A Class of Robust Principal Component Vectors

Hidehiko Kamiya

Okayama University, Okayama, Japan
E-mail: kamiya@e.okayama-u.ac.jp

and

Shinto Eguchi

The Institute of Statistical Mathematics, Tokyo, Japan
E-mail: eguchi@ism.ac.jp

Received March 18, 1999; published online February 5, 2001

This paper is concerned with a study of robust estimation in principal component analysis. A class of robust estimators which are characterized as eigenvectors of weighted sample covariance matrices is proposed, where the weight functions recursively depend on the eigenvectors themselves. Also, a feasible algorithm based on iterative reweighting of the covariance matrices is suggested for obtaining these estimators in practice. Statistical properties of the proposed estimators are investigated in terms of sensitivity to outliers and relative efficiency via their influence functions, which are derived with the help of Stein's lemma. We give a simple condition on the weight functions which ensures robustness of the estimators. The class includes, as a typical example, a method by the self-organizing rule in the neural computation. A numerical experiment is conducted to confirm a rapid convergence of the suggested algorithm.

AMS classifications: 62H25, 62F35.

Key words and phrases: asymptotic relative efficiency; gross-error sensitivity; influence function; principal component analysis; robustness against outliers.

1. INTRODUCTION

It is well known that the classical rule for principal component analysis (PCA) is quite sensitive to outliers (Huber [12]). In the statistical literature, this problem is well appreciated and a number of robust procedures have been developed (Munonna [13], Devlin, Gnanadesikan and Kettenring [6]). At the same time, since the works of Amari [1] and Oja [15], there has been an increasing interest in the study of connections between PCA and neural networks, and a number of self-organizing rules for PCA have been
proposed and studied. However, little attention has been paid to the problem of outliers in the neural network literature.

Recently, using the statistical physics approach, Xu and Yuille [18] generalized several commonly used self-organizing rules for PCA into robust versions. They claim that their rules can resist outliers very well, but their assertion is based only on computer experiments.

Higuchi and Eguchi [11] made theoretical investigation into a batch version of Xu and Yuille’s rules. In particular, they found the influence function of the principal component vector in an explicit form, and thereby studied the robustness of that rule.

In the present paper, we deal with the problem of estimating the first principal component vector with special attention to robustness against outliers. Specifically, we propose a class of robust estimators which includes Xu and Yuille’s rule as a particular case.

Let \( \rho(z) \) be a non-decreasing, concave function satisfying \( \rho(0) = 0 \) and \( \psi(0) = 1 \), where \( \psi = \partial \rho / \partial z \). Given a sample \( x_1, ..., x_n \in \mathbb{R}^p \), we consider the minimization of

\[
L_n(\gamma, \mu) = \frac{1}{n} \sum_{i=1}^{n} \rho\{z(\gamma, x_i - \mu)\}
\]

with respect to \( \mu \) and \( \gamma \), \( \| \gamma \| = 1 \), where

\[
z(\gamma, x) = \frac{1}{2} \{ \| x \|^2 - (\gamma^T x)^2 \}, \quad x \in \mathbb{R}^p.
\]

Note that the orthogonal projection of \( x \) onto \( \{ c \gamma \mid c \in \mathbb{R} \} \) is \( (\gamma^T x) \gamma \) so that the squared length of the residual \( x - (\gamma^T x) \gamma \) is \( \| x - (\gamma^T x) \gamma \|^2 = \| x \|^2 - (\gamma^T x)^2 = 2z \). Denote by \( (\hat{\gamma}_*, \hat{\mu}_*) \) a minimizer of this minimization problem, and we propose this \( \hat{\gamma}_* \) as an estimator of the first principal component vector. Then, a class of estimators is obtained by taking various \( \rho \) satisfying the above conditions. Hereafter, the estimators \( \hat{\gamma}_* \) defined in this way will be called \( \rho \)-estimators.

The above definition of \( \rho \)-estimators is based on the interpretation of PCA from the viewpoint of descriptive multivariate analysis, i.e., the orthogonal projection onto a one-dimensional affine space. In particular, when \( \rho \) is the identity function, it can be seen easily that \( \hat{\mu}_* = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \), so that \( \hat{\gamma}_* \) is the vector \( \gamma \) which minimizes the sum of squared lengths of the residuals

\[\| x_i - \bar{x} \|^2 - \{ \gamma^T (x_i - \bar{x}) \}^2, \]

i.e., \( \gamma_\star \) is the classical estimator. At the same time, this procedure is closely related to the \( M \)-estimation of covariance matrices (Maronna [13], Devlin, Gnanadesikan and Kettenring [6]). In general, estimation of
covariance matrices is the key to multivariate analysis, and robust estimation of these matrices yields robustness to the classical multivariate procedures (Campbell [3, 4]). However, our procedure is designed to work well for PCA, while the conventional M-estimation procedures are, originally, aimed at the estimation of the entire covariance matrix. As will be discussed in Section 7, the difference lies in whether the weighting of each $x_i$ is based on the residual ($\rho$-estimator) or the Mahalanobis distance ($M$-estimator), and it seems that for PCA, the residual is the better choice. This assertion is supported by a simulation study in Section 8.

The classical estimator is the best from the point of view of efficiency when the underlying distribution is normal. In terms of robustness, on the other hand, we can hope that a cleverer choice of $\rho$ will yield a $\rho$-estimator superior to the classical one. In this context, we propose taking $\rho$ which satisfies

$$\sup_{z \geq 0} \left\{ -z^{1/2} \frac{\partial \rho(z)}{\partial z} \right\} < \infty.$$  

(3)

As will be made clear in the sequel, the idea behind this condition is that a $\rho$-estimator with such a $\rho$ is qualitatively robust in the sense of boundedness of the influence function; See Hampel, Ronchetti, Rousseeuw and Stahel [9]. Xu and Yuille’s self-organizing rule for robust PCA can be considered a $\rho$-estimator for a particular choice of $\rho$ satisfying (3), using $\bar{x}$ in place of $\mu_*$. (See Section 2.)

In the course of the argument, it will be seen that $\hat{\beta}_*^\rho$ satisfies the key relation

$$S_* \hat{\beta}_* = (\hat{\beta}_*^T S_* \hat{\beta}_*) \hat{\beta}_*^\rho,$$

(4)

where

$$S_* = \frac{\sum_{i=1}^n \psi(z(\hat{\beta}_*, x_i - \mu_*)) (x_i - \mu_\rho)(x_i - \mu_\rho)^T}{\sum_{i=1}^n \psi(z(\hat{\beta}_*, x_i - \mu_*))}$$

(5)

and

$$\mu_* = \frac{\sum_{i=1}^n \psi(z(\hat{\beta}_*, x_i - \mu_*)) x_i}{\sum_{i=1}^n \psi(z(\hat{\beta}_*, x_i - \mu_*))}.$$

This relation suggests seeking for $\hat{\beta}_*^\rho$, in practice, by the following algorithm: For the $(t+1)$th iteration ($t = 0, 1, ...$), take $\gamma_{t+1} \parallel \gamma_{t+1} = 1$, to be the eigenvector of

$$S_* (t) = \frac{\sum_{i=1}^n \psi(z(\gamma_t, x_i - \mu_*)) (x_i - \mu_\rho)(x_i - \mu_\rho)^T}{\sum_{i=1}^n \psi(z(\gamma_t, x_i - \mu_*))}$$
corresponding to the largest eigenvalue, and put \( \mu_{t+1} \) as

\[
\mu_{t+1} = \frac{\sum_{i=1}^{n} \psi(z(y_i, x_i - \mu_t)) x_i}{\sum_{i=1}^{n} \psi(z(y_i, x_i - \mu_t))}.
\]

In fact, this rule guarantees a steady reduction of \( L_n(y, \mu) \), that is, \( L_n(t+1) < L_n(t) \), \( L_n(t) = L_n(y, \mu_t) \), for each iteration, no matter what the starting values \( y_0, \mu_0 \) are. (The proof will be given in the Appendix.) In this algorithm, the \( p \)-estimate is obtained by iteratively reweighting the sample covariance matrix. This can be carried out by computers as a slight modification of the case of the classical PCA. We will show in a numerical study that the convergence property is quite stable.

In addition, we can easily extend our method to the \( k \)th principal component vectors \( y^{(k)}, k = 2, \ldots, p-1 \), in the following way: Set \( y^{(1)} = \hat{\gamma}_*, \mu^{(1)} = \hat{\mu}_* \), and obtain \( y^{(k)}, \mu^{(k)} \), \( k = 2, \ldots, p-1 \), sequentially by applying our method to \( x^{(k)} \), where \( x^{(0)} = x_i - \mu^{(k-1)} \) and

\[
x^{(k)}_j = x^{(0)}_j - \sum_{j=1}^{k-1} (\gamma^{(j)}^T x^{(0)}_i) \gamma^{(j)}.
\]

However, for simplicity we confine our discussion to the first principal component vector.

Now, in general, by expressing an estimator in terms of a statistical functional, we can use the methodology of the theory of robust statistics (Huber [12], Hampel, Ronchetti, Rousseeuw and Stahel [9]). With this in mind, we define our estimator through a functional, and then derive the influence function of the functional in an explicit form. This allows us to investigate our estimator from two perspectives—robustness against outliers and asymptotic relative efficiency.

In terms of gross-error sensitivity, we derive condition (3) for the robustness of a \( p \)-estimator. As for its efficiency, we obtain the asymptotic relative efficiency by calculating its asymptotic covariance matrix.

The organization of this paper is as follows. In Section 2, we introduce a statistical functional which leads to a \( p \)-estimator mentioned above. Next, in Section 3, we establish some basic facts under normality. Using those results, we establish, in Section 4, the Fisher consistency of our statistical functional, and then derive its influence function in an explicit form. By examining the influence function, we investigate the robustness of our estimator via gross-error sensitivity. In Section 5, we study its asymptotic relative efficiency. Up to Section 5, we make the assumption of normality. In Section 6, however, we point out that essentially the same discussion about the Fisher consistency and the influence function carries over to the
general case of elliptically contoured distributions. Then, in Section 7, we
discuss a relation to $M$-estimation. Finally, in Section 8, we illustrate our
theoretical results with numerical experiments.

2. A ROBUST ESTIMATOR

By introducing a functional $T_\rho(G)$ of distributions $G$ on $\mathbb{R}^p$, we define,
in this section, an estimator of the first principal component vector $\gamma_1$.

As in Section 1, let $\rho(z)$ be a non-decreasing, concave function

$$
\psi(z) = \frac{\partial}{\partial z} \rho(z) \geq 0, \quad \frac{\partial^2}{\partial z^2} \psi(z) = \frac{\partial^2}{\partial z^2} \rho(z) \leq 0
$$

such that $\rho(0) = 0$, $\psi(0) = 1$. Note that the conditions $\rho(0) = 0$ and
$\psi(0) = 1$ are just a matter of normalization; what is essentially required is
that $0 < \psi(0) < \infty$.

Let $G$ be a distribution function on $\mathbb{R}^p$. Then we consider the problem
of minimizing

$$
L_G(\gamma, \mu) = E_G[\rho(z(\gamma, x - \mu))] \tag{6}
$$

with respect to $\mu$ and $\gamma \in S^{p-1}$, where $z(\cdot, \cdot)$ is as in (2). Note that when $G$
has finite second moments, $L_G(\gamma, \mu)$ is finite for any $\gamma$ and
$\mu$ since $\rho(z) \leq z$, $z \geq 0$, although for some choices of $\rho$s (e.g., $\rho_2$ below),
$L_G(\gamma; \mu)$ is always finite regardless of whether there exist finite second
moments. Note also that in this minimization problem, the optimal $\mu$ is
determined only up to the addition of a scalar multiple of the optimal $\gamma$,
since $L_G(\gamma; \mu)$ depends on $\gamma$ and $\mu$ only through the line $\{\mu + c\gamma | c \in \mathbb{R}\}$.
for some $c \in R$. We take, in particular, $c = 0$ in the last equation, and write the corresponding solution as $(\gamma, \mu) = \{ T_\mu(G), U_\mu(G) \}$. Equation (7) will be essential in establishing the Fisher consistency of $T_\mu$ (Theorem 4.1).

We are now in a position to define our estimator $\hat{\gamma}_\mu$. Let $\hat{F}_n$ be the empirical distribution of a sample $x_1, \ldots, x_n$. Using functionals $T_\mu$ and $U_\mu$, we define $\hat{\gamma}_\mu$ and $\hat{\mu}_\mu$ as

$$
\hat{\gamma}_\mu = T_\mu(\hat{F}_n), \quad \hat{\mu}_\mu = U_\mu(\hat{F}_n),
$$

respectively.

Note that $L_\gamma(\gamma; \mu)$ with $G = \hat{F}_n$ is just $L_n(\gamma; \mu)$ in (1); thus $(\hat{\gamma}_\mu, \hat{\mu}_\mu)$ defined here is a minimizer of $L_n(\gamma; \mu)$, in agreement with the definition in Section 1. In addition, when $G = \hat{F}_n$ and $(\gamma, \mu) = (\hat{\gamma}_\mu, \hat{\mu}_\mu)$, the $\Sigma_{\gamma, \mu} = \Sigma_{\gamma, \mu}(G)$ is equal to $S_n$ in (5) and relation (7) is reduced to (4) in Section 1.

Here we give some typical examples of $\rho$.

(i) Our main concern in this paper is the case $\rho(z) = \rho_0(z; \beta, \eta)$ with

$$
\rho_0(z; \beta, \eta) = \frac{1 + \exp(-\beta \eta)}{\beta} \log \frac{1 + \exp(-\beta(z - \eta))}{1 + \exp(\beta \eta)}, \quad \beta > 0, \quad \eta > 0.
$$

The corresponding $\psi(z)$ is a sigmoid, or logistic, function:

$$
\psi_0(z; \beta, \eta) = \frac{1 + \exp(-\beta \eta)}{1 + \exp(\beta(z - \eta))}.
$$

This choice of $\rho$ leads to Xu and Yuille’s self-organizing rule for robust PCA. See below.

(ii) The classical PCA corresponds to the choice $\rho(z) = \rho_1(z)$ with

$$
\rho_1(z) = \lim_{\beta \to 0} \rho_0(z; \beta, \eta) = \lim_{\eta \to \infty} \rho_0(z; \beta, \eta) = z,
$$

that is, $\rho_1 = \text{id}$ (the identity function).

(iii) $\rho(z) = \rho_2(z; \eta)$ with

$$
\rho_2(z; \eta) = \lim_{\beta \to \infty} \rho_0(z; \beta, \eta) = \min\{z, \eta\}, \quad \eta > 0.
$$

Since $\rho_2(z; \eta) \leq \eta$, we have finite $L_\gamma(\gamma; \mu)$ for arbitrary $G$. 

Note that $\rho_0$ and $\rho_2$ satisfy (3), while $\rho_1$ does not.

The function $\rho_0$ originates from a Gibbs distribution, which is proportional to
\[
    \exp \left[ -\beta \sum_{i=1}^{n} \{ u_i z(\gamma; x_i - \mu) + (1 - u_i) \eta \} \right],
\]
(8)
in the context of statistical physics (Xu and Yuille [18]). Here, the binary field $\{u_1, ..., u_n\}$ indicates whether the observations $x_i, i = 1, ..., n$, are outliers or not. Moreover, $\beta$ is the inverse of the temperature and $\eta$ the saturation parameter. Approximating the maximization of (8) with respect to $\gamma$ and $\{u_1, ..., u_n\}$ by that of the marginal distribution of $\gamma$, we obtain our $\rho_0$:
\[
    \sum_{u_i = 0, 1} \exp \left[ -\beta \sum_{i=1}^{n} \{ u_i z(\gamma; x_i - \mu) + (1 - u_i) \eta \} \right]
    \propto \exp \left[ -\beta \sum_{i=1}^{n} \rho_0(z(\gamma; x_i - \mu); \beta, \eta) \right].
\]

The above argument could be viewed as a search for the posterior mode. Here we give a slightly different interpretation of $\rho_0$ from the point of view of the frequentist. Consider the following distribution of $(u, x)$, $u = 0, 1$, $x \in \mathbb{R}^p$:
\[
    p(u, x; \mu, \Sigma, \pi_0, \pi_1) = \left\{ \pi_0 p_0(x - \mu; \Sigma) \right\}^{1 - u} \left\{ \pi_1 p_1(x - \mu; \Sigma) \right\}^u,
\]
(9)
where $\mu \in \mathbb{R}^p$, $\Sigma > 0$, $\pi_0 + \pi_1 = 1$, $\pi_0 > 0$, $\pi_1 > 0$ and
\[
    \frac{p_0(x - \mu; \Sigma)}{p_1(x - \mu; \Sigma)} = \frac{\pi_1}{\pi_0} \exp[\beta \{ z(x - \mu; \Sigma) - \eta \}]
\]
with $z(x; \Sigma) = \frac{1}{2} x^T \Sigma^{-1} x$. Under (9), the marginal distribution of $x$ is
\[
    p(x - \mu; \Sigma) = \pi_1 p_1(x - \mu; \Sigma)(1 + \exp[\beta \{ z(x - \mu; \Sigma) - \eta \}]).
\]
(10)
Thus, (10) is elliptically symmetric with density generator $f(z)$ if and only if so is $p_1(x - \mu; \Sigma)$ with generator proportional to $f(z)[1 + e^{\beta z - \eta}]$. In particular, when $f(z) \propto e^{-z^2}$, the distribution (10) reduces to an $\varepsilon$-contaminated normal model
\[
    (1 - \varepsilon) N(\mu, \Sigma) + \varepsilon N \left( \mu, \frac{1}{1-\beta} \Sigma \right), \quad \beta < 1.
\]
In general, the model (10) is multiplicative with the parametric part $1 + e^{\beta z - \eta}$ and the nonparametric part $p_1$. Under model (9), the conditional
probability of \(x\) being an outlier given \(x\), \(\Pr(u = 0 \mid x)\), is expressed in terms of a logistic distribution function:

\[
\Pr(u = 0 \mid x) = \frac{1}{1 + \exp(-\beta \{z(x - \mu; \Sigma) - \eta\})}.
\]

Note that this is independent of \(p_1\). The logarithm of (11) is essentially \(\log(1)\).

A gradient algorithm in the on-line way proposed in neural networks is

\[
\hat{\gamma}_{t+1} = \hat{\gamma}_t + \alpha \nabla(\hat{\gamma}_t),
\]

where

\[
\nabla(\gamma) = \sum_{i=-t+q(t)}^{t} \psi_0\{z_i(x_i - \hat{\mu}_i); \beta, \eta\} [\gamma^T(x_i - \hat{\mu}_i)(x_i - \hat{\mu}_i) - \{\gamma^T(x_i - \hat{\mu}_i)\}^2].
\]

with \(z_i(x) = \frac{1}{2} [\|x\|^2 - \{\gamma^T x\}^2]\). \(\alpha\) is a learning rate, and \(\hat{\mu}_i\) is supposed to be appropriately chosen. See Oja [15] and Xu and Yuille [18]. This algorithm uses temporary information from the on-line data, and \(q(t)\) must be determined according to the objective of the analysis.

Now we consider applying our algorithm to the on-line data by analogy with on-line learning algorithms in neural networks. Suppose the data are given in the on-line way as \(x_1, x_2, \ldots, x_t, \ldots\). Then the on-line version of our algorithm is:

\[
\hat{\mu}_{t+1} = \frac{\sum_{i=-t+q(t)}^{t} \psi\{z_i(x_i - \hat{\mu}_i)\} x_i}{\sum_{i=-t+q(t)}^{t} \psi\{z_i(x_i - \hat{\mu}_i)\}},
\]

and \(\hat{\gamma}_{t+1}\) is the first eigenvector of

\[
S_\hat{\gamma}(\hat{\gamma}_t) = \frac{\sum_{i=-t+q(t)}^{t} \psi\{z_i(x_i - \hat{\mu}_i)\}(x_i - \hat{\mu}_i)(x_i - \hat{\mu}_i)^T}{\sum_{i=-t+q(t)}^{t} \psi\{z_i(x_i - \hat{\mu}_i)\}}
\]

for \(t > q(t)\). Note that if \(q(t) = q\) is fixed to be larger than \(p - 2\), then in any step \(t > q\), nonsingularity of \(S_\hat{\gamma}(\hat{\gamma}_t)\) is guaranteed.

Here we see a connection between the above two algorithms for the on-line data: First note that in our algorithm, \(\hat{\gamma}_{t+1}\) is a solution of

\[
g(\gamma; \gamma^*) = S_\gamma(\gamma^*) \gamma - \{\gamma^T S_\gamma(\gamma^*) \gamma\} \gamma = 0
\]

with respect to \(\gamma\), with \(\gamma^* = \hat{\gamma}_t\). Then we find a formal relation that \(\nabla(\gamma)\) and \(g(\gamma; \gamma^*)\) with \(\psi(z) = \phi_0(z; \beta, \eta)\) are scalar multiples of each other if \(\hat{\mu}_i\) and \(\hat{\mu}_i\) are set equal. However, we will not give further observation about the case of the on-line data because the main aim of the present paper is to compare
our method with the classical PCA or the conventional robust PCA in the batch way.

Concerning the choice of the tuning parameters $\beta$ and $\eta$ in $\rho_0$, Eguchi and Higuchi [7] propose a selection rule based on the minimization of a risk function estimated by cross validation.

3. BASIC FACTS

In the present section, we establish some basic facts.

Let $x$ be distributed according to a $p$-variate normal distribution $N_p(\mu, \Sigma)$. Suppose without loss of generality that $\mu = 0$.

Assume $\Sigma$ has distinct eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, and write its spectral decomposition as

$$\Sigma = \Gamma \Lambda \Gamma^T, \quad \Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p) \in \mathcal{O}(p), \quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p), \quad (12)$$

where $\mathcal{O}(p)$ denotes the set of all $p \times p$ orthogonal matrices.

Define

$$a(x) = \{a_1(x), \ldots, a_p(x)\}^T = \Gamma^T x,$$

and partition $\Gamma$ and $a(x)$ as

$$\Gamma = (\gamma_1, \Gamma_2)$$

and

$$a(x) = \{a_1(x), a_{(2)}(x)^T\}^T = \{\gamma_1^T x, (\Gamma_2^T x)^T\}^T,$$

respectively, where $\Gamma_2 = (\gamma_2, \ldots, \gamma_p)$ and $a_{(2)}(x) = \{a_2(x), \ldots, a_p(x)\}^T$.

Now we define a weighted covariance matrix $\Sigma_*$. Using $\psi\{z(\gamma_1, \cdot)\}$ as a weight function, we define $\Sigma_*$ as $\Sigma_* = \Sigma_{\gamma_1} 0 \{N_p(0, \Sigma)\}$. Since $z(\gamma_1, x) = \frac{1}{2} \|a_{(2)}(x)\|^2$, we can write $\Sigma_*$ as

$$\Sigma_* = E[\tilde{\psi}\{z(\gamma_1, x)\} xx^T]$$

$$= I E[\tilde{\psi}\left\{\frac{1}{2} \|a_{(2)}(x)\|^2\right\} a(x) a(x)^T] \Gamma^T,$$

where $\tilde{\psi} = \psi/E[\psi(z)]$.

Put

$$A_* = E[\tilde{\psi}\left\{\frac{1}{2} \|a_{(2)}(x)\|^2\right\} a(x) a(x)^T].$$

Then, since $a(x) \sim N_p(0, A)$, we can see the following fact.
Proposition 3.1. Matrix $A_\star$ defined above is diagonal: $A_\star = \text{diag}(\lambda_1^\star, \lambda_2^\star, ..., \lambda_p^\star)$, where $\lambda_j^\star = \mathbf{E}[\psi_\varphi(\frac{1}{2}||a_\varphi(x)||^2)a_j(x)^2]$, $j = 1, 2, ..., p$.

It follows from this proposition that

$$\Sigma_\star = I A_\star I^T$$

is a spectral decomposition of $\Sigma_\star$. Note that the sets of eigenvectors for $\Sigma$ and $\Sigma_\star$ are the same, whereas the sets of eigenvalues are not. We see in particular that $\gamma_1$ satisfies

$$\Sigma_\star \gamma_1 = \lambda_1^\star \gamma_1.$$ 

4. INFLUENCE FUNCTION AND GROSS-ERROR SENSITIVITY

In Sections 4 and 5, let $\hat{F}_n$ be the empirical distribution of the i.i.d. sample $x_1, ..., x_n$ from $N_p(0, \Sigma)$. In the present section, we derive the influence function and look into the robustness of $\hat{\theta}_n = T_\rho(\hat{F}_n)$.

First we need the following result:

**Theorem 4.1.** For a function $\rho$ defined in Section 2, functional $T_\rho$ is Fisher consistent under $N_p(0, \Sigma)$:

$$T_\rho(N_p(0, \Sigma)) = \gamma_1.$$ 

The proof is given in the Appendix for general elliptically contoured distributions.

In general, the influence function, introduced by Hampel [8], measures the effect of an additional observation on a statistic. The influence function of $T_\rho$ at $G$ is defined as

$$\text{IF}(x; T_\rho, G) = \lim_{\varepsilon \to 0^+} \frac{T_{\rho_\varepsilon}(1 - \varepsilon) G + \varepsilon \delta_x - T_\rho(G)}{\varepsilon},$$

where $\delta_x$ is the point mass 1 at $x$ (Hampel, Ronchetti, Rousseeuw and Stahel [9], Subsection 4.2a).

Using Theorem 4.1, we can calculate the influence function.

**Theorem 4.2.** The influence function of $T_\rho$ at $N_p(0, \Sigma)$ is given by

$$\text{IF}(x; T_\rho, N_p(0, \Sigma)) = \psi\left(\frac{1}{2}||a_\varphi(x)||^2\right) a_1(x) \sum_{j=2}^{p} \frac{\lambda_j^\star a_j(x)}{\lambda_j^\star \gamma_j} \gamma_j.$$
This theorem can be proved with the help of the following lemma, which is essentially Stein’s lemma (Stein [17]).

**Lemma 4.1.** Suppose that $x$ is distributed as $N_p(0, \Sigma)$. Then we have

$$E \left[ \frac{\partial \tilde{\psi}}{\partial z} \left( \frac{1}{2} \|a_{2j}(x)\|^2 \right) a_j(x)^2 \right] = \frac{\hat{\lambda}_j^*}{\lambda_1} - \frac{\hat{\lambda}_j^*}{\lambda_j}, \quad j = 2, \ldots, p.$$

**Proof.** We have by integration by parts that

$$\hat{\lambda}_j^* = E(\tilde{\psi} a_j \cdot a_j) = \hat{\lambda}_j E \left( \tilde{\psi} + \frac{\partial \tilde{\psi}}{\partial z} a_j^2 \right).$$

The proof of Theorem 4.2 will be given in the Appendix. Note that when $\psi = \psi_1$, the theorem yields the well-known result in the classical case (Critchley [5]).

Having obtained the influence function, we can look into the robustness of our estimator. In general, the influence function is preferably bounded since otherwise an outlier might cause trouble. We define gross-error sensitivity (GES) of an estimator $T(F_n)\tilde{}$ by

$$\text{GES}_{\tilde{\psi}}^2 = \frac{1}{\sigma_1} \sup_{a_{2j}} \| I(F_n; T, N_p(0, \Sigma)) \|^2$$

as the worst possible influence of a gross error (Hampel, Ronchetti, Rousseau and Stahel [9], Subsection 4.2b).

Then we have for our estimator $\tilde{\psi} = T(\tilde{\psi})$ that

$$\text{GES}_{\tilde{\psi}}^2 = \sup_{\sigma_{2j}} \left\{ \tilde{\psi} \left( \frac{1}{2} \|a_{2j}(x)\| \right)^2 \sum_{j=2}^p \frac{\hat{\lambda}_j^* a_j^2}{\lambda_j^2 (\lambda_1 - \lambda_j)^2} \right\}$$

$$\leq 2 \rho \sum_{j=2}^p \frac{\hat{\lambda}_j^*}{\lambda_j^2 (\lambda_1 - \lambda_j)^2} \sup_{z \geq 0} \{z\psi(z)^2\} \left[ \frac{E \left( \tilde{\psi} \left( \frac{1}{2} \|a_{2j}(x)\| \right)^2 \right)}{E(\tilde{\psi})} \right]^2.$$

(13)

Thus we see that if $\rho$ satisfies (3), then the $\rho$-estimator has a finite GES. Hence, $\tilde{\psi} = T(\tilde{\psi})$ and $T(\tilde{\psi})$, for example, have finite GESs; more precisely, they satisfy

$$\text{GES}_{\tilde{\psi}}^2 \leq 2 \sum_{j=2}^p \frac{\hat{\lambda}_j^*}{\lambda_j^2 (\lambda_1 - \lambda_j)^2} \left( \frac{\zeta}{\beta} + \eta \right) \left[ E(\tilde{\psi}) \right]^2.$$
and
\[
\text{GES}^{2}_{\rho_{2}} \leq 2 \sum_{j=2}^{p} \frac{\lambda_{j}^{2}}{\lambda_{j}^{2}(\lambda_{1} - \lambda_{j})^{2}} \eta_{j} \{ E(\psi_{2}) \}^{2},
\]
where \( \zeta = \max_{x, x}(1 + e^{x}) \approx 0.278 \).

On the other hand, GES for \( \rho = \rho_{1} \), the classical case, is easily seen to be infinite.

5. ASYMPTOTIC RELATIVE EFFICIENCY

In this section, we derive the asymptotic distribution of \( \hat{\gamma}_{*} = T_{\rho}(\hat{F}_{n}) \) and thereby investigate its asymptotic relative efficiency. Throughout this section, we assume \( \rho \) is twice continuously differentiable.

Recall that \( T_{\rho} \) is Fisher consistent (Theorem 4.1). Then, according to a general result on robust statistics (Huber [12], Hampel, Ronchetti, Rousseeuw and Stahel [9]), \( n^{1/2}[T_{\rho}(\hat{F}_{n}) - T_{\rho}(N_{\rho}(0, \Sigma))] \) is asymptotically normal with mean vector 0 and covariance matrix \( V[T_{\rho}, N_{\rho}(0, \Sigma)] \) equal to that of \( IF\{x; T_{\rho}, N_{\rho}(0, \Sigma)\} \). Consequently,

\[
n^{1/2}(\hat{\gamma}_{*} - \gamma_{1}) \rightarrow N_{p}[0, V[T_{\rho}, N_{\rho}(0, \Sigma)] \text{ in distribution}
\]

with \( V[T_{\rho}, N_{\rho}(0, \Sigma)] = E[IF\{x; T_{\rho}, N_{\rho}(0, \Sigma)\} \mid IF\{x; T_{\rho}, N_{\rho}(0, \Sigma)\}^{T}] \) for \( x \sim N_{\rho}(0, \Sigma) \).

From the observation above, we obtain, as a corollary to Theorem 4.2, the asymptotic covariance matrix of \( \hat{\gamma}_{*} \).

**Corollary 5.1.** Let \( x_{1}, ..., x_{n} \) be independently and identically distributed according to \( N_{\rho}(0, \Sigma) \). Then the asymptotic covariance matrix of \( \hat{\gamma}_{*} = T_{\rho}(\hat{F}_{n}) \) is given by

\[
V[T_{\rho}, N_{\rho}(0, \Sigma)] = \lambda_{1} \sum_{j=2}^{p} \frac{\lambda_{j}^{**} \lambda_{j}^{2}}{\lambda_{j}^{2}(\lambda_{1} - \lambda_{j})^{2}} \gamma_{j} \gamma_{j}^{T},
\]

where \( \lambda_{j}^{**} = E[\{ \varphi(1) \mid \alpha(2) \}_{j}^{2} \alpha_{j}^{2}] \), \( j = 2, ..., p \).

The corollary above with \( \rho = \rho_{1} \) reduces to the standard result in multivariate analysis (Anderson [2, Theorem 13.5.1], Muirhead [14, Theorem 9.5.8]).

Note that the asymptotic distribution of \( n^{1/2}(\hat{\gamma}_{*} - \gamma_{1}) \) is concentrated on the orthogonal complement of \( \{ c \gamma_{1} \mid c \in R \} \), since \( \hat{\gamma}_{*} \) and \( \gamma_{1} \) are normalized to be of unit length. Changing the coordinate system accordingly, we get
a clearer view: \( n^{1/2} T_2 \hat{\gamma}_* \) is asymptotically \( N_{p-1} [0, V \{ T_2, N_p(0, \Sigma) \} ] \),

where

\[
V \{ T_2, N_p(0, \Sigma) \} T_2 = \Gamma_2^2 V \{ T_2, N_p(0, \Sigma) \} T_2 = \lambda_1 \text{diag} \left( \frac{\hat{\lambda}_j^* \hat{\lambda}_j^2}{\hat{\lambda}_j^* (\hat{\lambda}_1 - \hat{\lambda}_j)^2} \right)_{j=2, ..., p}.
\]

Now, by the Cauchy-Schwarz inequality, we have

\[
\hat{\lambda}_j^* \hat{\lambda}_j^2 = \{ E( a_j \cdot \hat{\psi}_j ) \}^2 \leq \lambda_j \hat{\lambda}_j^*
\]

for \( j = 2, ..., p \). It follows from this that

\[
V \{ T_2, N_p(0, \Sigma) \} T_2 = \lambda_1 \text{diag} \left( \frac{\lambda_j^* \hat{\lambda}_j^2}{\lambda_j^* (\lambda_1 - \lambda_j)^2} \right)_{j=2, ..., p}
\]

in the sense that the inequality holds for each pair of corresponding diagonal elements. This is the Cramér-Rao inequality for the asymptotic covariance matrix \( \hat{T}_* \), for its right-hand side is the asymptotic covariance matrix \( V \{ T_2, N_p(0, \Sigma) \} T_2 \) of the classical first principal component vector \( \hat{T}_* = T_2 \hat{\psi}_j \), i.e., the first normalized eigenvector of the usual sample covariance matrix

\[
S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T.
\]

Note that if \( \rho \) is different from \( \rho_1 = \text{id} \), then the inequalities in (14) and (15) are strict.

These arguments are summarized in the following corollary.

**Corollary 5.2.** Under the assumption of Corollary 5.1, we have

\[
V \{ T_2, N_p(0, \Sigma) \} T_2 \geq V \{ T_2, N_p(0, \Sigma) \} T_2
\]

in the sense mentioned above. Moreover, the equality holds if and only if

\[
\rho(z) = z, z \geq 0.
\]

Now we define in general the asymptotic relative efficiency \( \text{Eff}(\hat{\gamma}) \) of an arbitrary asymptotically normal estimator \( \hat{\gamma} \) of \( \gamma_1 \):

\[
n^{1/2}(\hat{\gamma} - \gamma_1) \rightarrow N_{p-1} [0, AV(\hat{\gamma})] \text{ in distribution.}
\]
Taking degeneracy of the asymptotic distributions into consideration, we define $\text{Eff}(\hat{\gamma})$ as

$$\text{Eff}(\hat{\gamma}) = \frac{\det \text{AV}(\hat{\gamma})_{F_2}}{\det \text{AV}(\hat{\gamma})_{F_1}},$$

where $\text{AV}(\hat{\gamma})_{F_2} = I_2^T \text{AV}(\hat{\gamma}) I_2$. Then, asymptotic relative efficiency of $\hat{\gamma}_*$ is obtained from (15) as follows.

**Corollary 5.3.** Under the same assumption as in Corollary 5.1, asymptotic relative efficiency of $\hat{\gamma}_*$ is given by

$$\text{Eff}(\hat{\gamma}_*) = \prod_{j=2}^p \frac{\lambda_j^* \lambda_j^{**}}{\lambda_j},$$

To illustrate our results, consider a case with $\lambda_j = 1 + \mu_j$, where $\lambda_j = 1 + \mu_j$. Take $\psi_j(z) = \psi_j(z; \beta) = e^{-\beta_j}$, and condition (3) is satisfied. Note that $\lim_{\beta \to 0} \rho_j(z; \beta) = \rho_j(1) = z$.

First, we have

$$L_{N_p(0, \Sigma)}(\gamma, \mu) = \frac{1}{\beta} \left[ 1 - E \left[ \exp \left\{ -\frac{\beta}{2} \left( |x - \mu|^2 - (\gamma^T \Gamma(x - \mu))^2 \right) \right\} \right] \right],$$

$$\gamma = \hat{\gamma},$$

where $\beta = \beta_j$, and $I$ is the identity matrix. Thus $U_j \{ N_p(0, \Sigma) \} = 0$, and minimizing $L_{N_p(0, \Sigma)}(\gamma, 0)$ is equivalent to maximizing $\gamma^T (\Gamma \Gamma^T) \gamma$ and hence $T_n \{ N_p(0, \Sigma) \} = \gamma_1$. 


Next, we calculate $\lambda_j^*, \lambda_{j*}, j = 2, \ldots, p$. Putting $A_2 = \text{diag}(\lambda_2, \ldots, \lambda_p)$, we get

$$E(\psi_3) \lambda_j^* = \int a_j^2 \exp \left( -\frac{\beta}{2} \|a_{(2)}\|^2 \right) N_{p-1}(0, A_2)(da_{(2)})$$

$$= \det A_2^{-1/2} \det(\beta I + A_2^{-1})^{-1/2} \int a_j^2 N_{p-1}(0, (\beta I + A_2^{-1})^{-1})(da_j)$$

$$= \det(I + \beta A_2)^{-1/2} \int a_j^2 N \left( 0, \frac{1}{\beta + 1/\lambda_j} \right)(da_j)$$

$$= \det(I + \beta A_2)^{-1/2} \frac{\lambda_j}{1 + \beta \lambda_j}$$

and

$$\{E(\psi_3)\}^2 \lambda_{j*} = \int a_j^2 \exp( -\beta \|a_{(2)}\|^2) N_{p-1}(0, A_2)(da_{(2)})$$

$$= \det(I + 2\beta A_2)^{-1/2} \frac{\lambda_j}{1 + 2\beta \lambda_j}$$

for $j \geq 2$.

By making use of these expressions, we can write $\text{IF}\{x; T_{p_3}, N_p(0, \Sigma)\}$ as

$$\text{IF}(x; T_{p_3}, N_p(0, \Sigma)) = e^{-(\beta/2) \|a_{(2)}(x)\|^2} a_1(x) \det(I + \beta A_2)^{1/2} \sum_{j=2}^p (1 + \beta \lambda_j) \frac{a_j(x)}{\lambda_1 - \lambda_j} \gamma_j.$$  

Concerning the gross-error sensitivity, we have

$$\text{GES}_{p_3} = \det(I + \beta A_2) \sup_{a_{(2)}} \left\{ \sum_{j=2}^p \frac{(1 + \beta \lambda_j)^2 a_j^2}{(\lambda_1 - \lambda_j)^2} e^{-\beta \|a_{(2)}\|^2} \right\}$$

$$\leq \det(I + \beta A_2) \sup_{a_{(2)}} \left\{ \frac{(1 + \beta \lambda_2)^2 \|a_{(2)}\|^2}{(\lambda_1 - \lambda_2)^2} e^{-\beta \|a_{(2)}\|^2} \right\}$$

$$= \det(I + \beta A_2) \frac{(1 + \beta \lambda_2)^2}{(\lambda_1 - \lambda_2)^2} \frac{1}{\beta}$$

$$\leq \frac{1}{(\lambda_1 - \lambda_2)^2} \frac{(1 + \beta \lambda_2)^{p+1}}{\beta}. $$
This upper bound is minimized when \( \beta = 1/(p\lambda_2) \) with minimum value 
\[ \frac{1}{(\lambda_1 - 2\lambda_2)} \times \lambda_2 \frac{(p + 1)p^{p+1/p}}{p}. \] Accordingly, we suggest choosing \( \beta = 1/(p\lambda_2) \) for robustness consideration. Note that the first inequality above is sharper than that of (13).

As for \( V(T_p, N_p(0, \Sigma)) \) and \( \text{Eff}(\hat{\beta}_p) \), we have
\[ V(T_p, N_p(0, \Sigma)) = \lambda_j \frac{\det(I + \beta A_2)}{\det(I + 2\beta A_2)} \left( \sum_{j=2}^{p} \frac{(1 + \beta \lambda_j)^2 \lambda_j}{(1 + 2\beta \lambda_j)(\lambda_1 - \lambda_2)} \right)^2, \]
and
\[ \text{Eff}(\hat{\beta}_p) = \left\{ \frac{\det(I + 2\beta A_2)^{1/2}}{\det(I + \beta A_2)} \right\}^{p-1} \prod_{j=2}^{p} \frac{1 + 2\beta \lambda_j}{(1 + \beta \lambda_j)^{p-1/2}}. \]

When \( \beta = 1/(p\lambda_2) \) as suggested above, the right-hand side equals
\[ \left\{ \frac{p(p + 2)}{(p + 1)^2} \right\}^{(p^2 - 1)/2}. \]

This is equal to 0.84, 0.77, 0.74 for \( p = 2, 3, 4 \), respectively, and goes to \( e^{-1/2} \approx 0.61 \) as \( p \to \infty \).

We finish up this section by briefly mentioning the estimation of \( \text{Eff}(\hat{\beta}_p) \).

We estimate \( \text{Eff}(\hat{\beta}_p) \) by
\[ \frac{\hat{\lambda}_1 \hat{\lambda}^{**}}{\hat{\lambda}^*} \left( \frac{\det S^*}{\det S} \right)^2 \left( \frac{\det S}{\det S^{**}} \right)^2, \]
where \( S^* \) and \( S^{**} \) are the weighted sample covariance matrices defined by
\[ S^* = \frac{\sum_{i=1}^{\hat{n} \lambda} \psi(z(\hat{x}, \hat{x} - \hat{x})) (x_i - \hat{x})(x_i - \hat{x})^T}{\sum_{i=1}^{\hat{n} \lambda} \psi(z(\hat{x}, \hat{x} - \hat{x}))}, \]
\[ S^{**} = \frac{\sum_{i=1}^{\hat{n} \lambda} \psi(z(\hat{x}, \hat{x} - \hat{x}))^2 (x_i - \hat{x})(x_i - \hat{x})^T}{\sum_{i=1}^{\hat{n} \lambda} \psi(z(\hat{x}, \hat{x} - \hat{x}))^2}, \]
and \( \hat{\lambda}_1, \hat{\lambda}^* \) and \( \hat{\lambda}^{**} \) are the largest eigenvalues of \( S, S^* \) and \( S^{**} \), respectively.

Construction of (16) is based on the following observations: (i) \( \hat{\lambda}^{**} = E[\hat{\psi}(\hat{a}(\hat{z})^2/2)^2 \sigma_j^2] \), \( j = 1, \ldots, p \), are indeed the eigenvalues of
\[ \Sigma_{*,*} = E[\tilde{\psi}(z|\gamma_1, x)]^2 xx^T, \] and (ii) \( \lambda_j^* > \lambda_1^* \) and \( \lambda_j^{**} > \lambda_j^{*} \) for \( j = 2, ..., p \).

Assertion (ii) follows from Lemma 4.1 and the equation
\[ 2E\left( \psi \frac{\partial \tilde{\psi}}{\partial z} a_j^{*} \right) = \frac{\lambda_j^{**}}{\lambda_j^{*}}, \quad j = 2, ..., p, \]
the proof of which in turn is similar to that of Lemma 4.1.

6. GENERALIZATION TO ELLIPTICAL CASE

In Section 4, we studied robustness of \( \hat{\psi}_n = T_n(\hat{F}_n) \) under the assumption that the underlying distribution of \( x \) is normal. However, essentially the same argument is possible for general elliptically contoured distributions. In this section, we briefly mention this point and state the results.

Consider an elliptically contoured distribution \( E_p(\mu, \Sigma, f) \) with density
\[ \frac{1}{(\det \Sigma)^{1/2}} f \left( \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right). \]

When \( E_p(\mu, \Sigma, f) \) has finite second moments, we agree that \( f \) is taken to be normalized so that the covariance matrix is \( \Sigma \). For an appropriate choice of \( \rho \) (e.g., \( \rho_2 \)), the objective function (6) is always finite and we can think of \( \Sigma \) in \( E_p(\mu, \Sigma, f) \) as a scatter matrix when \( E_p(\mu, \Sigma, f) \) does not have a covariance matrix. Also, without loss of generality, we assume \( \mu = 0 \) as in the case of normality. Now, \( a = a(x) \) is distributed with density \( (\det A)^{-1/2} f(a^T A^{-1} a) \), and Proposition 3.1 remains to hold true.

Furthermore, we assume \( f(z) > 0 \) for all \( z \geq 0 \). Then, the Fisher consistency is still valid, and the influence function of \( T_p \) is given by
\[ \text{IF}\{ x; T_p, E_p(0, \Sigma, f) \} = \tilde{\psi}\left( \frac{1}{2} \| a_{(2)}(x) \|^2 \right) a_1(x) \sum_{j=2}^{p} l_j a_j(x) \gamma_j, \]
where \( \tilde{\psi} = \psi/E(\psi) \) and \( l_j = \lambda_j^*/[\lambda_j^* (\lambda_1 - \lambda_j)] \), \( j = 2, ..., p \), with \( \lambda_j^* = E(\tilde{\psi} a_j^2) \), \( j = 1, ..., p \). The proofs of these facts are given in the Appendix.

The GES becomes
\[ \text{GES}_{\tilde{\psi}} = \sup_{a_{(2)}} \left\{ \tilde{\psi}\left( \frac{1}{2} \| a_{(2)} \|^2 \right)^2 \sum_{j=2}^{p} l_j^2 a_j^2 \right\} \]
\[ \leq 2 \left( \sum_{j=2}^{p} l_j^2 \right) \sup_{z \geq 0} \{ z \tilde{\psi}(z)^2 \}. \]
7. RELATION TO M-ESTIMATION

Here we discuss the relation between our method and $M$-estimation. Consider the problem of $M$-estimation defined by the minimization of

$$L^*_n(\mu, \Sigma) = \frac{1}{n} \sum_{i=1}^{n} \rho\left(\tilde{z}(x_i - \mu)\right) + \frac{1}{2c} \log \det \Sigma, \quad c > 0,$$

with respect to $\mu \in \mathbb{R}^p$ and $\Sigma > 0$, where

$$\tilde{z}(x) = \frac{1}{2} x^T \Sigma^{-1} x.$$

In the case of maximum likelihood estimation for elliptically contoured distribution (17), the $\rho$ in (18) is associated with density generator $f$ as

$$f(\tilde{z}) = f(0) \exp\{-c\rho(\tilde{z})\}, \quad \tilde{z} \geq 0,$$

with $c = -f'(0)/f(0)$ under the integrability condition

$$\int_0^\infty e^{-c\rho(\tilde{z})} \tilde{z}^{(p/2) - 1} d\tilde{z} < \infty. \quad (19)$$

Now, let $\Sigma$ be decomposed as in (12), and suppose we decide to set unknown $\lambda_1, ..., \lambda_p$ as $\lambda_1 = \cdots = \lambda_k = \infty, \lambda_{k+1} = \cdots = \lambda_p = 1$. Then, this minimization problem reduces to that of

$$\frac{1}{n} \sum_{i=1}^{n} \rho\{z_k(x_i - \mu)\}$$

with respect to $\mu$ and $\gamma_1, ..., \gamma_p$, where $z_k(x)$ is the $\tilde{z}(x)$ with $\lambda_j$s set as above. But since

$$\sum_{j=k+1}^{p} \gamma_j y_j^T = I - \sum_{j=1}^{k} \gamma_j y_j^T,$$

we obtain

$$z_k(x) = \frac{1}{2} \left\{ |x|^2 - \sum_{j=1}^{k} (y_j^T x)^2 \right\},$$
so that the minimization of (20) with respect to \( \gamma, \ldots, \gamma_p \) is equivalent to that of (20) with respect to \( \mu \) and \( \gamma, \ldots, \gamma_p \) only. The crucial point in this argument is that \( \lambda_{k+1}, \ldots, \lambda_p \) are set to be equal so that \( \gamma_{k+1}, \ldots, \gamma_p \) are not determined individually and thus are eliminated from the minimization problem. Now, the M-estimator of \( \gamma \) defined in this way with \( k = 1 \) is just our \( \rho \)-estimator. In general, this procedure provides an approach to the \( k \)-dimensional principal subspace analysis (the subspace spanned by \( \gamma, \ldots, \gamma_k \) is estimated directly) in contrast with the \( k \)-principal component analysis mentioned in Section 1 (each of \( \gamma, \ldots, \gamma_k \) is estimated individually).

As should be clear from the discussion above, the robustness obtained by introducing \( \rho \) in the present paper can be considered the same thing as the robustness by M-estimation. The difference between our method and M-estimation lies in the difference between \( \tilde{z}(x) \) and \( z^k(x) \). This difference is the problem concerning the definition of an outlier—whether the definition is based on the Mahalanobis distance or it is based on the residual in the orthogonal projection onto a \( k \)-dimensional subspace. A similar argument is found in He and Simpson [10] in the context of robust direction estimation. In that paper, M-estimators for directions are considered, not the Mahalanobis distance. Note, however, that \( \rho_0 \) and \( \rho_2 \) do not satisfy the integrability condition (19) so that \( \rho_0 \)- and \( \rho_2 \)-estimators cannot be regarded as maximum likelihood estimators under any elliptically contoured distributions, even if we ignore the difference between \( \tilde{z}(x) \) and \( z^k(x) \).

We note here that condition (3) for robustness remains the same for M-estimation, since we have by Huber [12] and Sibson [16] that the influence function of the M-estimator \( \hat{\gamma}_M \) of \( \gamma_1 \) is

\[
\psi \left( \frac{1}{2} \sum_{j=1}^{p} \frac{a_j^2}{\lambda_j} \right) a_1 \sum_{j=2}^{p} \frac{a_j}{\lambda_1 - \lambda_j} \psi_j,
\]

where \( c_\psi = p(p + 2) \left[ (p + 2) E_0 1 \{ \psi(\| y \|^2/2) \| y \|^2 \} + E_0 1 \{ (\partial^2 \psi / \partial \| y \|^2) \| y \|^2 / 2 \} \| y \|^4 \} \right] \) and \( \psi = \psi \left[ \sum_{j=1}^{p} a_j(x)^2/(2 \lambda_j) \right] \). What is important in the above expression of the influence function is that \( a_1 \) is included in the argument of \( \psi \). This implies that \( \hat{\gamma}_M \) with \( \rho \) satisfying (3) has the additional robustness in the direction of \( \gamma_1 \) : \( \sup_{a_2, \ldots, a_p} \| IF \| < \infty \) for fixed \( a_2, \ldots, a_p \), which is not shared by the \( \rho \)-estimator. We can show, moreover, that when \( \rho \) satisfies the stronger condition

\[
\sup_{z \geq 0} \left\{ \frac{\partial \rho(z)}{\partial z} \right\} < \infty,
\]
the influence function of \( \hat{\beta}_M \) is bounded as a function of all \( a_1, a_2, ..., a_p \):

\[
\sup_{a_1, ..., a_p} |IF| < \infty.
\]

The use of \( z_1(x) = z(\gamma_1, x) \) in our procedure is motivated by the interpretation of \( \text{PCA} \) as the orthogonal projection onto a direction. Also, from a theoretical point of view, \( z_1(x) \) seems to be the better choice if one is concerned exclusively with \( \text{PCA} \), i.e., the estimation of \( \gamma_1 \) only, and not with the estimation of the entire \( \Sigma \). This can be seen from the following consideration: Observations with direction near \( \gamma_1 \) work favorably for the estimation of \( \gamma_1 \) ("good" data), and the converse is true for observations with direction far from \( \gamma_1 \) ("bad" data). Consider, especially, the extreme case where \( \hat{\lambda}_1 \) is close to \( \hat{\lambda}_2, ..., \hat{\lambda}_p \) so that the Mahalanobis distance is almost the same as the ordinary Euclidean distance. Then, the common \( M \)-estimators handle both kinds of data in an equal way. That is, if "bad" data are rejected, then so are "good" data; if "good" data are accepted, then so are "bad" data. On the other hand, our \( \rho \)-estimators tend to discard "bad" data while employing "good" data, as can be seen from the definition of \( z_1(x) \). A simulation study in Section 8 gives support to our assertion.

8. NUMERICAL STUDY

We explore the robustness performance of the \( \rho_0 \)-estimator \( \hat{\gamma}_* = T_{\rho_0}(\hat{F}_n) \) by applying the algorithm suggested in Section 1. In this study, we consider the following distributional contamination model: \( (1-\epsilon) N_p(\mu, \Sigma) + \epsilon H \), where \( H \) is of Cauchy type with the same location-scatter structure as \( N_p(\mu, \Sigma) \). Precisely, \( x \sim H \) is generated by \( x = \mu + \Sigma^{1/2} y, \ y = (y_1, ..., y_p)^T \), with \( y_1, ..., y_p \sim 1/[\pi(1+y_i^2)] \), i.i.d.. For ease of computation, we ignore the problem of location and use \( x \) instead of \( \mu \).

Specifically, we set \( p = 7, \ \mu = 0, \ \Sigma = \text{diag}(5, 3, 2, 1, 0 \cdot 5, 0 \cdot 5, 0 \cdot 5)^2 \), and take 95 observations from \( N_7(0, \Sigma) \) and 5 from \( H \). If we analyze only the first 95 observations by the classical \( \text{PCA} \), the principal component vector, being calculated as \((-0 \cdot 9944, -0 \cdot 0927, -0 \cdot 0435, -0 \cdot 0148, 0 \cdot 0181, 0 \cdot 0028, -0 \cdot 0086 \), is seen to be almost equal to the first axis. However, when the 5 outliers are added to the data, the analysis breaks down with the principal component vector \( \hat{\gamma} = (0 \cdot 0481, -0 \cdot 0490, -0 \cdot 9976, -0 \cdot 0002, 0 \cdot 0016, -0 \cdot 0036, 0 \cdot 0087) \) considerably deviated from the original one, the inner product value being \(-0 \cdot 0091 \). The result is understandable, because the outliers are quite remote from the cloud of the typical observations. See Fig. 1.

Now we implement the algorithm to search for the \( \rho_0 \)-estimator \( \hat{\gamma}_* \) with tuning parameters \( \beta = 2 \) and \( \eta = 15 \). Then we observe a quick convergence of the algorithm, within several iterations, starting with the initial vector.
FIG. 1. Three-dimensional plot of 5 outliers and 95 observations in the subspace of minor components, where the 95 observations form a cloud around the origin.

FIG. 2. Plot of $\psi(\beta, \eta)$ for the 95 observations and the 5 outliers.
Moreover, we see that the $\tilde{\gamma}_*$ obtained is quite robust against atypical observations, for $\tilde{\gamma}_* = (0.9645, -0.2612, 0.0281, -0.0161, -0.0147, -0.0111, 0.0150)$ with the inner product with the original principal component vector being $-0.9847$. Fig. 2 shows that the resulting weight function properly assigns zeros to the outliers. We have plotted the residuals $z(\gamma, x_i - \bar{x})$, $i = 1, \ldots, 100$, for $\gamma = \tilde{\gamma}_*$ and $\hat{\gamma}$ in Fig. 3. Our $\tilde{\gamma}_*$ makes a neat job of properly yielding small residuals to well-organized observations and large residuals to outliers. In contrast, the classical $\hat{\gamma}$ is seriously influenced by the outliers and cannot do the same job. We have obtained similar results in a number of experiments. In particular, the good performance of our proposed estimator was confirmed for all $(\beta, \eta)$ we tried. See Fig. 4.

![FIG. 3. Plot of $z(\gamma, x_i - \bar{x})$s for $i = 1, \ldots, 100$. (a) The case of the $p_o$-estimator: $\gamma = \tilde{\gamma}_*$. (b) The case of the classical estimator: $\gamma = \hat{\gamma}.$](image-url)
In addition, we have studied an $M$-estimator in the above situation. We considered the $M$-estimator with the same $p_0$, and obtained it numerically by a reweighting algorithm defined similarly to our algorithm for $p_0$-estimator. Then, our observations are the following: First, the algorithm did not work very well, and we obtained its convergence only when we used good starting values—in contrast to the case of $p_0$-estimator. Thus, as a starting value, we had to employ the ordinary sample covariance matrix calculated without the outliers. Second, even when the algorithm did converge, the obtained estimate depended heavily on the choice of $(\beta, \eta)$. It is true that we observed nice behavior of the $M$-estimator under some values of $(\beta, \eta)$, but for other values, it gave quite poor results. In fact, the $M$-estimator with $(\beta, \eta)$ in some region showed as bad a performance (the inner product $-0.0091$) as that of the classical PCA.

**APPENDIX**

**Proof of Steady Reduction for the Algorithm in Section 1**

Here, we prove that

$$L_n(t+1) \leq L_n(t)$$
and that this inequality is strict as long as (i) $\gamma_{t+1} \neq \gamma_t$ or (ii) $\mu_{t+1} - \mu_t$ is not in the direction of $\gamma_{t+1}$.

Since $\rho$ is concave, we have

$$\rho\{z(\gamma_{t+1}, x_i - \mu_{t+1})\} - \rho\{z(\gamma_t, x_i - \mu_t)\} \leq \psi\{z(\gamma_{t+1}, x_i - \mu_{t+1})\} [z(\gamma_{t+1}, x_i - \mu_{t+1}) - z(\gamma_t, x_i - \mu_t)].$$

Thus, writing $\psi_{n} = \psi\{z(\gamma_{t+1}, x_i - \mu_{t+1})\}$, we obtain

$$L_n(t + 1) - L_n(t) \leq \frac{1}{n} \sum_{i=1}^{n} \psi_{n}\{z(\gamma_{t+1}, x_i - \mu_{t+1}) - z(\gamma_t, x_i - \mu_t)\}$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \psi_{n}(\|x_i - \mu_{t+1}\|^2 - \|x_i - \mu_t\|^2)$$

$$+ \frac{1}{2n} \sum_{i=1}^{n} \psi_{n}[\{\gamma_{t+1}^T(x_i - \mu_{t+1})\}^2 - \{\gamma_{t+1}^T(x_i - \mu_{t+1})\}^2].$$

Now, using $\mu_{t+1} = \sum_{i} \psi_{n} x_i / \sum_{i} \psi_{n}$, we can calculate the first term on the right-hand side of the above equation as

$$- \sum_{i} \psi_{n} \|\mu_{t+1} - \mu_t\|^2.$$

On the other hand, the second term can be computed as follows: Recalling the definition of $S_n(t)$, we can write it as

$$\frac{1}{2n} \sum_{i=1}^{n} \psi_{n}(x_i - \mu_{t+1})(x_i - \mu_{t+1})^T$$

$$= S_n(t) + (\mu_{t+1} - \mu_t)(\mu_{t+1} - \mu_t)^T$$

$$+ \left( \frac{1}{\sum_{i} \psi_{n}} \sum_{i} \psi_{n}(x_i - \mu_{t+1}) \right)(\mu_{t+1} - \mu_t)^T$$

$$+ (\mu_{t+1} - \mu_t) \left( \frac{1}{\sum_{i} \psi_{n}} \sum_{i} \psi_{n}(x_i - \mu_{t+1}) \right)^T$$

$$= S_n(t) - (\mu_{t+1} - \mu_t)(\mu_{t+1} - \mu_t)^T.$$
so (21) can be written as
\[ \frac{\psi_n}{2n} \left[ \gamma_i^T S_d(t) \gamma_i - \gamma_{i+1}^T S_d(t) \gamma_{i+1} + \frac{1}{2} (\gamma_{i+1} - \gamma_i)^2 \right]. \]

Therefore, we obtain
\[
L_n(t+1) - L_n(t) \leq - \frac{\psi_n}{2n} \left[ \|\gamma_{i+1} - \gamma_i\|^2 - \frac{1}{2} (\gamma_{i+1} - \gamma_i)^2 \right]
+ \frac{\psi_n}{2n} \left[ \gamma_i^T S_d(t) \gamma_i - \gamma_{i+1}^T S_d(t) \gamma_{i+1} \right] \leq 0.
\]

It is clear that the equality holds only when \( \gamma_{i+1} = \pm \gamma_i, \gamma_{i+1} = \gamma_i + \gamma_{i+1}, s \in R. \)

**Proof of Fisher Consistency**

We give a proof of the Fisher consistency of \( T_p \) for the general elliptical case. More precisely, we show \( T_p \{ E_p(0, \Sigma, f) \} = \gamma_1 \) for \( f > 0. \)

Since \( U_{f, f} \{ E_p(0, \Sigma, f) \} = \gamma_1 \), it suffices to prove that \( L_{E_p(0, \Sigma, f)}(\gamma, 0) \) is minimized by \( \gamma = \gamma_1 \). Now, recall that for a general \( G, (\gamma, \mu) = \{ T_p(G), U_{f, f}(G) \} \) must satisfy (7). From the discussion in Section 3, we can see that \( (\gamma, \mu) = (\gamma_j, 0) \) satisfies (7) with \( G = E_p(0, \Sigma, f) \). By the same reasoning, each \( (\gamma, \mu) = (\gamma_j, 0), j = 2, \ldots, p, \) also satisfies (7) with \( G = E_p(0, \Sigma, f) \). Furthermore, we can show that \( L_{E_p(0, \Sigma, f)}(\gamma_k, 0) < L_{E_p(0, \Sigma, f)}(\gamma_l, 0) \) for \( 1 \leq k \leq l \leq p \). On the other hand, as is shown below, \( (\gamma, 0), \gamma \in S^p-1, \) does not satisfy (7) with \( G = E_p(0, \Sigma, f) \) unless \( \gamma \in \{ \gamma_1, \ldots, \gamma_p \}. \)

Thus, \( L_{E_p(0, \Sigma, f)}(\gamma_1, 0) < L_{E_p(0, \Sigma, f)}(\gamma, 0) \) for all \( \gamma \in S^p-1, \gamma \neq \gamma_1 \).

So the proof will be finished if we verify that any \( h \in S^p-1 \) for which \( (\gamma, \mu) = (h, 0) \) satisfies relation (7) with \( G = E_p(0, \Sigma, f) \) is necessarily an eigenvector of \( \Sigma \).

Take \( H_2: p \times (p - 1) \) so that \( H = (h, H_2) \in C(p) \), and put
\[
K = H^T \Sigma^{-1} H
= \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}
= (k_{ij})
\]
and
\[ b = H^T x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \]
with \( K_{22} : (p-1) \times (p-1) \) and \( b_2 : (p-1) \times 1 \). Then, equation (7) with \( G = E_p(0, \Sigma, f) \) implies
\[ E[\psi(\frac{1}{2} \| b_2 \|^2) b_1 b_2] = 0. \] (22)
Here, \( b \) is distributed as \( E_p(0, K^{-1}, f) \), or with density
\[ (\det K)^{1/2} f(\frac{1}{2} b^T K b). \]

Using
\[ b^T K b = k_{11} \left( b_1 + \frac{k_{21}^T b_2}{k_{11}} \right)^2 + b_2^T \left( K_{22} - \frac{k_{21} k_{21}^T}{k_{11}} \right) b_2, \]
we can write (22) as
\[ 0 = \int_0^{\infty} \int \psi \left( \frac{1}{2} \| b_2 \|^2 \right) \left( c - \frac{k_{21}^T b_2}{k_{11}} \right) b_2 \\
\times f \left[ \frac{1}{2} \left( k_{11} c^2 + b_2^T \left( K_{22} - \frac{k_{21} k_{21}^T}{k_{11}} \right) b_2 \right) \right] dc \, db_2 \\
= - \left( \frac{2}{k_{11}} \right)^{3/2} \int \psi \left( \frac{1}{2} \| b_2 \|^2 \right) b_2 b_2^T \\
\times \tilde{f}(\nu) \left( K_{22} - \frac{k_{21} k_{21}^T}{k_{11}} \right) b_2 \, db_2 \, k_{21}, \]
where \( \tilde{f}(\nu) = \int_0^{\infty} f(c^2 + \frac{1}{2} \nu) \, dc \). Since
\[ \int \psi \left( \frac{1}{2} \| b_2 \|^2 \right) b_2 b_2^T \tilde{f}(\nu) \left( K_{22} - \frac{k_{21} k_{21}^T}{k_{11}} \right) b_2 \, db_2 \]
is positive definite, this yields \( k_{21} = 0 \).
Therefore, \( K \) is block-diagonal, and \( h \) is an eigenvector of \( \Sigma \).

Derivation of the Influence Function of \( T_p \)

We derive the expression of the influence function of \( T_p \), given in Section 6, for general elliptically contoured distributions \( E_p(0, \Sigma, f) \), \( f > 0 \). Moreover, we prove Theorem 4.2 for the normal case.
Let

\[ \gamma_*(e) = T_p\{ (1 - e) E_p(0, \Sigma, f) + e \delta_e \}, \]

\[ \mu_*(e) = U_p\{ (1 - e) E_p(0, \Sigma, f) + e \delta_e \}, \quad y \in \mathbb{R}^p. \]

We want to calculate

\[ \hat{\gamma}_*(0) = \frac{d}{de} \gamma_*(e) \bigg|_{e=0}, \]

which is nothing but IF \{ y; T_p, E_p(0, \Sigma, f) \}.

Since \( (\gamma_*, \mu_*) = (\gamma_*(e), \mu_*(e)) \) satisfies (7) with \( G = (1 - e) E_p(0, \Sigma, f) + e \delta_y \), we have

\[ (\Sigma_ - \lambda_1^* I) \hat{\gamma}_*(0) + \Sigma_*(0) \gamma_1 - \lambda_1^*(0) \gamma_1 = 0, \tag{23} \]

where

\[ \Sigma_*(0) = \Sigma_{\gamma_*(e), \mu_*(e)}\{ (1 - e) E_p(0, \Sigma, f) + e \delta_y \}, \tag{24} \]

\[ \lambda_1^*(0) = \gamma_*(e)^T \Sigma_*(0) \gamma_*(e), \tag{25} \]

and \( \Sigma_*(0) = (d/de) \Sigma_*(e) \). Here, we have used \( \gamma_*(0) = \gamma_1, \mu_*(0) = 0, \Sigma_*(0) = \Sigma_*, \) and \( \lambda_1^*(0) = \lambda_1^* \).

The second term on the left-hand side of (23) can be calculated as follows. By differentiating both sides of (24), we have

\[
E[\psi(z(\gamma_1, x))] \Sigma_*(0)
+
\left[-E[\psi(z(\gamma_1, x))] + \psi(z(\gamma_1, y))\right]
+
\left[\frac{d}{de} E[\psi(z(\gamma_*(e), x - \mu_*(e)))] \bigg|_{e=0}\right] \Sigma_*
= -E[\psi(z(\gamma_1, x))] \Sigma_* + \psi(z(\gamma_1, y)) y y^T
+
\left[\frac{d}{de} E[\psi(z(\gamma_*(e), x - \mu_*(e)))] \{x - \mu_*(e)\} x - \mu_*(e)\}^T \right]_{e=0}.
\tag{26}
\]

Now, since

\[
\left. \frac{d}{de} \psi(z(\gamma_*(e), x)) \right|_{e=0} = \frac{\partial \psi}{\partial z}(z(\gamma_1, x))(\gamma^T x) x^T \gamma_*(0)
= -\frac{\partial \psi}{\partial z}(x) x^T \gamma_*(0),
\]

\[
= -\frac{\partial \psi}{\partial z}(x) x^T \gamma_*(0),
\]

The second term on the left-hand side of (23) can be calculated as follows. By differentiating both sides of (24), we have

\[ E[\psi(z(\gamma_1, x))] \Sigma_*(0)
+
\left[-E[\psi(z(\gamma_1, x))] + \psi(z(\gamma_1, y))\right]
+
\left[\frac{d}{de} E[\psi(z(\gamma_*(e), x - \mu_*(e)))] \bigg|_{e=0}\right] \Sigma_*
= -E[\psi(z(\gamma_1, x))] \Sigma_* + \psi(z(\gamma_1, y)) y y^T
+
\left[\frac{d}{de} E[\psi(z(\gamma_*(e), x - \mu_*(e)))] \{x - \mu_*(e)\} x - \mu_*(e)\}^T \right]_{e=0}.
\tag{26}
\]
we obtain
\[
E \left[ \frac{d}{d \epsilon} \psi[z(\gamma_\epsilon(e), x)] \right]_{\epsilon=0} xx^T \gamma_1
\]
\[
= -E \left[ \frac{\partial \psi}{\partial z} a_1(x)^2 xx^T \right] \hat{\gamma}_\epsilon(0)
\]
\[
= -E(\psi) \Gamma A_\epsilon F^T \hat{\gamma_\epsilon}(0),
\]
where
\[
A_\epsilon = E \left[ \frac{\partial \tilde{\psi}}{\partial z} \left\{ \frac{1}{2} \|a_2(\epsilon)\|^2 \right\} a_1^2 a^T \right] = \text{diag} \left( E \left[ \frac{\partial \tilde{\psi}}{\partial z} \left\{ \frac{1}{2} \|a_2(\epsilon)\|^2 \right\} a_1^4 \right], E \left[ \frac{\partial \tilde{\psi}}{\partial z} \left\{ \frac{1}{2} \|a_2(\epsilon)\|^2 \right\} a_1^2 a_2^2 \right], \ldots, E \left[ \frac{\partial \tilde{\psi}}{\partial z} \left\{ \frac{1}{2} \|a_2(\epsilon)\|^2 \right\} a_2^4 \right] \right).
\]
Next, we can show
\[
\frac{d}{d \epsilon} E\left[ \psi[z(\gamma_\epsilon(e)), x-\mu_\epsilon(e)) \{x-\mu_\epsilon(e)\}^T \right]_{\epsilon=0} \gamma_1 = 0,
\]
and noting \[\gamma_1^T \hat{\gamma}_\epsilon(0) = 0\], we can also verify
\[
\frac{d}{d \epsilon} E\left[ \psi[z(\gamma_\epsilon(e)), x-\mu_\epsilon(e)) \right]_{\epsilon=0} = 0.
\]
Hence, we see from (26) that \[\hat{\Sigma}_\epsilon(0) \gamma_1\] can be written as
\[
\hat{\Sigma}_\epsilon(0) \gamma_1 = -\lambda_1 \tilde{\psi} \gamma_1 + \tilde{\psi} a_1(y) y - \Gamma A_\epsilon F^T \hat{\gamma}_\epsilon(0).
\]
(27)
On the other hand, concerning the third term on the left-hand side of (23), we have by differentiating (25) that
\[
\hat{\gamma}_\epsilon(0) = 2\gamma_1^T \Sigma_\epsilon \hat{\gamma}_\epsilon(0) + \gamma_1 \hat{\Sigma}_\epsilon(0) \gamma_1.
\]
Here, we have
\[
2\gamma_1^T \Sigma_\epsilon \hat{\gamma}_\epsilon(0) = 2\lambda_1 \gamma_1^T \hat{\gamma}_\epsilon(0) = 0
\]
and
\[
\gamma_1^T \hat{\Sigma}_\epsilon(0) \gamma_1 = -\lambda_1 \tilde{\psi} + \tilde{\psi} a_1(y)^2
\]
by (27), and thus obtain
\[ \hat{\lambda}_q^*(0) = -\hat{\lambda}_q^* \hat{\psi} + \hat{\psi} a_1(y)^2. \] (28)

Using (27) and (28), we can write (23) as
\[
0 = (\Sigma* - \hat{\lambda}_q^* I) \hat{\gamma}_q(0) \\
+ [-\hat{\lambda}_q^* \hat{\psi} \gamma_1 + \hat{\psi} a_1(y) y - \Gamma A* \Gamma^T \hat{\gamma}_q(0)] \\
- [-\hat{\lambda}_q^* \hat{\psi} \gamma_1 + \hat{\psi} a_1(y)^2 \gamma_1] \\
= \Gamma (A* - \hat{\lambda}_q^* I - A*) \Gamma^T \hat{\gamma}_q(0) \\
- \hat{\psi} a_1(y) \{ a_1(y) \gamma_1 - y \}. \] (29)

Now, since
\[
A* - \hat{\lambda}_q^* I - A* = \begin{pmatrix}
- E \left[ \frac{\partial \hat{\psi}}{\partial z} \left( \frac{1}{2} \|a_{(2)}(x)\|^2 \right) a_1(x)^2 \right] & 0^T \\
0 & -L^{-1}
\end{pmatrix},
\]
where
\[
L = \text{diag}(l_2, ..., l_p),
\]
\[
l_j = \frac{1}{\hat{\lambda}_q^* - \hat{\lambda}_q^* - E \left[ \frac{\partial \hat{\psi}}{\partial z} \left( \frac{1}{2} \|a_{(2)}(x)\|^2 \right) a_1(x)^2 a_j(x)^2 \right]}, \quad j = 2, ..., p,
\]
equation (29) yields
\[
\hat{\gamma}_q(0) = \hat{\psi} \left( \frac{1}{2} \|a_{(2)}(y)\|^2 \right) a_1(y) \Gamma (A* - \hat{\lambda}_q^* I - A*)^{-1} \Gamma^T \{ a_1(y) \gamma_1 - y \} \\
= -\hat{\psi} \left( \frac{1}{2} \|a_{(2)}(y)\|^2 \right) a_1(y) \gamma_1 \Gamma_2 \\
\times \left( \begin{array}{c}
1 \\
E \left[ \frac{\partial \hat{\psi}}{\partial z} \left( \frac{1}{2} \|a_{(2)}(y)\|^2 \right) a_1(y)^2 \right] \\
0 \end{array} \right) \begin{pmatrix} 0 \\ \gamma_2 \end{pmatrix} \\
= \hat{\psi} \left( \frac{1}{2} \|a_{(2)}(y)\|^2 \right) a_1(y) \Gamma_2 L a_{(2)}(y) \\
= \hat{\psi} \left( \frac{1}{2} \|a_{(2)}(y)\|^2 \right) a_1(y) \sum_{j=2}^{p} l_j a_j(y) \gamma_j,
\]
and we have obtained the desired expression.
In the particular case where \( f(z) = (2\pi)^{p/2} e^{-z^2/2} \), or \( E_p(0, \Sigma, f) = N_p(0, \Sigma) \), we can write \( \lambda_j, j = 2, \ldots, p \), as

\[
\lambda_j = \frac{\hat{\lambda}_j}{\sum_{i=1}^{*}(\hat{\lambda}_i - \lambda_i)}
\]

by Lemma 4.1, and this leads to Theorem 4.2.

ACKNOWLEDGMENTS

We thank Professor Shizuhiko Nishisato for his kind advice and useful suggestions. We also thank the referees for their valuable comments which led to much improvement on the earlier version of the manuscript.

REFERENCES

