Correspondence estimation of the source profiles in receptor modeling

Byron J Gajewski, University of Kansas Medical Center
Clifford H Spiegelman, Texas A&M University
Correspondence estimation of the source profiles in receptor modeling

Byron J. Gajewski1,* and Clifford H. Spiegelman2

1Assistant Professor, Schools of Allied Health and Nursing, The University of Kansas Medical Center, Kansas City, KS 66160, U.S.A.
2Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

SUMMARY

This article considers the estimation of source profiles from pollution data collected at one receptor site. At this receptor site, varying metrological conditions can cause errors that are possibly a mixture of distributions. A standard estimator utilizes a least squares approach because of its optimal properties under normally distributed errors and consistency under many other distributions. In contrast, we study the behavior of least squares relative to our new approach, which is better suited for dealing with errors having a mixture of distributions. The estimator loses efficiency under normal errors, but in turn gains efficiency while in the presence of a mixture of distributions. The new alternative has a tuning constant that determines the level of efficiency, which we show using asymptotic theory for large samples and simulation for small samples. An example from Houston, TX, U.S.A. is considered.

KEY WORDS: constrained modified L_2 error; constrained nonlinear least squares; asymptotic inefficiency; robust factor analysis

1. INTRODUCTION

Proper analysis of pollution data collected at receptor sites can lead to a better understanding of the pollution sources. To that end, it is important to establish good estimators for source profiles in receptor modeling (e.g. Park et al., 2000a, 2000b; Billheimer, 2001; Christensen and Sain, 2002; Park, Spiegelman, and Henry 2002). The type of estimation procedure should depend on our assumptions. For example, some assumptions rely on a probability distribution function that remains the same throughout the analysis time frame. However, due to the nature of receptor pollution data, the distribution tends to change in part due to meteorological conditions. Because of this, the pollution data may be a mixture of distributions. Least squares based estimators are sensitive to the mixture of distributions and may not be a good estimator for the profiles associated with the general population. Therefore, to estimate the source profiles associated with the general population we consider an alternative estimator that is motivated by robust estimation. We focus our attention on the development...
of the new method and study its theoretical properties as compared to least squares and other alternatives.

For data in which least squares are appropriate, Park et al. (2002) present an approach to receptor modeling from a statistical point of view. One of their most important results is estimating source profiles and providing its statistical properties. In this article, we extend their presentation by considering a robust alternative. Similar to Park et al.’s least squares approach, we provide a discussion of ours from a statistical point of view. In addition, we discuss traditional statistical views found in factor analysis (FA). FA is a view of receptor modeling in which one interprets the constrained loadings as source profiles.

The robust FA literature provides an array of potential estimators (Campbell, 1980; Balderjahn, 1986; Ammann, 1989; Brown, 1990; Browne, 1990; Kano and Ihara, 1994; Walczak and Massart, 1995; Kosfeld, 1996; Meijer and Mooijaart, 1996; Curran et al., 1996; Paatero, 1997; Hawkins and McLachlan, 1997; Lee and Press, 1998; Pison et al., 2002; Croux et al., 2002). Arguably, the most widely used estimator of robust loadings in FA is the minimum volume ellipsoid (MVE) method popularized by Rousseeuw (1985). This method, and variations on it (like minimum covariance determinant, MCD), is designed to find an ellipsoid with minimum volume covering half the data points. The literature demonstrates the merits of MVE in robust FA (Naga and Antille, 1990; Filzmoser, 1999; Marden, 1999).

We are not particularly interested in classic robustness in FA. Our interest is more specific; we want good estimators for the source profiles in receptor modeling. In this article we consider an alternative estimator we call Constrained Modified L2Error (CML2E). In this article we consider an alternative estimator we call Constrained Modified L2Error (CML2E). The CML2E has a tuning parameter which determines the level of weighting. The approach is conceptually close to Park et al.’s (2002) least squares approach in that it accounts for the type of constraints. It extends the least squares approach because it can be used to find good estimators of the source profiles in the setting of a mixture of distributions. We demonstrate this advantage as applied to receptor modeling.

In this article, we introduce the notation, general model and constraints in Section 2. In Sections 3 and 4, we present our alternative to least squares along with asymptotic properties which details a strategy for selecting a tuning constant used in practice. We also explore the small sample properties with simulation that compares our approach to other robust alternatives. We demonstrate our CML2E estimator on data that simulates a mixture of two source profiles. We apply our method to data collected at a receptor site in Houston, Texas, U.S.A. We make concluding remarks in Section 5.

2. BACKGROUND

2.1. Model

Pollution data collected at one site provide adequate information for model building and form a basis for estimating the number of major pollution sources, the amount of pollution from each source, and the proportion of species from each source. Over time, species data are stored in the matrix \( Y \). The foundation of receptor modeling is decomposing \( Y \) into specific bilinear model parameters (see Linder and Sundberg, 1998, for a bilinear model from a calibration standpoint).

Viewed as a chemical mass balance equation, the matrix \( Y \) is a sum of all the pollution sources over time, plus error (Henry et al., 1984, 1994; Hopke, 1985, 1991; Spiegelman and Dattner, 1993). Using the notation in Park et al. (2002), the chemical mass balance equation can be viewed as a factor analysis
(FA) model (Anderson, 1984; Mardia et al., 1995; Johnson and Wichern, 1998) with the equation 

\[ Y_i = \alpha_i P + \varepsilon_i. \]

The scores (\( \alpha_i \)) are source contributions at time \( i \), and the loadings (\( P \)) hold the source profiles having dimension \( q \) by \( p \). The number of sources is \( q \) (assumed known), and the number of variables is \( p \). We assume that \( \varepsilon_i \) is the error, and the contributions (\( \alpha_i \)) are assumed to be a fixed set of parameters. By defining \( Y_i, \alpha_i \) and \( \varepsilon_i \) to be the rows of \( Y, A \) and \( E \), respectively, we have the matrix version of the FA model as \( Y = AP + E \). The profiles (loadings) contain information regarding the percentage of a particular species in a major pollution source. The profiles are fingerprints for the sources in that they are a basis for providing specific information, and their estimation is the focus of this article.

2.2. Constraints

Natural constraints on the profiles can occur in practice. The parameters \( A \) and \( P \) are not identified in FA without these constraints. Park et al. (2002) define conditions on the profiles (loadings, \( P \)). As in Park et al. (2002), we choose to zero some of the elements in the profile matrix. This specifically is displayed as condition one (C1) to condition three (C3).

C1. There are at least \( q-1 \) zero elements in each row of \( P \).

C2. The columns of \( P \) containing the assigned 0s in the \( k \)th row, with those assigned 0s deleted, is called \( P^{(k)} \). The rank of \( P^{(k)} \) is \( q-1 \). These 0s are preset.

C3. We assume that each row of \( P \) sums to one and that each element is non-negative.

We say \( P \subseteq C \), to mean that the loadings are in the constrained set \( C \), defined from conditions C1–C3. We choose to interpret the source profiles such that each element is the proportion of species that contributes to a particular source. The third condition preserves this interpretation. We demonstrate these conditions with the matrix

\[
P = \begin{pmatrix}
0.1 & 0 & 0 & 0.9 \\
0.8 & 0.2 & 0 & 0 \\
0.5 & 0 & 0.5 & 0
\end{pmatrix}
\]

Since each \( P^{(k)} \) is a matrix composed of the columns containing the assigned 0s in the \( k \)th row with those assigned 0s deleted, we obtain

\[
P^{(1)} = \begin{pmatrix}
0.2 & 0 \\
0 & 0.5
\end{pmatrix}
\]

Similarly,

\[
P^{(2)} = \begin{pmatrix}
0 & 0.9 \\
0.5 & 0
\end{pmatrix}
\]

and

\[
P^{(3)} = \begin{pmatrix}
0 & 0.9 \\
0.2 & 0
\end{pmatrix}
\]

The constraints are directly incorporated into the model for unique parameter estimates.

3. THEORY

3.1. Objective functions

The term ‘constrained’ is used throughout the discussion of competing estimators to emphasize that all estimators presented are calculated with constraints placed on the profiles in order to obtain meaningful parameter interpretations.
Park et al. (2002) present the least squares approach to estimation (a similar approach to that of Bartlett, 1937). Taken directly from the definition of least squares, when identifiability conditions C1–C3 are applied, the constrained nonlinear least squares (CNLS) objective function is

$$
\sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij} - \alpha_iP) = \text{tr}[(Y - AP)'(Y - AP)]
$$

(1)

Define $P = (I - P'(PP')^{-1})$. By plugging the estimated scores $\hat{A} = YP'(PP')^{-1}$ into (1) and minimizing with respect to $P$, the CNLS estimator is

$$
\hat{P}_{C} = \arg \min_{P \subseteq C} \sum_{i=1}^{n} Y_iP_Y'
$$

(2)

where the ‘C’ will be short for the CNLS approach and $Y_i$ is the $i$th row of $Y$. Notice that $P = (I - P'(PP')^{-1})$ projects $P$ onto a matrix of zeros. We define $P_0$ to be the true unknown profiles (loadings).

Based on the L2 error (L2E) found in Scott (1998), we develop a competitor to CNLS. Let $\theta_0$ be the unknown parameter that determines a probability density function (pdf), and let $\theta$ be the estimator for $\theta_0$. In the univariate parametric model, Scott finds $\hat{\theta}$ by minimizing the L2E objective function to get the estimator: $\hat{\theta} = \arg \min \{ f(x \mid \theta)^2 dx - (2/n) \sum_{i=1}^{n} f(x_i \mid \theta) \}$, where $x_1, x_2, x_3, \ldots, x_n \sim \text{iid } f(x_i \mid \theta_0)$. We develop the constrained modified L2E (CML2E) using Scott’s concept. Recall in FA that $Y = AP_0 + \varepsilon$, where $\varepsilon_i = Y_i - \alpha_iP_0 \sim \text{iid } N(0, \sigma^2I_p)$ and $\alpha_i$ is the $i$th row of $A$. Therefore, using these residuals, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})$, we calculate the L2E objective function:

$$
L_2E(\alpha_i) = \int f(y \mid \alpha_i, P)^2 dy - 2\sum_{i=1}^{n} f(Y_i \mid \alpha_i, P)/n
$$

(3)

where $f(y \mid \alpha_i, P) = \exp(-(y - \alpha y)'(2\sigma^2)^{1/2}(2\pi)^{p/2})$. The integral in (3) is calculated as $\int f(y \mid \alpha_i, P)^2 dy = 1/(2\pi)^{p/2}\sigma^p$. This calculation leads to the L2E objective function $F(A, P) = 1/(2\pi)^{p/2}\sigma^p - 2\sum_{i=1}^{n} f(Y_i \mid \alpha_i, P)/n$. To facilitate calculations, we modify the L2E objective function with the least squares estimator for the scores ($A$). The least squares solution to the scores is then $\hat{A} = YP'(PP')^{-1}$; hence, the L2E objective function, while applying identifiability conditions C1–C3 during estimation, is

$$
F(P) = 1/(2\pi)^{p/2}\sigma^p - 2\sum_{i=1}^{n} \exp[-Y_i(I - P'(PP')^{-1})Y_i/(2\sigma^2)]/(2\pi)^{p/2}\sigma^p n
$$

(4)

Unlike CNLS, the L2E has a nuisance parameter, $\sigma^2$, from the error structure (strategies for estimating $\sigma^2$ are presented in Section 4.2). We choose to modify the L2E by placing multiples of $\sigma^2$, called the tuning parameter ($c$), into (4). This substitution leads to the constrained modified L2E (CML2E) estimator

$$
\hat{P}_{R} = \arg \min_{P \subseteq C} \left\{ -\sum_{i=1}^{n} \exp[-Y_i(I - P'(PP')^{-1})Y_i/c] \right\}
$$

(5)
where ‘$R$’, for robust (correspondence), will stand for the CML$_2$E estimator. Notice that $c = 2\sigma^2$ corresponds to (4). We will show that, by adjusting $c$, the resulting estimator will have properties similar to CNLS for large $c$ and different properties from CNLS for smaller $c$. One can view our estimator as an extension of the univariate estimation case found in Basu et al. (1998). We emphasize that Park et al. (2002) has established the validity of the CNLS, but there are particular distribution structures where the CML$_2$E is better than CNLS. Two theorems and a simulation will illustrate this fact in addition to assisting us in the choice of the tuning parameter $c$.

3.2. Main theorems

Two main theorems provide insight regarding the properties of CNLS and CML$_2$E. These theorems hold true, assuming a particular structure on the model. Initially, the errors are normal, independent, and identically distributed, but this assumption is relaxed in later theory. Similar to a quasi-structural problem proposed by Gleser (1983) (see also Park et al., 2002), we assume that the first two moments of the scores converge.

The base model assumptions are:

1. $E[Y_i] = \alpha_i P_0$
2. $\varepsilon_i \sim \text{iid } N(0, \sigma^2 I_p)$

<table>
<thead>
<tr>
<th>$1 - \Delta$</th>
<th>aveBias</th>
<th>aveMSE</th>
<th>Max</th>
<th>Corr</th>
<th>Med</th>
<th>%BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.04</td>
<td>0.06</td>
<td>59.32</td>
<td>0.9997</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>1.91</td>
<td>3.13</td>
<td>422.07</td>
<td>0.9904</td>
<td>1.07</td>
<td>94</td>
</tr>
<tr>
<td>0.2</td>
<td>3.35</td>
<td>9.66</td>
<td>760.58</td>
<td>0.9785</td>
<td>3.22</td>
<td>95</td>
</tr>
<tr>
<td>0.3</td>
<td>4.68</td>
<td>17.66</td>
<td>1013.88</td>
<td>0.9656</td>
<td>6.42</td>
<td>95</td>
</tr>
<tr>
<td>0.4</td>
<td>5.24</td>
<td>22.61</td>
<td>1128.85</td>
<td>0.9577</td>
<td>7.57</td>
<td>100</td>
</tr>
<tr>
<td>0.5</td>
<td>6.05</td>
<td>29.53</td>
<td>1274.34</td>
<td>0.9468</td>
<td>10.78</td>
<td>99</td>
</tr>
<tr>
<td>0.6</td>
<td>6.47</td>
<td>34.81</td>
<td>1358.92</td>
<td>0.9384</td>
<td>11.97</td>
<td>100</td>
</tr>
<tr>
<td>0.7</td>
<td>7.08</td>
<td>42.02</td>
<td>1463.46</td>
<td>0.9336</td>
<td>14.05</td>
<td>92</td>
</tr>
<tr>
<td>0.8</td>
<td>7.15</td>
<td>45.74</td>
<td>1482.84</td>
<td>0.9277</td>
<td>14.47</td>
<td>88</td>
</tr>
<tr>
<td>0.9</td>
<td>7.96</td>
<td>52.50</td>
<td>1607.94</td>
<td>0.9223</td>
<td>15.44</td>
<td>92</td>
</tr>
<tr>
<td>CML$_2$E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.01</td>
<td>0.09</td>
<td>71.16</td>
<td>0.9995</td>
<td>0.03</td>
<td>1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.02</td>
<td>0.09</td>
<td>68.29</td>
<td>0.9996</td>
<td>0.03</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.09</td>
<td>0.10</td>
<td>75.48</td>
<td>0.9996</td>
<td>0.04</td>
<td>2</td>
</tr>
<tr>
<td>0.3</td>
<td>0.08</td>
<td>0.11</td>
<td>76.38</td>
<td>0.9995</td>
<td>0.04</td>
<td>4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.13</td>
<td>0.12</td>
<td>82.24</td>
<td>0.9995</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>0.5</td>
<td>0.16</td>
<td>0.13</td>
<td>84.41</td>
<td>0.9985</td>
<td>0.05</td>
<td>2</td>
</tr>
<tr>
<td>0.6</td>
<td>0.03</td>
<td>0.16</td>
<td>94.14</td>
<td>0.9995</td>
<td>0.06</td>
<td>6</td>
</tr>
<tr>
<td>0.7</td>
<td>0.40</td>
<td>0.28</td>
<td>126.19</td>
<td>0.9986</td>
<td>0.10</td>
<td>19</td>
</tr>
<tr>
<td>0.8</td>
<td>0.40</td>
<td>1.91</td>
<td>207.28</td>
<td>0.9141</td>
<td>0.60</td>
<td>57</td>
</tr>
<tr>
<td>0.9</td>
<td>6.68</td>
<td>45.10</td>
<td>1345.61</td>
<td>0.9105</td>
<td>12.04</td>
<td>93</td>
</tr>
</tbody>
</table>
Remark 1. The base model assumes that each variable has the same variance, i.e. \( \varepsilon_i \sim N^{\text{iid}}(0, \sigma^2 I_p) \). Park et al. (2002) provides a remedy to the case when has a covariance matrix \( \Sigma \) that is not equal to \( \sigma^2 I_p \).

Remark 2. Assumption 6 is required in order to minimize the variance inflation factor (VIF). The number ten comes from VIF for regression, i.e. Hocking (1996).

Remark 3. In assumption 5, if the source contributions behave like normal variates, then Graybill (1976) provides the basis of the calculations when \( \alpha_i \sim N(\alpha_0, K_0) \). In our case, 
\[
E(\exp(-\alpha_i P_0 P'_0 \alpha_i/(c(1 + 2\sigma^2/c))) = \Psi(P_0, \alpha_0, K_0)
\]
where the form of \( \Psi(P_0, \alpha_0, K_0) \) is found using the integral presented in Graybill (1976, p.48). By the weak law of large numbers,
\[
n^{-1} \sum_{i=1}^n \exp(-\alpha_i P_0 P'_0 \alpha_i/(c(1 + 2\sigma^2/c))) \xrightarrow{\text{a.s.}} \Psi(P_0, \alpha_0, K_0)
\]
Consequently, this demonstrates the assumptions with ‘normally behaved’ parameters. By using the same argument, we can extend to a finite mixture of ‘normally behaved’ parameters.

Prior to the Theorems, it is necessary to clearly define some notation about the profiles (loadings, \( P_0 \)). The column vector of the free parameters of \( P_0 \) is \( \theta_0 \) and is on a compact set since each element of \( P_0 \) is non-negative and each row sums to one. Let \( r \) be the length of \( \theta_0 \). Throughout this paper we assume that \( \theta_0 \) is in the interior of its parameter space. We use this and the base model assumptions to prove the first of two main Theorems. We prove all Theorems in the Appendix of this article. The values \( \hat{\theta}_R \) and \( \hat{\theta}_C \) are, respectively, the CML2E and CNLS estimators of \( \theta_0 \).

Theorem 1. Assume that the data are from the base model. Then, as \( n \to \infty \), \( \hat{\theta}_R \xrightarrow{\text{p}} \theta_0 \),
\[
\sqrt{n}(\hat{\theta}_R - \theta_0) \xrightarrow{\text{d}} N(0, a\{B^{-1} AB^{-1} + B^{-1}\}), \quad \hat{\theta}_C \xrightarrow{\text{p}} \theta_0
\]
and
\[
\sqrt{n}(\hat{\theta}_C - \theta_0) \xrightarrow{\text{d}} N(0, B^{-1} AB^{-1} + B^{-1})
\]
where
\[
a = (1 + 2\sigma^2/c)^{p-q+2}/(1 + 4\sigma^2/c)^{(p-q)/2+1}
\]
\[
A = L(P_0 \otimes (P_0 P'_0)^{-1})L', \quad B = L(P_0 \otimes (K_0 + \alpha'_0 \alpha_0))L'/\sigma^2
\]
and
\[
L = [\text{Vec}(\partial P_0/\partial \theta_1)' \text{Vec}(\partial P_0/\partial \theta_2)' \cdots \text{Vec}(\partial P_0/\partial \theta_r)']'
\]
Remark 4. The variance result for $\hat{\theta}_C$ in the FA is a special case of the variance calculation in Park et al. (2002). The error structure is normal as opposed to more general.

Theorem 1 shows that CNLS and CML$_2$E consistently estimate the true loadings and it describes their asymptotic distributions. This expands on a similar theorem in Park et al. (2002) to include properties for CML$_2$E. While providing a comparison for the two estimators, Theorem 1 is limited to normally distributed errors. Under these conditions, for any linear combination of $C_{18}$ provide a case where CML$_2$E has a lower asymptotic MSE than CNLS.

We use a simple mixture of normal errors as an alternative to normal errors for two reasons. The first is that the simple mixture of normal errors broadens the class of distributions to data that are not normally distributed. The second reason for using this class of distributions is that we simplify the expectation of powers of quadratic forms from non-normal data, specifically referencing a set of papers from Subrahmaniam (1968) to express the difficulty in these calculations.

As an alternative to normally distributed data, we model mixture distributions in the form of heavy tails and mean shifts. The first of the two mixtures is

$$M1: Y_i \sim \Delta \mathcal{N}(\alpha_i P_0, \sigma^2 I) + (1 - \Delta)\mathcal{N}(\alpha_i P_0, (K\sigma)^2 I), \text{ where } 0 \leq \Delta \leq 1$$

This mixture models heavy tailed, but symmetric, distributions. The ‘heaviness’ is based on large $K$ for fixed $\Delta$, where $\Delta$ is the mixing parameter. We will later show that the estimating metrics in this article produce unbiased estimators for finite $K$ (scalar). The second mixture is

$$M2: Y_i \sim \Delta \mathcal{N}(\alpha_i P_0, \sigma^2 I) + (1 - \Delta)\mathcal{N}(\gamma_i P_2, (K\sigma)^2 I), \text{ where } 0 \leq \Delta \leq 1.$$

This mixture models changes in the mean and have heavy tails. We can easily show that the estimating metrics under this distribution produce biased estimates. However, simulation shows that CML$_2$E has lower MSE than CNLS. Asymptotic theory addresses the properties of the estimators under $M1$; simulation addresses the properties of the estimators under $M2$.

The results of Theorem 2 indicate that both CML$_2$E and CNLS estimators are consistent and asymptotically normal when the mix manifests itself symmetrically in the variance of the error. The asymptotic inefficiency (AI) is the trace of the variance of CML$_2$E divided by the trace of the variance of the CNLS.

Theorem 2. Under the assumptions in Theorem 1 but with the error distribution presented in $M1$, the asymptotic inefficiency (AI) of CML$_2$E to CNLS is

$$AI = \left\{ \left\{ \Delta a_1 + (1 - \Delta) a_K K^4 \right\} + \left\{ \Delta a_1 + (1 - \Delta) a_K K^2 \right\} \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \right\} / \left\{ \left\{ \Delta + (1 - \Delta) K^4 \right\} + \left\{ \Delta + (1 - \Delta) K^2 \right\} \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \right\}^{-1} \right\} / \left\{ \Delta b_1 + (1 - \Delta) b_K \right\}^2,$$

where $a_K = (1/(1 + 4\sigma^2 K^2/c))^{(p-q)/2+1}$ and $b_K = (1/(1 + 2\sigma^2 K^2/c))^{(p-q)/2+1}$. We define matrices $A$ and $B$ in Theorem 1. CML$_2$E and CNLS produce estimators that are consistent with the true loadings, and are both asymptotically normal.

A small simulation demonstrates the impact of Theorem 2. We ran all simulations on a 1400 MHz PC with 128 MB of RAM, operated on Windows 2000. We generated random numbers from Matlab 6.1. We estimate profiles (loadings) with CML$_2$E and CNLS using a nonlinear constrained minimization procedure in Matlab called ‘constr.m.’ This m-file minimizes functions of parameters with equality and inequality constraints.

The scores are generated once from the lognormal(1,1), in other words from a mean one, variance one normal distribution followed by an exponential transformation. We use the profiles (loadings) shown below:

\[ P = \begin{pmatrix} 0.47 & 0 & 0 & 0.20 & 0.21 & 0.12 \\ 0.34 & 0.02 & 0.37 & 0 & 0 & 0.27 \\ 0.23 & 0.32 & 0.14 & 0 & 0.31 \end{pmatrix} \]

C1–C3 are the identifiability constraints applied to this loadings matrix. The number of factors is \( q = 3 \), and the number of variables is \( p = 6 \). The errors are generated from \( N(0,0.15^2) \). We assume that \( n = 200, c = 1, \Delta = 0.7 \) and \( K = 7 \). We repeat the simulation 200 times. The trace of the mean squared error (MSE) of \( \hat{\theta} \) is simply \( \text{tr}\{\text{MSE}(\hat{\theta})\}n = \text{tr}\{E_{\theta_0}[((\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^\top)]\}n \). The CNLS results in a simulated and asymptotic \( \text{tr}\{\text{MSE}(\hat{\theta}_C)\}n \) of 0.2893 and 0.3046, respectively. The CML2E has a simulated and asymptotic \( \text{tr}\{\text{MSE}(\hat{\theta}_E)\}n \) of 0.0333 and 0.0329, respectively. This corresponds to an AI 0.1151 in the simulated case and an AI of 0.1080 in the theoretical case.

The predetermined value \( c = 1 \) produced a better estimate of the loadings using CML2E relative to CNLS, but this choice is not necessarily optimal for all values of the tuning parameter. Before estimation of the source profiles, the value \( c \) is usually not given; therefore we discuss in the next Section a strategy for choosing \( c \) in practice. This strategy is justified by the asymptotic theorems and a more extensive simulation study.

4. DISCUSSION AND SIMULATIONS

4.1. Choosing the tuning parameter \( c \)

This section illustrates how the theorems are used to select the tuning parameter \( c \). Figure 1 represents the asymptotic inefficiency (AI) versus the tuning parameter \( c \), where \( c \) is in multiples of \( \sigma^2 \). The AI is the trace of the MSE of the CML2E estimator divided by the trace of the MSE of the CNLS estimator. As \( c \) tends to infinity, the AI reduces down to one. Thus, we gain efficiency in the normal case but we show later that we lose robustness in the mixture case. Therefore, we must select an intermediate value of \( c \). If one chooses \( c = 10\sigma^2 \) then under normality the AI for 12 variables and 3 factors is 1.1676, and the AI for 6 variables and 3 factors is 1.0730. Also, by increasing the number of factors from \( q = 3 \) to \( q = 6 \), one decreases the AI. For example, the AI for 12 variables and 6 factors is 1.193, and the AI for 6 variables and 6 factors is 1.0286. We choose \( c = 10\sigma^2 \) because the AI is between the AI of the ‘6% trim mean’ (1.04) and the median (1.57) for the univariate location problem (Hampel et al., 1986, p. 29).

Using \( c = 10\text{Var}(\varepsilon) \), we now wish to investigate the efficiency under heavy tailed error using the range of \( \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \) given in Theorem 2. Note that \( A \) and \( B \) from Theorem 2 are positive definite, and thus \( 0 < \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) < \infty \). Also note that the AI from Theorem 2 is monotone with respect to \( \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \). This is advantageous because the knowledge of \( K_0, \alpha_0 \) and \( P_0 \) is not required at the endpoints. Case 1 refers to the largest endpoint, and case 2 corresponds to the lowest endpoint.

- Case 1 is \( \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \rightarrow \infty \). Therefore, \( AE \rightarrow 1/\{\Delta b_1 + (1 - \Delta)b_K\}^2 \).
- Case 2 is \( \text{tr}(B^{-1})/\text{tr}(B^{-1}AB^{-1}) \rightarrow 0 \). Therefore,

\[ AE \rightarrow \left\{ \frac{\{(\Delta a_1 + (1 - \Delta)a_K K^4)/(\{\Delta + (1 - \Delta)K^4\}\}^{-1}\}}{\{\Delta b_1 + (1 - \Delta)b_K\}^2} \right\} \]
The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \sigma^2 \) for FA, then the \( AI \) for 12 variables is 1.1676 and the \( AI \) for six variables is 1.0730. We choose \( c = 10 \sigma^2 \) as the value of our tuning parameter for FA.

4.2. Scale parameter (\( \sigma^2 \))

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.

The choice of \( c = 10 \text{Var} (\varepsilon_{ij}) \) is asymptotically justified by plotting the \( AI \) of the endpoints versus the tuning parameter, \( c \). If one chooses \( c = 10 \text{Var} (\varepsilon_{ij}) \), then the CML\(_2\)E has lower variance, shown in Figure 2, for all heavy tailed values for \( \Delta = 0.5 \) and \( 1.5 \leq K < \infty \). The parameter \( \Delta = 0.5 \) represents a 50% contamination in the case with \( q = 3 \) and \( p = 12 \). We construct a similar plot for other choices of \( p \) and \( q \). This is true for any feasible \( P_0 \), \( A \) and \( \sigma^2 \).

We expand the comparison of the two estimators under other distributions and for small sample sizes through simulations that are motivated from actual data. In practice, estimation of the scale parameter, \( \sigma^2 \), is needed before such a discussion is made.
The minimum volume ellipsoid (MVE) estimator has the highest possible asymptotic breakdown (50%). We estimate $\sigma^2$ with the first $q$ singular values of the estimated covariance matrix. We call this estimator MVEstd.

We now focus on our main issue of model correspondence.

4.3. Simulations

Receptor modeling provides a reasonable forum for contrasting the estimator from CNLS with the estimator from CML$_2$E in practice. Under-specification of the number of sources of pollution in receptor modeling is equivalent to under-specification of the number of factors in the FA model. Under-specification can occur in pollution data when there are weather changes in the environment around the receptor site. Many air sheds have a prevailing wind pattern that will cause the receptor site to read pollution data under these conditions. Typically collectors use a wind meter during data gathering. However, there are instances where the wind meter is stuck, or the meter does not reflect the true wind pattern. It is difficult for the analyst to group the data according to the prevailing wind patterns, causing the data to have more sources of pollution than selected for the model. In this case, we illustrate through simulation that CNLS yields poor estimates relative to CML$_2$E.

We motivate the simulation from an actual pollution analysis in the literature (Henry et al., 1997). Using these results, we first set up a simulation using the base model assumptions in order to further
illustrate Theorem 1. Following this initial case, the base model assumptions are perturbed to form the basis for the second simulation, which has a mix of two sets of data each with three source profiles. The final simulation illustrates the robustness of CML$_2$E relative to CNLS with the mix of profiles from three sources and a set of profiles from six sources, and demonstrates the under-specified number of sources.

4.3.1. Simulations for three source profiles

Henry et al. (1997) report estimation of loadings from pollution data collected at the Houston Ship Channel in Houston (Clinton Drive), TX. The data have 56 volatile organic compounds (species). The authors estimate the loadings for 12 ($p = 12$) scientifically important species using CNLS on 183 hourly averaged data corresponding to southern winds. The location of the receptor site relative to pollution sources provides the reason for choosing such directions.

Our simulations are motivated from this data set and analysis. The data provides a comparison between CNLS and CML$_2$E under the mixture of two sets of pollution sources, where each set has three sources, generated from the profiles $P_0$ (the main profiles) and $P_2$ (the minor profiles), with a distribution form M2. Park et al.’s (2002) analysis provides the main profiles as well as motivation for generation of the scores (A), which we generate once from the lognormal$(3.6,0.5^2)$ distribution. We determine the minor matrix of profiles ($P_2$) to be close to orthogonal to the main matrix of profiles ($P_0$), in vector form, in order to provide vastly different pollution data. To illustrate the mixture of two sets of pollution data each with $q = 3$ factors and $p = 12$ variables, multivariate data are generated 100 times from $Y_i \sim \Delta N(\alpha_iP_0, \sigma^2 I) + (1 - \Delta)N(\alpha_iP_2, \sigma^2 I)$. The levels of the simulation are $1 - \Delta = 0.0$ and 0.1, as well as $n = 20, 50, 100, 200, 500$ and 1000. The variance is $\sigma^2 = 0.25^2$.

We study the mean squared error in detail. For each simulation we define $z_s = \sum_{k=1}^{r} (\hat{\theta}_{ks} - \theta_{ks})^2$, where $\theta_{00}$ is the $k$th element of the feasible parameters from the main profiles, and $\hat{\theta}_{ks}$ is the estimate of $\theta_{00}$ from the $s$th simulation. Therefore, an estimate of the trace of the mean squared error would be $E(z_s) = \text{tr}(\text{MSE} (\hat{\theta})) = \sum_{i=1}^{nreps} \sum_{k=1}^{r} (\hat{\theta}_{ks} - \theta_{00})^2 / nreps$, and the bias is estimated with simulation with $E(x_s) = \sum_{i=1}^{nreps} \text{Bias}(\hat{\theta}) = \sum_{i=1}^{nreps} \sum_{k=1}^{r} (\hat{\theta}_{ks} - \theta_{00}) / nreps$, where $nreps$ is the number of replications (100 unless otherwise noted).

Figure 3 provides a summary of the MSE and the bias for the estimators with normally distributed errors (1-$\Delta = 0.0$). We define the tuning parameter, $c$, and set it to $10^6\sigma^2$ before each iteration of the simulation. An exact $c = 100\text{Var}(\varepsilon_{ij})$ is not possible since $\text{Var}(\varepsilon_{ij})$ is estimated before each simulation using MADstd. In relation to the true variance, the tuning parameter varied from 5.66$c^2$ to 6.90$c^2$. This caused the asymptotic MSE to vary as the sample size increases (‘asymptotic MSE’ is the MSE derived from Theorem 1). Figure 3(a) provides the simulated MSE, plus or minus two times the simulation standard error, as a function of the sample size. The $\text{tr}(\text{MSE}(\hat{\theta}_C))$ is significantly less than $\text{tr}(\text{MSE}(\hat{\theta}_R))$ for all sample sizes ($n = 20, 50, 100, 200, 500$ and 1000). This result is consistent with the asymptotic theory. The asymptotic MSE for CNLS is not significantly different from its simulated calculation, $\text{tr}(\text{MSE}(\hat{\theta}_C))$, when $n = 100, 200, 500$ and 1000, and the asymptotic MSE for CML$_2$E is not significantly different from its simulated calculation, $\text{tr}(\text{MSE}(\hat{\theta}_R))$, when $n = 100, 200, 500$ and 1000. Figure 3(b) provides the simulated bias, plus or minus two times the simulation standard error, as a function of the sample size. There appears to be no significant bias.

Figure 4 provides a summary of the MSE and the bias for the estimators when $1 - \Delta = 0.1$. This case is a mixture of two normal distributions with different source profiles ($P_0$ and $P_2$). The tuning parameter, $c$, again was $10^6c^2$. Figure 4(a) provides the simulated MSE, plus or minus two times the simulation standard error, as a function of the sample size. The asymptotic MSE for CML$_2$E is relatively close to $\text{tr}(\text{MSE}(\hat{\theta}_R))$ for all sample sizes, but the asymptotic MSE for CNLS is
significantly different from $\text{tr}(\text{MSE}(\hat{\theta}_C))n$ for all sample sizes. The $\text{tr}(\text{MSE}(\hat{\theta}_R))n$ is significantly less than $\text{tr}(\text{MSE}(\hat{\theta}_C))n$ for all sample sizes. Figure 4(b) provides the simulated bias, plus or minus two times the simulation standard error, as a function of the sample size. This plot shows that the change in MSE from CNLS to CML$_2$E is due to the significant bias of $\hat{\theta}_C$. It also shows that $\hat{\theta}_R$ is approximately unbiased for all sample sizes.

Figure 3. The first plot provides the simulated MSE, plus or minus two times the simulation standard error, as a function of the sample size. The second plot provides the simulated bias, plus or minus two times the simulation standard error, as a function of the sample size. The '*'s represent the asymptotic values. $1 - \Delta = 0.0$

Figure 4. The first plot provides the simulated MSE, plus or minus two times the simulation standard error, as a function of the sample size. The second plot provides the simulated bias, plus or minus two times the simulation standard error, as a function of the sample size. The '*'s represent the asymptotic values. $1 - \Delta = 0.1$
4.3.2. Simulation of a mix with an under-specification of sources. Despite its clear strengths under the previous simulation experiment, our ultimate goal is to illustrate the CML2E’s correspondence for the case of an under-specified number of sources. The use of the mixture model forms the basis of a final simulation.

To illustrate the mixture of two sets of pollution data with one having \( q = 3 \) factors, the other having \( q = 6 \) factors, and both having \( p = 12 \) variables, multivariate data were generated 100 times from

\[
Y_i \sim \Delta N(\alpha_i P_0, \sigma^2) + (1 - \Delta)N(\alpha_i P_0 + 0.75 \beta_i P_2, \sigma^2)
\]

where the parameters are the same as the previous simulation except \( 1 - \Delta = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \) and \( 0.9 \), and \( n = 100 \). The variance is \( \sigma^2 = 0.5^2 \), and \( \beta_{ik} \sim \text{lognormal}(3.6, 0.5^2) \).

Thus, the user specifies three sources of pollution despite the partial presence of six sources. There are six sources in the mix because \( P_0 \) is 3 by 12 and \( P_2 \) is 3 by 12, resulting in the source profile \( [P_0/P_2] \) being 6 by 12.

In this simulation we consider a representative methodology to compare from the literature. A natural choice is the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD). We report a comparison of MVE and MCD to CML2E and CNLS, and display with simulation. However, we need to address the conversion from the robust covariance matrix to that of the source profiles.

### Table 2. Measures of estimated profiles from MVE and MCD for simulated data in the FA case. The data are a function of a mixture of two source profiles determined by the mixture parameter \( 1 - \Delta \). All measures are measured by \( 10^4 \), except Corr and \%BP

<table>
<thead>
<tr>
<th>( 1 - \Delta )</th>
<th>aveBias</th>
<th>aveMSE</th>
<th>Max</th>
<th>Corr</th>
<th>Med</th>
<th>%BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>-0.20</td>
<td>0.21</td>
<td>110.09</td>
<td>0.9990</td>
<td>0.08</td>
<td>9</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.25</td>
<td>0.20</td>
<td>109.40</td>
<td>0.9933</td>
<td>0.07</td>
<td>8</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.71</td>
<td>0.24</td>
<td>118.17</td>
<td>0.9980</td>
<td>0.09</td>
<td>15</td>
</tr>
<tr>
<td>0.3</td>
<td>-1.80</td>
<td>0.40</td>
<td>149.17</td>
<td>0.9945</td>
<td>0.13</td>
<td>24</td>
</tr>
<tr>
<td>0.4</td>
<td>-4.03</td>
<td>1.15</td>
<td>258.83</td>
<td>0.9887</td>
<td>0.25</td>
<td>55</td>
</tr>
<tr>
<td>0.5</td>
<td>-5.23</td>
<td>2.54</td>
<td>401.29</td>
<td>0.9795</td>
<td>0.51</td>
<td>82</td>
</tr>
<tr>
<td>0.6</td>
<td>-5.45</td>
<td>4.28</td>
<td>517.09</td>
<td>0.9828</td>
<td>0.77</td>
<td>90</td>
</tr>
<tr>
<td>0.7</td>
<td>-4.55</td>
<td>8.02</td>
<td>713.67</td>
<td>0.9616</td>
<td>1.27</td>
<td>97</td>
</tr>
<tr>
<td>0.8</td>
<td>-3.00</td>
<td>11.53</td>
<td>880.91</td>
<td>0.9399</td>
<td>1.65</td>
<td>99</td>
</tr>
<tr>
<td>0.9</td>
<td>-1.06</td>
<td>15.09</td>
<td>1009.42</td>
<td>0.9641</td>
<td>2.33</td>
<td>100</td>
</tr>
<tr>
<td>MCD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>-0.44</td>
<td>0.25</td>
<td>294.97</td>
<td>0.9992</td>
<td>0.09</td>
<td>11</td>
</tr>
<tr>
<td>0.1</td>
<td>-0.06</td>
<td>0.25</td>
<td>275.14</td>
<td>0.9991</td>
<td>0.08</td>
<td>14</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.41</td>
<td>0.22</td>
<td>248.63</td>
<td>0.9992</td>
<td>0.08</td>
<td>14</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.68</td>
<td>0.20</td>
<td>230.24</td>
<td>0.9994</td>
<td>0.07</td>
<td>9</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.26</td>
<td>0.20</td>
<td>247.91</td>
<td>0.9993</td>
<td>0.06</td>
<td>12</td>
</tr>
<tr>
<td>0.5</td>
<td>-2.50</td>
<td>0.41</td>
<td>447.68</td>
<td>0.9980</td>
<td>0.10</td>
<td>24</td>
</tr>
<tr>
<td>0.6</td>
<td>-5.25</td>
<td>1.49</td>
<td>1312.44</td>
<td>0.9787</td>
<td>0.19</td>
<td>64</td>
</tr>
<tr>
<td>0.7</td>
<td>-6.20</td>
<td>5.15</td>
<td>1399.88</td>
<td>0.9691</td>
<td>0.53</td>
<td>84</td>
</tr>
<tr>
<td>0.8</td>
<td>-6.30</td>
<td>12.44</td>
<td>1656.45</td>
<td>0.9509</td>
<td>1.35</td>
<td>100</td>
</tr>
<tr>
<td>0.9</td>
<td>-5.80</td>
<td>16.15</td>
<td>1865.00</td>
<td>0.9558</td>
<td>2.05</td>
<td>100</td>
</tr>
</tbody>
</table>

We address two choices for estimating the source profiles given the robust covariance matrix. Both rely on the fact that $S_R \xrightarrow{P} \sigma^2 I + P'K_0P$ when $Y_i \sim (\alpha_iP_0, \sigma^2 I)$, $n^{-1} \sum_{i=1}^n (\alpha_i - \bar{\alpha}) \xrightarrow{\sigma^2} K_0$. The first approach is to do the estimation by minimizing the objective function $0.5tr((S_R - \sigma^2 I - P'K_0P)^2)$ under the proper constraints, where $S_R$ is a robust estimate of the covariance matrix. We find this approach to be unstable in our simulations; consequently we chose a more stable alternative approach based on the following argument. As long as the variability in the contributions, $K_0$, is large relative to $1/(1 - \sigma^2)$, then

$$(I - (\sigma^2 I + P'K_0P))^{-1} = (I - P'(K_0^{-1}(1 - \sigma^2) + (PP')^{-1})P)/(1 - \sigma^2)$$

In our application this assumption is valid. Based on the above approximation, a natural estimator for the projection matrix is $(I - S_R)^{-1}(1 - \sigma^2)$. Given the proper equality constraints, we obtain unique source profiles. Specifically, let $M$ be the matrix of equality constraints, and $g$ be the row vector. Then $MP^v = g$ and $(1 - \sigma^2)I_q \otimes (I - S_R)^{-1}P^v = 0$ or

$$\left(\frac{M}{(1 - \sigma^2)I_q \otimes (I - S_R)^{-1}}\right)P^v = 0$$

Using the steps provided above, we then estimate the source profiles from the robust covariance matrix using MVE or MCD. We use a computational intensive approximate Monte Carlo approach to both in practice (Rousseeuw and van Zomeren, 1990). We use a freeware program named R that calls the original program in Fortran, to calculate the MVE and MCD.

We extract the simulation formed in MATLAB and calculate MVE for each iteration as a function of the contamination parameter, $1 - \Delta$, using MVEstd to estimate scale. We compare the source profiles from the MVE (MCD) to CML$_2$E and CNLS using six different measures of the simulation results. The measures are the average bias (aveBias), mean squared error (MSE), maximum of absolute difference (Max) and the median squared error (Med), averaged over all feasible parameters. The final two measures Corr (minimum correlation coefficient between true profiles and estimated profiles) and %BP (percentage of bad points) use all of the parameters associated with the 100 simulation replications.

We summarize the comparison between CML$_2$E, CNLS and MVE (MCD) in Tables 1 and 2 and represent it graphically in Figure 5. We use MVEstd as input for $c = 10\sigma^2$ into CML$_2$E.

The normal theory measures are the aveBias and the MSE. To demonstrate the theory derived in the last section, we simulated the CML$_2$E for 1000 replications in the $1 - \Delta = 0.0$ case. The simulated MSE and asymptotic MSE were 0.0895 and 0.0819, respectively (values multiplied by $10^4$). Additionally, the zero contaminated $(1 - \Delta = 0.0)$ simulation reflects that the CNLS has a lower MSE than that of CML$_2$E, MVE and MCD. This inequality reverses under the most modest of mixture $(1 - \Delta = 0.1)$, where the CML$_2$E has a lower MSE than MVE and MCD, which is below CNLS. This trend continues until $1 - \Delta = 0.9$, where MVE is actually best. The max measure behaves qualitatively similar to the MSE but the difference is sharper; it spikes to above 0.1 when $1 - \Delta = 0.4$ in the CNLS case, but is around 0.01 for CML$_2$E, and 0.01 to 0.06 for MVE and MCD. This trend continues until they all perform poorly at $1 - \Delta = 0.9$. The max measure favors CML$_2$E for moderate contamination.
The Corr measures the worst-case profile as compared to the true profile because it is the minimum correlation. It has similar patterns to that of the max and MSE measures.

The final measure used to compare estimators is %BP. A bad point occurs when that estimated profile is on the boundary of the parameter space, specifically when at least one feasible parameter is either a 0 or a 1. The %BP measure is either best or no worse for the CML\_2E relative to all estimators.

A comparison of MCD to MVE shows a consistency with the literature. For all intermediate values of contamination MCD outperforms MVE in terms of all measures except Max. For the Max measure there is a reversal in the two comparisons. This states that once in a while MCD can have some problems but still outperforms MVE in an average sense.

To summarize, the six measures equally favor the MVE, MCD and CML\_2E over the CNLS for small contamination. For moderate contamination, the CML\_2E is favored over the MVE and MCD. For close to 100% contamination (1 – \(\Delta\) = 1) none of the estimators performs well.

### 4.4. Application

By mixing the 183 Clinton Drive observations presented in the previous subsection, we illustrate our methodology using actual data. These data correspond to southern winds (180–190°). We choose to
mix these data with 75 observations that correspond to southeast winds (110–120°). We fit these mixed data using CNLS and CML₂E \(c = 10\text{Var}(\varepsilon_i)\). The correlation between the fit by Henry et al. (1997) (with the southern data only) and the CNLS (with the southern data and the contamination data) is 0.85, and the correlation with CML₂E is 0.97. We summarize the goodness of fit with principal component plots in Figure 6. These scatter plots represent the first three principal components of the normalized \(Y(o)\), which is superimposed with the estimated loadings from CNLS (*) and CML₂E (+). When the bulk of the data lie within the convex hull of the coordinates of an estimate, we observe a good fit. Notice that CNLS fails to cover a large bulk of data.

In addition, 50 repeats of ‘case resampling’ lead to an estimate of the average bias (as compared to the CNLS estimated profile in the southern direction) over the feasible parameters. This bias was \(-0.0400\) for the CNLS estimator and \(-0.0114\) for the CML₂E estimator. The average MSE was 0.0024 for CML₂E, as compared to 0.0346 for its CNLS counterpart. Finally, the maximum bias was 0.1222 for the CML₂E, compared to 0.5275 for the CNLS. This measure showed the largest discrepancy, and clearly demonstrates the weakness of the CNLS estimator in the mixed setting.
5. CONCLUSION

The main presentation in this article focuses on properties of an alternative to the CNLS estimator, which provide correspondence to the main source profiles (loadings) in receptor modeling or factor analysis, in a setting of mixed models. The alternative, CML$_2$E, is a function of a tuning parameter $c$.

In general, a reasonable tuning parameter is selected before estimation by solving for the equation

$$a = \left(1 + 2\sigma^2/c\right)^{(p-q)/2+1}$$

after the user chooses the desired efficiency ($a$) in the normal case. Asymptotic theory and simulation justify the choice of $c = 10\text{Var}(\varepsilon_{ij})$ when $p = 12$, $q = 3$ and the user chooses $a = 1.167$. By thoughtful selection of $c$, the CML$_2$E is a useful alternative to CNLS in constrained FA problems. We specifically address a particular kind of FA problem found in receptor modeling.

The United States Environmental Protection Agency (USEPA) calls for alternatives in the Photochemical Assessment Monitoring Stations (PAMS) workbook located at:
http://www.epa.gov/airprogm/oar/oaps/pams/analysis/receptor/factblrectxtsac.html#definition.

Throughout estimation using both CNLS and CML$_2$E, the user can make decisions regarding the quality and possible ‘main’ patterns in pollution data. However, since the CNLS is more efficient under normality, we do not completely dismiss it. Rather, we advise simultaneous estimation of both CNLS and the CML$_2$E by adjusting the tuning parameter to obtain variations of the estimator.

APPENDIX: PROOFS OF THEOREMS

Proof of Theorem 1

We prove the Theorem in two parts. We first prove that $\hat{P} \xrightarrow{p} P_0$ and then prove asymptotic normality. The proof of consistency and asymptotic normality are found in Park et al. (2002) when $\hat{P} = \hat{P}_C$.

Moment calculations of the CML$_2$E complete the proof. Recall that the true profile is $P_0$, where $P$ is an arbitrary profile:

$$Y_i/\sigma \sim N(\alpha_i P_0/\sigma, I)$$

Since $P = (I - P'(PP')^{-1}P)$ is idempotent,

$$Y_iP_i/\sigma^2 \sim \chi^2_{p-q, \lambda}$$

where the non-centrality parameter is $\lambda_i = \alpha_i P_0 P_0'/\sigma^2$. We use the moment generating function of the non-central chi-square distribution to assist us in calculating the moments for the CML$_2$E. When $z$ is a non-central chi-square random variable, the moment generating function is

$$E[\exp(iz)] = 1/(\sqrt{1-2t})^\theta \exp(2t\lambda/(1-2t))$$
Therefore the expectation of the CML$_2$E is
\[
E[-\exp(-Y_i'P_0/c)] = E[-\exp(tz)]|_{t=-\sigma/c} \\
= -1/(\sqrt{1-2t})^{d_f} \exp(2t\lambda/(1-2t))|_{t=-\sigma/c} \\
= -1/(1 + 2\sigma^2/c)^{(p-q)/2} \exp(-(1/c)\alpha_0P_0'\alpha_0'/ (1 + 2\sigma^2/c))
\] (10)

We now calculate the second moment of the CML$_2$E used for the variance, and apply the weak law of large numbers (WLLN). Again, we use the moment generating function of the non-central chi-square,
\[
E\left[-\exp(-Y_i'P_0/c)^2\right] = E[\exp(-Y_i'P_0/c)/(c/2)]\\
= E[\exp(tz)]|_{t=-2\sigma^2/c} \\
= 1/(\sqrt{1-2t})^{d_f} \exp(2t\lambda/(1-2t))|_{t=-2\sigma^2/c} \\
= 1/(1 + 4\sigma^2/c)^{(p-q)/2} \\
\times \exp(-(2/c)\alpha_0P_0'\alpha_0'/ (1 + 4\sigma^2/c))
\] (11)

These moment calculations help begin the consistency argument. Recall that $\theta_0$, a column vector, contains only the ‘free parameters of $P_0$’. Let $\hat{\theta}_n$ be the value that minimizes the CML$_2$E. We prove consistency of this value by contradiction. Suppose that the estimated profiles do not converge in probability to the true profiles; that is, suppose $\hat{\theta}_n \nrightarrow_{n \rightarrow \infty} \theta^* \neq \theta_0$. Using (10) and (11), Theorem assumptions 1–4, and since $0 < \exp(-t) \leq 1$, for $t \geq 0$, then
\[
\text{Var}\left[-n^{-1} \sum_{i=1}^{n} \exp(-Y_i'P_0'/c)\right] = n^{-2} \sum_{i=1}^{n} \text{Var}[\exp(-Y_i'P_0'/c)] \\
\leq n^{-1} \left[1/(1 + 4\sigma^2/c)^{(p-q)/2}\right] \\
\xrightarrow{n \rightarrow \infty} 0
\] (12)

Hence, by the WLLN,
\[
-n^{-1} \sum_{i=1}^{n} \exp(-Y_i'P_0'/c) + n^{-1} \sum_{i=1}^{n} \exp(-\alpha_iP_0'\alpha_0'/(c(1 + 2\sigma^2/c)) \xrightarrow{p} 0
\] (13)

Using assumption 5 of the Theorem we obtain
\[
-n^{-1} \sum_{i=1}^{n} \exp(-Y_i'P_0'/c) \xrightarrow{n \rightarrow \infty} -\Psi(P, \alpha_0, K_0)/(1 + 2\sigma^2/c)^{(p-q)/2}
\] (14)
Since \((-t)\) is convex, we have
\[
0 < \exp\left( -n^{-1} \sum_{i=1}^{n} \alpha_i P_0 P_0' \alpha_i' / [c(1 + 2\sigma^2/c)] \right) \\
\leq n^{-1} \sum_{i=1}^{n} \exp\left( -\alpha_i P_0 P_0' \alpha_i' / [c(1 + 2\sigma^2/c)] \right) \leq 1
\]
(15)

Now take \(n\) to infinity, so that
\[
0 < \exp( -\text{tr}\{ (K_0 + \alpha_0' \alpha_0) P_0 P_0' \} / [c(1 + 2\sigma^2/c)] ) \leq \Psi(P, \alpha_0, K_0) \leq 1
\]
(16)

This is true by the base model assumptions 3 and 5. Since \(K_0\) is full rank, then \((K_0 + \alpha_0' \alpha_0) P_0\) span the row space of \(P_0\). The fact that \(P\) is unique implies that \((K_0 + \alpha_0' \alpha_0) P_0 P_0'\) is uniquely minimized when \(P = P_0\). This occurs when \(P = R P_0\), but recall that, under the conditions C1–C3, \(R = I\). When \(P = P_0\) it uniquely maximizes the term to the right of zero in (15) (Park et al., 2002). Since this is true for all \(n\), then by a sandwich of the inequality, \(\Psi(P, \alpha_0, K_0)\) is uniquely maximized when \(P = P_0\), where \(\Psi(P, \alpha_0, K_0) = 1\).

We now form a Taylor series expansion of the objective function about the limit of the above subsequence, \(\theta^*\). We obtain
\[
-m^{-1} \sum_{i=1}^{m} \exp( -Y_i P_n Y_i' / c ) = -m^{-1} \sum_{i=1}^{m} \exp( -Y_i P' Y_i' / c ) \\
+ \left[ m^{-1} \sum_{i=1}^{m} \exp( -Y_i P Y_i' / c ) Y_i D_i(P) Y_i' / c \right]_{\theta = \tilde{\theta}} \\
\times \left( \tilde{\theta}_m - \theta^* \right),
\]
(17)

where \(\tilde{\theta} = \tilde{\theta}_m \lambda + \theta^* (1 - \lambda)\) and \(D_i(P) = \partial P / \partial \theta_r\).

We next show that each element in the derivative row vector has a bound. That is, if \(| m^{-1} \sum_{i=1}^{m} \exp( -Y_i P Y_i' / c ) Y_i D_i(P) Y_i' / c | \leq W\) in probability, we will get the contradiction. By the triangle inequality, and the fact that quadratic forms are linear combinations of chi-squared distributions, when the original data are normally distributed, we get
\[
| m^{-1} \sum_{i=1}^{m} \exp( -Y_i P Y_i' / c ) Y_i D_i(P) Y_i' / c | \leq m^{-1} \sum_{i=1}^{m} \exp( -Y_i P Y_i' / c ) \sum_{j=1}^{p} \omega_j z_{ij} \\
\leq m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{p} \exp( -Y_i P Y_i' / c ) | \omega_j | z_{ij} \\
\leq m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{p} | \omega_j | z_{ij}
\]
(18)

where the \(\omega_j\)'s are proportional to the eigenvalues of \(D_i(P)\), and the \(z_{ij}\) are distributed chi-squared. Therefore, the double sum on the right hand side has a bound with probability one. Since \(\tilde{\theta}_m \stackrel{m \to \infty}{\longrightarrow} \theta^*\) this implies...
The right hand side is not the minimum; therefore, we have a contradiction. Thus $\hat{P}_R \overset{p}{\rightarrow} P_0$.

It is shown in Park et al. (2002) that consistently estimated profiles are in the interior of the parameter space, with probability approaching one as $n$ goes to infinity, if we assume $\theta_0$ is in the interior of the parameter space (which is assumed).

Since both estimators are consistent, we express asymptotic normality with a Taylor series argument (Serfling, 1980). We write the specific variances of the estimators in the following general terms. Suppose $\hat{\theta}_*$ is a consistent estimator for $\theta_0$ and is calculated by minimizing the objective function $F_*$. Therefore $\sqrt{n}(\hat{\theta}_* - \theta_0) \overset{d}{\rightarrow} N(0, \{B_*^{-1}A_*B_*^{-1} + B_*^{-1}\})$, where $\partial^2 F_*/(\partial \theta_j \partial \theta_f) \mid _{\theta = \theta_0} \overset{p}{\rightarrow} B_{*j}$. These explicit values are facilitated by the fact that $E_{\theta_0}[\partial^2 F_*/(\partial \theta_j \partial \theta_f)] \mid _{\theta = \theta_0} = \partial E_{\theta_0}[F_*/(\partial \theta_j)] \mid _{\theta = \theta_0}$ and (10). Many pages of linear algebra show that $B_C^{-1}A_CB_C^{-1} + B_C^{-1} = B_A^{-1}B_1^{-1} + B^{-1}$ and $B_R^{-1}A_RB_R^{-1} + B_R^{-1} = a(B_1^{-1}AB_1^{-1} + B^{-1})$, where $A$, $B$ and $a$ are mentioned in the Theorem statement. We refer the reader to Gajewski (2000) for the tedious and long calculations for $A$ and $B$.

Using the tools from Theorem 2, the proofs of Theorems 3 and 4 are straightforward. We present more of the details for Theorem 4, which is a generalization of the proof for Theorem 3.

**Proof of Theorem 2**

We easily extend the techniques from the previous theorem to the case of consistency for the source profiles under a simple mixture of normals. For example, suppose we have the simple mixture M1, which is

$$Y_i \sim \Delta N(\alpha_1P_0, \sigma^2) + (1 - \Delta)N(\alpha_0P_0, (K\sigma)^2)$$

Then, using the techniques from before,

$$-n^{-1} \sum_{i=1}^{n} \exp(-Y_iP_0^T/c) \overset{p}{\rightarrow}$$

$$-\Delta \Psi_1(P, \alpha_0, K_0)/(1 + 2\sigma^2/c)^{(p-q)/2} - (1 - \Delta)\Psi_2(P, \alpha_0, K_0)/(1 + 2(\sigma K)^2/c)^{(p-q)/2}$$

Both $\Psi_1(P, \alpha_0, K_0)$ and $\Psi_2(P, \alpha_0, K_0)$ have a maximum of one when $P = P_0$, and thus a unique maximum at $P = P_0$.

We argue for asymptotic normality using the same techniques as used in Theorem 1. Also, by assuming the mixture M1, we can calculate the asymptotic efficiency of the trace of the variance of the CML2E.

**REFERENCES**


SOURCE PROFILE ESTIMATION IN RECEPTOR MODELING

633


