Linear unmixing of multivariate observations

a structural model

Linear Unmixing of Multivariate Observations: A Structural Model

by

Marcel Wolbers and Werner Stahel

Research Report No. 106
August 2002

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Linear Unmixing of Multivariate Observations: A Structural Model

Marcel Wolbers and Werner Stahel

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

August 2002

Abstract

In many fields of science there are multivariate observations which are generated by a (physical) linear mixing process of contributions from different sources. If it is assumed that the composition of the sources is constant for different observations, these observations are, up to measurement error, non-negative linear combinations of a fixed set of so-called source profiles which characterize the sources. The goal of linear unmixing is to recover both the source profiles and the source activities (also called scores) from a multivariate dataset.

We present a new parametric mixing model which assumes a multivariate lognormal distribution for the scores. This model is proved to be identifiable. To calculate the MLE we propose the combination of two variants of the MCEM algorithm. The proposed model is applied to simulated datasets and to air pollution measurements from Zurich. In addition to the basic model we discuss several extensions.

Key Words. linear mixing model, source apportionment, latent variables, identifiability, MCEM algorithm

Heading: Linear Unmixing of Multivariate Observations: A Structural Model

1 Introduction

The problem of explaining multivariate observations as mixtures of certain sources occurs in many fields of science. To illustrate this type of problems it is best to look at a specific example first:

Example 1 *Six daily measurements of 13 VOC (volatile organic compounds) were automatically recorded at a monitoring station in Wallisellen, a suburb of Zurich from October 1996 until February 1997. After removal of missing values, the dataset consists of 749 measurements of the 13 compounds. This dataset is a subset of a larger dataset discussed in detail in Locher (1999).*

The measured data are believed to stem from emission sources such as exhaust from gasoline driven cars, evaporation of gasoline, solvents of chemicals used in production, etc. The following assumptions about the sources are a reasonable approximation to reality:

- *The number of different sources is substantially smaller than the number of recorded variables.*
- *An emission source emits the compounds in constant proportions. The vector of proportions which characterizes a source is called its profile.*
- *The contributions of one source to an observation can be obtained by a scalar multiplication of the profile vector with the activity of that source at the time of the observation. This implies that chemical reactions of the compounds in the air are slow compared to the transport time between emission and measurement site.*

The assumptions imply that the measured concentrations at a certain time are, up to measurement error, a sum of the different source profiles scaled with the activities of the corresponding sources at that time. Since little is known a priori about the composition of the source profiles, one is interested in simultaneously estimating the source profiles as well as the source activities.

In mathematical formulas, the model described in Example 1 is (up to measurement error)

$$\mathbf{X}_i \approx \mathbf{C}\mathbf{S}_i \quad (i = 1, \ldots, n). \tag{1}$$

In this notation, the m-vectors $\mathbf{X}_i$ model the observations, the $m \times p-$matrix $\mathbf{C}$ contains the profiles of the p sources as its column vectors and the p-vectors $\mathbf{S}_i$ are the source activities or scores. In order to get physically meaningful results, it is reasonable to require that the source profiles matrix $\mathbf{C}$ as well as the scores $\mathbf{S}_i$ are non-negative. If both $\mathbf{C}$ and the scores $\mathbf{S}_i$ are unknown or only partially known, model (1) together with the non-negativity constraints is called a *linear mixing model*. The task of estimating the unknown quantities $\mathbf{C}$ and $\mathbf{S}_i$ is called *linear unmixing*. Often, the number p of different sources is unknown and therefore has to be estimated as well.

Linear mixing models have been applied in many fields including air quality studies (as in Example 1 above), chromatography and spectroscopy, geology and hydrochemical studies of natural catchments. A nice interdisciplinary introduction to mixing models which also provides references to applications is Akerjord & Christophersen (1996). Coverage of

linear mixing models in statistical journals has been rather sparse. Notable exceptions are Bandeen-Roche (1994), Renner (1993) and Park, Guttorp & Henry (2001). Most of the available methodology was published in the more quantitative journals of the application areas.

A fundamental problem of linear mixing models is that they are usually *non-identifiable* unless there are additional constraints imposed: Let $\mathbf{T}$ be a regular $p \times p-$matrix. Setting $\mathbf{C}^* := \mathbf{CT}$ and $\mathbf{S}_i^* := \mathbf{T}^{-1}\mathbf{S}_i$ leads to an equivalent model, i.e. $\mathbf{CS}_i = \mathbf{C}^*\mathbf{S}_i^*$ if $\mathbf{C}^*$ and $\mathbf{S}_i^*$ also satisfy the non-negativity constraints.

One possibility to ensure identifiability is to make *distributional assumptions* about the scores, thus leading to a *structural model* in contrast to functional models which treat the scores as unknown (incidental) parameters. Bandeen-Roche (1994) gives general conditions on the distribution of the scores to ensure that the model (1) (without measurement error and with compositional data) is identifiable. We suggest a specific *parametric model* for the scores distribution, namely the multivariate lognormal distribution, and a lognormal multiplicative error. This model, which we call the *lognormal structural mixing model*, is introduced in more detail in Section 2. In Section 3 we prove identifiability of the proposed model (with measurement error) and review asymptotic results for the MLE obtained in the first authors PhD thesis, see Wolbers (2002). To actually calculate the MLE of the lognormal structural mixing model, two variants of the Monte Carlo EM algorithm are suggested in Section 4. They treat the scores and the measurement errors, respectively, as latent variables and are shown to nicely complement each other. A favorable feature of the lognormal mixing model is that it is easily extended: In Section 5 we show that it is straightforward to extend the model as well as the proposed EM algorithms to allow for covariates. We also discuss methods to deal with zero-measurements as well as a variant of the model for compositional data. The MLE and the proposed algorithm to compute it are shown to be successful both in a simulation (Section 6) and in an application to the VOC measurements of Example 1 (Section 7). We conclude by surveying possibilities for further research.

2 The structural lognormal mixing model

A natural choice for the scores distribution of the linear mixing model is the *multivariate lognormal distribution*. It assumes that the logarithm of the scores is multivariate normally distributed. More details about this distribution are given in Crow & Shimizu (1988). This distribution is quite flexible and a natural adaptation of the multivariate normal distribution to the non-negativity constraints.

The observations in linear mixing models are often concentrations. For such data, Tukey's idea of "first aid transformations" (Mosteller & Tukey 1977) would suggest taking logarithms. Since the transformed observations no longer satisfy a linear mixing model, the idea cannot be directly applied. Instead we assume the error to be *multiplicative and lognormally distributed* to obtain a realistic model.

Combining the assumptions for the scores and the error we arrive at the following model:

$$\mathbf{X}_i = \mathbf{CS}_i \circ \mathbf{E}_i \quad (i = 1, \ldots, n), \tag{2}$$

where $\circ$ stands for the elementwise product and we assume that both the scores and the

error are multivariate lognormally distributed: $\mathbf{S}_i \sim \Lambda(\boldsymbol{\mu}, \boldsymbol{\Psi})$ and $\mathbf{E}_i \sim \Lambda(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ is diagonal. The assumption of independence of different components of the error is common for linear mixing models. Without this constraint, the covariance structure of the error is confounded with the subspace spanned by the source profiles (on which the observations without measurement error lie). Finally, we assume that both scores and error are i.i.d. and independent of each other.

Model (2) *resembles the factor analysis model*. Indeed, this model is a natural adaptation of the factor analysis model to the non-negativity constraints and the fact that in linear mixing models there exists a meaningful coordinate origin. The structural mixing model has *two special features*: First, while there has been a considerable debate in psychology whether the latent factors postulated by factor analysis have any real existence, the linear mixing model is based on physical laws in most applications. Second, since it will be shown below that the structural mixing model is identifiable under some regularity conditions, this model can do without the rather arbitrary rotation step.

Two *other structural mixing models* have been suggested in the literature. Unlike our model, both of them model *compositional* data. Bandeen-Roche & Ruppert (1991) and Bandeen-Roche (1994) assume a Dirichlet distribution for the scores and the error. The Dirichlet class has some elegant mathematical properties. However, Dirichlet assumptions imply strong independence conditions for the scores and are considerably less flexible than the lognormal assumptions of our model. Aitchison (1987) provides an extensive comparison of the Dirichlet and the logistic normal distribution (an adaptation of the lognormal distribution to the compositional case, see Aitchison (1987)) which clearly favors the logistic normal distribution as a modeling tool. The second model was developed independently of our work by Billheimer (2001). His model assumes that the scores and the errors are logistic normally distributed. Therefore it is identical to a variant of our model adapted to the compositional case, see Section 5.3. However, contrary to our work, he treats the model in a Bayesian framework.

3 Identifiability and asymptotics

The source profiles are only identifiable up to a scaling factor as is easily seen. Thus we always assume that the columns of $\mathbf{C}$ are *standardized*, i.e. that they are constrained to sum to 1, and that they are linearly independent. In addition, we have a problem similar to what Redner & Walker (1984) named the "label switching problem" for the analysis of mixtures. Thus we need a slightly specialized definition of identifiabilty:

Definition 1 *The structural lognormal mixing model (2) is* identifiable, *if for each pair* $\boldsymbol{\theta}_1 = (\mathbf{C}_1, \boldsymbol{\mu}_1, \boldsymbol{\Psi}_1, \boldsymbol{\Sigma}_1)$ *and* $\boldsymbol{\theta}_2 = (\mathbf{C}_2, \boldsymbol{\mu}_2, \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}_2)$ *of parameters determining densities* $f_1(\mathbf{x}|\boldsymbol{\theta}_1)$ *and* $f_2(\mathbf{x}|\boldsymbol{\theta}_2)$ *one has* $f_1(\mathbf{x}|\boldsymbol{\theta}_1) = f_2(\mathbf{x}|\boldsymbol{\theta}_2)$ *for almost all* $\mathbf{x}$ *if and only if there exists a* $p \times p$ *permutation matrix* $\mathbf{P}$ *such that* $\mathbf{C}_2 = \mathbf{C}_1\mathbf{P}$, $\boldsymbol{\mu}_2 = \mathbf{P}^t\boldsymbol{\mu}_1$, $\boldsymbol{\Psi}_2 = \mathbf{P}^t\boldsymbol{\Psi}_1\mathbf{P}$, *and* $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1$.

It is easy to see that the model (2) *without measurement error* is identifiable: If we assume that $\boldsymbol{\Psi}$ is nondegenerate, then the columns of $\mathbf{C}$ are identified (up to permutations and scaling) as the vertices of the support of the distribution of $\mathbf{CS}$ (since the support of the distribution of $\mathbf{S}$ is $\mathbb{R}_+^p$, the support of $\mathbf{CS}$ is a cone). Moreover, $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ are identified

as mean vector and covariance matrix of the distribution of $\log(\mathbf{S}) = \log(\mathbf{C}^+(\mathbf{CS}))$ where $\mathbf{C}^+$ is any $p \times m$-matrix satisfying $\mathbf{C}^+\mathbf{C} = \mathbf{I}$.

If we allow for measurement error, additional regularity conditions are required:

Theorem 1 *Assume the following set of regularity conditions:*

1. *The columns of* $\mathbf{C}$ *are linearly independent and the entries of each of its columns sum to 1.*
2. *No row of* $\mathbf{C}$ *contains only zeros.*
3. *If any row of* $\mathbf{C}$ *is deleted, there remain two disjoint submatrices of rank* p*.*
4. $\boldsymbol{\Psi}$ *is non-singular.*

Then the structural lognormal model is identifiable.

The proof is given in the appendix.

Remark. The proof relies on the fact that the columns of $\mathbf{C}$ are identified as the vertices of the support of the distribution of $\mathbf{CS}$ which is a cone. However, if observations close to the facets of the cone are improbable under the true model, reconstruction of the cone becomes hard. In such cases theoretical identifiability still holds but for reasonable sample sizes one has to expect practical identifiability problems. This happens e.g. if the components of $\boldsymbol{\mu}$ are of different orders of magnitudes or if an eigenvalue of $\boldsymbol{\Psi}$ is small.

The conditions of Theorem 1 imply that $m > 2p$. In this case the problem is generically identifiable, i.e. it is identifiable except for a set of Lebesgue measure 0 of the parameter space. In practice it always seems advisable to try to model with a small p for parsimony reasons; $p < m/2$ as well as the conditions of Theorem 1 will usually be satisfied. In theory, the conditions of Theorem 1 are generally too strong: Indeed, it can easily be shown that for the case of one source ($p = 1$), the parameters are identifiable from the first two moments of the log-data if $\mathbf{C}$ contains no zeros and at least two variables are recorded. (Note that in this simple case, the log-data is multivariate normally distributed.)

For asymptotics as well as for the calculation of the MLE it is easier to work with the log-observations denoted by $\mathbf{Z}_i := \log(\mathbf{X}_i)$ and the log scores $\mathbf{V}_i := \log(\mathbf{S}_i)$. This leads to the following form of the density function:

$$f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\theta}) = \int_{\mathbb{R}^p} f_{\mathbf{Z}|\mathbf{V}=\mathbf{v}}(\mathbf{z}; \boldsymbol{\theta}) \cdot f_{\mathbf{V}}(\mathbf{v}; \boldsymbol{\theta}) d\mathbf{v} \tag{3}$$

where $f_{\mathbf{Z}|\mathbf{V}=\mathbf{v}}$ and $f_{\mathbf{V}}$ are the densities of $\mathcal{N}(\log(\mathbf{C}\exp(\mathbf{V})), \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi})$, respectively.

Standard methods (see e.g. van der Vaart (1998)) are used to prove *consistency and asymptotic normality* of the MLE of the structural lognormal mixing model. However, the likelihood cannot be written in closed form and thus the conditions are hard to check. Rigorous proofs of asymptotic properties can be found in Wolbers (2002). Here, we only summarize these results: Consistency of the MLE is proved under two weak assumptions: identifiability and the existence of a lower bound for the measurement error variances. Due to analytical complications, asymptotic normality (if the parameters are inner points of the parameter space) is only proved for two special cases: either $p \leq 2$ or $\boldsymbol{\Psi}$ is assumed to be diagonal.

4 Calculation of the MLE

A major problem in determining the MLE of the structural lognormal model (2) is that the likelihood is not available in closed form since the density (3) is an integral. Instead of trying to maximize Monte Carlo or other numerical approximations to the likelihood, we propose to use two variants of the *Monte Carlo EM algorithm* of Wei & Tanner (1990).

Rather than performing one hard optimization, the EM algorithm approaches the problem at hand by augmenting the observed data with latent (unobserved) data and carrying out a series of simpler maximizations, see e.g. Dempster, Laird & Rubin (1977). We present two variants of the Monte Carlo EM algorithm for the problem at hand which differ in what they regard as the latent data.

4.1 Monte Carlo EM algorithm – variant 1

In this section, we view the log-scores $\mathbf{v}_i$ as latent data, leading to the complete data $(\mathbf{z}_i, \mathbf{v}_i)$ (where $\mathbf{z}_i$ are the log-observations). For the complete data, the likelihood L^c is available in closed form:

$$\begin{aligned}\log L^c(\boldsymbol{\theta}; \mathbf{z}, \mathbf{v}) &= -\frac{m+p}{2}\log(2\pi) - \sum_{j=1}^{m}\log(\sigma_j) - \frac{1}{2}\log(\det(\boldsymbol{\Psi})) \\ &\quad -\frac{1}{2}\sum_{j=1}^{m}\Big[z^{(j)} - \log\Big(\sum_{k=1}^{p} c_{jk}\exp(v^{(k)})\Big)\Big]^2/\sigma_j^2 - \frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^t\boldsymbol{\Psi}^{-1}(\mathbf{v}-\boldsymbol{\mu})\end{aligned}$$

Each iteration of the EM algorithm consists of an E-step (expectation) and a M-step (maximization). Let the current parameters be $\hat{\boldsymbol{\theta}}^{(q)}$. Then the $(q+1)$th iteration has the following form:

E-step: Compute $Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(q)}, (\mathbf{z}_i)) = \sum_{i=1}^{n}\widetilde{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}, \mathbf{z}_i)$ where

$$\begin{aligned}\widetilde{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(q)}, \mathbf{z}_i) &= \mathbf{E}\left[\log L^c(\boldsymbol{\theta}; \mathbf{z}_i, \mathbf{v})\Big|\mathbf{z}_i; \hat{\boldsymbol{\theta}}^{(q)}\right] \\ &= \int_{\mathbb{R}^p}\log L^c(\boldsymbol{\theta}; \mathbf{z}_i, \mathbf{v}) f_{\mathbf{V}|\mathbf{Z}=\mathbf{z}_i}(\mathbf{v}; \hat{\boldsymbol{\theta}}^{(q)})d\mathbf{v}\end{aligned}$$

is the expected complete data log-likelihood and the expectation is with respect to the conditional distribution of $\mathbf{V}$ given $\mathbf{Z} = \mathbf{z}_i$ and the old parameters $\hat{\boldsymbol{\theta}}^{(q)}$.

M-step: Determine $\hat{\boldsymbol{\theta}}^{(q+1)} := \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(q)}, (\mathbf{z}_i))$.

As is shown in Dempster et al. (1977), an EM-iteration always increases the log-likelihood. However, convergence to the maximum likelihood estimator can be slow.

If the conditional expectation in the E-step cannot be written in closed form, as is the case here, the expectation can be estimated by Monte Carlo methods leading to the Monte Carlo EM (MCEM) algorithm, see e.g. Wei & Tanner (1990).

We propose to use a *multivariate t importance sampler* to approximate the E-step (see e.g. Evans & Swartz (2000) for details): We note that $f_{\mathbf{V}|\mathbf{Z}=\mathbf{z}_i}(\mathbf{v}; \hat{\boldsymbol{\theta}}^{(q)}) \propto f_{\mathbf{Z},\mathbf{V}}(\mathbf{z}_i, \mathbf{v}; \hat{\boldsymbol{\theta}}^{(q)})$ and expand $\mathbf{v} \mapsto \log f_{\mathbf{Z},\mathbf{V}}(\mathbf{z}_i, \mathbf{v}; \hat{\boldsymbol{\theta}}^{(q)})$ in a Taylor series around its maximum. This leads to the expectation (maximizer) $\boldsymbol{\xi}_i$ and the covariance matrix (minus the inverse of the Hessian at the maximizer) $\boldsymbol{\Phi}_i$ of an approximating distribution which we choose to be a

multivariate t-distribution with λ degrees of freedom. (By default we set $\lambda = 10$.) Generate $\mathbf{v}_{i\ell} \sim t_p(\lambda, \boldsymbol{\xi}_i, (1-2/\lambda)\boldsymbol{\Phi}_i)$ and preliminary weights $w_{i\ell} = f_{\mathbf{Z},\mathbf{V}}(\mathbf{z}_i, \mathbf{v}_{i\ell}; \hat{\boldsymbol{\theta}}^{(q)})/g(\mathbf{v}_{i\ell})$ where g is the density of the $t-$distribution generating the $\mathbf{v}_{i\ell}$. Since $\mathbf{v} \mapsto f_{\mathbf{V}|\mathbf{Z}=\mathbf{z}_i}(\mathbf{v}; \hat{\boldsymbol{\theta}}^{(q)})$ is only known up to a normalizing constant, the weights are standardized and then the importance sampling estimator

$$\hat{Q}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(q)}, \mathbf{z}_i) = \sum_{\ell=1}^{r} \tilde{w}_{i\ell} \log L^c(\boldsymbol{\theta}; \mathbf{z}_i, \mathbf{v}_{i\ell}) \tag{4}$$

is used, where we have set $\tilde{w}_{i\ell} := (\sum_{k=1}^{r} w_{ik})^{-1} w_{i\ell}$. Equation (4) avoids the calculation of the normalizing constant of $f_{\mathbf{V}|\mathbf{Z}=\mathbf{z}_i}(\mathbf{v}; \boldsymbol{\theta})$. In addition, this estimate often has a smaller mean squared error than the unbiased standard importance sampling estimator, see Liu (2001).

If we use the approximation (4), the M-step has an explicit form for some of the parameters:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{(q+1)} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{r} \tilde{w}_{i\ell} \mathbf{v}_{i\ell} \\
\hat{\boldsymbol{\Psi}}^{(q+1)} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{r} \tilde{w}_{i\ell} (\mathbf{v}_{i\ell} - \hat{\boldsymbol{\mu}}^{(q+1)})^t (\mathbf{v}_{i\ell} - \hat{\boldsymbol{\mu}}^{(q+1)})
\end{aligned}$$

Moreover, the rows $\hat{\mathbf{C}}_{j\cdot}^{(q+1)}$ of $\hat{\mathbf{C}}^{(q+1)}$ can be obtained by minimizing

$$\mathbf{c} \mapsto \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{r} \tilde{w}_{i\ell} \Big(z_i^{(j)} - \log \Big(\sum_{k=1}^{p} c_{jk} \exp(\mathbf{v}_{i\ell}^{(k)}) \Big) \Big)^2$$

subject to non-negativity constraints: $\hat{\mathbf{C}}_{j\cdot}^{(q+1)}$ is the minimizer and $\hat{\sigma}_j^{(q+1)}$ is the square root of the minimal function value. The standardization of $\hat{\mathbf{C}}^{(q+1)}$ was not taken into account in this minimization. Otherwise, the rows of $\hat{\mathbf{C}}^{(q+1)}$ could not be estimated independently. We prefer instead to switch to an equivalent standardized model after calculation of all parameters $\hat{\boldsymbol{\theta}}^{(q+1)}$, which can easily be done.

4.2 Monte Carlo EM algorithm – variant 2

Dempster et al. (1977) argue that the EM algorithm converges rapidly if the information loss due to incompleteness is small. Intuitively, the latent log-scores carry a lot of information about $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$, and the information loss about these parameters due to non-observing the scores will be large resulting in slow convergence of the algorithm. To complement the MCEM algorithm of Section 4.1, we thus developed another variant of the EM algorithm. This second variant treats the observations without measurement error or, equivalently, the measurement errors as latent variables. Since the observations without measurement error lie exactly on a p-dimensional subspace, this second variant keeps the subspace spanned by the current source profiles fixed and only aims at finding an improved estimate of the source profiles within this subspace. This variant can thus be used to complement variant 1 but not as a stand-alone algorithm.

More precisely: Let the current parameters be $\hat{\boldsymbol{\theta}}^{(q)}$. The $m \times p$ matrix $\mathbf{B}$ contains as its columns a basis of the linear subspace spanned by the current source profiles, i.e. we can set $\mathbf{B} := \hat{\mathbf{C}}^{(q)}$. For this variant we replace the parameter $\mathbf{C}$ by a $p \times p-$matrix $\mathbf{T}$ (non-singular with $\sum_k t_{k\ell} = 1$ for $1 \leq \ell \leq p$) such that $\mathbf{C} = \mathbf{BT}$. Thus, we use the parameter set $\tilde{\boldsymbol{\theta}} = (\mathbf{T}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ and set $\tilde{\boldsymbol{\theta}}^{(q)} = (\mathbf{I}, \hat{\boldsymbol{\mu}}^{(q)}, \hat{\boldsymbol{\Psi}}^{(q)}, \hat{\boldsymbol{\Sigma}}^{(q)})$. The *latent observations* $\mathbf{u}_i$ are the coordinates of the observations without measurement error $\hat{\mathbf{x}}_i$ expressed in the basis $\mathbf{B}$, i.e. $\hat{\mathbf{x}}_i = \mathbf{C}\mathbf{s}_i = \mathbf{B}\mathbf{u}_i$. The relation between the scores $\mathbf{s}_i$ and $\mathbf{u}_i$ is therefore given by $\mathbf{s}_i = \mathbf{T}^{-1}\mathbf{u}_i$. Thus, the complete data density is

$$f_{\mathbf{Z},\mathbf{F}}(\mathbf{z}, \mathbf{u}; \tilde{\boldsymbol{\theta}}, \mathbf{B}) = f_{\mathbf{Z}|\mathbf{S}=\mathbf{T}^{-1}\mathbf{u}}(\mathbf{z}; \tilde{\boldsymbol{\theta}}, \mathbf{B}) \cdot \frac{1}{|\det(\mathbf{T})|} f_{\mathbf{S}}(\mathbf{T}^{-1}\mathbf{u}; \tilde{\boldsymbol{\theta}}, \mathbf{B})$$

The *MC E-step* of the MCEM algorithm then takes the following form: Compute

$$\hat{Q}(\tilde{\boldsymbol{\theta}}; \tilde{\boldsymbol{\theta}}^{(q)}, \mathbf{B}, (\mathbf{z}_i)) = \sum_{i=1}^{n} \sum_{\ell=1}^{r} \tilde{w}_{i\ell} \log L^c(\tilde{\boldsymbol{\theta}}; \mathbf{z}_i, \mathbf{u}_{i\ell}, \mathbf{B})$$

where we again use importance sampling: The distribution of $\mathbf{u}_i$ given $\mathbf{z}_i, \tilde{\boldsymbol{\theta}}^{(q)}$ and $\mathbf{B}$ is the same as that of $\exp(\mathbf{v}_i)$ given $\mathbf{z}_i$ and $\hat{\boldsymbol{\theta}}^{(q)}$. Therefore, the weights $\tilde{w}_{i\ell}$ and the $\mathbf{v}_{i\ell}$ can be generated as described in Section 4.1, and $\mathbf{u}_{i\ell} = \exp(\mathbf{v}_{i\ell})$.

Explicitly, the complete data log-likelihood is

$$\begin{aligned}
\log L^c(\tilde{\boldsymbol{\theta}}; \mathbf{z}, \mathbf{u}, \mathbf{B}) &= -\frac{m+p}{2}\log(2\pi) - \sum_{j=1}^{m} \log(\sigma_j) \\
&\quad -\frac{1}{2}(\mathbf{z} - \log(\mathbf{B}\mathbf{u}))^t \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \log(\mathbf{B}\mathbf{u})) \\
&\quad -\log(|\det(\mathbf{T})|) - \sum_{k=1}^{p} \log(\mathbf{T}^{-1}\mathbf{u})^{(k)} - \frac{1}{2}\log(\det(\boldsymbol{\Psi})) \\
&\quad -\frac{1}{2}(\log(\mathbf{T}^{-1}\mathbf{u}) - \boldsymbol{\mu})^t \boldsymbol{\Psi}^{-1}(\log(\mathbf{T}^{-1}\mathbf{u}) - \boldsymbol{\mu})
\end{aligned}$$

as long as $\mathbf{T}^{-1}\mathbf{u} > 0$ and $-\infty$ otherwise.

The *M-step* computes $\tilde{\boldsymbol{\theta}}^{(q+1)} = \arg\max_{\tilde{\boldsymbol{\theta}}} \hat{Q}(\tilde{\boldsymbol{\theta}}; \tilde{\boldsymbol{\theta}}^{(q)}, \mathbf{B}, (\mathbf{z}_i))$. The error standard deviations are easy to update:

$$\hat{\sigma}_j^{(q+1)} = \Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{r} \tilde{w}_{i\ell}(z_i^{(j)} - \log(\mathbf{B}\mathbf{u}_{i\ell})^{(j)})^2\Big)^{1/2}.$$

Moreover, straightforward calculations show that

$$\begin{aligned}
\hat{\mathbf{T}}^{(q+1)} &= \arg\max_{\mathbf{T}} \Big(-n\log(|\det(\mathbf{T})|) - n\sum_{k=1}^{p} [\hat{\boldsymbol{\mu}}^{(q+1)}]^{(k)} \\
&\quad -\frac{1}{2} n \log(\det(\hat{\boldsymbol{\Psi}}^{(q+1)}))\Big) \qquad (5)
\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{(q+1)} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{r}\tilde{w}_{i\ell}\log(\mathbf{T}^{-1}\mathbf{u}_{i\ell}) \\ \hat{\boldsymbol{\Psi}}^{(q+1)} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{r}\tilde{w}_{i\ell}(\log(\mathbf{T}^{-1}\mathbf{u}_{i\ell})-\hat{\boldsymbol{\mu}}^{(q+1)})(\log(\mathbf{T}^{-1}\mathbf{u}_{i\ell})-\hat{\boldsymbol{\mu}}^{(q+1)})^t\end{aligned}$$

Therefore, the only hard calculation in the M-step is the optimization (5).

If the E-step could be performed exactly, the cone spanned by the columns of $\hat{\mathbf{C}}^{(q)}$ would always be a subset of the cone spanned by the columns of $\hat{\mathbf{C}}^{(q+1)}$. To see this property, note that the support of the conditional distribution of $\mathbf{u}$ given $\mathbf{z}_i$ and the old parameters ($\hat{\mathbf{T}}^{(q)}=\mathbf{I}, \mathbf{B}=\hat{\mathbf{C}}^{(q)},\ldots$) is $\mathbb{R}^p_+$. If $\hat{\mathbf{C}}^{(q+1)}=\mathbf{B}\hat{\mathbf{T}}^{(q+1)}$ did not possess the claimed property, $(\hat{\mathbf{T}}^{(q+1)})^{-1}\mathbf{u}$ would contain negative entries for some $\mathbf{u}\in\mathbb{R}^p_+$ thus leading to $Q(\tilde{\boldsymbol{\theta}}^{(q+1)};\tilde{\boldsymbol{\theta}}^{(q)},\mathbf{B},(\mathbf{z}_i))=-\infty$. The MCEM-iteration only requires that $(\hat{\mathbf{T}}^{(q+1)})^{-1}\mathbf{u}_{i\ell}>0$ for all simulated $\mathbf{u}_{i\ell}$.

4.3 Combining the two variants

Here we discuss how the two variants of the MCEM algorithm may be combined and how to determine the Monte Carlo sample size. For the choice of reasonable starting values we refer to Wolbers (2002).

Note that it is relatively cheap to calculate a Monte Carlo approximation of the log-likelihood, once an approximate E-step has been performed: The expectations $\boldsymbol{\xi}_i$ and variances $\boldsymbol{\Phi}_i$ of the approximating t-distributions (and possibly also the samples $\mathbf{v}_{i\ell}$) can be reused to get an approximation of the likelihood. Only the importance sampling weights have to be changed. To decrease the variance of estimated differences of log-likelihood values at different parameters it proved useful to base all log-likelihood calculations on the same standard multivariate t-sample of size $n\times r$. However, surprisingly to us, using the same standard t-sample for all MC E-steps performed disadvantageously in simulations (see Section 6) and is thus not recommended.

The *choice between the two EM-variants* is then performed as follows: We start with an EM-step variant 1 followed by an EM-step variant 2. At each iteration, the increase (or decrease) in the log-likelihood is recorded. In the following, we adaptively switch between the two variants depending on which one has lead to a larger increase of the log-likelihood when it was performed the last time.

Regarding the *Monte Carlo sample size*, simulations revealed that it is possible to start with a small r such as $r=10$, since already small sizes lead to quite reasonable fits in examples. The automated rule for increasing the MC sample size of MCEM algorithms in Booth & Hobert (1999), which is based on constructing confidence bands for the maximizer of each approximate E-step, could also be implemented here. We use a less sophisticated rule which was easier to implement: Since we already need to estimate the log-likelihood to choose between the two EM-variants, we increase r, if the approximation of the log-likelihood decreases for (say) two consecutive iterations.

5 Extensions and variants of the structural lognormal mixing model

5.1 Dealing with zero measurements

The structural lognormal mixing model (2) is clearly not adequate for modeling values of zero in observations. A simple amendment is to include a small non-negative vector $\boldsymbol{\tau}$ into the lognormal mixing model, leading to

$$\mathbf{X}_i + \boldsymbol{\tau} = (\mathbf{C}\mathbf{S}_i + \boldsymbol{\tau}) \circ \mathbf{E}_i \quad (i = 1, \ldots, n),$$

This model allows for zeros. It has the additional advantage of slightly increasing the measurement error variance to $\mathrm{Var}\,(\mathbf{X}_i|\mathbf{S}_i = \mathbf{s}_i) = (\mathbf{C}\mathbf{s}_i + \boldsymbol{\tau})^2 \circ \mathrm{Var}\,(\mathbf{E}_i)$ which is more realistic for small observations.

The algorithms of Section 4 can easily be enhanced to allow for an additional parameter $\boldsymbol{\tau}$. In principle, $\boldsymbol{\tau}$ could also be estimated, but our implementation only allows to include $\boldsymbol{\tau}$ as a fixed parameter into the model.

A more sophisticated alternative would be to introduce the idea of a *detection limit*: Zero measurements are obtained because the true value is below the detection limit of the measuring device. If the detection limit of a device is known, zero measurements could be modeled as censored observations.

5.2 The lognormal structural mixing model with covariates

Sometimes covariates for the scores are available. For example, for air pollution data it would be more realistic to let the scores depend on meteorological data and time. Model (2) can be modified to allow for covariates by assuming the log-scores to follow a multivariate linear regression model with q-dimensional covariates $\mathbf{u}_i$:

$$\log(\mathbf{S}_i) \sim \mathcal{N}\left(\boldsymbol{\mu}_i = \boldsymbol{\beta}^t \mathbf{u}_i, \boldsymbol{\Psi}\right)$$

where $\boldsymbol{\beta}$ is a $q \times p$-matrix. It is straightforward to genaralize the algorithms of Section 4 to allow for this extension. Moreover, if it is a priori clear that different components of the scores depend on different covariates identification of $\mathbf{C}$ will become easier.

5.3 A mixing model for compositional data

In some applications, the observations $\mathbf{x}_i$ are naturally in compositional form. In such cases, the logistic normal distribution described in Aitchison (1987) naturally substitutes the lognormal distribution as a modeling tool: A random vector $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(p)})$ with values in the strictly positive compositional simplex $\mathcal{S}_+^{(p-1)} = \{\mathbf{y} \in \mathbb{R}_+^p : \sum_{j=1}^p y^{(j)} = 1\}$ has a *logistic normal distribution* with parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\Phi}$, $\mathbf{Y} \sim \mathcal{L}(\boldsymbol{\zeta}_{p-1}, \boldsymbol{\Phi}_{(p-1)\times(p-1)})$, if $\mathbf{Z} := (\log(y^{(1)}/y^{(p)}), \ldots, \log(y^{(p-1)}/y^{(p)}))$ has a multivariate normal distribution with the same parameters, $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\zeta}, \boldsymbol{\Phi})$.

Following Aitchison (1987), we define a compositional constraining operator $\mathcal{C} : \mathbb{R}_+^m \to \mathcal{S}^{(m-1)}$ by $\mathcal{C}(\mathbf{x}) := \mathbf{x}/\sum_{j=1}^m x^{(j)}$ and a perturbation operator $\odot : \mathbb{R}_+^m \times \mathbb{R}_+^m \to \mathcal{S}^{(m-1)}$ by $\mathbf{x} \odot \mathbf{e} := \mathcal{C}(\mathbf{x} \circ \mathbf{e})$.

A variant of the structural lognormal mixing model is then given by the model

$$\tilde{\mathbf{X}}_i = \mathbf{C}\tilde{\mathbf{S}}_i \odot \tilde{\mathbf{E}}_i \quad (i = 1, \ldots, n), \tag{6}$$

where the assumptions are the same as for the structural lognormal mixing model (2), except that we assume that the scores and the errors are logistic normally distributed: $\tilde{\mathbf{S}}_i \sim \mathcal{L}(\boldsymbol{\zeta}_{p-1}, \boldsymbol{\Phi}_{(p-1)\times(p-1)})$ and $\tilde{\mathbf{E}}_i \sim \mathcal{L}(\mathbf{0}, \boldsymbol{\Sigma}'_{(m-1)\times(m-1)})$ where the components of $\tilde{\mathbf{E}}_i$ are "independent except for the summation constraint", i.e. $\boldsymbol{\Sigma}'$ is of the form

$$(\boldsymbol{\Sigma}')_{j\ell} = \begin{cases} \sigma_j^2 + \sigma_m^2 & \text{if } j = \ell \\ \sigma_m^2 & \text{if } j \neq \ell \end{cases} \tag{7}$$

(Equivalently, a lognormal error $\tilde{\mathbf{E}}_i \sim \Lambda(\mathbf{0}, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2))$ could be specified.) As has been mentioned, this model was independently developed and treated in a Bayesian context by Billheimer (2001).

Model (6) is not only analogous to the structural model (2), but *the two models are connected* by the following fact:

Lemma 1 *If* $\mathbf{X}$ *follows a structural lognormal model with parameters* $\boldsymbol{\theta} = (\mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ *then* $\tilde{\mathbf{X}} = \mathcal{C}(\mathbf{X})$ *follows a structural logistic normal model with parameters* $\tilde{\boldsymbol{\theta}} = (\mathbf{C}, \boldsymbol{\zeta} = \mathbf{F}\boldsymbol{\mu}, \boldsymbol{\Phi} = \mathbf{F}\boldsymbol{\Psi}\mathbf{F}^t, \boldsymbol{\Sigma}')$ *where* $\mathbf{F}$ *is as giben by* $\mathbf{F} = \begin{pmatrix} & & -1 \\ & \mathbf{I}_{p-1} & \vdots \\ & & -1 \end{pmatrix}$ *and* $\boldsymbol{\Sigma}'$ *is as in (7).*

Lemma 1 follows directly by using Aitchison (1987, Property 6.1, p. 117) and by noting that $\mathcal{C}(\mathbf{CS} \circ \mathbf{E}) = (\mathbf{C} \cdot \mathcal{C}(\mathbf{S})) \odot \mathcal{C}(\mathbf{E})$.

While they have not been implemented, it is clear that similar algorithms to those of Section 4 for the structural lognormal model could be developed to calculate the MLE of the logistic normal mixing model.

6 Simulation

We study the performance of the proposed algorithms to compute the MLE for an examples with $p = 3$ sources and $m = 10$ variables. The sample size is $n = 250$ and the log-scores have different variances and some strong correlations:

$$\mathbf{s}_i \sim \Lambda(\boldsymbol{\mu}, \boldsymbol{\Psi}) \text{ with } \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Psi} = \begin{pmatrix} 1 & 0.5 & 1 \\ 0.5 & 2 & 0 \\ 1 & 0 & 1.5 \end{pmatrix}.$$

The chosen source profiles matrix can be seen as lines in Figure 2 and the log-measurement error standard deviation was set to 0.1 for the first 5 variables and to 0.2 for the others corresponding to about 10% and 20% relative error, respectively.

6.1 Comparison of different combinations of the MCEM algorithms

Ten datasets were generated. Four different combinations of the MCEM algorithm variants introduced in Sections 4.1 and 4.2 were tried out: Automated switching between the two variants as described in Section 4.3, alternating between the two variants, using only variant 1 and finally using only variant 1 and using the same standard t-sample for all MC E-steps. For all MC E-steps, $r = 100$ MC samples per observation were used. All log-likelihood approximations were based on the same standard multivariate t-sample with $r = 1000$ MC-samples per observation for higher accuracy. In all 10 cases either automated switching (6 times) or alternating between the two variants lead to the highest log-likelihood. Using only MCEM variant 1 and the same $t-$sample performed worst in 8 cases.

Figure 1 displays 50 MCEM iterations for all four combinations and two datasets. (The log-likelihoods of the starting values and those after the first iteration are not shown since they are much lower than the others.) In the first case, all combinations perform comparably while in the second, the combinations involving only variant 1 perform significantly worse. In two of the 10 simulations, using only MCEM variant 1 and the same $t-$sample even lead to log-likelihood curves which started decreased slightly after a couple of successful iterations at the beginning, a pattern of overadaptation to the $t-$sample which was surprising to us. Based on these results, we choose the automated MCEM algorithm of Section 4.3 for the following investigations.

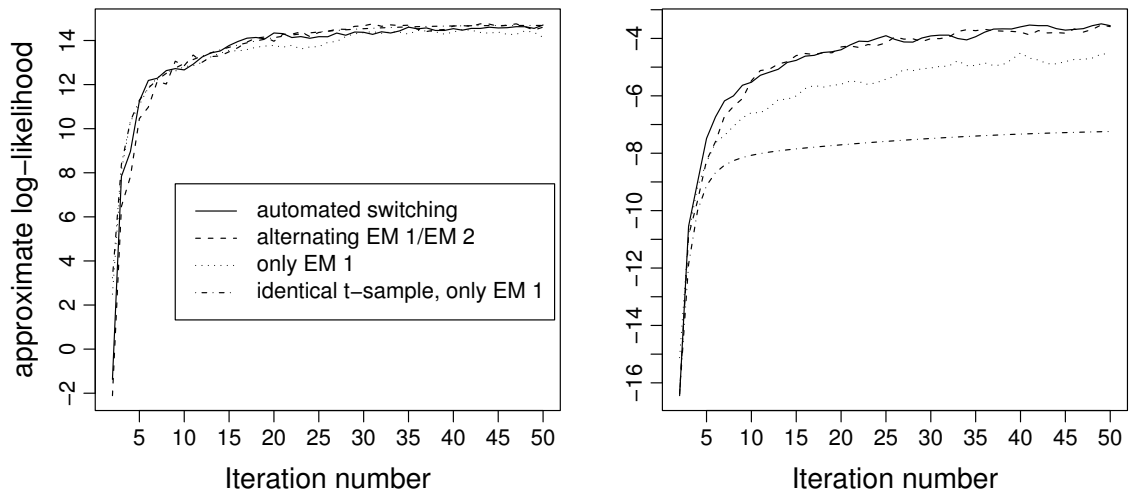

Figure 1: Log-likelihoods after iterations 2-50 of two simulations and different combinations of the MCEM algorithms.

6.2 Results for the automated MCEM algorithm

The automated MCEM algorithm of Section 4.3 was run for 100 simulated datasets. The number of MC samples per observation was successively increased along the sequence 10-40-160-640-1000. Finally, 10 alternating MCEM iterations (again with $r = 1000$) were

performed. Altogether, about 100 MCEM iterations were needed for an estimate. Computing time per run was approximately 110 minutes on a Linux computer with a Intel Xeon chip (1500 Mhz) which is quite long but partly due to the fact that the algorithm was completely programmed with the statistical software R (publicly available from CRAN (1997 ff.)) and not with a compilable language such as C.

Figure 2 displays parameter estimates for 100 simulated datasets. Source 2 is very accurately estimated while estimates of the sources 1 and 3 are less precise. This is also illustrated by Figure 3 which shows the true standardized scores (i.e. $\mathcal{C}(\mathbf{S}_i)$ in the notation of Section 5.3) as well as true and (projected) estimated sources for the first 4 datasets. For one dataset (plot on the lower left), source 1 is not very accurately estimated.

7 Application to the VOC measurements

The dataset explained in Example 1 contains a few zero measurements. Therefore the variant of the structural model which allows for zeros discussed in Section 5.1 was chosen and $\boldsymbol{\tau}$ was set to 0.07 ppbC for the sum of 2- and 3-methyl pentane and to 0.05 ppbC for all other compounds. (This corresponds to the magnitude of absolute measurement error as described in Locher (1999, p. 137).)

The automated MCEM algorithm described in Section 4.3 was run for models with $p = 2 \ldots 6$ sources. Monte Carlo sample size r was increased along the sequence 10-30-90-250. Finally, 25 MCEM steps (variant 1) were performed for $r = 500$ and $r = 1000$ each. As a rough guide for choosing p, the maximum log-likelihoods of these models are displayed in Figure 4. In order to see how much of the variability of the dataset is explained by the model it is also useful to look at the estimated log-error standard deviations for different p. They are also displayed in Figure 4. From these plots it is clear that $p = 2$ is too small. Three or possibly four sources seem appropriate. Fortunately, the corresponding source profiles for different p agree quite well. Thus, the choice of p is somewhat less crucial.

We decided to use $p = 3$ for closer examination. For $p = 4, 5, 6$, the estimated log-scores covariance matrices $\hat{\boldsymbol{\Psi}}$ become closer to singular: For $p = 4$, the correlation between log-scores components 1 and 4 is 0.93 and the condition number of $\hat{\boldsymbol{\Psi}}$ is 100.66. For $p = 5$ and 6, the condition numbers of the estimated $\hat{\boldsymbol{\Psi}}$ are 700 and 1122, respectively.

We chose a resampling procedure to assess the *uncertainty of the parameter estimates*. Because both the scores and the residuals show significant temporal dependence, the blockwise bootstrap was chosen, see Künsch (1989). Each bootstrap sample consisted of 62 blocks of length 12 (corresponding to two days each) and one block of length 5, sampled with replacement from the dataset. Figure 5 shows boxplots of parameter estimates obtained from 100 bootstrap samples. Lines indicate estimates from the original dataset. It can be seen that the estimates are quite precise.

In addition to the measured VOC, the dataset from Wallisellen contains additional covariates which have not been included into the analysis above: Temperature, relative humidity, air pressure, wind speed and direction. A simple regression model for the estimated scores of the form

$$\log(\hat{s}^{(k)}) \sim \texttt{hour*weekend} + \texttt{temp} + \texttt{rel.hum.} + \texttt{airpress} + \texttt{windspeed}$$

(where `hour*weekend` is an abbriviation for main effects plus an interaction of the factors hour of measurement (6 levels) and an indicator of whether the day of measurement was

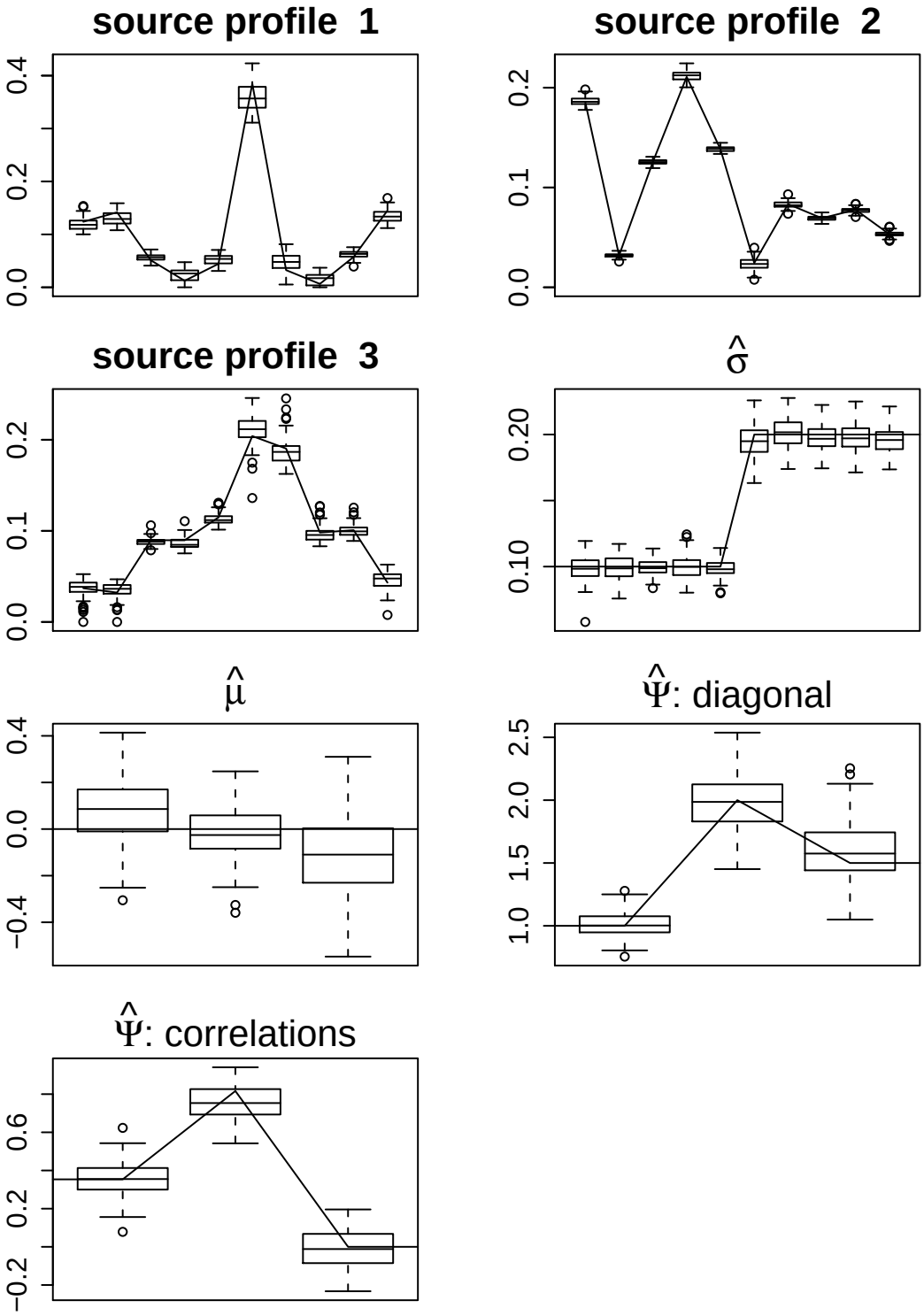

Figure 2: Results of the automated MCEM algorithm for the simulation of Section 6. Boxplots are based on 100 simulated datasets. Lines indicate true values.

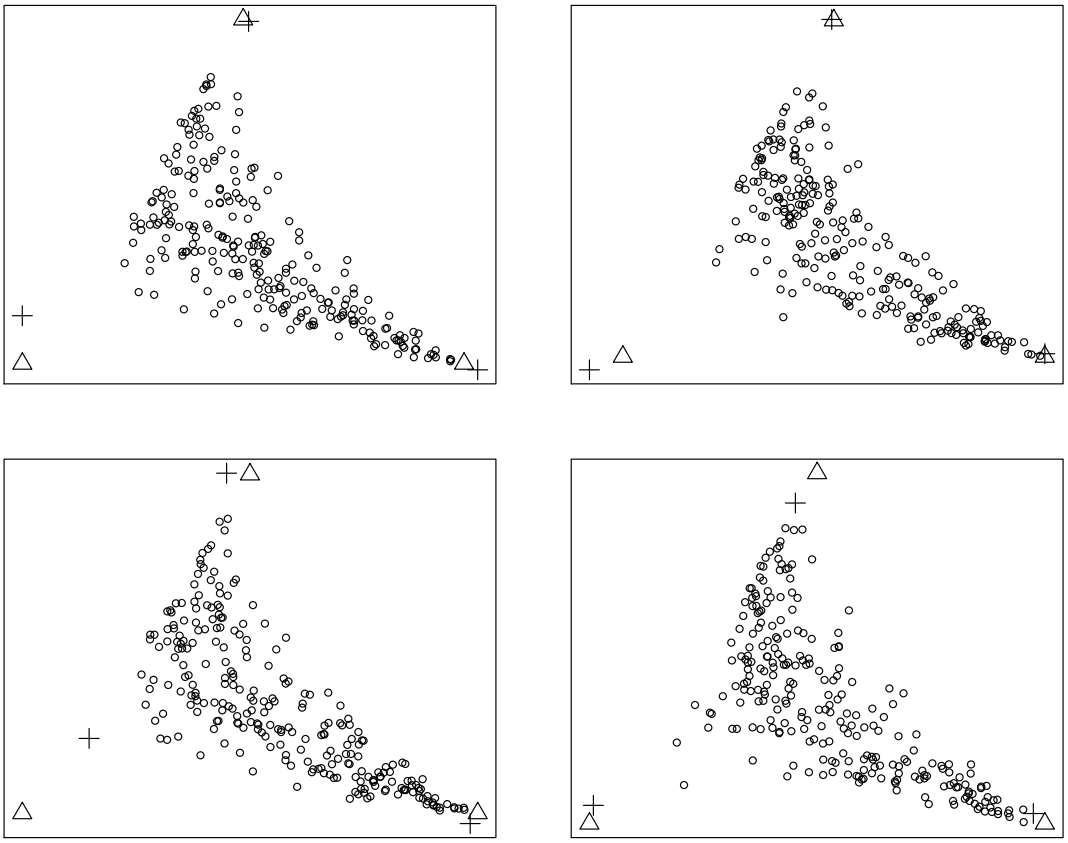

Figure 3: True standardized scores as well as true (△) and (projected) estimated sources (+) for the first 4 simulations of Simulation 6.

a weekday) explains about 55% of the variation of scores components 1 and 2 and 15% of component 3. (An additional term `winddirection*windspeed`, where `winddirection` is a factor with 8 levels, improved the fit only slightly and was therefore not included.) The MLE of the *structural lognormal mixing models with covariates* described in Section 5.2 was determined for the regression model described above and $p = 2 \ldots 4$. In all three cases, the estimated source profiles were very close to those of the model without covariates and thus confirmed the profiles obtained in the simpler analysis.

8 Conclusions

The *structural lognormal mixing model* is a natural adaptation of the factor analysis model to the non-negativity constraints and the fact that in linear mixing models there exists a meaningful coordinate origin. Since it models the scores by a parametric distribution, it is amenable to standard statistical theory and desirable statistical properties such as identifiability and consistency of the MLE can be proved. The assumed multivariate lognormal distribution of the scores is a flexible parametric model. Nevertheless, a major

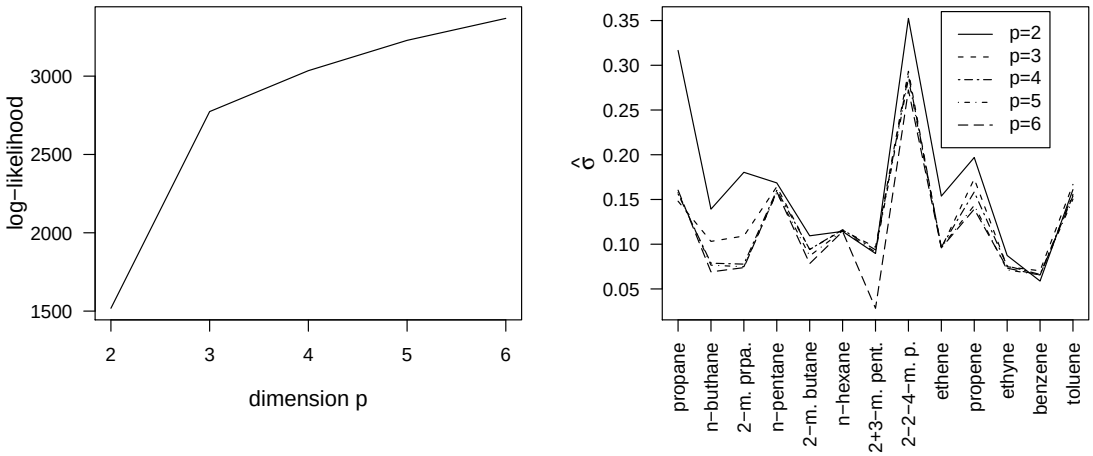

Figure 4: Maximum log-likelihoods and estimated log-error standard deviations for the Wallisellen dataset and different dimensions p.

disadvantage of the structural lognormal model is that the assumption of i.i.d. lognormal scores is an oversimplification in many applications. However, we believe that the structural lognormal model is a fine starting point for yet more realistic models which go well beyond the relatively simple assumption of i.i.d. lognormal scores. Here is a list of possible extensions which provide opportunities for further research:

1. It is conceptually simple to allow the scores to depend on *covariates*. This extension has already been discussed in Section 5.2.

2. Observations below the *detection limit* of the measuring device require special treatment. As has been mentioned in Section 5.1, it would be fruitful to extend the structural model to allow for left-censored observations.

3. In some applications, both the scores and the errors should be modeled as *multivariate time series*. A first publication which allows for time dependence in linear mixing models is Park et al. (2001). However, these authors do not include non-negativity constraints of the scores and identifiability of the model is assured by pre-specified zeros in the source profiles matrix. Neither does their model allow for covariates. It would thus be useful to combine the ideas in Park et al. (2001) with the structural lognormal model with covariates.

4. Simulations in Wolbers (2002) show that the MLE of the structural lognormal mixing model may perform badly if the data contains gross errors. It would thus be desirable to robustify the MLE. A first step in this direction which provides some degree of robustness to outliers and which should be relatively easy to implement would be to model the log-errors with a t-distribution and to calculate the MLE in this new model.

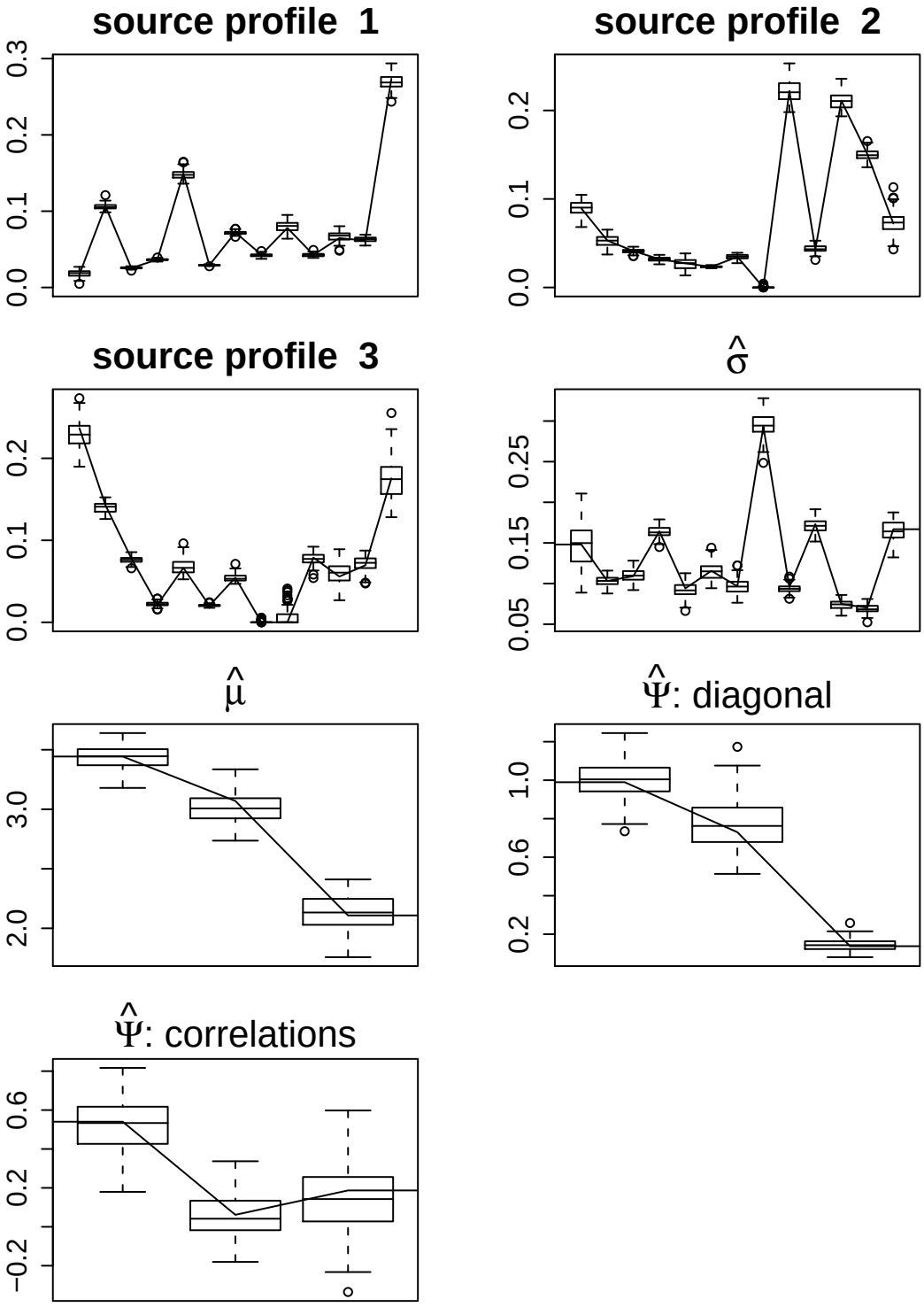

Figure 5: Parameter estimates for 100 bootstrap samples of the Wallisellen dataset. Lines indicate estimates from the original dataset.

5. It is straightforward to include simple constraints on the parameter values into the structural lognormal model. However, in some applications more complex *prior knowledge* is available. This calls for Bayesian or penalized likelihood methods.

Fortunately, it is relatively simple to write down models which extend the structural lognormal mixing model with the features discussed above. However, a detailed examination of these models would be both worthwhile and challenging. One particular challenge is to design reliable and fast algorithms for computing the MLE or the posterior distribution of these extended models.

A Appendix: Proof of Theorem 1

Our method of proof uses a known result concerning uniqueness of a special matrix decomposition.

Theorem 2 *Let* $\mathbf{M}$ *be a positive definite* $m \times m$ *matrix. Suppose there exists a decomposition of* $\mathbf{M}$ *of the form*

$$\mathbf{M} = \mathbf{N}\mathbf{N}^t + \mathbf{D}$$

where $\mathbf{N}$ *is a* $m \times p$ *matrix of rank* $p < m$ *and* $\mathbf{D}$ *is a diagonal matrix with positive diagonal entries. Then the following condition is sufficient for ensuring uniqueness of this decomposition (up to multiplication of* $\mathbf{N}$ *from the right by an orthogonal matrix): If any row of* $\mathbf{N}$ *is deleted there remain two disjoint submatrices of rank* p*.*

For a proof, see Anderson & Rubin (1956, Theorem 5.1) who also apply this result to discuss identifiability of the factor analysis model.

Proof of Theorem 1. We first show that $\mathbf{\Sigma}$ can be identified from the second moments of $\mathbf{X}$: Note that

$$\mathbf{E}\left[E^{(j)}E^{(\ell)}\right] = \begin{cases} \exp(\frac{1}{2}(\sigma_j^2 + \sigma_\ell^2)) : & j \neq \ell \\ \exp(2\sigma_j^2) \quad\quad\quad\quad : & j = \ell \end{cases}$$

If we let $\mathbf{F} = \mathrm{diag}(\exp(\frac{1}{2}\sigma_1^2), \ldots, \exp(\frac{1}{2}\sigma_m^2))$, then straightforward calculations show that

$$\begin{aligned} \mathbf{E}\left[(\mathbf{CS}\circ\mathbf{E})(\mathbf{CS}\circ\mathbf{E})^t\right] &= \mathbf{C}\,\mathbf{E}\left[\mathbf{S}\mathbf{S}^t\right]\mathbf{C}^t \circ \mathbf{E}\left[\mathbf{E}\mathbf{E}^t\right] \\ &= \mathbf{F}\mathbf{C}\,\mathbf{E}\left[\mathbf{S}\mathbf{S}^t\right]\mathbf{C}^t\mathbf{F}^t + \mathbf{D} \\ &= \mathbf{L} + \mathbf{D} \end{aligned} \tag{8}$$

where we have set $\mathbf{L} = \mathbf{F}\mathbf{C}\,\mathbf{E}\left[\mathbf{S}\mathbf{S}^t\right]\mathbf{C}^t\mathbf{F}^t$ and $\mathbf{D}$ is a diagonal matrix with diagonal entries $\mathbf{D}_{jj} = \exp(2\sigma_j^2)(\mathbf{C}\,\mathbf{E}\left[\mathbf{S}\mathbf{S}^t\right]\mathbf{C}^t)_{jj} - \mathbf{L}_{jj}$.

We assumed that if any row of $\mathbf{C}$ is deleted there remain two disjoint submatrices of rank p. It is easy to check that the matrix $\mathbf{A} := \mathbf{F}\mathbf{C}\mathbf{R}^\mathbf{t}$ has the same property if $\mathbf{R}$ is a regular $p \times p$-matrix with $\mathbf{R}^t\mathbf{R} = \mathbf{E}\left[\mathbf{S}\mathbf{S}^t\right]$ (e.g. its Choleski decomposition). Moreover $\mathbf{A}\mathbf{A}^t = \mathbf{L}$. Thus Theorem 2 implies uniqueness of the decomposition (8) into the two matrices $\mathbf{L}$ and $\mathbf{D}$. A simple calculation shows that the variances σ_j^2 can now be identified as $\sigma_j^2 = \log(\frac{\mathbf{D}_{jj}+\mathbf{L}_{jj}}{\mathbf{L}_{jj}})$.

Since the distribution of the error $\mathbf{E}$ is identified and independent of $\mathbf{S}$, we can recover the distribution $\mathbf{CS}$ (for example through the characteristic function of $\log(\mathbf{CS})$ which can be recovered from that of $\log(\mathbf{X}) = \log(\mathbf{CS}) + \log(\mathbf{E})$). The model without measurement error is identified by the reasoning directly before Theorem 1 in the main text. □

References

Aitchison, J. (1987), *The Statistical Analysis of Compositional Data*, Chapman & Hall.

Akerjord, M.-A. & Christophersen, N. (1996), 'Assessing mixing models within a common framework', *Environmental Science and Technology* **30**, 2105–2112.

Anderson, T. W. & Rubin, H. (1956), Statistical inference in factor analysis, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. 5, pp. 111– 150.

Bandeen-Roche, K. (1994), 'Resolution of additive mixtures into source components and contributions: A compositional approach', *Journal of the American Statistical Association* **89**, 1450–1458.

Bandeen-Roche, K. & Ruppert, D. (1991), 'Source apportionment with one source unknown (disc:pp.185-187)', *Chemometrics and Intelligent Laboratory Systems* **10**, 169–184.

Billheimer, D. (2001), 'Compositional receptor modeling', *Environmetrics* **12**, 451–467.

Booth, J. G. & Hobert, J. P. (1999), 'Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm', *Journal of the Royal Statistical Society, Ser. B* **61**(1), 265–285.

CRAN (1997 ff.), *The **C**omprehensive **R** **A**rchive **N**etwork*, `http://cran.R-project.org/`.

Crow, E. L. & Shimizu, K. E. (1988), *Lognormal Distributions: Theory and Applications*, Marcel Dekker.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Ser. B* **39**, 1–37.

Evans, M. & Swartz, T. (2000), *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press.

Künsch, H. R. (1989), 'The jackknife and the bootstrap for general stationary observations', *The Annals of Statistics* **17**, 1217–1241.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics, Springer.

Locher, R. (1999), VOC-Immissionsmessungen in den Kantonen Zürich, Schaffhausen und Luzern 1993-1998, Umwelt-Materialien Nr. 118 - Luft, Bundesamt für Umwelt, Wald und Landschaft, Bern.

Mosteller, F. & Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley.

Park, E. S., Guttorp, P. & Henry, R. C. (2001), 'Multivariate receptor modeling for temporally correlated data by using MCMC', *Journal of the American Statistical Association* **96**, 1171–1183.

Redner, R. A. & Walker, H. F. (1984), 'Mixture densities, maximum likelihood and the EM algorithm', *SIAM Review* **26**, 195–202.

Renner, R. M. (1993), 'The resolution of a compositional data set into mixtures of fixed source compositions', *Applied Statistics* **42**, 615–631.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

Wei, G. C. G. & Tanner, M. A. (1990), 'A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms', *Journal of the American Statistical Association* **85**, 699–704.

Wolbers, M. (2002), Linear unmixing of multivariate observations, PhD thesis, ETH Zürich (Diss. ETH 14723).