Robust High-Dimensional Modeling with the Contaminated Gaussian Distribution

Antonio Punzo          Paul D. McNicholas

Abstract

The contaminated Gaussian distribution represents a simple robust elliptical generalization of the Gaussian distribution; differently from the often-considered $t$-distribution, it also allows for automatic detection of outliers, spurious points, or noise (collectively referred to as bad points herein). Starting from this distribution, we propose the contaminated Gaussian factor analysis model as a method for robust data reduction and detection of bad points in high-dimensions. A mixture of contaminated Gaussian factor analyzers model follows therefrom, and extends the recently proposed mixtures of contaminated Gaussian distributions to high-dimensional data, i.e., where $p$ (number of dimensions) is large relative to $n$ (sample size). The number of free parameters is controlled through the dimension of the latent factor space. For each discussed model, we outline a variant of the classical expectation-maximization algorithm for parameter estimation. Various implementation issues are discussed, and we use real data for illustration.

Keywords: Contaminated Gaussian distribution; EM algorithm; factor analysis models; mixture models; model-based clustering.

1 Introduction

For $p$-variate data assumed to arise from a continuous random variable, statistical inference is commonly focused on elliptical distributions (Cambanis et al., 1981); in this class, the Gaussian distribution is the most widely considered because of its computational and theoretical convenience. However, for many applied problems, the tails of the Gaussian distribution are lighter than required. The $t$ distribution, thanks to the degrees of freedom, provides a common way to broaden the Gaussian tails (Lange et al., 1989 and Kotz and Nadarajah, 2004). A further elliptical alternative is represented by the contaminated Gaussian distribution, a two-component Gaussian mixture in which one of the components, with a large prior probability, represents the “good” observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the “bad” observations (Aitkin and Wilson, 1980). It constitutes a common and simple theoretical model for the occurrence of outliers, spurious points, or noise (collectively referred to as “bad” points herein).

The contaminated Gaussian distribution is the core of this paper. Firstly, for maximum likelihood (ML) estimation of its parameters, we propose an expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993) which stems from the characterization of the contaminated Gaussian distribution as a Gaussian scale mixture model. In contrast with the ECM algorithm illustrated in Punzo and McNicholas (2014), in each CM step of the new algorithm we have the advantage that the mean and the covariance matrix (for the good observations) are updated independently from the other two parameters (proportion of good observations and
inflation parameter). Secondly, we introduce the contaminated Gaussian factor analysis model, a robust extension of the classical (Gaussian) factor analysis model obtained by adopting the contaminated Gaussian distribution for the errors and the latent factors. Our proposal advantageously embeds the (Gaussian) factor analysis model in a larger model with two additional parameters that: (1) afford protection against elliptical non-Gaussianity (of errors and/or latent factors) and (2) allow for automatic detection of bad points. Thirdly, Punzo and McNicholas (2014) have recently proposed mixtures of contaminated Gaussian distributions both as a robust generalization of mixtures of Gaussian distributions, and as an improvement of mixtures of $t$ distributions in terms of automatic detection of bad points in a clustering perspective. However, the mixture of contaminated Gaussian distributions, with unrestricted component-covariance matrices of the good observations, say $\Sigma_g$, is a highly parametrized model with $p(p+1)/2$ parameters for each $\Sigma_g, g = 1, \ldots, G$. To introduce parsimony, Punzo and McNicholas (2014) also define thirteen variants of the general model obtained, as in Celeux and Govaert (1995), via eigen-decomposition of $\Sigma_1, \ldots, \Sigma_G$. But if $p$ is large relative to the sample size $n$, it may not be possible to use this decomposition to infer an appropriate model for $\Sigma_1, \ldots, \Sigma_G$. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of $\Sigma_g$ when $p$ is large relative to $n$. To address this problem, following the literature on the adoption of factor analyzers within mixture models (see, among many others, McLachlan and Peel, 2000, Chapter 8, McLachlan et al., 2003, McNicholas and Murphy, 2008, Zhao and Yu, 2008, Montanari and Viroll, 2011, and Subedi et al., 2013), we propose mixtures of contaminated Gaussian factor analyzers, where a contaminated Gaussian factor analysis model is used for each mixture component. The result is a means of fitting mixtures of contaminated Gaussian distributions in situations where $p$ would be sufficiently large relative to the sample size $n$ to cause potential problems with singular or near-singular estimates of $\Sigma_1, \ldots, \Sigma_G$. The number of free parameters is controlled through the dimension of the latent factor space.

The paper is organized as follows. Section 2 contextualizes the contaminated Gaussian distribution as a Gaussian scale mixture model. The contaminated Gaussian factor analysis model is introduced in Section 3 while mixtures of contaminated Gaussian factor analyzers are presented in Section 4. In particular, in each section, hence for each model: (1) identifiability is discussed, (2) an EM-based algorithm is outlined for ML estimation of the parameters, (3) computational details are given, and (4) a real data analysis is discussed to appreciate the advantages of the model. Computation is done in the R software environment for statistical computing and graphics (R Core Team, 2013). The paper concludes with a discussion in Section 5.

## 2 The Contaminated Gaussian Distribution

For robustness sake, one of the most common ways to generalize the Gaussian distribution is represented by the Gaussian scale mixture model

$$
\int_0^\infty \phi (x; \mu, \Sigma/w) \, dH(w),
$$

where $H(\cdot)$ is a probability distribution function and $\phi (x; \mu, \Sigma/w)$ denotes the density of a $p$-variate Gaussian random vector $X$ with mean $\mu$ and covariance matrix $\Sigma/w$. Model (1) is unimodal, symmetrical, and guarantees heavier tails than those of the Gaussian distribution (see, e.g., Watanabe and Yamaguchi, 2003). It also includes some well-known models; for example, if $W \sim \text{gamma}(\nu/2, \nu/2)$, then we obtain the $t$ distribution with location parameter $\mu$, positive definite inner product matrix $\Sigma$, and $\nu$ degrees of freedom.

For our aim, it is important to note that if we focus on the dichotomous random variable $W = \begin{cases} 1 & \text{with probability } \alpha \\ 1/\eta & \text{with probability } 1 - \alpha \end{cases}$,
with probability mass function
\[ p_C(w; \alpha, \eta) = \alpha^{\frac{w-1}{\eta}} (1 - \alpha)^{\frac{1-w}{\eta}}, \]
then, from model (1), we obtain the contaminated Gaussian distribution
\[ p_{CN}(x; \mu, \Sigma, \alpha, \eta) = \alpha \phi(x; \mu, \Sigma) + (1 - \alpha) \phi(x; \mu, \eta \Sigma), \tag{2} \]
where \( \alpha \in (0, 1) \) and \( \eta > 1 \) (cf. Punzo and McNicholas, 2014). In summary,
\[ X \mid w \sim N_p(\mu, \Sigma/w), \tag{3} \]
\[ W \sim C(\alpha, \eta), \tag{4} \]
and
\[ X \sim CN_p(\mu, \Sigma, \alpha, \eta). \tag{5} \]
As we can see in (2), a contaminated Gaussian distribution is a two-component Gaussian mixture in which one of the components, typically with a large prior probability \( \alpha \), represents the “good” observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix \( \eta \Sigma \), represents the “bad” observations (Aitkin and Wilson, 1980).

An advantage of (2) with respect to the often used \( t \) distribution is that, once the parameters in \( \vartheta = \{ \mu, \Sigma, \alpha, \eta \} \) are estimated, say \( \hat{\vartheta} = \{ \hat{\mu}, \hat{\Sigma}, \hat{\alpha}, \hat{\eta} \} \), we can establish whether a generic observation \( x_i \) is either good or bad via the a posteriori probability. That is, compute
\[ P \left( x_i \text{ is good} \mid \hat{\vartheta} \right) = \frac{\hat{\alpha} \phi(x_i; \hat{\mu}, \hat{\Sigma})}{p_{CN}(x_i; \hat{\vartheta})}, \tag{6} \]
and \( x_i \) will be considered good if \( P \left( x_i \text{ is good} \mid \hat{\vartheta} \right) > 1/2 \), while it will be considered bad otherwise.

### 2.1 Maximum likelihood estimation via the ECME algorithm

Given an observed random sample \( x_1, \ldots, x_n \) from \( X \sim CN_p(\mu, \Sigma, \alpha, \eta) \), we consider now the application of the expectation-conditional maximization either (ECME) algorithm of Liu and Rubin (1994) for maximum likelihood (ML) estimation of \( \vartheta = \{ \mu, \Sigma, \alpha, \eta \} \); here, based on (3), (4), and (5), \( \vartheta \) is partitioned as \( \vartheta = \{ \vartheta_1, \vartheta_2 \} \), where \( \vartheta_1 = \{ \mu, \Sigma \} \) and \( \vartheta_2 = (\alpha, \eta)' \).

The ECME algorithm is an extension of the ECM algorithm of Meng and Rubin (1993). With this extension, some or all of the CM-steps of the ECM algorithm are replaced by steps that conditionally directly maximize the observed-data log-likelihood function, and not the expectation of the complete-data log-likelihood. Anyway, both these algorithms are variants of the classical expectation-maximization (EM) algorithm (Dempster et al., 1977), which is a natural approach for ML estimation when data are incomplete. In our case, incompleteness arises from the characterization of the contaminated Gaussian distribution given by (3), (4), and (5). The complete data are taken to be \( (x_1', \ldots, x_n', w_1, \ldots, w_n)' \), and the complete-data likelihood can be written
\[ L_c(\vartheta) = \prod_{i=1}^{n} \phi(x_i; \mu, \Sigma/w_i) p_C(w_i; \alpha, \eta). \]

Accordingly, the complete-data log-likelihood can be written
\[ l_c(\vartheta) = l_{c1}(\vartheta_1) + l_{c2}(\vartheta_2), \]
where

\[ l_{1c}(\vartheta_1) = -\frac{n p}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| + \frac{p}{2} \sum_{i=1}^{n} \log w_i - \frac{1}{2} \sum_{i=1}^{n} w_i (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \]

and

\[ l_{2c}(\vartheta_2) = \log \alpha \sum_{i=1}^{n} \frac{w_i - 1/\eta}{1 - 1/\eta} + \log (1 - \alpha) \sum_{i=1}^{n} \frac{1 - w_i}{1 - 1/\eta}. \]

2.1.1 E-step

The E-step on the \((k+1)\)th iteration requires the calculation of

\[ Q(\vartheta; \vartheta^{(k)}) = E_{\vartheta^{(k)}}[ l_{1c}(\vartheta) | x_1, \ldots, x_n] \]

\[ = E_{\vartheta^{(k)}}[ l_{1c}(\vartheta_1) | x_1, \ldots, x_n] + E_{\vartheta^{(k)}}[ l_{2c}(\vartheta_2) | x_1, \ldots, x_n] \]

where the expectation, as it can be noted by the subscript, is taken using the current fit \(\vartheta^{(k)}\) for \(\vartheta\). Here, we replace \(w_i\) by

\[ E_{\vartheta^{(k)}}[W_i|x_i] = w_i^{(k)} \]

where

\[ w_i^{(k)} = \frac{\alpha^{(k)} \phi \left(x_i; \mu^{(k)}, \Sigma^{(k)}\right) + \frac{1 - \alpha^{(k)}}{\eta^{(k)}} \phi \left(x_i; \mu^{(k)}, \eta^{(k)} \Sigma^{(k)}\right)}{\alpha^{(k)} \phi \left(x_i; \mu^{(k)}, \Sigma^{(k)}\right) + \left(1 - \alpha^{(k)}\right) \phi \left(x_i; \mu^{(k)}, \eta^{(k)} \Sigma^{(k)}\right)}. \]

2.1.2 CM-step 1

The first CM-step on the \((k+1)\)th iteration requires the calculation of \(\vartheta_1^{(k+1)}\) by maximizing \(Q_1(\vartheta_1; \vartheta^{(k)})\) with \(\vartheta_2\) fixed at \(\vartheta_2^{(k)}\). This yields

\[ \mu^{(k+1)} = \frac{\sum_{i=1}^{n} w_i^{(k)} x_i}{\sum_{i=1}^{n} w_i^{(k)}} \quad (7) \]

and

\[ \Sigma^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} w_i^{(k)} (x_i - \mu^{(k+1)}) (x_i - \mu^{(k+1)})' \]

2.1.3 CM-step 2

The updating of \(\vartheta_2\), with \(\vartheta_1\) fixed at \(\vartheta_1^{(k+1)}\), cannot be directly derived based on the \(Q_2\)-function because it is meaningless to estimate \(\eta\) when \(w_i^{(k)}\) is given. To solve this issue, the updating of \(\alpha\) and \(\eta\) is directly performed based on the observed-data log-likelihood; this choice leads to the ECME algorithm. In particular, the second CM-step, on the \((k+1)\)th iteration, requires the calculation of \(\vartheta_2^{(k+1)}\) by maximizing the observed-data log-likelihood

\[ l_{CN}(\vartheta_2 | \vartheta_1^{(k+1)}) = \sum_{i=1}^{n} \log \left[ p_{CN}(x_i; \mu^{(k+1)}, \Sigma^{(k+1)}, \alpha, \eta) \right], \]

under the constraints \(\eta > 1\) and \(\alpha \in (\alpha^*, 1)\); the latter constraint is justified by the fact that, for practical purposes, one could require that the proportion of good data is at least equal to a pre-determined value \(\alpha^*\), the most natural choice being \(\alpha^* = 1/2\); this choice is justified because, in robust statistics, it is usually assumed that at least half of the points are good (cf. Hennig, 2002, p. 250).
2.2 Computational considerations and details

The second CM-step is operationally performed using the `optim()` function in the `stats` package for R. The method of Nelder and Mead (1965) is considered for the numerical search of the maximum.

### 2.2.1 Initialization

The choice of the starting values for the ECM algorithm constitutes an important issue. The standard initialization consists of selecting a value for $\vartheta(0)$. In particular, a random initialization is usually repeated $t$ times, from different random positions, and the solution maximizing the observed-data log-likelihood $l(\vartheta)$ among these $t$ runs is selected (see Biernacki et al., 2003, Karlis and Xekalaki, 2003, and Bagnato and Punzo, 2013 for more complicated strategies).

Instead of selecting $\vartheta(0)$ randomly, we suggest the following technique. The Gaussian distribution can be seen as nested in the corresponding contaminated Gaussian distribution. In particular, the former can be obtained from the latter when $\alpha \to 1^{-}$ and $\eta = \eta^*$. Then, the closed-form ML estimates of $\mu$ and $\Sigma$ for the Gaussian distribution, along with the constraints $\alpha = \alpha^*$ (with $\alpha^* \to 1^{-}$) and $\eta = \eta^*$ (with $\eta^* \to 1^{+}$), can be used as $\vartheta(0)$; in the analysis of Section 2.3, we fix $\alpha^* = 0.999$ and $\eta^* = 1.001$. From an operational point of view, thanks to the monotonicity property of the ECM algorithm (see, e.g., McLachlan and Krishnan, 2007, p. 28), this also guarantees that the observed-data log-likelihood of the contaminated Gaussian distribution will be always greater than or equal to the observed-data log-likelihood of the corresponding Gaussian distribution. This is a fundamental consideration for the use of likelihood-based model selection criteria, and likelihood ratio tests, for choosing/assessing between a Gaussian distribution and a contaminated Gaussian distribution.

### 2.2.2 Convergence Criterion

The Aitken acceleration (Aitken, 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the ECM algorithm. Based on this estimate, we can decide whether or not the algorithm has reached convergence; i.e., whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration $k + 1$ is given by

$$ a^{(k+1)} = \frac{l^{(k+2)} - l^{(k+1)}}{l^{(k+1)} - l^{(k)}} $$

where $l^{(k)}$ is the observed-data log-likelihood value from iteration $k$. Then, the asymptotic estimate of the log-likelihood at iteration $k + 2$ is given by

$$ l^{(r+2)} = l^{(r+1)} + \frac{1}{1 - a^{(r+1)}} \left( l^{(r+2)} - l^{(r+1)} \right), $$

cf. Böhning et al. (1994). The ECM algorithm can be considered to have converged when $l^{(k+2)} - l^{(k+1)} < \epsilon$. In the analysis of Section 2.3 we fix $\epsilon = 0.001$.

2.3 Real data analysis

The bivariate data considered here relate 298 daily returns of two financial indexes (DAX 30 and FTSE MIB) spanning the period from July 24th, 2012, to September 30th, 2013 (the share prices used to compute the daily returns are downloadable from http://finance.yahoo.com/). A scatter plot of the data is provided in Figure 1. Mardia’s test of multivariate symmetry, as implemented by the `mardia()` function of the `psych` package (Revelle, 2014), produces a $p$-value of 0.102, leading us to not reject the null at the commonly considered significance levels.
In the class of symmetric bivariate distributions we focus on the classical Gaussian distribution, that we recall to be nested in the contaminated Gaussian distribution. On data at hand, these distributions can be statistically compared via the likelihood-ratio (LR) statistic

$$LR = -2 \left[ l_N (\hat{\mu}, \hat{\Sigma}) - l_{CN} (\hat{\mu}, \hat{\Sigma}, \hat{\alpha}, \hat{\eta}) \right],$$

where the hat denotes the ML estimate of the underlying parameter, while $l_N (\cdot)$ and $l_{CN} (\cdot)$ are the log-likelihood functions for the Gaussian and the contaminated Gaussian distributions, respectively. Under the null of bivariate Gaussianity (versus the alternative of bivariate contaminated Gaussianity), $LR$ is asymptotically distributed as a $\chi^2$ with two degrees of freedom, corresponding to the difference in the number of free parameters of the models under the two hypotheses. The resulting $p$-value is $9.724 \times 10^{-6}$, which leads to the rejection of the null, in favor of the alternative, at any reasonable significance level.

Figure 2 shows the graphical results from the ML estimation of the contaminated Gaussian distribution; bad points are represented by black bullets and contour lines from the estimated distribution are superimposed. The ML estimates for $\alpha$ and $\eta$ are $\hat{\alpha} = 0.5$ and $\hat{\eta} = 3.602$, respectively; in some sense, such result indicates that we are far to infer that data at hand arise from a single Gaussian distribution.

## 3 The Contaminated Gaussian Factor Analysis Model

### 3.1 The Model

The (Gaussian) factor analysis model (Spearman, 1904) is a well-known, and widely used, data reduction tool aiming to find latent factors that explain the variability in the data. The model (see Bartholomew et al., 2011, Chapter 3) assumes that the $p$-variate random vector $X$ is modeled using a $q$-variate vector of factors $U \sim N_q (0_q, I_q)$ where $q \ll p$. The model is

$$X = \mu + \Lambda U + e, \quad \text{(8)}$$

where $\Lambda$ is a $p \times q$ matrix of factor loadings and $e \sim N_p (0_p, \Psi)$ is the error term, with $\Psi = \text{diag} (\psi_1, \ldots, \psi_p)$. It follows from (8) that $X \sim N_p (\mu, \Lambda \Lambda' + \Psi)$. 
The factor analysis model is, however, sensitive to bad points as it adopts the Gaussian distribution for errors and latent factors. To improve its robustness, for data having longer than Gaussian tails or bad points, McLachlan et al. (2007) introduce the \( t \)-factor analysis model which considers the multivariate \( t \) for the distributions of the errors and the latent factors (see also McLachlan and Krishnan, 2007, Section 5.14.4). Although the \( t \)-factor analysis model robustifies the classical factor analysis model, once applied to data at hand, it does not allow for automatic detection of bad points. To solve this problem, recalling (6), we introduce the contaminated Gaussian factor analysis model.

Based on (8), the contaminated Gaussian factor analysis model generalizes the corresponding Gaussian factor analysis model by assuming

\[
\begin{pmatrix}
X \\
U
\end{pmatrix} \sim CN_{p+q}(\mu^*, \Sigma^*, \alpha, \eta),
\]

(9)

where

\[
\mu^* = \begin{pmatrix}
\mu \\
o_q
\end{pmatrix}
\quad \text{and} \quad
\Sigma^* = \begin{pmatrix}
\Lambda\Lambda' + \Psi & \Lambda \\
\Lambda' & I_q
\end{pmatrix}.
\]

Using the characterization of the contaminated Gaussian distribution discussed in Section 2, the joint density of \( X \) and \( U \), given \( W = w \), can be written

\[
\begin{pmatrix}
X \\
U
\end{pmatrix} \mid w \sim CN_{p+q}(\mu^*, \Sigma^*/w),
\]

(10)

with \( W \sim C(\alpha, \eta) \). Thus,

\[
X \mid w \sim N_p(\mu, (\Lambda\Lambda' + \Psi)/w),
\]

\[
U \mid w \sim N_q(0_q, I_q/w),
\]

\[
e \mid w \sim N_p(0_p, \Psi/w),
\]

so that

\[
X \sim CN_p(\mu, \Lambda\Lambda' + \Psi, \alpha, \eta),
\]

\[
U \sim CN_q(0, I_q, \alpha, \eta),
\]

\[
e \sim CN_p(0_p, \Psi, \alpha, \eta).
\]
The factors and error terms are no longer independently distributed as in the normal-based model for factor analysis; however, they are uncorrelated. To see this, we have from [9] that conditional on \( w, U \) and \( e \) are uncorrelated, and hence, unconditionally uncorrelated.

### 3.2 Identifiability and number of free parameters

Literally speaking, the number of parameters of the contaminated Gaussian factor analysis model is \( p + pq + p + 2 \): we have \( p \) values in \( \mu \), \( pq \) values in \( \Lambda \), \( p \) values in \( \Psi \), one \( \alpha \), and one \( \eta \). However, for identifiability sake when \( q > 1 \), we have to impose \( q(q-1)/2 \) constraints for \( \Lambda \) to be uniquely defined (cf. McLachlan and Peel, 2000, p. 241 and Bartholomew et al., 2011, p. 64); in fact, there is an infinity of choices for \( \Lambda \) because the model is still satisfied if we replace \( U \) by \( HU \) and \( \Lambda \) by \( \Lambda H' \), where \( H \) is an orthogonal matrix of order \( q \). The number \( m \) of free parameters for the model is then

\[
p + \left[ pq - \frac{1}{2}q(q-1) \right] + p + 2. \tag{11}
\]

### 3.3 Maximum likelihood estimation via the AECM algorithm

To find ML estimates for the parameters \( \vartheta = \{ \mu, \Lambda, \Psi, \alpha, \eta \} \) of the contaminated Gaussian factor analysis model, we consider the application of the alternating expectation-conditional maximization (AECM) algorithm of Meng and van Dyk (1997). The AECM algorithm is an extension of the ECM algorithm, where the specification of the complete data is allowed to be different on each CM-step. To apply the AECM algorithm, we partition \( \vartheta \) as \( \vartheta = \{ \vartheta_1, \vartheta_2 \} \), where \( \vartheta_1 = \{ \mu, \alpha, \eta \} \) and \( \vartheta_2 = \{ \Lambda, \Psi \} \). For this application of the AECM algorithm, the \((k+1)\)th iteration consists of two cycles, and there is one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of \( \vartheta \) into \( \vartheta_1 \) and \( \vartheta_2 \).

#### 3.3.1 First cycle

The first cycle of the \((k+1)\)th iteration of the AECM algorithm is practically equivalent to the \((k+1)\)th iteration of the ECME algorithm for the contaminated Gaussian distribution (see Section 2.1). The only difference is that we do not update \( \Sigma \), that is \( \Lambda \Lambda' + \Psi \), but we only update \( \mu \) according to (7) and \( \alpha \) and \( \eta \) according to what described in Section 2.1.3. At the end of the first cycle, we set \( \vartheta^{(k+1/2)} = \{ \vartheta_1^{(k+1)}, \vartheta_2^{(k)} \} \).

#### 3.3.2 Second cycle

In the second cycle of the \((k+1)\)th iteration of the AECM algorithm, we update \( \vartheta_2 \) by specifying the missing data to be the factors \( u_i, \ldots, u_n \) and the weights \( w_1, \ldots, w_n \). From (9) we have that

\[
X_i|u_i, w_i \sim N_p(\mu + \Lambda u_i, \Psi/w_i). 
\]

Thus, the complete data are \((x'_1, \ldots, x'_n, u'_1, \ldots, u'_n, w_1, \ldots, w_n)'\), and the complete-data likelihood can be factored as

\[
L_{c2}(\vartheta_2) = \prod_{i=1}^{n} \phi(x_i; \mu^{(k+1)} + \Lambda u_i, \Psi/w_i) \phi(u_i; 0_q, I_q/w_i) p_C \left(w_i; \alpha^{(k+1)}, \eta^{(k+1)} \right). 
\]
The complete-data log-likelihood is
\[
l_{c2}(\theta_2) = -\frac{n}{2}(p + q) \log (2\pi) + \frac{1}{2} (p + q) \sum_{i=1}^{n} \log w_i - \frac{1}{2} \sum_{i=1}^{n} w_i u_i u_i' - \frac{n}{2} \log |\Psi|
\]
\[+ \log \alpha^{(k+1)} \sum_{i=1}^{n} w_i - 1/\eta^{(k+1)} + \log \left(1 - \alpha^{(k+1)}\right) \sum_{i=1}^{n} \frac{1 - w_i}{1 - 1/\eta^{(k+1)}} \]
\[-\frac{1}{2} \text{tr} \left[ \Psi^{-1} \sum_{i=1}^{n} w_i (x_i - \mu^{(k+1)}) (x_i - \mu^{(k+1)})' \right] \]
\[+ \sum_{i=1}^{n} w_i (x_i - \mu^{(k+1)})' \Psi^{-1} \Lambda u_i - \frac{1}{2} \text{tr} \left( \Lambda' \Psi^{-1} \Lambda \sum_{i=1}^{n} w_i u_i u_i' \right), \tag{12}
\]
where \( \text{tr}(\cdot) \) denotes the trace operator.

The E-step on the second cycle of the \((k+1)\)th iteration requires the calculation of
\[Q_2(\theta_2; \vartheta^{(k+1/2)}) = E_{\theta^{(k+1/2)}}[l_{c2}(\theta_2) | x_1, \ldots, x_n]. \]

In addition, we update \(w_i\) to
\[w_i^{(k+1/2)} = \frac{\alpha^{(k+1)} \phi \left( x_i; \mu^{(k+1)}, \Sigma^{(k)} \right) + 1 - \alpha^{(k+1)} \eta^{(k+1)}}{\alpha^{(k+1)} \phi \left( x_i; \mu^{(k+1)}, \eta^{(k+1)} \Sigma^{(k)} \right) + (1 - \alpha^{(k+1)}) \phi \left( x_i; \mu^{(k+1)}, \eta^{(k+1)} \Sigma^{(k)} \right)}, \]
where \( \Sigma^{(k)} = \Lambda^{(k)} \Lambda^{(k)}' + \Psi^{(k)} \), due to the last two rows of (12) we also need to calculate
\[E_{\theta^{(k+1/2)}}(W_iU_i|x_i, w_i) \text{ and } E_{\theta^{(k+1/2)}}(W_iU_iU_i'|x_i, w_i). \] From (10) we obtain that
\[U_i|x_i, w_i \sim N_q \left( \gamma' (x_i - \mu), (I_q - \gamma' \Lambda') / w_i \right), \tag{13}
\]
where \( \gamma = (\Lambda \Lambda' + \Psi)^{-1} \Lambda \). Hence, from (10) and (13) we have that
\[E_{\theta^{(k+1/2)}}(W_iU_i|x_i, w_i) = w_i^{(k+1/2)} (I_q - \gamma' \Lambda') \gamma^{(k)} (x_i - \mu^{(k+1)}) \]
and
\[E_{\theta^{(k+1/2)}}(W_iU_iU_i'|x_i, w_i) = I_q - \gamma^{(k)} \Lambda^{(k)} + w_i^{(k+1/2)} (I_q - \gamma^{(k)} \Lambda^{(k)}) (x_i - \mu^{(k+1)})' \gamma^{(k)}, \]
where \( \gamma^{(k)} = (\Lambda^{(k)} \Lambda^{(k)}' + \Psi^{(k)})^{-1} \Lambda^{(k)} \). Starting from (12), the expected complete-data log-likelihood of second cycle is
\[Q_2(\theta_2; \vartheta^{(k+1/2)}) = C - \frac{n}{2} \log |\Psi| - \frac{n}{2} \text{tr} \left( \Psi^{-1} S^{(k+1/2)} \right) \]
\[+ n \text{tr} \left( \Lambda' \Psi^{-1} \Lambda R^{(k+1/2)} \right) - \frac{n}{2} \text{tr} \left( \Lambda' \Psi^{-1} \Lambda R^{(k+1/2)} \right), \]
where
\[S^{(k+1/2)} = \frac{1}{n} \sum_{i=1}^{n} w_i^{(k+1/2)} (x_i - \mu^{(k+1)}) (x_i - \mu^{(k+1)})', \]
\[R^{(k+1/2)} = I_q - \gamma^{(k)} \Lambda^{(k)} + \gamma^{(k)} S^{(k+1/2)} \gamma^{(k)}, \]
and \(C\) includes the terms that do not depend on \(\vartheta_2\).
The CM-step on this second cycle of the \((k + 1)\)th iteration is implemented by the maximization of \(Q_2 (\vartheta_2; \vartheta^{(k + 1/2)})\) over \(\vartheta_2\) with \(\vartheta_1\) set equal to \(\vartheta^{(k+1)}_1\). After some algebra, this yields the updated estimates
\[
\Lambda^{(k+1)} = S^{(k+1/2)} \gamma^{(k)} \left( R^{(k+1/2)} \right)^{-1}
\]
and
\[
\Psi^{(k+1)} = \text{diag} \left( S^{(k+1/2)} - \Lambda^{(k+1)} \gamma^{(k)} S^{(k+1/2)} \right).
\]

3.4 Computational details

In the second cycle of the \((k + 1)\)th iteration of the AECM algorithm, we need to compute \(\gamma^{(k)}\) which, in turn, requires the inversion of the \(p \times p\) matrix \(\Lambda^{(k)} \Lambda^{(k)\prime} + \Psi^{(k)}\). This inversion can be slow for large values of \(p\). To ease it, we use the Woodbury identity (Woodbury, 1950)
\[
\left( \Lambda^{(k)} \Lambda^{(k)\prime} + \Psi^{(k)} \right)^{-1} = \left( \Psi^{(k)} \right)^{-1} - \left( \Psi^{(k)} \right)^{-1} \Lambda^{(k)} \left( I_q + \Lambda^{(k)\prime} \Psi^{(k)\prime} \Lambda^{(k)} \right)^{-1} \Lambda^{(k)\prime} \left( \Psi^{(k)} \right)^{-1},
\]
which requires the simpler inversions of the diagonal \(p \times p\) matrix \(\Psi^{(k)}\) and the \(q \times q\) matrix \(I_q + \Lambda^{(k)\prime} \Psi^{(k)\prime} \Lambda^{(k)}\). This leads to a significant speed-up when \(q \ll p\).

Based on the idea of Section 2.2.1, the AECM algorithm is initialized with the estimates of \(\mu\), \(\Lambda\) and \(\Psi\) provided by a Gaussian factor analysis model, along with the constraints \(\alpha = \alpha^*\) (with \(\alpha^* \to 1^-\)) and \(\eta = \eta^*\) (with \(\eta^* \to 1^+\)); in the analysis of Section 3.5, the (preliminary) Gaussian factor analysis model is estimated by the \texttt{pgmmEM()} function of the \texttt{pgmm} package for \texttt{R} (McNicholas et al., 2011). The \texttt{pgmmEM()} function implements an AECM algorithm to obtain ML estimates. Finally, the Aitken acceleration is considered as convergence criterion (see Section 2.2.2 for details).

3.5 Real data analysis

We illustrate the contaminated Gaussian factor analysis model on the \texttt{state.x77} data set available on the \texttt{datasets} package for \texttt{R}. This data set is a compilation of data about the \(n = 50\) US states put together from the 1977 \textit{Statistical Abstract of the United States} (available for free online at \url{http://www.census.gov/compendia/statab/}), with the actual measurements mostly made a few years before. The \(p = 8\) variables, included into the data set, are:

- Population: population estimate as of July 1, 1975;
- Income: per capita income in dollars (1974);
- Illiteracy: illiteracy (1970, percent of population);
- Life Exp: life expectancy in years (1969–71);
- Murder: murder and non-negligent manslaughter rate per 100,000 population (1976);
- HS Grad: percent high-school graduates (1970);
- Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city;
- Area: land area in square miles.

The scatterplot matrix of the data is displayed in Figure 3.

To reduce the dimensionality of the data, so to have a lower number \(q\) of latent factors that explain their variability, we compare the Gaussian, and the contaminated Gaussian, factor analysis models. With regards to the former, the \texttt{pgmmEM()} function of the \texttt{pgmm} package is run to obtain ML estimates. Two values for \(q\) are considered, \(q = 1\) and \(q = 2\), and the Bayesian information criterion (BIC; Schwarz, 1978)
\[
\text{BIC} = 2 \ell \left( \hat{\vartheta} \right) - m \log n,
\]
where \( m \) is the overall number of free parameters, is used to select the best one. For both the models, the best BIC values (−4538.511 for the Gaussian factor analysis model, and −4532.987 for the contaminated Gaussian factor analysis model) correspond to \( q = 1 \). As the models are nested, the BIC also suggests that the selected contaminated Gaussian factor analysis model is better. Nevertheless, we prefer to have a \( p \)-value quantifying this choice. Following the lines of Section 2.3, we perform an LR test to compare the best BIC models. The asymptotic \( \chi^2 \) distribution has, also in this case, two degrees of freedom and the \( p \)-value is 0.001; this result leads us to reject, at the usual significance levels, the null hypothesis that a Gaussian factor analysis model works well on these data, in favor of the alternative contaminated Gaussian factor analysis model. The advantage of this choice is that not only can we reduce dimensionality in the presence of bad points, but we can identify them. When we view the results of our analysis of the state.x77 data in this way, we see from Figure 4 that there is an anomalous point (black bullet). Note also that the ML estimates of \( \alpha \) and \( \eta \) are \( \hat{\alpha} = 0.955 \) and \( \hat{\eta} = 3.510 \), respectively.

### 4 Mixtures of Contaminated Gaussian Factor Analyzers

To robustify the classical mixture of Gaussian distributions to the occurrence of bad points, and also to allow for their automatic detection, Punzo and McNicholas (2014) propose the mixture of contaminated Gaussian distributions

\[
p(x; \vartheta) = \sum_{g=1}^{G} \pi_g p_{CN}(x; \mu_g, \Sigma_g, \alpha_g, \eta_g)
\]  

(15)
where, for the \( g \)th mixture component, \( \pi_g \) is its mixing proportion, with \( \pi_g > 0 \) and \( \sum_{g=1}^{G} \pi_g = 1 \), while \( \mu_g, \Sigma_g, \alpha_g \) and \( \eta_g \) are defined as in (2).

In (15), there are \( p (p + 1) / 2 \) parameters for each \( \Sigma_g, \ g = 1, \ldots, G \). This means that as the number of components \( G \) grows, the total number of parameters can quickly become very large relative to the sample size \( n \), leading to overfitting. To model high-dimensional data, and to add parsimony, we consider the contaminated Gaussian factor analysis model of Section 3 in each mixture component; this leads to the mixture of contaminated Gaussian factor analyzers given by (15) but with component covariance matrices given by (16).

\[
\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g.
\]  

### 4.1 Identifiability and number of free parameters

Intuitively, the identifiability of the family of mixtures of contaminated Gaussian factor analyzers requires the identifiability of the family of mixtures of contaminated Gaussian distributions, as well as the identifiability of the family of factor analysis models. Since the identifiability of the class of contaminated Gaussian distributions has been established (see Punzo and McNicholas, 2014), this leaves the question of the identifiability of the family of factor analysis models; in other words, it requires the identifiability of the class of covariance structures defined in (16). Unfortunately, the topic of identification may itself deserve a separate research project. In this paper, we will not attempt to establish general rules for the identification of the proposed mixture models. However, based on the considerations leading to (11), we can say that the overall

**Figure 4:** state.x77 data: scatterplot matrix with the detected bad point denoted by •.

![Scatterplot Matrix](image-url)
number \( m \) of free parameters for the model is

\[
(G - 1) + Gp + G \left[ pq - \frac{1}{2} q (q - 1) \right] + Gp + 2G.
\]

### 4.2 Maximum likelihood estimation via the AECM algorithm

To find ML estimates for the parameters \( \vartheta = \{ \pi_g, \mu_g, \Lambda_g, \Psi_g, \alpha_g, \eta_g \}_{g=1}^G \) of the mixture of contaminated Gaussian factor analyzers model, we consider the AECM algorithm. We partition \( \vartheta \) as \( \{ \vartheta_1, \vartheta_2, \vartheta_3 \} \), where \( \vartheta_1 = \{ \pi_g, \mu_g \}_{g=1}^G \), \( \vartheta_2 = \{ \alpha_g, \eta_g \}_{g=1}^G \), and \( \vartheta_3 = \{ \Lambda_g, \Psi_g \}_{g=1}^G \). The \((k+1)\)th iteration of the algorithm consists of three cycles, and there is one E-step and one CM-step for each cycle. Concerning the specification of the incomplete data, it is useful to introduce the component-indicator vector \( z_i = (z_{i1}, \ldots, z_{iG})' \), where \( z_{ig} \) is one or zero according to whether \( x_i \) belongs or does not belong to the \( g \)th component, with \( i = 1, \ldots, n \) and \( g = 1, \ldots, G \).

#### 4.2.1 First cycle

For the first cycle of the \((k+1)\)th iteration of the AECM algorithm, we specify the missing data to be the component-indicator vectors \( z_1, \ldots, z_n \) and the weights \( w_1, \ldots, w_n \), with \( w_i = (w_{i1}, \ldots, w_{iG})' \). Thus, the complete data are \( (x_1', \ldots, x_n', w_1', \ldots, w_n', z_1', \ldots, z_n')' \) and the complete-data likelihood can be factored as

\[
L_{c1}(\vartheta_1) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \pi_g \phi \left( x_i; \mu_g, \Sigma_g \right) / w_{ig} \right] p_C \left( w_{ig}; \alpha_g, \eta_g \right) \right] z_{ig}.
\]

Accordingly, the complete-data log-likelihood can be written as

\[
l_{c1}(\vartheta_1) = l_{c1} \left( \{ \pi_g \}_{g=1}^G \right) + l_{c2} \left( \{ \mu_g \}_{g=1}^G \right),
\]

where

\[
l_{c1} \left( \{ \pi_g \}_{g=1}^G \right) = \sum_{g=1}^{G} n_g \log \pi_g , \tag{17}
\]

and

\[
l_{c2} \left( \{ \mu_g \}_{g=1}^G \right) = \frac{np}{2} \log (2\pi) + \frac{p}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log w_{ig} - \frac{1}{2} \sum_{g=1}^{G} n_g \log \left| \Sigma_g \right| - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} w_{ig} (x_i - \mu_g)' (\Sigma_g)^{-1} (x_i - \mu_g), \tag{18}
\]

with \( n_g = \sum_{i=1}^{n} z_{ig} \).

The E-step for the first cycle of the \((k+1)\)th iteration requires the calculation of

\[
Q_{c1}(\vartheta_1; \hat{\vartheta}^{(k)}) = E_{\vartheta^{(k)}} \left[ l_{c1}(\vartheta_1) | x_1, \ldots, x_n \right].
\]

This E-step can be effected by first taking the expectation of \( l_{c1}(\vartheta_1) \) conditional on \( x_1, \ldots, x_n \) and \( z_1, \ldots, z_n \), and then finally over the \( z_i \) given \( x_i \). It can be seen from \((17)\) and \((18)\) that in
order to do this, we need to calculate $E_{g^{(k)}} (Z_{ig} | x_i) = z_{ig}^{(k)}$ and $E_{\theta^{(k)}} (W_{ig} | x_i, z_i) = w_{ig}^{(k)}$, where

$$z_{ig}^{(k)} = \frac{\pi_{g}^{(k)} p_{CN} \left( x_i; \mu_{g}^{(k)}, \Sigma_{g}^{(k)}, \alpha_{g}, \gamma_{g} \right)}{\sum_{j=1}^{G} \pi_{j}^{(k)} p_{CN} \left( x_i; \mu_{j}^{(k)}, \Sigma_{j}^{(k)}, \alpha_{j}, \gamma_{j} \right)}$$

and

$$w_{ig}^{(k)} = \frac{\alpha_{g}^{(k)} \phi \left( x_i; \mu_{g}^{(k)}, \Sigma_{g}^{(k)} \right) + \frac{1 - \alpha_{g}^{(k)}}{\gamma_{g}^{(k)}} \phi \left( x_i; \mu_{g}^{(k)}, \gamma_{g}^{(k)} \Sigma_{g}^{(k)} \right)}{\frac{\alpha_{g}^{(k)} \phi \left( x_i; \mu_{g}^{(k)}, \Sigma_{g}^{(k)} \right) + \frac{1 - \alpha_{g}^{(k)}}{\gamma_{g}^{(k)}} \phi \left( x_i; \mu_{g}^{(k)}, \gamma_{g}^{(k)} \Sigma_{g}^{(k)} \right)}{\alpha_{g}^{(k)} \phi \left( x_i; \mu_{g}^{(k)}, \Sigma_{g}^{(k)} \right) + \frac{1 - \alpha_{g}^{(k)}}{\gamma_{g}^{(k)}} \phi \left( x_i; \mu_{g}^{(k)}, \gamma_{g}^{(k)} \Sigma_{g}^{(k)} \right)}}.$$

Using these results we have that

$$Q_{c1} \left( \vartheta_1; \vartheta^{(k)} \right) = Q_{1c1} \left( \left\{ \pi_{g}^{(k)} \right\}_{g=1}^{G} \right) + Q_{2c1} \left( \left\{ \mu_{g}^{(k)} \right\}_{g=1}^{G} \right),$$

where

$$Q_{1c1} \left( \left\{ \pi_{g}^{(k)} \right\}_{g=1}^{G} ; \vartheta^{(k)} \right) = \sum_{g=1}^{G} n_{g}^{(k)} \log \pi_{g}^{(k)} ,$$

and

$$Q_{2c1} \left( \left\{ \mu_{g}^{(k)} \right\}_{g=1}^{G} ; \vartheta^{(k)} \right) = C - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig}^{(k)} w_{ig}^{(k)} (x_i - \mu_{g})' \left( \Sigma_{g}^{(k)} \right)^{-1} (x_i - \mu_{g}) ,$$

with $n_{g}^{(k)} = \sum_{i=1}^{n} z_{ig}^{(k)}$ and $C$ including the terms that do not depend on $\mu_1, \ldots, \mu_G$.

The CM-step for the first cycle of the $(k+1)$th iteration requires the maximization of $Q_{c1} \left( \vartheta_1; \vartheta^{(k)} \right)$. The solutions for $\pi_{g}^{(k+1)}$ and $\mu_{g}^{(k+1)}$ exist in closed form and are

$$\pi_{g}^{(k+1)} = \frac{n_{g}^{(k)}}{n}$$

and

$$\mu_{g}^{(k+1)} = \sum_{i=1}^{n} z_{ig}^{(k)} w_{ig}^{(k)} x_i / \sum_{i=1}^{n} z_{ig}^{(k)} w_{ig}^{(k)} .$$

At the end of this cycle, we write $\vartheta^{(k+1/3)} = \left\{ \vartheta_1^{(k+1)}, \vartheta_2^{(k+1)}, \vartheta_3^{(k+1)} \right\}$.

### 4.2.2 Second cycle

For the second cycle of the $(k+1)$th iteration of the AECM algorithm, for the updating of $\vartheta_2$, we specify the missing data to be only $z_1, \ldots, z_n$. The complete-data likelihood is

$$L_{c2} \left( \vartheta_2 \right) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \pi_{g}^{(k+1)} p_{CN} \left( x_i; \mu_{g}^{(k+1)}, \Sigma_{g}^{(k)}, \alpha_{g}, \gamma_{g} \right) \right]^{z_{ig}} ,$$

where $\Sigma_{g}^{(k)} = \Lambda_{g}^{(k)} \Lambda_{g}^{(k)'} + \Psi_{g}^{(k)}$. Accordingly, the complete-data log-likelihood is

$$l_{c2} \left( \vartheta_2 \right) = C + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \left[ p_{CN} \left( x_i; \mu_{g}^{(k+1)}, \Sigma_{g}^{(k)}, \alpha_{g}, \gamma_{g} \right) \right] ,$$

where $C$ includes the terms that do not depend on $\vartheta_2$. 

14
The E-step on the second cycle of the \((k+1)\)th iteration requires the calculation of
\[
Q_{c2} \left( \vartheta_2; \vartheta^{(k+1/3)} \right) = E_{\vartheta^{(k+1/3)}} [ l_{c2} ( \vartheta_2 ) | \mathbf{x}_1, \ldots, \mathbf{x}_n ].
\]

Based on (19), in order to do this, we need to calculate \( E_{\vartheta^{(k+1/3)}} ( Z_{ig} | \mathbf{x}_i ) = z_{ig}^{(k+1/3)} \), where
\[
z_{ig}^{(k+1/3)} = \frac{\pi_g^{(k+1/3)} p_{CN}( \mathbf{x}_i; \mu_g^{(k+1/3)}, \Sigma_g^{(k+1/3)}, \alpha_g^{(k)}, \eta_g^{(k)} )}{\sum_{j=1}^{G} \pi_j^{(k+1/3)} p_{CN}( \mathbf{x}_i; \mu_j^{(k+1/3)}, \Sigma_j^{(k)}, \alpha_j^{(k)}, \eta_j^{(k)} )}
\]

Using this result we have that
\[
Q_2 \left( \vartheta_2; \vartheta^{(k+1/3)} \right) = \sum_{g=1}^{G} Q_{g2} \left( \alpha_g, \eta_g; \vartheta^{(k)} \right),
\]
where
\[
Q_{g2} \left( \alpha_g, \eta_g; \vartheta^{(k+1/3)} \right) = C_g + \sum_{i=1}^{n} z_{ig}^{(k+1/3)} \log \left[ p_{CN}( \mathbf{x}_i; \mu_g^{(k+1/3)}, \Sigma_g^{(k)}, \alpha_g, \eta_g ) \right],
\]
with \( C_g \) including the terms that do not depend on \( \alpha_g \) and \( \eta_g \). Maximizing (20) with respect to \( \vartheta_2 \), under the constraints on these parameters, is equivalent to independently maximizing each of the \( G \) expressions (21) over \( \alpha_g^{(k+1/3)} \) and \( \eta_g^{(k+1/3)} \), under the constraints \( \eta_g > 1 \) and \( \alpha_g \in (\alpha^*, 1) \), \( g = 1, \ldots, G \). This maximization is equivalent to the numerical maximization problem discussed in Section 2.1.2 for the contaminated Gaussian distribution, with the only difference being that each observation \( \mathbf{x}_i \) contributes to the log-likelihood with a known weight \( z_{ig}^{(k+1/3)} \). At the end of this cycle, we write \( \vartheta^{(k+2/3)} = \{ \vartheta_1^{(k+1)}, \vartheta_2^{(k+1)}, \vartheta_3^{(k)} \} \).

### 4.2.3 Third cycle

For the third cycle of the \((k+1)\)th iteration of the AECM algorithm, for the updating of \( \vartheta_3 \), we specify the missing data to be \( \mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n \), and \( \mathbf{w}_1, \ldots, \mathbf{w}_n \). Thus, the complete data are \( \{ \mathbf{x}_1', \ldots, \mathbf{x}_n', \mathbf{w}_1', \ldots, \mathbf{w}_n', \mathbf{z}_1', \ldots, \mathbf{z}_n', \mathbf{u}_1', \ldots, \mathbf{u}_n \}' \) and, according to the definition of the contaminated Gaussian factor analysis model given in Section 3, the complete-data likelihood can be factored as
\[
L_{c3} ( \vartheta_3 ) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[ \pi_g^{(k+1)} \phi \left( \mathbf{x}_i; \mu_g^{(k+1)} + \mathbf{A}_g \mathbf{u}_{ig}, \mathbf{\Psi}_g / w_{ig} \right) \phi \left( \mathbf{u}_{ig}; \mathbf{0}_g, \mathbf{I}_g / w_{ig} \right) p_C \left( \mathbf{w}_{ig}; \alpha_g^{(k+1)}, \eta_g^{(k+1)} \right) \right] z_{ig}.
\]

Accordingly, the complete-data log-likelihood is
\[
l_{c3} ( \vartheta_3 ) = \sum_{g=1}^{G} n_g \log \pi_g^{(k+1)} - \frac{p + q}{2} \log (2\pi) + \frac{p + q}{2} \sum_{g=1}^{G} n_g \sum_{i=1}^{n} z_{ig} \log w_{ig} - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} w_{ig} \mathbf{u}_{ig}' \mathbf{u}_{ig}
\]
\[
+ \sum_{g=1}^{G} \log \pi_g^{(k+1)} \sum_{i=1}^{n} z_{ig} w_{ig} - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{n} \frac{z_{ig} w_{ig}}{1 - \eta_g^{(k+1)}} + \sum_{g=1}^{G} \log \left( 1 - \alpha_g^{(k+1)} \right) \sum_{i=1}^{n} z_{ig} \frac{1 - w_{ig}}{1 - 1/\eta_g^{(k+1)}}
\]
\[
- \frac{1}{2} \sum_{g=1}^{G} n_g \log |\mathbf{\Psi}_g| - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} w_{ig} \left( \mathbf{x}_i - \mu_g^{(k+1)} \right)' \mathbf{\Psi}_g^{-1} \left( \mathbf{x}_i - \mu_g^{(k+1)} \right)
\]
\[
+ \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} w_{ig} \left( \mathbf{x}_i - \mu_g^{(k+1)} \right)' \mathbf{\Psi}_g^{-1} \mathbf{A}_g \mathbf{u}_{ig} - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} w_{ig} \mathbf{u}_{ig}' \mathbf{\Psi}_g^{-1} \mathbf{A}_g \mathbf{u}_{ig},
\] (22)
where \( n_g = \sum_{i=1}^{n} z_{ig} \)

The E-step on the third cycle of the \((k + 1)\)th iteration requires the calculation of

\[
Q_3 \left( \vartheta_3; \vartheta^{(k+2/3)} \right) = E_{\vartheta^{(k+2/3)}} \left[ l_{c3} (\vartheta_3) | x_1, \ldots, x_n \right].
\]

In addition to update \( z_{ig} \) to

\[
z^{(k+2/3)}_{ig} = \frac{\pi^{(k+1)}_g \text{PCN} \left( x_i; \mu^{(k+1)}_g, \Sigma^{(k)}_g, \alpha^{(k+1)}_g, \eta^{(k+1)}_g \right)}{\sum_{j=1}^{G} \pi^{(k+1)}_j \text{PCN} \left( x_i; \mu^{(k+1)}_j, \Sigma^{(k)}_j, \alpha^{(k+1)}_j, \eta^{(k+1)}_j \right)},
\]

and \( w_{ig} \) to

\[
w^{(k+2/3)}_{ig} = \frac{\alpha^{(k+1)}_g \phi \left( x_i; \mu^{(k+1)}_g, \Sigma^{(k)}_g \right) + \frac{1 - \alpha^{(k+1)}_g}{\eta^{(k+1)}_g} \phi \left( x_i; \mu^{(k+1)}_g, \eta^{(k+1)}_g \Sigma_g \right)}{\alpha^{(k+1)}_g \phi \left( x_i; \mu^{(k+1)}_g, \Sigma^{(k)}_g \right) + \phi \left( x_i; \mu^{(k+1)}_g, \eta^{(k+1)}_g \Sigma_g \right)},
\]

where \( \Sigma^{(k)}_g = \Lambda^{(k)}_g \Lambda^{(k)'}_g + \Psi^{(k)}_g \), due to the last two rows of \([22]\) we also calculate

\[
E_{\vartheta^{(k+2/3)}} \left( Z_{ig} W_{ig} U_{ig} | x_i, w_{ig} \right) = w^{(k+2/3)}_{ig} \gamma^{(k)}_g \left( x_i - \mu^{(k+1)}_g \right)
\]

and

\[
E_{\vartheta^{(k+2/3)}} \left( Z_{ig} W_{ig} U_{ig} U^{(k+2/3)}_{ig} | x_i, w_{ig} \right) = I_q - \gamma^{(k)'}_g \Lambda^{(k)}_g + w^{(k+2/3)}_{ig} \gamma^{(k)}_g \left( x_i - \mu^{(k+1)}_g \right) \left( x_i - \mu^{(k+1)}_g \right)'
\]

where \( \gamma^{(k)}_g = \left( \Lambda^{(k)}_g \Lambda^{(k)'}_g + \Psi^{(k)}_g \right)^{-1} \Lambda^{(k)}_g \). Hence, starting from \([22]\), the expected complete-data
log-likelihood of the third cycle is

\[
Q_3 \left( \vartheta_3; \vartheta^{(k+2/3)} \right) = \sum_{g=1}^{G} Q_{3g} \left( \Lambda_g, \Psi_g; \vartheta^{(k+2/3)} \right),
\]

with

\[
Q_{3g} \left( \Lambda_g, \Psi_g; \vartheta^{(k+2/3)} \right) = C_g - \frac{1}{2} n^{(k+2/3)}_g \log |\Psi_g| - \frac{1}{2} n^{(k+2/3)}_g \text{tr} \left( \Psi_g^{-1} S^{(k+2/3)}_g \right)
\]

\[
+ n^{(k+2/3)}_g \text{tr} \left( \Psi_g^{-1} \Lambda_g \gamma^{(k)}_g S^{(k+2/3)}_g \right) - \frac{1}{2} n^{(k+2/3)}_g \text{tr} \left( \Lambda_g \Psi_g^{-1} \Lambda_g R^{(k+2/3)}_g \right),
\]

where

\[
S^{(k+2/3)}_g = \frac{1}{n^{(k+2/3)}_g} \sum_{i=1}^{n^{(k+2/3)}_g} \left( x_i - \mu^{(k+1)}_g \right) \left( x_i - \mu^{(k+1)}_g \right)'
\]

and

\[
F^{(k+2/3)}_g = I_q - \gamma^{(k)'}_g \Lambda^{(k)}_g + \gamma^{(k)}_g S^{(k+2/3)}_g \gamma^{(k)}_g,
\]

and \( C_g \) includes the terms that do not depend on \( \Lambda_g \) and \( \Psi_g, \, g = 1, \ldots, G \). After some algebra, maximization of \( Q_{3g} \left( \Lambda_g, \Psi_g; \vartheta^{(k+2/3)} \right) \) over \( \Lambda_g \) and \( \Psi_g \) yields the updated estimates

\[
\Lambda^{(k+1)}_g = S^{(k+2/3)}_g \gamma^{(k)}_g \left( R^{(k+2/3)}_g \right)^{-1}
\]

and

\[
\Psi^{(k+1)}_g = \text{diag} \left( S^{(k+2/3)}_g - \Lambda^{(k+1)}_g \Lambda^{(k)'}_g S^{(k+2/3)}_g \right),
\]

16
4.3 Computational details

Analogously to the estimation of the contaminated Gaussian distribution, the `optim()` function is used to maximize each $Q_{gc2}(\alpha_g, \eta_g; \theta^{(k+1/3)})$, $g = 1, \ldots, G$, and the Woodbury identity, given in [14], is considered to compute $\gamma^{(k)}_g$, $g = 1, \ldots, G$. Moreover, based on Section 3.4 the AECM algorithm is initialized with the estimates of $\pi_g$, $\mu_g$, $\Lambda_g$, and $\Psi_g$ provided by a mixture (with $G$ components) of Gaussian factor analyzers, along with the constraint $\alpha_g = \alpha^*$ (with $\alpha^* \rightarrow 1^-$) and $\eta_g = \eta^*$ (with $\eta^* \rightarrow 1^+$), $g = 1, \ldots, G$. In the analysis of Section 4.4 the (preliminary) mixture of Gaussian factor analyzers is estimated by the `pgmmEM()` function of the pgmm package for R. The Aitken acceleration is considered as convergence criterion (see Section 2.2.2 for details).

4.4 Real data analysis

The `f.voles` data set, detailed in Flury (1997, Table 5.3.7) and available in the Flury package for R (Flury, 2012), consists of measurements on female voles from two species, *M. californicus* and *M. ochrogaster*. The data refer to $n = 86$ observations for which we have a binary variable `Species`, denoting the species (45 *M. ochrogaster* and 41 *M. californicus*), as well as the following $p = 7$ continuous variables (the names of the variables are the same as in the original analysis of this data set by Airoldi and Hoffmann, 1984):

- Age: Age measured in days;
- L2.Condylo: Condylar incisive length;
- L9.Inc.Foramen: Incisive foramen length;
- L7.Alveolar: Alveolar length of upper molar tooth row;
- B3.Zyg: Zygomatic width;
- B4.Interorbital: Interorbital width;
- H1.Skull: Skull height.

All of the variables related to the skull are measured in units of 0.1 mm. The scatterplot matrix of these data, with the clustering induced by `Species`, is shown in Figure 5.

For our purposes, we assume that data are unlabelled with respect to `Species` and that our interest is in evaluating, with respect to mixtures of Gaussian factor analyzers, clustering and robustness to anomalous points of our proposed model. Following the scheme adopted by Peel and McLachlan (2000) on the well-known crabs data, 36 “perturbed” data sets are created by substituting the original value 122 of `Age` for the 81th point (highlighted by a black bullet in Figure 5) with atypical values ranging from 450 to 800, with increments of 10. The two competing techniques are run for $G \in \{1, 2, 3\}$ and $q \in \{1, 2, 3\}$; the best model is selected by BIC.

As concerns the mixture of Gaussian factor analyzers model, apart from the perturbations 530 and 640, where the selected values are $G = 3$ and $q = 2$, in the former case, and $G = 1$ and $q = 3$ in the latter case, the best values are $G = 1$ and $q = 2$. Hence, the value $G = 2$ is never selected and the obtained clustering is far from that induced by `Species`. When the number of mixture components is fixed to $G = 2$, in a sort of model-based classification analysis, the best BIC model has always a single factor ($q = 1$) and the classification on the unperturbated 85 observations is shown in Table 1. As we can see, there are four misclassified observations.

|                  | 1    | 2    |
|------------------|------|------|
| *M. californicus*| 38   | 3    |
| *M. ochrogaster* | 1    | 43   |

Table 1: Classification results for the mixture of Gaussian factor analyzers model on the perturbed variables of the `f.voles` data set. Classification is only evaluated for the unperturbed observations.
corresponding to a misclassification rate of 0.047.

With regards to the mixture of contaminated Gaussian factor analyzers model, for all perturbed data sets, the best BIC values are \( G = 2 \) and \( q = 2 \), always leading to the classification results in Table 2. In contrast with the clustering results obtained for the mixture of Gaussian

Table 2: Classification results for the mixture of contaminated Gaussian factor analyzers model on the perturbed variants of the \textit{f.voles} data set.

|                | 1  | 2  | Bad |
|----------------|----|----|-----|
| \textit{M. californicus} | 41 |    |     |
| \textit{M. ochrogaster}   |    | 44 | 1   |

factor analyzers model, the true number of groups is always selected; furthermore, the selected model is robust to these perturbations, with the number of misallocated observations being null regardless of the particular value perturbed. Interestingly, we can also note how the model is able to detect the introduced bad point (see the last column of Table 2). Finally, by recalling that the original value of \( \text{age} \) for the 81th point was 122, it is also interesting to note that the estimated value of \( \eta_g \) (in the group containing the outlier) increases almost linearly as the value of this point further departs from its true value (cf. Figure 6).

5 Discussion

In this paper, the factor analysis model has been extended to the contaminated Gaussian distribution. Methodological contributions have been contextualized in the high-dimensional setting and have involved the definition of both the contaminated Gaussian factor analysis
model (cf. Section 3) and the mixture of contaminated Gaussian factor analyzers model (cf. Section 4). In one sense, these models can be respectively viewed as a generalization of the (Gaussian) factor analysis model, and of the mixture of (Gaussian) factor analyzers model, that accommodates outlying observations, spurious observations, or noise, which we have collectively referred to as bad points. Although approaches for high-dimensional data such as the $t$-factor analysis model and mixtures of $t$-factor analyzers can be used for data comprising bad points, they “assimilate” bad points rather than separating them out. Computational contributions have concerned the detailed illustration of AECM algorithms for fitting the above models, as well as the definition of a new version, with respect to Punzo and McNicholas (2014), of the ECM algorithm for maximum likelihood parameter estimation of the contaminated Gaussian distribution; the new version adopts the characterization of this distribution as a member of the Gaussian scale mixtures family.

When applied to real data, our models have shown their supremacy with respect to their Gaussian counterparts. On the bivariate “symmetric” financial data of Section 2.3 we have pointed out that, with respect to the contaminated Gaussian distribution, there were about the fifty percent of bad points creating departure from normality, while on the state.x77 data of Section 3.5 we have shown how the contaminated Gaussian factor analysis model can reduce dimensionality and simultaneously detect bad points. Finally, in the application to the f.voles data, our mixture of contaminated Gaussian factor analyzers model performed better in terms of clustering/classification and gave consistent results regardless of the extent of the perturbation (cf. Section 4.4).

Future work will focus on the development of an R package to facilitate dissemination of our approaches. In addition, high-dimensional contamination of non-elliptical densities will be explored and once realized, will lead to an even more flexible modelling paradigm.

References

Airoldi, J. and Hoffmann, R. (1984). Age variation in voles (*Microtus californicus, M. ochrogaster*) and its significance for systematic studies. Occasional papers of the Museum of Natural History 111, University of Kansas, Lawrence, KS.
Aitken, A. C. (1926). On Bernoulli’s numerical solution of algebraic equations. In Proceedings of the Royal Society of Edinburgh, volume 46, pages 289–305.

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. Technometrics, 22(3), 325–331.

Baganato, L. and Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Computational Statistics, 28(4), 1571–1597.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). Latent Variable Models and Factor Analysis: A Unified Approach, volume 899 of Wiley Series in Probability and Statistics. Wiley, United Kingdom, third edition.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis, 41(3-4), 561–575.

Böhm, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals of the Institute of Statistical Mathematics, 46(2), 373–388.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. Journal of Multivariate Analysis, 11(3), 368–385.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition, 28(5), 781–793.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38.

Flury, B. (1997). A First Course in Multivariate Statistics. Springer, New York.

Flury, B. (2012). Flury: Data Sets from Flury, 1997. R package version 0.1-3.

Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison. Journal of Classification, 19(2), 249–276.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics & Data Analysis, 41(3–4), 577–590.

Kotz, S. and Nadarajah, S. (2004). Multivariate t-Distributions and Their Applications. Cambridge University Press, Cambridge.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84(408), 881–896.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika, 81(4), 633–648.

McLachlan, G. and Krishnan, T. (2007). The EM algorithm and extensions, volume 382 of Wiley Series in Probability and Statistics. John Wiley & Sons, New York, second edition.

McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. John Wiley & Sons, New York.

McLachlan, G. J., Peel, D., and Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics & Data Analysis, 41(3), 379–388.

McLachlan, G. J., Bean, R. W., and Ben-Tovim Jones, L. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Computational Statistics & Data Analysis, 51(11), 5327–5338.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. Statistics and Computing, 18(3), 285–296.

20
McNicholas, P. D., Jampani, K. R., McDaid, A. F., Murphy, T. B., and Banks, L. (2011). *pgmm*: Parsimonious Gaussian mixture models. R package version 1.0.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2), 267–278.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3), 511–567.

Montanari, A. and Viroli, C. (2011). Maximum likelihood estimation of mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 55(9), 2712–2723.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *The Computer Journal*, 7(4), 308–313.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4), 339–348.

Punzo, A. and McNicholas, P. D. (2014). Robust clustering via parsimonious mixtures of contaminated Gaussian distributions. arXiv.org e-print 1305.4669, available at: [http://arxiv.org/abs/1305.4669](http://arxiv.org/abs/1305.4669).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. (2014). *psych*: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.4.3.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7(1), 5–40.

Watanabe, M. and Yamaguchi, K. (2003). *The EM Algorithm and Related Statistical Models*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Woodbury, M. A. (1950). Inverting modified matrices. Technical Report 42 of the Statistical Research Group, Princeton University, Princeton, New Jersey.

Zhao, J.-H. and Yu, P. L. H. (2008). Fast ML estimation for the mixture of factor analyzers via an ECM algorithm. *IEEE Transactions on Neural Networks*, 19(11), 1956–1961.