (\mu_o, \Sigma_o) \). If \( P \) is just elliptically symmetric with center \( \mu_o \) and scatter matrix \( \Sigma_o \), then \( \mu(P) = \mu_o \) and \( \Sigma(P) = \gamma \Sigma_o \) for some \( \gamma > 0 \).

### 2.3 Symmetrized \( M \)-functionals

Suppose that \( P \) is (approximately) elliptically symmetric with unknown center \( \mu_o \) and unknown scatter matrix \( \Sigma_o \). If we are mainly interested in the shape matrix \( \det(\Sigma_o)^{-1/q} \Sigma_o \), i.e. a multiple of \( \Sigma_o \) with determinant 1, then one may get rid of the nuisance location parameter \( \mu_o \) by replacing \( P' = P \) or \( P' = \hat{P} \) with its symmetrization \( P' \ominus P' := L(X' - X'') \) with independent \( X', X'' \sim P' \).

Indeed, \( P \ominus P \) is (approximately) elliptically symmetric with center 0 and the same shape matrix \( \det(\Sigma_o)^{-1/q} \Sigma_o \). Consequently, if we define \( \Sigma(P') \) to be the minimizer of

\[
\int [\rho(x^\top \Sigma^{-1} x) - \rho(x^\top x)] (P' \ominus P')(dx) + \log \det(\Sigma)
\]

with respect to \( \Sigma \), then the shape matrix of \( \Sigma(\hat{P}) \) is a plausible estimator of \( \det(\Sigma_o)^{-1/q} \Sigma_o \).

This symmetrization has a second, even more important advantage: Consider an arbitrary distribution \( P \), i.e. it is no longer assumed to be (approximately) elliptically symmetric. Now suppose that a random vector \( X \sim P \) may be written as \( X = [X_1^\top, X_2^\top]^\top \) with independent components \( X_1 \in \mathbb{R}^{q(1)}, X_2 \in \mathbb{R}^{q(2)} \). Then \( \Sigma(P) \) is block-diagonal in the sense that

\[
\Sigma(P) = \begin{bmatrix} \Sigma_1(P) & 0 \\ 0 & \Sigma_2(P) \end{bmatrix}
\]

with symmetric matrices \( \Sigma_i(P) \in \mathbb{R}^{q(i) \times q(i)} \).

### 2.4 The general settings

Let \( Q \) be a probability distribution on \( \mathbb{R}^q \). Now we seek to minimize a certain target functional \( L(\cdot, Q) \) on the space \( \mathbb{R}^{q \times q}_{\text{sym},>0} \) of symmetric and positive definite matrices in \( \mathbb{R}^{q \times q} \), where \( L(\cdot, \cdot) \) and \( Q \) have to satisfy certain conditions:

**Setting 0.** We assume that \( Q(\{0\}) = 0 \), and for \( \Sigma \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) we define

\[
L_0(\Sigma, Q) := q \int \log \left( \frac{x^\top \Sigma^{-1} x}{x^\top x} \right) Q(dx) + \log \det(\Sigma).
\]
Moreover, we assume that
\[ Q(\mathcal{V}) < \frac{\dim(\mathcal{V})}{q} \]
for any linear subspace \( \mathcal{V} \) of \( \mathbb{R}^q \) with \( 1 \leq \dim(\mathcal{V}) < q \).

**Setting 1.** Let \( \rho : [0, \infty) \to \mathbb{R} \) be twice continuously differentiable such that \( \rho' > 0 \geq \rho'' \).

Further we assume that \( \psi'(s) := s\rho'(s) \) satisfies the following two properties: \( \psi' > 0 \) and \( q < \psi(\infty) := \lim_{s \to \infty} \psi(s) < \infty \). For \( \Sigma \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) we define
\[ L_\rho(\Sigma, Q) := \int \left[ \rho(x^\top \Sigma^{-1} x) - \rho(x^\top x) \right] Q(dx) + \log \det(\Sigma). \]

Moreover, we assume that
\[ Q(\mathcal{V}) < \frac{\psi(\infty) - q + \dim(\mathcal{V})}{\psi(\infty)} \]
for any linear subspace \( \mathcal{V} \) of \( \mathbb{R}^q \) with \( 0 \leq \dim(\mathcal{V}) < q \).

Note that for \( \nu > 0 \), \( \rho = \rho_{\nu,q} \) satisfies the conditions of Setting 1 with \( \psi(s) = (\nu+q)s/(\nu+s) \).

Hence \( \psi(\infty) = \nu + q \), and \( Q \) has to satisfy
\[ Q(\mathcal{V}) < \frac{\nu + \dim(\mathcal{V})}{\nu + q} \]
for proper linear subspaces \( \mathcal{V} \) of \( \mathbb{R}^q \).

Note also that Setting 0 is similar to Setting 1 if we define \( \rho := \rho_{0,q} \) as in \( \Xi \). The main difference to Setting 1 is that \( L_0(t\Sigma, Q) = L_0(\Sigma, Q) \) for arbitrary \( \Sigma \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) and \( t > 0 \). In what follows we often write \( L(\Sigma, Q) \) for \( L_0(\Sigma, Q) \) or \( L_\rho(\Sigma, Q) \).

The assumptions on \( \rho \) and \( Q \) imply that the functional \( L(\cdot, Q) \) has essentially a unique minimizer:

**Theorem 1** (DPS 2013). In Setting 0 there exists a unique matrix \( \Sigma_0(Q) \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) such that
\[ L_0(\Sigma_0(Q), Q) \leq L_0(\cdot, Q) \quad \text{and} \quad \det(\Sigma_0(Q)) = 1. \]

In Setting 1 there exists a unique matrix \( \Sigma_\rho(Q) \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) such that
\[ L_\rho(\Sigma_\rho(Q), Q) \leq L_\rho(\cdot, Q). \]

### 3 Analytical properties of \( L(\cdot, Q) \)

As shown in Dümbgen et al. (2013), the functionals \( L_0(\cdot, Q) \) and \( L_\rho(\cdot, Q) \) are smooth, strictly convex and coercive in a certain sense. To make this precise, we utilize the matrix-valued exponential
For \( A \in \mathbb{R}^{q \times q} \) let

\[
\exp(A) := \sum_{k=0}^{\infty} \frac{1}{k!} A^k.
\]

In case of \( A = A^\top \) we may write \( A = U \text{diag}(\lambda) U^\top \) with an orthogonal matrix \( U \in \mathbb{R}^{q \times q} \) and some vector \( \lambda = (\lambda_i)_{i=1}^{q} \in \mathbb{R}^{q} \). Then

\[
\exp(A) = U \text{diag}(\exp(\lambda)) U^\top
\]

with \( \exp(\lambda) := (\exp(\lambda_i))_{i=1}^{q} \). Moreover,

\[
\log \det(\exp(A)) = \text{tr}(A).
\]

If \( A \in \mathbb{R}_{\text{sym},>0}^{q \times q} \), i.e. \( \lambda \in (0, \infty)^{q} \), then \( A = \exp(\log(A)) \) with

\[
\log(A) := U \text{diag}(\log(\lambda)) U^\top
\]

and \( \log(\lambda) := (\log \lambda_i)_{i=1}^{q} \).

By means of the matrix-valued exponential function and logarithm, we can describe the behavior of \( L(\cdot, Q) \) in a neighborhood of any matrix \( \Sigma \in \mathbb{R}_{\text{sym},>0}^{q \times q} \) quite elegantly. To this end we write \( \Sigma = BB^\top \) for some nonsingular matrix \( B \in \mathbb{R}^{q \times q} \), e.g. \( B = \Sigma^{1/2} \). Note that

\[
\{ B \exp(A)B^\top : A \in \mathbb{R}_{\text{sym}}^{q \times q} \} = \mathbb{R}_{\text{sym},>0}^{q \times q}.
\]

In case of \( \det(\Sigma) = 1 \),

\[
\{ B \exp(A)B^\top : A \in \mathbb{R}_{\text{sym}}^{q \times q}, \text{tr}(A) = 0 \} = \{ \Gamma \in \mathbb{R}_{\text{sym},>0}^{q \times q} : \det(\Gamma) = 1 \}.
\]

Here is a basic expansion of \( L(B \exp(\cdot)B^\top, Q) \) around 0:

**Theorem 2 (DPS 2013).** For a nonsingular matrix \( B \in \mathbb{R}^{q \times q} \) define \( Q_B := \mathcal{L}(B^{-1} X) \) with \( X \sim Q \). Then for \( A \in \mathbb{R}_{\text{sym}}^{q \times q} \),

\[
L(B \exp(A)B^\top, Q) - L(BB^\top, Q) = L(\exp(A), Q_B) = G(A, Q_B) + \frac{1}{2} H(A, Q_B) + o(\|A\|^2)
\]

as \( A \to 0 \), where

\[
G(A, Q_B) := \langle A, I_q - \Psi(Q_B) \rangle,
\]

\[
H(A, Q_B) := \langle A^2, \Psi(Q_B) \rangle + \int \rho''(\|x\|^2)(x^\top A x)^2 Q_B(dx).
\]
and

\[ \Psi(Q_B) := \int \rho'(\|x\|^2) x x^\top Q_B(\text{d}x). \]

Moreover, \( H(A, Q_B) \) is continuous in \( B \), and

\[
H(A, Q_B) \begin{cases} 
\geq 0, & \text{in Setting 0, if } A \neq 0, \, \text{tr}(A) = 0, \\
> 0 & \text{in Setting 1, if } A \neq 0.
\end{cases}
\]

One consequence of Theorem 2 is that the gradient of \( L(B \exp(\cdot) B^\top, Q) \) at \( 0 \in \mathbb{R}^{q \times q}_{\text{sym}} \) is given by the matrix

\[
G(Q_B) := I_q - \Psi(Q_B) \in \mathbb{R}^{q \times q}_{\text{sym}}.
\]

Note also that \( \Psi(Q_B) \) is positive definite, because otherwise \( Q \) would be concentrated on a proper linear subspace of \( \mathbb{R}^q \).

Moreover, the second derivative of the function \( L(B \exp(\cdot) B^\top, Q) \) at \( 0 \in \mathbb{R}^{q \times q}_{\text{sym}} \) corresponds to the quadratic form

\[
\mathbb{R}^{q \times q}_{\text{sym}} \times \mathbb{R}^{q \times q}_{\text{sym}} \ni (A, A') \mapsto \langle A', H(Q_B) A \rangle
\]

with the self-adjoint linear operator \( H(Q_B) : \mathbb{R}^{q \times q}_{\text{sym}} \to \mathbb{R}^{q \times q}_{\text{sym}} \) given by

\[
H(Q_B)A := 2^{-1} (\Psi(Q_B)A + A \Psi(Q_B)) + \int \rho''(\|x\|^2) x^\top A x x^\top Q_B(\text{d}x).
\]

Theorem 2 implies that this operator is positive definite in Setting 1. In Setting 0,

\[
\Psi(Q_B) = q \int \|x\|^{-2} x x^\top Q_B(\text{d}x),
\]

\[
H(Q_B)A = 2^{-1} (\Psi(Q_B)A + A \Psi(Q_B)) - q \int \|x\|^{-4} x^\top A x x^\top Q_B(\text{d}x),
\]

and one easily verifies that \( H(Q_B)I_q = 0 \) and \( \text{tr}(H(Q_B)A) = 0 \) for any \( A \in \mathbb{R}^{q \times q}_{\text{sym}} \). Hence in both settings one may view \( H(Q_B) \) as a self-adjoint and positive definite linear operator from the set

\[
\mathbb{M} := \begin{cases} 
\{ A \in \mathbb{R}^{q \times q}_{\text{sym}} \colon \text{tr}(A) = 0 \} & \text{in Setting 0} \\
\mathbb{R}^{q \times q}_{\text{sym}} & \text{in Setting 1}
\end{cases}
\]

onto itself. In particular, \( H(Q_B)^{-1} \) stands for the corresponding inverse mapping.

An important consequence of Theorem 2 is a convexity property of \( L(\cdot, Q) \):

**Corollary 3.** For any nonsingular \( B \in \mathbb{R}^{q \times q} \) and \( A \in \mathbb{R}^{q \times q}_{\text{sym}} \), the mapping

\[
t \mapsto L(B \exp(tA) B^\top, Q)
\]

is twice continuously differentiable and convex on \( \mathbb{R} \). In Setting 0 it is strictly convex if \( A \neq 0 \) and \( \text{tr}(A) = 0 \). In Setting 1 it is strictly convex if \( A \neq 0 \).
This corollary implies that $\Sigma = BB^\top$ minimizes $L(\cdot, Q)$ if, and only if, the gradient $G(Q_B)$ equals 0, i.e.

$$\Psi(Q_B) = I_q. \quad (3)$$

This is equivalent to the fixed-point equation

$$\Sigma = \int \rho'(x^\top \Sigma^{-1} x) xx^\top Q(dx). \quad (4)$$

## 4 Algorithms

### 4.1 Fixed-point and gradient algorithms

Tyler (1987) and Kent and Tyler (1991) proposed algorithms based on iterations of the fixed-point equation (4). That means, if $\Sigma \in \mathbb{R}_{\text{sym},>0}^{q \times q}$ is our current candidate for a minimizer of $L(\cdot, Q)$, then we replace it with

$$\int \rho'(x^\top \Sigma^{-1} x) Q(dx).$$

When implementing this method we realized that it is slightly more efficient to utilize the formulation (3) directly: If $\Sigma = BB^\top$ for some nonsingular matrix $B \in \mathbb{R}^{q \times q}$, then

$$\int \rho'(x^\top \Sigma^{-1} x) Q(dx) = B \Psi(Q_B) B^\top.$$

Now we use some factorization $\Psi(Q_B) = CC^\top$ with nonsingular $C \in \mathbb{R}^{q \times q}$ and replace $B$ with $BC$.

Replacing $\Sigma$ with $B \Psi(Q_B) B^\top$ yields always an improvement, because

$$L(B \Psi(Q_B) B^\top, Q) - L(BB^\top, Q) < 0$$

unless $\Psi(Q_B) = I_q$; (5)

see DPS (2013). Here are descriptions of the two versions of fixed-point algorithms:

**Algorithm FP**. Choose an arbitrary matrix $\Sigma_0 \in \mathbb{R}_{\text{sym},>0}^{q \times q}$. Then for $k = 0, 1, 2, \ldots$ define

$$\Sigma_{k+1} := \int \rho'(x^\top \Sigma_k^{-1} x) xx^\top Q(dx).$$

**Algorithm FP**. Choose an arbitrary matrix $\Sigma_0 = B_0 B_0^\top$ with nonsingular $B_0 \in \mathbb{R}^{q \times q}$, and let $Q_0 := Q_{B_0}$.

Suppose that for some integer $k \geq 0$ we have already determined a nonsingular matrix $B_k \in \mathbb{R}^{q \times q}$. Writing $Q_k := Q_{B_k}$, we compute

$$\Psi_k := \Psi(Q_k) = \int \rho'(\|x\|^2) xx^\top Q_k(dx).$$
Then we write $\Psi_k = C_k C_k^\top$ for some nonsingular $C_k \in \mathbb{R}^{q \times q}$ and define

$$B_{k+1} := B_k C_k.$$ 

This corresponds to the new candidate $\Sigma_{k+1} := B_{k+1} B_{k+1}^\top = B_k \Psi_k B_k^\top$.

An important fact is that under the conditions of Theorem 1 the sequence $(\Sigma_k)_{k=0}^\infty$ converges to a minimizer of $L(\cdot, Q)$, no matter which starting point $\Sigma_0$ has been chosen; see also Theorem 5 later. In our algorithms we typically use $\Sigma_0 := \int xx^\top Q(dx)$.

As to the factorization $\Psi_k = C_k C_k^\top$, we utilize $C_k = U_k \text{diag}(\phi_k)^{1/2}$, where $\phi_k \in (0, \infty)^q$ contains the eigenvalues of $\Psi_k$ and $U_k$ is an orthogonal matrix of corresponding eigenvectors. In practice, of course, we need a stopping criterion for Algorithm FP, and the one we utilize is that $\|I_q - \Psi_k\| < \delta$ for some given small number $\delta > 0$.

One may view the fixed-point algorithm as an approximate gradient method: Note that with the gradient $G_k := G(Q_k)$ of $L(B_k \exp(\cdot)B_k^\top, Q_k)$ at $0 \in \mathbb{R}^{q \times q}$,

$$\Sigma_{k+1} = B_k \Psi_k B_k^\top = B_k(I_q - G_k) B_k^\top = B_k \exp(-G_k)B_k^\top + O(\|G_k\|^2).$$

**Suboptimality of Algorithm FP.** As shown later, the steps performed in Algorithm FP are clearly suboptimal, at least when $\Sigma_k$ is already close to the limit $\Sigma(Q)$. To understand this thoroughly and to devise improvements we first provide a corollary to Theorem 2.

**Corollary 4.** Let $\Sigma = BB^\top$ for a nonsingular matrix $B \in \mathbb{R}^{q \times q}$. Further let $Q_* := Q_{\Sigma(Q)^{1/2}}$. If we write $B = \Sigma^{1/2} V$ with an orthogonal matrix $V \in \mathbb{R}^{q \times q}$, then for any $A \in \mathbb{R}^{q \times q}$,

$$L(B \exp(A)B^\top, Q) - L(BB^\top, Q) = L(\exp(A), Q_B)$$

$$= G(A, Q_B) + \frac{1}{2} H(A, Q_B) + r(B, A)\|A\|^2$$

$$= G(A, Q_B) + \frac{1}{2} H(V^\top AV, Q_*) + r_*(B, A)\|A\|^2,$$

where

$$|r(B, A)| + |r_*(B, A)| \to 0 \text{ as } BB^\top \to \Sigma(Q) \text{ and } A \to 0.$$ 

Moreover,

$$H(V^\top AV, Q_*) = \|A\|^2 + \int \rho''(\|x\|^2)(x^\top V^\top AV x)^2 Q_*(dx).$$
Now let us apply this corollary to Algorithm FP. We write $B_k = \Sigma_k^{1/2} V_k$ for some orthogonal matrix $V_k \in \mathbb{R}^{q \times q}$. If we fix an arbitrary constant $C > 1$, then uniformly in $A \in \mathbb{R}^{q \times q}_{\text{sym}}$ with $\|A\| \leq C\|G_k\|$, we have

$$L(B_k \exp(A)B_k^T, Q) - L(B_k B_k^T, Q) = L(\exp(A), Q_k)$$

$$= \langle A, G_k \rangle + \frac{1}{2} H(V_k^T AV_k, Q_*) + r_s(B_k, A)\|A\|^2$$

$$= \langle A, G_k \rangle + \frac{1}{2} H(V_k^T AV_k, Q_*) + o(\|G_k\|^2).$$

In particular, if we choose $A = -t_k G_k$ with a bounded sequence $(t_k)_k$ in $[1, \infty)$,

$$L(\exp(-t_k G_k), Q_k) = \|G_k\|^2 \left( -t_k + \frac{t_k^2}{2} \frac{H(V_k^T G_k V_k, Q_*)}{\|G_k\|^2} + o(1) \right).$$

Consequently, an approximately optimal choice of $t_k$ would be

$$t_k^* = \frac{\|G_k\|^2}{H(V_k^T G_k V_k, Q_*)},$$

$$= \left( 1 + \int \rho''(||x||^2) \frac{(x^T V_k^T G_k V_k x)^2}{\|G_k\|^2} Q_*(dx) \right)^{-1} \min_{M \in \text{Mat}} \left[ \int \rho''(||x||^2) (x^T M x)^2 Q_*(dx) \right]^{-1} L_\min H(Q_*)^{-1}].$$

The upper bound involves the minimal eigenvalue of the symmetric operator $H(Q_*) : \mathbb{M} \to \mathbb{M}$. The lower bound follows from $\rho'' \leq 0$ and is typically strictly larger than 1, for instance if $\rho = \rho_{\text{var}}$ as defined in (1) or (2). Hence the steps performed during the fixed-point algorithm tend to be too short!

**Algorithm G.** One could easily fix this deficiency as follows: As a proxy for $t_k^*$, which involves the unknown quadratic form $H(\cdot, Q_*)$, we compute in the $k$-th iteration the number

$$t_k = \frac{\|G_k\|^2}{H(G_k, Q_k)} = t_k^* (1 + o(1)).$$

The latter equality follows from Corollary 4. Indeed, the latter corollary implies that we obtain

$$L(\exp(-t_k G_k), Q_k) = -\|G_k\|^2 / (2H(G_k, Q_k))(1 + o(1)) \leq -\|G_k\|^2 / (2(1 + o(1))).$$

Thus we check whether

$$L(\exp(-t_k G_k), Q_k) \leq -\|G_k\|^2 / 4. \quad (6)$$

If yes, we replace $B_k$ with $B_{k+1} = B_k C_k$, where $C_k C_k^\top = \exp(-t_k G_k)$. Otherwise we perform a usual fixed-point step as described before. The number 4 in (6) could be replaced with any number $C > 2$.

Implementing this gradient method yielded already a substantial reduction of computation time. But in view of Theorem 2, it is certainly tempting to try a Newton-Raphson procedure.

11
4.2 (Partial) Newton-Raphson procedures

Suppose that our current candidate for \( \Sigma(Q) \) is \( \Sigma = BB^\top \). In view of Corollary 4 we should replace \( \Sigma \) with

\[
\tilde{\Sigma} = B \exp\left(-H(Q_B)^{-1}G(Q_B)\right)B^\top,
\]

because \( H(Q_B)^{-1}G(Q_B) \) is the unique minimizer of

\[
\mathcal{M} \ni A \mapsto G(A,Q_B) + \frac{1}{2}H(A,Q_B) = \langle A, G(Q_B) \rangle + \frac{1}{2}\langle A, H(Q_B)A \rangle.
\]

A problem with this promising update \( \tilde{\Sigma} \) is that the computation of the inverse operator \( H(Q_B)^{-1} \) may be too computer- or memory-intensive. Indeed, we implemented a full Newton-Raphson algorithm, and it required only very few iterations, as expected. But the running time was even longer than with Algorithm FP, because the computation and inversion of \( H(Q_B) \), which may be represented by a symmetric matrix in \( \mathbb{R}^{\dim(M) \times \dim(M)} \), was too time-consuming. Note that \( \dim(M) \) equals \( q(q + 1)/2 - 1 \) in Setting 0 and \( q(q + 1)/2 \) in Setting 1.

Much more successful was the following partial Newton-Raphson approach: Consider the matrix \( \Psi(Q_B) \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) and its spectral decomposition,

\[
\Psi(Q_B) = U \operatorname{diag}(\phi)U^\top
\]

with an orthogonal matrix \( U \in \mathbb{R}^{q \times q} \) whose columns are eigenvectors of \( \Psi(Q_B) \) and a vector \( \psi \in (0, \infty)^q \) containing the corresponding eigenvalues. Now we concentrate on the functional

\[
\mathbb{R}^q \ni a \mapsto L(B \exp(\operatorname{diag}(a))U^\top B^\top, Q) - L(BB^\top, Q).
\]

That means, we consider only matrices \( \Sigma = B\Gamma B^\top \) with \( \Gamma \in \mathbb{R}^{q \times q}_{\text{sym},>0} \) such that the columns of \( U \) are eigenvectors of \( \Gamma \), too. Now the Taylor expansion in Theorem 2 may be rewritten as follows:

\[
L(B \exp(\operatorname{diag}(a))U^\top B^\top, Q) - L(BB^\top, Q)
= L(\exp(\operatorname{diag}(a)), Q_{BU}) = \tilde{G}(Q_{BU})^\top a + \frac{1}{2}a^\top \tilde{H}(Q_{BU})a + o(\|a\|^2),
\]

where

\[
\tilde{G}(Q_{BU}) := 1_q - \int \rho'(|x|^2)s(x)Q_{BU}(dx) = 1_q - \phi \in \mathbb{R}^q,
\]

\[
\tilde{H}(Q_{BU}) := \operatorname{diag}(\phi) + \int \rho''(|x|^2)s(x)s(x)^\top Q_{BU}(dx) \in \mathbb{R}^{q \times q}_{\text{sym}}
\]

with \( 1_q := (1)^q_{j=1} \) and

\[
s(x) := (x_j^2)_{j=1}^q \quad \text{for} \; x = (x_j)_{j=1}^q \in \mathbb{R}^q.
\]
In Setting 1, \( \tilde{H}(Q_B U) \) is a positive definite matrix, and
\[
\arg\min_{a \in \mathbb{R}^q} \left( \tilde{G}(Q_B U)^\top a + \frac{1}{2} a^\top \tilde{H}(Q_B U)a \right) = -\tilde{H}(Q_B U)^{-1}\tilde{G}(Q_B U).
\]
In Setting 0, the matrix \( \tilde{H}(Q_B U) \) satisfies \( \tilde{H}(Q_B U)1_q = 0 \) and \( a^\top \tilde{H}(Q_B U)a > 0 \) whenever \( a \neq 0, 1_q^\top a = 0 \). Moreover, \( 1_q^\top \tilde{G}(Q_B U) = 0 \). Thus we may write
\[
\arg\min_{a \in \mathbb{R}^q} \left( \tilde{G}(Q_B U)^\top a + \frac{1}{2} a^\top \tilde{H}(Q_B U)a \right) = -\left( \tilde{H}(Q_B U) + c 1_q 1_q^\top \right)^{-1}\tilde{G}(Q_B U)
\]
for any constant \( c > 0 \).

Algorithm PN. Choose an arbitrary matrix \( \Sigma_0 = B_0 B_0^\top \) with nonsingular \( B_0 \in \mathbb{R}^{q \times q} \), and let \( Q_0 := Q_{B_0} \).

Suppose that for some integer \( k \geq 0 \) we have already determined a nonsingular matrix \( B_k \in \mathbb{R}^{q \times q} \). Writing \( Q_k := Q_{B_k} \), we compute
\[
\Psi_k := \Psi(Q_k) = \int \rho^*(\|x\|^2) x x^\top Q_k(dx).
\]
Then we write \( \Psi_k = U_k \text{diag}(\phi_k)U_k^\top \) with an orthogonal matrix \( U_k \in \mathbb{R}^{q \times q} \) and a vector \( \phi_k \in (0, \infty)^q \). Next we define
\[
\tilde{Q}_k := (Q_k)U_k = Q_{B_k}U_k
\]
and
\[
a_k := \begin{cases} -\tilde{H}(\tilde{Q}_k)^{-1}\tilde{G}(\tilde{Q}_k) & \text{in Setting 1,} \\ -(\tilde{H}(\tilde{Q}_k) + c 1_q 1_q^\top)^{-1}\tilde{G}(\tilde{Q}_k) & \text{in Setting 0.} \end{cases}
\]
We expect that replacing \( B_k \) with \( B_k \exp(\text{diag}(a_k/2)) \) results in a change of \( L(\cdot, Q) \) of about \( a_k^\top \tilde{G}(\tilde{Q}_k)/2 < 0 \). Now we check whether
\[
L(\exp(\text{diag}(a_k)), \tilde{Q}_k) \leq a_k^\top \tilde{G}(\tilde{Q}_k)/4. \tag{7}
\]
If yes, we define
\[
B_{k+1} := B_k U_k \exp(\text{diag}(a_k/2))
\]
which corresponds to the new candidate \( \Sigma_{k+1} := B_{k+1} B_{k+1}^\top = B_k \exp(\text{diag}(a_k)) B_k^\top \). If \( \text{(7)} \) is violated we just perform a step of the fixed-point algorithm and set \( B_{k+1} := B_k U_k \exp(\text{diag}(\phi_k)^{1/2}) \), i.e. our new candidate is \( \Sigma_{k+1} := B_{k+1} B_{k+1}^\top = B_k \exp(\text{diag}(\phi_k)) B_k^\top \). Again, the number 4 in \( \text{(7)} \) could be replaced by any number \( C > 2 \).

The new Algorithm PN is also guaranteed to converge to a minimizer of \( L(\cdot, Q) \):

**Theorem 5.** For any starting point \( \Sigma_0 \in \mathbb{R}^{q \times q}_{\text{sym}, \succ 0} \) and in both Settings 0 and 1, Algorithm FP as well as Algorithm PN yield a sequence \( \langle \Sigma_k \rangle_k \) converging to a minimizer \( \Sigma_\ast \) of \( L(\cdot, Q) \).
Table 1: Pseudo-code for the $M$-estimator.

Explicit pseudo-code. Suppose that $Q = \sum_{i=1}^{n} w_i \delta_{x_i}$ with a certain weight vector $w = (w_i)_{i=1}^{n}$ in $(0,1)^n$ such that $\sum_{i=1}^{n} w_i = 1$ and a data matrix $X = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n \times q}$. Then our Algorithm PN may be formulated as in Table 1.

Symmetrized $M$-estimators. Suppose that

$$Q = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \delta_{x_i - x_j}$$

for a certain data matrix $X = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n \times q}$. In principle one could utilize the algorithm just described with $N = \binom{n}{2}$ in place of $n$ and $X$ replaced by a data matrix $\tilde{X}$ containing all $N$ differences $x_i - x_j$. For large $n$, however, this may require too much computer memory, and one should avoid the explicit storage of such a large data matrix $\tilde{X}$.

In any case it turned out that the computation time can be reduced substantially if we first
compute the $M$-estimator $\Sigma(\hat{Q})$ for the surrogate distribution
\[
\hat{Q} := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_{\pi(i)} - X_{\pi(i+1)}}
\]
with a randomly chosen permutation $\pi$ of $\{1, 2, \ldots, n\}$ and $\pi(n+1) := \pi(1)$. Then we use this estimator $\Sigma(\hat{Q})$ as a starting parameter in Algorithm PN.

Table 2 contains pseudo-code for the computation of the symmetrized $M$-estimator without using a large data matrix $\tilde{X}$. Instead it utilizes auxiliary programs to compute the following objects:

- $\text{RPermute}(n) \rightarrow$ a random permutation of $\{1, 2, \ldots, n\}$,
- $\text{Psi}(X) \rightarrow \frac{1}{N} \sum_{1 \leq i < j \leq n} \rho'(\|x_i - x_j\|^2)(x_i - x_j)(x_i - x_j)^\top$,
- $H(\phi, X) \rightarrow \text{diag}(\phi) + \frac{1}{N} \sum_{1 \leq i < j \leq n} \rho''(\|x_i - x_j\|^2)s(x_i - s_j)s(x_i - x_j)^\top$,
- $\text{DL}(X, Y, a) \rightarrow \frac{1}{N} \sum_{1 \leq i < j \leq n} \left[ \rho(\|y_i - y_j\|^2) - \rho(\|x_i - x_j\|^2) \right] + \frac{q}{N} \sum_{k=1}^{q} a_k$.

### 4.3 Numerical examples and comparisons

To compare various algorithms, we simulated 500 times a data matrix $X$ with independent entries $X_{ij} \sim \text{Exp}(1) - 1$, i.e. centered standard exponential random variables, where $n = 1000$ and $q = 50$. Then we computed the corresponding scatter estimator $\Sigma(\hat{P})$ with $\rho = \rho_{2,50}$. Iterations were stopped when $\|G_k\|_F \leq 10^{-7}$. Table 3 contains the median number of iterations and the median computing times (in seconds) using R (2013) on a MacBook Pro. In brackets the corresponding inter quartile ranges are recorded as well. In the last four rows one sees the relative efficiencies of Algorithm PN compared with Algorithms FP$_0$, FP, G and CG, respectively.

The algorithms used are FP$_0$, FP, G, CG and PN. Here CG stands for a conjugate gradient method which is not described here to save space. Roughly saying, in each step one is optimizing over a two-dimensional space spanned by the current and preceding gradient, $G_k$ and $G_{k-1}$.

Algorithm FP is substantially more efficient than Algorithm FP$_0$ with respect to computation time, despite the fact that it uses spectral decompositions of $\Psi_k$. Algorithm G is even better, both in terms of iteration numbers and time. Algorithm CG requires fewer iterations than Algorithm G, but each iteration requires more time, so there is little improvement with respect to time. Algorithm PN is clearly superior to all others.
\[ \Sigma \leftarrow \text{AlgorithmPN.symm}(X, \delta) \]
\[ \pi \leftarrow \text{RPermute}(n) \]
\[ X^0 \leftarrow [x_{\pi(1)} - x_{\pi(2)}, x_{\pi(2)} - x_{\pi(3)}, \ldots, x_{\pi(n)} - x_{\pi(1)}]^\top \]
\[ B \leftarrow \text{AlgorithmPN}(X^0, (1/n)_{i=1}^n, \delta)^{1/2} \]
\[ Y \leftarrow XB^{-1} \]
\[ \Psi \leftarrow \text{Psi}(Y) \]
\[ (U, \phi) \leftarrow \text{Eigen}(\Psi) \]

\[ \text{while} \ |1_q - \phi| > \delta \text{ do} \]
\[ B \leftarrow BU \]
\[ Y \leftarrow YU \]
\[ \tilde{H} \leftarrow H(\phi, Y) \text{ (in Setting 0)} \]
\[ a \leftarrow \tilde{H}^{-1}(\phi - 1_q) \]
\[ Z \leftarrow Y \exp(-\text{diag}(a)/2) \]
\[ DL \leftarrow DL(Y, Z, a) \]
\[ DL_0 \leftarrow a^\top(1_q - \phi)/4 \]
\[ \text{if} \ DL \leq DL_0 \text{ then} \]
\[ B \leftarrow B \exp(\text{diag}(a)/2) \]
\[ Y \leftarrow Z \]
\[ \text{else} \]
\[ B \leftarrow B \text{diag}(\phi)^{1/2} \]
\[ Y \leftarrow Y \text{diag}(\phi)^{-1/2} \]
\[ \text{end if} \]
\[ \Psi \leftarrow \text{Psi}(Y) \]
\[ (U, \phi) \leftarrow \text{Eigen}(\Psi) \]
\[ \text{end while} \]
\[ \Sigma \leftarrow BB^\top \]
\[ \text{return } \Sigma \]

Table 2: Pseudo-code for the symmetrized \( M \)-estimator.

| Algorithm | FP\(_0\) | FP | G | CG | PN |
|-----------|--------|----|---|----|----|
| iterations | 326 \((2)\) | 326 \((2)\) | 171 \((1)\) | 60 \((5)\) | 7 \((1)\) |
| time (sec) | 3.541 \((0.066)\) | 1.715 \((0.035)\) | 1.511 \((0.123)\) | 1.455 \((0.123)\) | 0.052 \((0.010)\) |
| time factors | 1.00 | 2.06 \(1.00\) | 2.34 \(1.14\) | 2.43 \(1.04\) | 68.10 \(32.98\) |

Table 3: Medians (IQRs) of computation costs in a simulation experiment.
5 Proofs

Proof of Corollaries 3 and 4. For $t \in \mathbb{R}$ define $F(t) := L(B \exp(tA)B^T, Q)$ and $B(t) := B \exp((t/2)A)$. Note that $B(t)$ is nonsingular with $B(0) = B$. For $u \in \mathbb{R}$,

$$F(t + u) - F(t) = L(B(t)\exp(uA)B(t)^T, Q) - L(B(t)B(t)^T, Q)$$

$$= L(\exp(uA), Q_{B(t)})$$

$$= uG(A, Q_{B(t)}) + \frac{u^2}{2}H(A, Q_{B(t)}) + o(u^2)$$

as $u \to 0$. Consequently, $F'(t) = G(A, Q_{B(t)})$ and for $A \in \mathbb{R}^{q \times q}_{\text{sym}},$

$$F''(t) = H(A, Q_{B(t)}) \begin{cases} 
\geq 0, \\
> 0 & \text{in Setting 0 if } A \neq 0, \text{tr}(A) = 0, \\
> 0 & \text{in Setting 1 if } A \neq 0.
\end{cases}$$

Since $H(A, Q_B)$ is continuous in $B$, and since $\mathbb{R} \ni t \mapsto B(t) \in \mathbb{R}^{q \times q}$ is continuous, we see that $F$ is twice continuously differentiable and convex. It is even strictly convex unless $\text{tr}(A) = 0$ in Setting 0 or $A = 0$ in Setting 1.

To verify Corollary 4 we utilize the same auxiliary function $F = F(\cdot | B, A)$ and write $L(B \exp(A)B^T, Q) - L(BB^T, Q)$ as

$$F(1) - F(0) = F'(0) + \int_0^1 (1 - t)F''(t) \, dt = G(A, Q_B) + \int_0^1 (1 - t)H(A, Q_{B(t)}) \, dt.$$ 

Now let $B = \Sigma^{1/2}V$ with an orthogonal matrix $V \in \mathbb{R}^{q \times q}$, and define

$$C(t) := B(t)V^T = \Sigma^{1/2}V \exp((t/2)A)V^T.$$

Then

$$r(B, A) = \|A\|^{-2} \int_0^1 (1 - t)(H(A, Q_{B(t)}) - H(A, Q_B)) \, dt$$

$$= \|A\|^{-2} \int_0^1 (1 - t)(H(V^T AV, Q_{C(t)}) - H(V^T AV, Q_{\Sigma^{1/2}})) \, dt;$$

$$r_*(B, A) = \|A\|^{-2} \int_0^1 (1 - t)(H(A, Q_{B(t)}) - H(V^T AV, Q_*)) \, dt$$

$$= \|A\|^{-2} \int_0^1 (1 - t)(H(V^T AV, Q_{C(t)}) - H(V^T AV, Q_*)) \, dt,$$

so $|r(B, A)| + |r_*(B, A)|$ is no larger than 3/2 times the supremum of

$$|H(A', Q_{\Sigma^{1/2}V, \exp(A_v)\Sigma^\ast}) - H(A', Q_*)|.$$
over all \( A', A_o \in \mathbb{R}^{d \times q} \) with \( \| A' \| \leq 1 \), \( \| A_o \| \leq \| A \| / 2 \) and all orthogonal matrices \( V_o \in \mathbb{R}^{q \times q} \).

But this converges to zero as \( \Sigma = BB^\top \rightarrow \Sigma(Q) \) and \( A \rightarrow 0 \), because then
\[
\| \Sigma^{1/2} V_o \exp(A_o) V_o^\top - \Sigma(Q)^{1/2} \| \leq \| \Sigma^{1/2} \| \| V_o \exp(A_o) V_o^\top - I_q \| + \| \Sigma^{1/2} - \Sigma(Q)^{1/2} \|
\]
\[
= \| \Sigma^{1/2} \| \| \exp(A_o) - I_q \| + \| \Sigma^{1/2} - \Sigma(Q)^{1/2} \|
\]
\[
\rightarrow 0.
\]

Finally, because \( G(Q)_* = I_q - \Psi(Q)_* = 0 \), we may write
\[
H(V^\top AV, Q)_* = \langle (V^\top AV)^2, I_q \rangle + \int \rho''(\|x\|^2)(x^\top V^\top AVx)^2 Q_*(dx)
\]
\[
= \| A \|^2 + \int \rho''(\|x\|^2)(x^\top V^\top AVx)^2 Q_*(dx).
\]

Proof of Theorem \cite{5}. Dropping the index \( k \) for the moment, suppose that \( \Sigma = BB^\top \) is our current candidate parameter. Then one step of Algorithm FP replaces \( \Sigma \) with
\[
B \Psi(Q_B) B^\top = \int \rho'(x^\top \Sigma^{-1} x) x x^\top Q(dx).
\]

Hence \( L(\Sigma, Q) \) changes by
\[
\delta_1(\Sigma) := L(B \Psi(Q_B) B^\top, Q) - L(\Sigma, Q) = L(\Psi(Q_B), Q_B) \leq 0,
\]
and the inequality is strict unless \( \Sigma \) minimizes \( L(\cdot, Q) \) already, see \cite{5}. Note also that \( \delta_1(\Sigma) \) is a continuous function of \( \Sigma \).

Algorithm PN is slightly more difficult to quantify, because the eigenmatrix \( U \) in the representation \( \Psi(Q_B) = U \text{diag}(\phi) U^\top \) is not unique. However,
\[
\min_{a \in \mathbb{R}^r} \left( \tilde{G}(Q_{BU})^\top a + \frac{1}{2} a^\top \tilde{H}(Q_{BU}) a \right) \leq \min_{a \in \text{span}(\tilde{G}(Q_{BU}))} \left( \tilde{G}(Q_{BU})^\top a + \frac{1}{2} a^\top \tilde{H}(Q_{BU}) a \right)
\]
\[
= \frac{-\| \tilde{G}(Q_{BU}) \|^2}{2 \tilde{G}(Q_{BU})^\top \tilde{H}(Q_{BU}) \tilde{G}(Q_{BU})}
\]
\[
= \frac{-\| G(Q_{BU}) \|^2}{2 G(Q_{BU})^\top G(Q_{BU})}
\]
\[
= \frac{-\| G(Q_{\Sigma^{1/2}}) \|^2}{2 H(G(Q_{\Sigma^{1/2}}), Q_{\Sigma^{1/2}})}.
\]

In the last step we utilized that fact that \( BU = \Sigma^{1/2} W \) for some orthogonal matrix \( W \in \mathbb{R}^{q \times q} \), and that \( G(Q_{BU}) = W^\top G(Q_{\Sigma^{1/2}}) W, H(G(Q_{BU}), Q_{BU}) = H(G(Q_{\Sigma^{1/2}}), Q_{\Sigma^{1/2}}) \). Consequently, the change of \( L(\Sigma, Q) \) with Algorithm PN is at least
\[
\delta_2(\Sigma) := \max \left( \delta_1(\Sigma), \frac{-\| G(Q_{\Sigma^{1/2}}) \|^2}{4 H(G(Q_{\Sigma^{1/2}}), Q_{\Sigma^{1/2}})} \right) \leq 0,
\]

18
again a continuous function of \( \Sigma \), and the inequality is strict unless \( \Sigma \) minimizes \( L(\cdot, Q) \).

In Setting 1, the minimizer \( \Sigma_\rho(Q) \) is unique, and we may utilize the following standard arguments: Suppose that \( (\Sigma_k)_k \) does not converge to \( \Sigma_\rho(Q) \). We know that \( L(\Sigma_k, Q) \) is decreasing in \( k \geq 0 \), and all \( \Sigma_k \) belong to the compact set \( \{ \Sigma : L(\Sigma, Q) \leq L(\Sigma_0, Q) \} \). Hence there would exist a subsequence \( (\Sigma_{k(\ell)})_\ell \) with limit \( \Sigma_* \neq \Sigma_\rho(Q) \). But then continuity of \( L(\cdot, Q) \) and \( \delta_j(\cdot) \) would imply that

\[
L(\Sigma_*, Q) = \lim_{\ell \to \infty} L(\Sigma_{k(\ell)}, Q) \\
= \lim_{\ell \to \infty} L(\Sigma_{k(\ell)+1}, Q) \\
\leq \lim_{\ell \to \infty} \left( L(\Sigma_{k(\ell)}, Q) + \delta_j(\Sigma_{k(\ell)}) \right) \\
= L(\Sigma_*, Q) + \delta_j(\Sigma_*) \\
< L(\Sigma_*, Q).
\]

In Setting 0, note first that \( L(\Sigma, Q), \Psi(Q_B) \) and \( H(Q_B) \) remain unchanged if we replace \( (\Sigma, B) \) with \( (t\Sigma, t^{1/2}B) \) for some number \( t > 0 \). Hence, with the same arguments as in Setting 1, we may conclude that \( t_k \Sigma_k \to \Sigma_0(Q) \) as \( k \to \infty \), where \( t_k := \det(\Sigma_k)^{-q/2} \).

Now in case of Algorithm FP an elementary calculation shows that the matrices \( M_k := \Sigma_0(Q)^{-1/2} \Sigma_k \Sigma_0(Q)^{-1/2} \) satisfy the equation

\[
M_{k+1} = \int \frac{q}{x^T M_k^{-1} x} x x^T Q \Sigma_0(Q)^{1/2} (dx).
\]

Together with the equation \( \Psi(Q \Sigma_0(Q)^{1/2}) = I_q \) this implies that

\[
\lambda_{\min}(M_{k+1}) \geq \lambda_{\min}(M_k) \quad \text{and} \quad \lambda_{\max}(M_{k+1}) \leq \lambda_{\max}(M_k).
\]

Hence the sequence \( (M_k)_k \) converges to a multiple of the identity matrix. In other words, \( (\Sigma_k)_k \) converges to a multiple of \( \Sigma_0(Q) \).

The definition of Algorithm PN implies that for sufficiently large \( k \), the new candidate \( \Sigma_{k+1} \) is given by \( B_k \exp(\text{diag}(a_k))B_k^T \) with \( a_k \in \mathbb{R}^q \) satisfying \( 1_q^T a_k = 0 \). Hence \( \det(\Sigma_{k+1}) = \det(\Sigma_k) \) for sufficiently large \( k \). Consequently \( (\Sigma_k)_k \) converges to a multiple of \( \Sigma_0(Q) \).

References

[1] O. Arslan, J. Kent, and P. Constable, Convergence behaviour of the EM algorithm for the t-distribution, Communications in Statistics, Series A: Theory and Methods, 24 (1995), pp. 2981–3000.
[2] L. Dümbgen, *On Tyler’s M-functional of scatter in high dimension*, Ann. Inst. Statist. Math., 50 (1998), pp. 471–491.

[3] L. Dümbgen, M. Pauly, and T. Schweizer, *M-functionals of multivariate location and scatter*, Tech. Rep. 77, University of Bern, 2013.

[4] J. T. Kent and D. E. Tyler, *Redescending M-estimates of multivariate location and scatter*, Ann. Statist., 19 (1991), pp. 2102–2119.

[5] K. Nordhausen, H. Oja, and E. Ollila, *Robust independent component analysis based on two scatter matrices*, Austrian J. Statist., 37 (2008), pp. 91–100.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[7] S. Sirkiä, S. Taskinen, and H. Oja, *Symmetrised M-estimators of multivariate scatter*, J. Multivar. Anal., 98 (2007), pp. 1611–1629.

[8] D. E. Tyler, *A distribution-free M-estimator of multivariate scatter*, Ann. Statist., 15 (1987), pp. 234–251.

[9] D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja, *Invariant coordinate selection (with discussion)*, J. Royal Statist. Soc. B, 71 (2009), pp. 549–592.