Statistical matching of non-Gaussian data

Daniel Ahfock\textsuperscript{1}, Saumyadipta Pyne\textsuperscript{2,3}, and Geoffrey J. McLachlan\textsuperscript{1}

\textsuperscript{1}School of Mathematics and Physics, University of Queensland, Brisbane, Australia
\textsuperscript{2}Public Health Dynamics Laboratory, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{3}Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

The statistical matching problem is a data integration problem with structured missing data. The general form involves the analysis of multiple datasets that only have a strict subset of variables jointly observed across all datasets. The simplest version involves two datasets, labelled A and B, with three variables of interest X, Y, and Z. Variables X and Y are observed in dataset A and variables X and Z are observed in dataset B. Statistical inference is complicated by the absence of joint (Y, Z) observations. Parametric modelling can be challenging due to identifiability issues and the difficulty of parameter estimation. We develop computationally feasible procedures for the statistical matching of non-Gaussian data using suitable data augmentation schemes and identifiability constraints. Nearest-neighbour imputation is a common alternative technique due to its ease of use and generality. Nearest-neighbour matching is based on a conditional independence assumption that may be inappropriate for non-Gaussian data. The violation of the conditional independence assumption can lead to improper imputations. We compare model based approaches to nearest-neighbour imputation on a number of flow cytometry datasets and find that the model based approach can address some of the weaknesses of the nonparametric nearest-neighbour technique.

1 Introduction

The statistical matching problem is a data integration scenario where the set of available datasets shares only a strict subset of variables. The pattern of missingness is such that some dimensions are wholly unobserved in each dataset (D'Orazio et al., 2006). The simplest case involves two datasets, dataset A and dataset B, and three groups of variables, labelled X, Y, and Z. We assume the data generating process is governed by a parametric model \( f(x, y, z; \theta) \), for some parameter \( \theta \in \Omega \subset \mathbb{R}^p \). Dataset A contains \( n_A \) independent observations from the marginal distribution \( f(x, y; \theta) \) and dataset B consists of \( n_B \) independent observations from the marginal distribution \( f(x, z; \theta) \). Table 1 represents the missing data pattern. A common objective in the statistical matching problem is to impute the missing observations in each dataset so complete data techniques can be used in downstream analyses (R"assler, 2002). The statistical matching problem occurs in flow cytometry analysis due to technological limitations, and there has been recent interest in
developing appropriate statistical methods for integrating cytometry datasets \cite{Pedreira2008, O'Neill2015}. Most previous work on the statistical matching problem for continuous data assumes a multivariate normal model. Flow cytometry data typically have characteristics that make a multivariate Gaussian model inappropriate. Observations are from a number of heterogeneous cell subpopulations, and observations within a cell subtype often exhibit skewness and heavy tails. Finite mixtures of skew-normal and skew-$t$ distributions have shown excellent performance in the analysis of flow cytometry data \cite{Pyne2009, Lee2016}. We study theoretical aspects of the statistical matching of non-Gaussian data and propose new computationally efficient algorithms for the statistical matching of skew-normal data, mixtures of Gaussians and mixtures of skew-normal distributions.

A central issue in the statistical matching problem is that the lack of joint observations on the $Y$ and $Z$ variables renders most statistical models nonidentifiable. For example, if $f(x, y, z; \theta)$ is a multivariate normal distribution, the $Y$, $Z$ covariance parameters are not identifiable. In a parametric framework, it is common to adopt restrictions on the parameters so that the generative model remains identifiable. Under identification constraints it is possible to impute the missing data using a parametric modelling strategy. Let $\hat{\theta}$ denote the maximum likelihood estimate of $\theta$. The missing $Z$ values in dataset A can be sampled from $f(z|x, y; \hat{\theta})$, and the missing $Y$ values in dataset $B$ can be sampled from $f(y|x, z; \hat{\theta})$.

Parametric strategies for Gaussian data typically involve reparameterising the model so that the likelihood can then be factored into a mathematically tractable form \cite{D'Orazio2006}. Reparameterisation and factorisation are powerful strategies in missing data problems, as it can greatly simplify maximum likelihood estimation \cite[Chapter 7]{Little2002}. The reparameterisation for the statistical matching of Gaussian data involves expressing the observed-data likelihood in terms of a multiple response regression model. We show how the regression model specification can be extended to cover skew-normal distributions, mixtures of Gaussians, and mixtures of skew-normal distributions through appropriate data augmentation schemes. Factorisation of the complete-data log likelihood allows the EM algorithm \cite{Dempster1977} to be used for computationally efficient parameter estimation.

An alternative method for imputation is nearest-neighbour matching, a nonparametric method based on the assumption that the $Y$ and $Z$ variables are conditionally independent given the $X$ variables. Nearest-neighbour imputation has been widely applied in the statistical matching problem due to its ease of implementation and generality \cite{D'Orazio2006}. In particular, nearest-neighbour matching has been advocated for the file matching of flow cytometry data \cite{Pedreira2008}. An important finding is that the $Y$ and $Z$ variables are likely to be conditionally dependent given the $X$ variables if the generative model is a mixture model or a skewed distribution. The violation of the conditional independence assumption can lead to nearest-neighbour imputation.

| Dimensions | $X$ | $Y$ | $Z$ |
|------------|-----|-----|-----|
| Dataset A  | ✓   | ✓   | ✓   |
| Dataset B  | ✓   | ✓   | ✓   |

Table 1: Missing data pattern in the statistical matching problem.
exhibiting undesirable behaviour.

The conditional independence assumption can be adopted for any data generating process, and is the dominant approach to resolve the pathological missing data pattern in the statistical matching problem. This conditional independence assumption is difficult to justify for a range of latent variable models, and model based approaches can impose more coherent constraints. We compare nearest-neighbour matching to model based imputation on a number of real datasets. The model based approach showed a large improvement over the nearest-neighbour method, however there can be issues when the identification constraints are violated. Parametric imputation strategies are a flexible and computationally feasible option for statistical matching.

2 Background

2.1 Notation

We assume \( n_A \) observations in dataset A and \( n_B \) observations in dataset B, for a total of \( n = n_A + n_B \) records. These datasets are combined to form a single data matrix, where rows 1 to \( n_A \) contain the records from dataset A on the \( X \) and \( Y \) variables and rows \( n_B \) rows contain the observations from dataset B on the \( X \) and \( Z \) variables. Let \( d_X, d_Y \) and \( d_Z \) give the dimension of the \( X \), \( Y \) and \( Z \) variables respectively and set \( d = d_X + d_Y + d_Z \).

We assume the data generating process consists of \( n \) independently and identically distributed observations from the parametric model \( (X_i^T, Y_i^T, Z_i^T)^T \sim f(x_i, y_i, z_i; \theta) \) for \( i = 1, \ldots, n \). Let \( x_i, y_i \) and \( z_i \) represent the realised values for the \( X_i, Y_i \) and \( Z_i \) random vectors respectively for \( i = 1, \ldots, n \). Observations \( z_i \) are missing for \( i = 1, \ldots, n_A \). Observations \( y_i \) are missing for \( j = n_A + 1, n_A + 2, \ldots, n_A + n_B \). The joint \( n \times d \) data matrix of interest is represented in the display (1). Observed values are shaded, with white cells indicating missing values in the joint data matrix. Let \( x_{\text{obs}} = (x_1, x_2, \ldots, x_n) \) represent the observed \( X \) values over the \( n \) records. Let \( x_{A\text{obs}} = (x_1, \ldots, x_{n_A}) \) represent the \( n_A \) observed \( X \) values on dataset A.

\[
\begin{pmatrix}
  x_1 & y_1 & z_1 \\
  \vdots & \vdots & \vdots \\
  x_{n_A} & y_{n_A} & z_{n_A} \\
  x_{n_A+1} & y_{n_A+1} & z_{n_A+1} \\
  \vdots & \vdots & \vdots \\
  x_{n_A+n_B} & y_{n_A+n_B} & z_{n_A+n_B}
\end{pmatrix}
\]

(1)

Let \( x_{B\text{obs}} = (x_{n_A+1}, \ldots, x_{n_A+n_B}) \) denote the \( n_B \) observed \( X \) values in dataset B. Let \( y_{\text{obs}} = (y_1, \ldots, y_{n_A}) \) represent the \( n_A \) observed values \( Y \) variables in dataset A, and let \( z_{\text{obs}} = (z_{n_A+1}, \ldots, z_{n_A+n_B}) \) denote the \( n_B \) observed \( Z \) values in dataset B. Likewise, let \( z_{\text{mis}} = (z_1, \ldots, z_{n_A}) \) represent the \( n_A \) missing \( Z \) values in dataset A, and let \( y_{\text{mis}} = (y_{n_A+1}, \ldots, y_{n_A+n_B}) \) represent the \( n_B \) missing \( Y \) values in dataset B.
2.2 Gaussian data

The statistical matching problem has been explored in depth under the assumption of a multivariate Gaussian generative model. Suppose that for \( i = 1, \ldots, n \) we have observations from the multivariate normal distribution

\[
\begin{bmatrix}
X_i \\
Y_i \\
Z_i
\end{bmatrix}
\sim N
\left( \begin{bmatrix}
\mu_X \\
\mu_Y \\
\mu_Z
\end{bmatrix},
\begin{bmatrix}
\Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\
\Sigma_{XY} & \Sigma_{YY} & \Sigma_{YZ} \\
\Sigma_{XZ} & \Sigma_{YZ} & \Sigma_{ZZ}
\end{bmatrix}
\right),
\]

(2)

where the mean and covariance parameters have been partitioned in an obvious fashion. With the missing data pattern represented in (1), the only non-identifiable parameter is \( \Sigma_{YZ} \). The most common identifiability constraint is that \( \Sigma_{YZ} = \Sigma_{XY} \Sigma_{XX}^{-1} \Sigma_{XZ} \).

Parameter estimation under the identification constraint \( \Sigma_{YZ} = \Sigma_{XY} \Sigma_{XX}^{-1} \Sigma_{XZ} \) has a long history in the literature (Lord, 1955; Anderson, 1957; Moriarity and Scheuren, 2003). As discussed in D’Orazio et al. (2006, Chapter 2), the multivariate normal statistical matching problem has a useful connection to linear regression modelling that can be used to obtain closed form maximum likelihood estimates. The conditional distribution of \( Y_i \) given \( X_i = x_i \) can be represented as a regression model

\[
Y_i | X_i = x_i \sim N(\alpha_Y + \beta_Y x_i, \Omega_Y),
\]

(3)

for \( i = 1, \ldots, n \) where

\[
\beta_Y = \Sigma_{XY} \Sigma_{XX}^{-1},
\]

(4)
\[
\alpha_Y = \mu_Y - \beta_Y \mu_X,
\]

(5)
\[
\Omega_Y = \Sigma_{YY} - \Sigma_{XY} \Sigma_{XX}^{-1} \Sigma_{XY}.
\]

(6)

The same holds for the conditional distribution of \( Z_i \) given \( X_i = x_i \)

\[
Z_i | X_i = x_i \sim N(\alpha_Z + \beta_Z x_i, \Omega_Z),
\]

(7)

for \( i = 1, \ldots, n \) where

\[
\beta_Z = \Sigma_{ZX} \Sigma_{XX}^{-1},
\]

(8)
\[
\alpha_Z = \mu_Z - \mu_X \beta_Z,
\]

(9)
\[
\Omega_Z = \Sigma_{ZZ} - \Sigma_{ZX} \Sigma_{XX}^{-1} \Sigma_{XZ}.
\]

(10)

Now let \( \eta = (\eta_X, \eta_Y, \eta_Z) \), where \( \eta_X = (\mu_X, \Sigma_{XX}) \), \( \eta_Y = (\alpha_Y, \beta_Y, \Omega_Y) \) and \( \eta_Z = (\alpha_Z, \beta_Z, \Omega_Z) \). The benefit of the regression parameterisation is that that the likelihood for each observation factors into three components with distinct parameter blocks. We can write \( f(x_i, y_i, z_i; \eta) = f(x_i; \eta_X) f(y_i|x_i; \eta_Y) f(z_i|x_i; \eta_Z) \). The observed-data likelihood can be expressed as a product
of three terms,

\[
f(x_{\text{obs}}, y_{\text{obs}}, Z_{\text{obs}}; \eta) = \prod_{i=1}^{n_A} f(x_i, y_i; \eta) \prod_{k=n_A+1}^{n} f(x_k, z_k; \eta)
\]

\[
= \prod_{i=1}^{n} f(x_i; \eta_X) \prod_{j=1}^{n_A} f(y_j|x_j; \eta_Y) \prod_{k=n_A+1}^{n} f(z_k|x_k; \eta_Z)
\]

\[
= f(x_{\text{obs}}; \eta_X)f(y_{\text{obs}}|x_{\text{obs}}^A; \eta_Y)f(z_{\text{obs}}|x_{\text{obs}}^B; \eta_Z).
\]

The first likelihood block \(f(x_{\text{obs}}; \eta_X)\) involves the \(n\) observed \(X\) variables. The second likelihood block \(f(y_{\text{obs}}|x_{\text{obs}}^A; \eta_Y)\) involves the conditional likelihood for dataset \(A\) and the third likelihood block \(f(z_{\text{obs}}|x_{\text{obs}}^B; \eta_Z)\) involves the conditional likelihood for dataset \(B\). Maximisation of the observed-data likelihood

\[
f(x_{\text{obs}}, y_{\text{obs}}, Z_{\text{obs}}; \eta) = f(x_{\text{obs}}; \eta_X)f(y_{\text{obs}}|x_{\text{obs}}^A; \eta_Y)f(z_{\text{obs}}|x_{\text{obs}}^B; \eta_Z)
\]

is straightforward, as we now have three separate maximisation problems over \(\eta_X, \eta_Y\) and \(\eta_Z\). Both datasets are used to estimate \(\eta_X\) using the fully observed \(X\) variables. Dataset \(A\) is used to estimate \(\eta_Y\) and dataset \(B\) is used to estimate \(\eta_Z\).

Using the regression specifications (3) and (7), we can express the likelihoods \(f(y_{\text{obs}}|x_{\text{obs}}^A; \eta_Y)\) and \(f(z_{\text{obs}}|x_{\text{obs}}^B; \eta_Z)\) as multiple response regression models. Recall the data structure in display (1). Let \(Y_A\) give the \(n_A \times d_Y\) matrix of responses from dataset \(A\). Row \(i\) in \(Y_A\) is given by \(y_i^T\) for \(i = 1, \ldots, n_A\). Let \(Z_B\) give the \(n_B \times d_Z\) from dataset \(B\). Row \(i\) in \(Z_B\) is given by \(z_i^T\) for \(i = n_A + 1, \ldots, n_A + n_B\). Let \(B_A\) represent the design matrix for dataset \(A\) and let \(B_B\) give the design matrix for dataset \(B\). Specifically,

\[
B_A = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots \\ 1 & x_{n_A}^T \end{bmatrix}, \quad B_B = \begin{bmatrix} 1 & x_{n_A+1}^T \\ 1 & x_{n_A+2}^T \\ \vdots \\ 1 & x_{n_A+n_B}^T \end{bmatrix}.
\]

Let \(\Gamma_A\) and \(\Gamma_B\) contain the regression coefficients for (3) and (7) respectively: \(\Gamma_A = [\alpha_Y \beta_Y]^T\) and \(\Gamma_B = [\alpha_Z \beta_Z]^T\). We will use the matrix normal distribution introduced by Dawid (1981) to specify the regression models compactly. A random \(n \times p\) matrix \(M\) is said to have the matrix normal distribution \(MN(I_n, \Sigma)\) if each row is a draw from a \(p\)-variate normal distribution \(N(0, \Sigma)\). The conditional regression models for each dataset can be written as

\[
Y_A = B_A \Gamma_A + \epsilon_A,
\]

\[
Z_B = B_B \Gamma_B + \epsilon_B,
\]

where \(\epsilon_A \sim MN(I_{n_A}, \Omega_Y)\) and \(\epsilon_B \sim MN(I_{n_B}, \Omega_Z)\). Equation (16) is a representation of the likelihood factor \(f(y_{\text{obs}}|x_{\text{obs}}^A; \eta_Y)\) and equation (17) is a representation of the likelihood factor \(f(z_{\text{obs}}|x_{\text{obs}}^B; \eta_Z)\). The sufficient statistics for the regressions are \(Y_A^TY_A, Z_B^T Z_B, B_A^T Y_A, B_B^T Z_B, B_A^T B_A\).
and $B^T_B B_B$. The maximum likelihood estimators of the regression parameters $\Gamma_A = [\alpha_Y \beta_Y]^T$ and $\Gamma_B = [\alpha_Z \beta_Z]^T$ are given by

$$\hat{\Gamma}_A = (B^T_A B_A)^{-1} B^T_A Y_A,$$
$$\hat{\Gamma}_B = (B^T_B B_B)^{-1} B^T_B Z_B.$$

The maximum likelihood estimators of the error covariance matrices $\Omega_Y$ and $\Omega_Z$ are given by

$$\hat{\Omega}_Y = \frac{1}{n_A} (Y_A - B_A \hat{\Gamma}_A)^T (Y_A - B_A \hat{\Gamma}_A),$$
$$\hat{\Omega}_Z = \frac{1}{n_B} (Z_B - B_B \hat{\Gamma}_B)^T (Z_B - B_B \hat{\Gamma}_B).$$

These results follow from general results on multiple response regression models, for example see Rencher and Christensen (2012, Chapter 10). The maximum likelihood estimators of the $\eta_X$ parameters are given by the sample mean and the sample covariance of the $n$ observed $X$ values

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_X)(x_i - \hat{\mu}_X)^T.$$ 

Inverting the transformations defined in equations (5) to (10) gives the maximum likelihood estimates of the original parameters $\mu$ and $\Sigma$. We thus obtain closed form maximum likelihood estimators under the identification restriction $\Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ}$.

3 Extensions

3.1 Overview

A parametric approach in the statistical matching problem requires the identification of suitable identifiability constraints, and a feasible procedure for maximum likelihood estimation. One approach is to introduce the missing data $y_{obs}, z_{obs}$ into the model as latent variables. In general, we can form the complete-data likelihood $f(x_{obs}, y_{obs}, z_{obs}, y_{mis}, z_{mis}; \theta)$ and define appropriate EM iterations for parameter estimation. Lee et al. (2011) consider the statistical matching of flow cytometry data with mixtures of PCA models and take this approach. Introduction of the missing observations $y_{mis}$ and $z_{mis}$ into the complete data log-likelihood can be computationally demanding on flow cytometry datasets with a large number of observations. Due to the non-identifiability of the model, the fitted mixture model is only used to assign cluster labels to the observations. Lee et al. propose to use nearest-neighbour matching within each group of labelled points.

We propose a different method, where we introduce appropriate parameter constraints so that $y_{mis}$ and $z_{mis}$ do not need to be included in the complete-data log likelihood. This is to obtain a more computationally efficient EM algorithm, and to avoid the identifiability issues with fitting a completely unconstrained model. Secondly, we propose to impute the missing data using the constrained fitted model. The strategy involves extending the linear regression connection that was used for Gaussian data.
3.2 Skew-normal

There are many forms of the skew-normal distribution, we will work with the same multivariate version as in [Pyne et al. 2009]. For compact notation, let the vector \( \mathbf{w}^i = (x^i, y^i, z^i)^T \). The density of the skew-normal distribution is

\[
f(x, y, z; \mu, \Sigma, \delta) = 2\phi_p(x, y, z; \mu, \Lambda) \Phi(\alpha^T(w - \mu)),
\]

where \( \Lambda = \Sigma + \delta \delta^T \) and \( \alpha^T = \delta^T \Lambda^{-1} / (1 - \delta^T \Lambda^{-1} \delta) \). Let \( TN(\mu, \sigma^2, a) \) denote a lower truncated normal distribution where \( \mu \) and \( \sigma^2 \) give the mean and variance of the underlying normal distribution and \( a \) gives the lower truncation bound. We say \( U \sim TN(\mu, \sigma^2, a) \) if \( U \overset{d}{=} [W | W > a] \) where \( W \sim N(\mu, \sigma^2) \). The model (18) has a hierarchical representation,

\[
\begin{bmatrix}
X_i \\
Y_i \\
Z_i
\end{bmatrix} = 
\begin{bmatrix}
\mu_X \\
\mu_Y \\
\mu_Z
\end{bmatrix} + \begin{bmatrix}
\delta_X \\
\delta_Y \\
\delta_Z
\end{bmatrix} U_i + V_i,
\]

where \( U_i \sim TN(0, 1, 0) \) and \( V_i \sim N(0, \Sigma) \), where \( \mu \) and \( \Sigma \) are partitioned as in (2). The only non-identifiable parameter is \( \Sigma_{YZ} \). The regression model specification also enables efficient inference for the skew-normal distribution under the identification restriction \( \Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ} \). The statistical matching problem for skew-normal data can be expressed in terms of a multiple response regression model with latent unobserved variables in the conditional mean function. Using the hierarchical model (19), conditional on knowing \( U_i = u_i \), it holds that,

\[
\begin{bmatrix}
X_i \\
Y_i \\
Z_i
\end{bmatrix} \mid U_i = u_i \sim 
\begin{bmatrix}
\mu_X + \delta_X u_i \\
\mu_Y + \delta_Y u_i \\
\mu_Z + \delta_Z u_i
\end{bmatrix} + V_i,
\]

where \( V_i = N(0, \Sigma) \), for \( i = 1, \ldots, n \). Let \( \beta_Y = \Sigma_{YX} \Sigma_{XX}^{-1} \) and \( \beta_Z = \Sigma_{ZX} \Sigma_{XX}^{-1} \). The conditional mean of \( Y_i \) given \( X_i \) and \( U_i \), and the conditional mean of \( Z_i \) given \( X_i \) and \( U_i \) are given respectively by:

\[
E[Y_i | X_i = x_i, U_i = u_i] = \mu_Y + \delta_Y u_i + \beta_Y (x_i - \mu_X - \delta_X u_i),
\]

\[
E[Z_i | X_i = x_i, U_i = u_i] = \mu_Z + \delta_Z u_i + \beta_Z (x_i - \mu_X - \delta_X u_i).
\]

We have that \( \text{var}(Y_i | X_i, U_i) = \Sigma_{YY} - \Sigma_{XY} \Sigma_{XX}^{-1} \Sigma_{XY} \) and \( \text{var}(Z_i | X_i, U_i) = \Sigma_{ZZ} - \Sigma_{ZX} \Sigma_{XX}^{-1} \Sigma_{XZ} \). The augmented likelihood for dataset A and the augmented likelihood for dataset B can be expressed as conditional regression models. Let \( \alpha_Y = \mu_Y - \beta_Y \mu_X \), \( \Omega_Y = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \) and \( \lambda_Y = \delta_Y - \beta_Y \delta_X \). Collecting terms in (21), the distribution of \( Y_i \) given \( X_i \) and the latent scaling variable \( U_i \) can be represented as the regression model:

\[
Y_i | X_i = x_i, U_i = u_i \sim N(\alpha_Y + \lambda_Y u_i + \beta_Y x_i, \Omega_Y).
\]
Let $\beta_Z = \Sigma_{ZX} \Sigma_{XX}^{-1}, \alpha_Z = \mu_Z - \beta_Z \mu_X, \Omega_Z = \Sigma_{ZZ} - \Sigma_{ZX} \Sigma_{XX}^{-1} \Sigma_{XZ}$ and $\lambda_Z = \delta_Z - \beta_Z \delta_X$. Similarly, collecting terms in the conditional distribution of $Z_i$ given $X_i$ and the latent scaling variable $U_i$ can be represented as a regression model:

$$Z_i | X_i = x_i, U_i = u_i \sim N(\alpha_Z + \lambda_Z u_i + \beta_Z x_i, \Omega_Z).$$

(24)

Let $\Gamma_A$ and $\Gamma_B$ contain the regression parameters for $\{\alpha_Y, \lambda_Y, \beta_Y\}^T$ and $\Gamma_B = [\alpha_Z, \lambda_Z, \beta_Z]^T$. The regressions for dataset A and dataset B can be written as

$$Y_A = B_A \Gamma_A + \epsilon_A,$$

$$Z_B = B_B \Gamma_B + \epsilon_B,$$

(25)

(26)

where $\epsilon_A \sim MN(I_{n_A}, \Omega_Y)$ and $\epsilon_B \sim MN(I_{n_B}, \Omega_Z)$. The design matrices for the regressions now include the latent $U_i$ terms. The complete-data design matrices are given by

$$B_A = \begin{bmatrix} 1 & u_1 & x_1^T \\ 1 & u_2 & x_2^T \\ \vdots \\ 1 & u_{n_A} & x_{n_A}^T \end{bmatrix}, B_B = \begin{bmatrix} 1 & u_{n_A+1} & x_{n_A+1}^T \\ 1 & u_{n_A+2} & x_{n_A+2}^T \\ \vdots \\ 1 & u_{n_A+n_B} & x_{n_A+n_B}^T \end{bmatrix}.$$  

(27)

The complete-data sufficient statistics are again $Y_A^T Y_A, Z_B^T Z_B, B_A^T Y_A, B_B^T Z_B, B_A^T B_A$ and $B_B^T B_B$. Suppose that we impose the identification restriction $\Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ}$. This corresponds to a conditional independence assumption on the components of the latent random variable $V$ in the hierarchical skew-normal model. Under the the identification restriction, the $Y$ and $Z$ variables are conditionally independent given $X$ and the latent $U$. The augmented data likelihood for a single observation then has the structure

$$f(x_i, y_i, z_i, u_i; \theta) = f(x_i, y_i, z_i | u_i; \theta)f(u_i)$$

$$= f(x_i | u_i; \theta)f(y_i | x_i, u_i; \theta)f(z_i | u_i; \theta)f(u_i).$$

(28)

Using this property we can show the complete-data likelihood has a similar structure to the Gaussian likelihood $[14]$. Let $\eta = (\eta_X, \eta_Y, \eta_Z)$ represent the parameters for each likelihood block. We have $\eta_X = (\mu_X, \Sigma_{XX}, \delta_X), \eta_Y = (\alpha_Y, \beta_Y, \Omega_Y, \lambda_Y)$ and $\eta_Z = (\alpha_Z, \beta_Z, \Omega_Z, \lambda_Z)$. Let $u$ contain the latent skewing variable for each observation, so $u = (u_1, \ldots, u_n)$. Using the conditional regression models, the complete-data likelihood can be expressed as

$$f(x_{obs}, y_{obs}, z_{obs}, u; \eta) = f(x_{obs} | u; \eta_X)f(y_{obs} | x_{obs}^A, u; \eta_Y)f(z_{obs} | x_{obs}^B, u; \eta_Z)f(u).$$

(29)

Equation (25) is a representation of the likelihood factor $f(y_{obs} | x_{obs}^A, u; \eta_Y)$ and equation (26) is a representation of the likelihood factor $f(z_{obs} | x_{obs}^B, u; \eta_Z)$. The E-step involves calculating the expected value of the sufficient statistics given the current parameter estimates, and can be carried out using the results in Pyne et al. (2009). The M-step is simplified due to choice of parameterisation for the regression model. There are three separate maximisation tasks over $\eta_X, \eta_Y$ and $\eta_Z$.
that have closed form solutions. The key point is that we do not have to introduce the missing observations \( y_{mis} \) and \( z_{mis} \) to obtain a workable EM algorithm under the identification restriction \( \Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ} \). The algorithm is given in Web Appendix A.

### 3.3 Mixture of Gaussians

We can also obtain an efficient EM algorithm for statistical matching of Gaussian mixture data using the conditional regression specifications and appropriate identification restrictions. The matching problem for Gaussian mixture data can be represented as a mixture of regression problem (Quandt 1972). Recall the hierarchical model for the mixture model discussed in Section 4.2. For independent observations \( i = 1, \ldots, n \):

\[
X_i, Y_i, Z_i|S_i = h \sim f(x_i, y_i, z_i; \theta_h) \\
S_i \sim \text{Categorical}(\pi_1, \ldots, \pi_g).
\]

Here \( f(x_i, y_i, z_i; \theta_h) \) denotes a multivariate normal density with parameters \( \theta_h = (\mu^{(h)}, \Sigma^{(h)}) \). The vector \( \mu^{(h)} \) gives the mean vector for component \( h \) and the matrix \( \Sigma^{(h)} \) represent the covariance matrix of the \( h \)th component distribution for \( h = 1, \ldots, g \). We can partition each component mean \( \mu^{(h)} \) and each component covariance matrix \( \Sigma^{(h)} \) as in (2). An identification restriction is then

\[
\Sigma^{(h)} = \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1} \Sigma_{XY}^{(h)}, \quad \text{for } h = 1, \ldots, g.
\]  

(30)

Consider the augmented likelihood for a single observation given the latent cluster indicator \( S_i \). Conditional on the latent cluster indicator \( S_i \), we can again use the regression model specification from Section 2.2. Specifically,

\[
Y_i|X = x_i, S_i = h \sim N(\alpha_y^{(h)} + \beta_y^{(h)} x_i, \Omega_y^{(h)}),
\]

(31)

where \( \beta_y^{(h)} = \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1} \), \( \alpha_y^{(h)} = \mu_y^{(h)} - \beta_y^{(h)} \mu_X \) and \( \Omega_y^{(h)} = \Sigma_{YY}^{(h)} - \Sigma_{YY}^{(h)} \Sigma_{XX}^{-1} \Sigma_{XY}^{(h)} \). The conditional distribution of \( Z_i \) given \( X_i \) obeys a similar regression equation:

\[
Z_i|X = x_i, S_i = h \sim N(\alpha_z^{(h)} + \beta_z^{(h)} x_i, \Omega_z^{(h)}),
\]

where \( \beta_z^{(h)} = \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1} \), \( \alpha_z^{(h)} = \mu_z^{(h)} - \beta_z^{(h)} \mu_X \) and \( \Omega_z = \Sigma_{ZZ} - \Sigma_{ZX} \Sigma_{XX}^{-1} \Sigma_{XZ} \). To describe the complete-data log likelihood first let \( s \) represent the vectors of latent cluster indicators, so \( s = (s_1, \ldots, s_n) \), where \( s_i \in \{1, \ldots, g\} \) for \( i = 1, \ldots, n \). Let \( \pi = (\pi_1, \ldots, \pi_g) \) be a vector containing the mixing proportions. We can show that complete-data likelihood has a similar structure to what we obtained in for the Gaussian likelihood (13). For Gaussian mixture data, the complete-data likelihood has the structure

\[
f(x_{obs}, y_{obs}, z_{obs}, s; \eta, \pi) = f(x_{obs}|s; \eta_X) f(y_{obs}|x_{obs}, s; \eta_Y) f(z_{obs}|x_{obs}^B, s; \eta_Z) f(s; \pi).
\]

(32)

The complete-data likelihood is the product of four components. The first likelihood block \( f(x_{obs}|s; \eta_X) \) is a standard Gaussian likelihood. The second and third likelihood blocks \( f(y_{obs}|x_{obs}^A, s; \eta_Y) \) and
The complete-data likelihood has the structure
\[ f(z_{\text{obs}}|x_{\text{obs}}^R, s; \eta Z) \] correspond to regression likelihoods. The final likelihood contribution from \( f(s; \pi) \) is the usual multinomial likelihood seen in a finite mixture. The E-step and M-step for \( f(x_{\text{obs}}|s; \eta X) \) follow from standard results on Gaussian mixture models. The E-step and M-steps for \( f(y_{\text{obs}}|x_{\text{obs}}^A, s; \eta Y) \) and \( f(z_{\text{obs}}|x_{\text{obs}}^B, s; \eta Z) \) follow from results on mixtures of regression models given in [Jones and McLachlan (1992)]. The details are given in Web Appendix B. We avoid introducing the missing observations \( y_{\text{mis}} \) and \( z_{\text{mis}} \) into the complete-data log likelihood by using the component-wise restrictions (30).

### 3.4 Mixtures of skew-normal

The statistical matching of data from a mixture of skew-normal distributions as a mixture of regressions problem, with latent variables in the design matrix. For independent observations \( i = 1, \ldots, n \) we have that
\[
X_i, Y_i, Z_i | S_i = h \sim f(x_i, y_i, z_i; \theta_h) \\
S_i \sim \text{Categorical}(\pi_1, \ldots, \pi_g).
\]

Here \( f(x_i, y_i, z_i; \theta_h) \) denotes a skew-normal density with parameters \( \theta_h = (\mu^{(h)}, \Sigma^{(h)}, \delta^{(h)}) \). The vector \( \mu^{(h)} \) gives the mean vector for component \( h \), the matrix \( \Sigma^{(h)} \) represents the scale matrix for component \( h \) and \( \delta^{(h)} \) gives the skewness vector for component \( h \) for \( h = 1, \ldots, g \). We can partition each component mean \( \mu^{(h)} \) and skewness vector \( \delta^{(h)} \) as in (19). Each component scale matrix \( \Sigma^{(h)} \) can be partitioned as in (2). An identification restriction is then
\[
\Sigma_Y^{(h)} = \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1} \Sigma_{XY}^{(h)}, \text{ for } h = 1, \ldots, g. \quad (33)
\]

Introducing the latent \( U_i \) and \( S_i \) into the augmented likelihood, the complete-data model can be expressed in terms of conditional regression specifications. Let \( \beta_Y^{(h)} = \Sigma_{YX}^{(h)} \Sigma_{XX}^{-1}, \alpha_Y^{(h)} = \mu_Y^{(h)} - \beta_Y^{(h)} \mu_X^{(h)}, \Omega_Y^{(h)} = \Sigma_{YY}^{(h)} - \Sigma_{YX}^{(h)} \Sigma_{XX}^{-1} \Sigma_{XY}^{(h)} \) and \( \lambda_Y^{(h)} = \delta_Y^{(h)} - \beta_Y^{(h)} \delta_X^{(h)} \). Then
\[
Y_i | X_i = x_i, U_i = u_i, S_i = h \sim N(\alpha_Y^{(h)} + \lambda_Y^{(h)} u_i + \beta_Y^{(h)} x_i, \Omega_Y^{(h)}) \quad (34)
\]

Let \( \beta_Z^{(h)} = \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1}, \alpha_Z^{(h)} = \mu_Z^{(h)} - \beta_Z^{(h)} \mu_X^{(h)}, \Omega_Z^{(h)} = \Sigma_{ZZ}^{(h)} - \Sigma_{ZX}^{(h)} \Sigma_{XX}^{-1} \Sigma_{ZX}^{(h)} \) and \( \lambda_Z^{(h)} = \delta_Z^{(h)} - \beta_Z^{(h)} \delta_X^{(h)} \). Similarly, collecting terms in (22) the conditional distribution of \( Z_i \) given \( X_i \) and the latent scaling variable \( U_i \) can be represented as a regression model
\[
Z_i | X_i = x_i, U_i = u_i, S_i = h \sim N(\alpha_Z^{(h)} + \lambda_Z^{(h)} u_i + \beta_Z^{(h)} x_i, \Omega_Z^{(h)}). \quad (35)
\]

The complete-data likelihood has the structure
\[
f(x_{\text{obs}}, y_{\text{obs}}, z_{\text{obs}}, s; \eta, \pi) = f(x_{\text{obs}}|s, u; \eta_X) f(y_{\text{obs}}|x_{\text{obs}}^A, s, u; \eta_Y) f(z_{\text{obs}}|x_{\text{obs}}^B, s, u; \eta_Z) f(s; \pi) f(u) \quad (36)
\]

The first likelihood block \( f(x_{\text{obs}}|s, u; \eta_X) \) corresponds to a skew-normal likelihood. The second and third likelihood blocks are regression likelihoods with the latent \( u \) variables in the design matrices.
We can again define an EM algorithm for parameter estimation under the identification constraint \(33\). The E-step is carried out using existing results for the skew-normal distribution. The M-step is again simplified by the choice of parameterisation. We have three separate maximisation tasks over \(\eta_X\), \(\eta_Y\) and \(\eta_Z\). The algorithm is given in full in Web Appendix C. Once again we do not introduce the missing observations \(y_{\text{mis}}\) and \(z_{\text{mis}}\) into complete-data likelihood by using the identification restriction \(33\).

4 Nearest-neighbour matching

Nearest-neighbour imputation is a popular alternative to parametric methods in data fusion problems (Aluja-Banet et al., 2007; Saporta, 2002). Nearest-neighbour methods rely on the conditional independence assumption, that is \(Y\) and \(Z\) are conditionally independent given \(X\). The nearest-neighbour method matches observations in dataset A and dataset B based on the Euclidean distance measured using the common \(X\) dimensions. Missing values are imputed by taking values from the nearest-neighbour in the donor set. More formally, the missing \(Z\) values for observation \(i\) in dataset A are set as

\[z_i = z_k, \quad \text{where} \quad k = \arg\min_{k \in \{n_A+1, \ldots, n_A+n_B\}} ||x_i - x_k||_2,\]

for \(i = 1, \ldots, n_A\). Similarly, the missing \(Y\) values for observation \(i\) in dataset B are set as

\[y_i = y_k, \quad \text{where} \quad k = \arg\min_{k \in \{1, \ldots, n_A\}} ||x_i - x_k||_2,\]

for \(i = n_A+1, \ldots, n_A+n_B\). As discussed by Räsler (2002), each of the observations in imputed dataset can be viewed as exchangeable draws from some distribution \(g(x_i; y_i, z_i)\). The distribution \(g(x_i; y_i, z_i)\) is useful to characterise the behaviour of the nearest-neighbour method. Marella et al. (2008) consider the asymptotic form of the nearest-neighbour imputation distribution and show that as the size of the donor set tends to infinity,

\[g(x_i; y_i, z_i) \rightarrow f(x_i; \theta)f(y_i|x_i; \theta)f(z_i|x_i; \theta). \quad (37)\]

In equation \(37\), \(f(x_i; \theta), f(y_i|x_i; \theta)\) and \(f(z_i|x_i; \theta)\) are the marginal and conditional distributions from the true generative model \(f(x_i, y_i, z_i)\). The nearest-neighbour strategy will only produce the correct joint distribution if \(Y\) and \(Z\) are conditionally independent given \(X\). For multivariate-normal data, the asymptotic form of the nearest-neighbour imputation scheme is

\[g(x_i, y_i, z_i) = \mathcal{N}_p \left( \mu = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix} \right). \quad (38)\]

The distribution \(g(x_i, y_i, z_i)\) is equivalent to the distribution recovered by the maximum likelihood parametric approach under the identification constraint \(\Sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ}\). The asymptotic equivalence between nearest-neighbour matching and a parametric methods may not necessarily
hold for non-Gaussian data.

It can be difficult to justify the conditional independence assumption for non-Gaussian data as there may be no parameter \( \theta^* \) such that \( f(x, y, z; \theta^*) = f(x; \theta^*) f(y|x; \theta^*) f(z|x; \theta^*) \). This problem is likely to arise when there is some latent structure in the generative model that links the \( Y \) and \( Z \) variables. In these situations, the nearest-neighbour imputation distribution \( g(x_i, y_i, z_i) = f(x; \theta) f(y|x; \theta) f(z|x; \theta) \) may not be a good approximation for the true generative model \( f(x, y, z; \theta) \). Even in situations where the generative model is identifiable, the nearest-neighbour scheme may fail to produce statistically sound imputations. We examine this issue for skew-normal distribution and Gaussian mixtures.

### 4.1 Skew-normal

In general, the imputation of skewed data can be challenging relative to the multivariate normal case (Sterne et al., 2009), and the statistical matching of skew-normal data presents some difficulties. Assuming that all variables have some skewness, so all elements of \( \delta \) are nonzero, it is not possible for \( Y \) and \( Z \) to be conditionally independent given \( X \) (Azzalini and Capitanio, 1999). Contrary to the normal case, there is no parameter constraint on \( \mu, \Sigma, \delta \) such that \( f(x_i, y_i, z_i; \mu, \Sigma, \delta) = f(x_i; \mu, \Sigma, \delta) f(y_i|x_i; \mu, \Sigma, \delta) f(z_i|x_i; \mu, \Sigma, \delta) \). The absence of a conditional independence constraint poses problems for nearest-neighbour imputation. To illustrate, we generated data from a three-dimensional skew-normal distribution. The parameters were set as \( \mu = 0, \Sigma = I \) and \( \delta = (1, 3, 5)^T \). The first, second and third dimensions were labelled as \( X, Y \) and \( Z \) respectively. We generated two datasets with \( n_A = n_B = 5000 \). Panels (a) and (b) in Figure 1 compare the true joint \((Y, Z)\) distribution to the nearest-neighbour asymptotic distribution. The nearest-neighbour distribution shows a much weaker linear association between the \( Y \) and \( Z \) variables compared to the true distribution. The contours in panel (b) are more square shaped than the skewed ellipse in panel (a). We applied nearest-neighbour imputation to the simulated dataset to check the correspondence with the theoretical distribution in (b). Panel (c) shows a smoothed scatter plot of the imputed data along with contours from a kernel density estimate as red dashed lines. The contours of the density estimate closely resemble the theoretical contours in (b), and are again more square shaped than the contours of the true \((Y, Z)\) distribution in (a).

### 4.2 Finite mixture models

Finite mixture models are well suited for the statistical analysis of flow cytometry data. To describe the model, let \( \Psi \) contain the mixing proportions \( \pi_1, \ldots, \pi_g \) and parameters \( \theta_1, \ldots, \theta_g \) for each component. The \( j \)th component distribution can be represented as \( f(x, y, z; \theta_j) \) for \( j = 1, \ldots, g \). The distribution function of the \( g \)-component mixture model is given by

\[
 f(x_i, y_i, z_i; \Psi) = \sum_{j=1}^{g} \pi_j f(x_i, y_i, z_i; \theta_j). \tag{39}
\]

It will be helpful to conceptualise the mixture model in terms of a latent cluster indicator \( S_i \sim \text{Categorical}(\pi_1, \ldots, \pi_g) \) for each observation \( i = 1, \ldots, n \). The generative model has the hierarchical
Figure 1: Analysis of nearest-neighbour imputation for skewed data. Panel (a) shows contours of the true \((Y, Z)\) joint distribution. Panel (b) shows contours of the nearest-neighbour imputation distribution \(g(y, z)\). Panel (c) gives a smoothed scatter plot of the imputed data in the simulation. The red-dashed lines given contours from a kernel density estimate using the imputed data. Nearest-neighbour imputation underestimates the \(Y, Z\) correlation.

representation:

\[
X_i, Y_i, Z_i | S_i = h \sim f(x_i, y_i, z_i; \theta_h) \\
S_i \sim \text{Categorical}(\pi_1, \ldots, \pi_g).
\]

It is unlikely for the conditional independence assumption to be appropriate when the generative model is a finite mixture. Given a mixture model with well-separated components, it is not possible to find an appropriate restrictions on the mixture parameters \(\Psi\) such that \(f(x_i, y_i, z_i; \Psi) = f(x_i; \Psi)f(y_i|x_i; \Psi)f(z_i|x_i; \Psi)\). Conditional on \(X\), the remaining \(Y\) and \(Z\) variables are almost certainly dependent due to the latent cluster indicator \(S\). The violation of the conditional independence assumption for finite mixture models means that the nearest-neighbour imputation may have undesirable behaviour. The tendency for nearest-neighbour imputation to produce spurious clusters has been demonstrated empirically in statistical matching problems in flow cytometry analysis (Lee et al., 2011; O’Neill et al., 2015).

From the analysis of the asymptotic model, spurious clusters emerge when the nearest-neighbour match originates from a different mixture component to the query point \(x_i\). The probability of observing spurious clusters is related to how informative the common \(X\) variables are for classification. If the clusters are poorly separated using the \(X\) variables, then there is a high probability of mismatching observations across clusters and giving improper imputations.

To illustrate, we generated data from an equally weighted two component Gaussian mixture model. The first, second and third dimensions were labelled as the \(X\), \(Y\) and \(Z\) variables respectively. Component 1 had a mean of \((-0.1, 0, 0)^T\) and component 2 had a mean of \((0.1, 1, 1)^T\). Each component had covariance matrix \(0.01 I\). The components are well separated in the marginal \(Y\) and \(Z\) dimensions and there is a large amount of overlap in the \(X\) dimension. We generated \(n_A = n_B = 500\)
Figure 2: Analysis of nearest-neighbour imputation for Gaussian mixture data. Panel (a) shows contours of the true \((Y, Z)\) joint distribution. Panel (b) shows contours of the nearest-neighbour imputation distribution \(g(y, z)\). Panel (c) gives a shows the imputed data from the simulation. Nearest-neighbour imputation introduces spurious clusters. Observations in each dataset. Panels (a) and (b) in Figure 2 compare the true distribution on the \(Y\) and \(Z\) variables to the asymptotic nearest neighbour distribution. The nearest-neighbour distribution shows four clusters in the joint distribution when there should only be two. This is because the \(X\) variables contain limited information for clustering. Panel (c) shows the results of applying nearest-neighbour matching to the simulated dataset. The introduction of two spurious clusters is consistent with the predicted behaviour from panel (b).

5 Examples

5.1 Skew-normal

We return to the skew-normal example in Section 4.1. We simulated another two datasets with \(n_A = n_B = 500\). The parameters were again set as \(\mu = 0\), \(\Sigma = I\) and \(\delta = (1, 3, 5)^T\). The first, second and third dimensions were labelled as \(X\), \(Y\) and \(Z\) respectively. We applied nearest-neighbour matching and parametric imputation using the constraint \(\Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ}\). Parameters were estimated using maximum likelihood. The results are shown in Figure 3. In this example, there is a large difference between the nearest-neighbour imputations (b) and the parametric imputations (c). The true sample correlation between the \(Y\) and \(Z\) variables is \(\hat{\rho}_{YZ} = 0.83\). Nearest-neighbour imputation gives an underestimate, with \(\hat{\rho}_{YZ} = 0.14\). Parametric imputation gives much better estimate \(\hat{\rho}_{YZ} = 0.85\). The \((Y, Z)\) distribution is recoverable in this scenario, as the generative model satisfies the identification restriction \(\Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XZ}\).
Figure 3: Comparison of imputation strategies on skew-normal data.

6 Data application

6.1 Iris dataset

We analysed a Anderson’s iris dataset [Anderson 1935, Fisher 1936] using a Gaussian mixture model. Figure 4 shows a pairs plot of the dataset with colour and shape giving the species labels.

We considered two matching scenarios, under the assumption that the species labels were unavailable. These are described in Table 2. Nearest-neighbour matching was compared to parametric imputation using the same identification constraints in Section 4.2. An important difference between the two scenarios is that in Scenario 1, the $X$ single variable sepal does not give enough information to reliably classify the observations into species. In Scenario 2, the sepal length and sepal width together give enough information to reliably classify the observations into species.

Figure 5 shows the results for Scenario 1. The asymptotic analysis showed that when the $X$ variables are uninformative for clustering, we expect nearest-neighbour imputation to produce spurious clusters. In this example, we encounter mismatched species labels across datasets A and B when using sepal width to find nearest-neighbours. The parametric strategy does not create spurious clusters. The restrictions on the model based clustering approach encode a different set of assumptions that are more reasonable for a Gaussian mixture. As such, the parametric model imputed data shows more fidelity with the true data.

|   | $X$               | $Y$            | $Z$            |
|---|-------------------|----------------|----------------|
| 1 | Sepal Width       | Petal Length   | Petal Width    |
| 2 | Sepal Width, Sepal Length | Petal Length | Petal Width    |

Table 2: Matching scenarios using the iris dataset. Clustering using the $X$ variables is substantially easier in Scenario 2.

Figure 6 shows the results for Scenario 2. In this case nearest-neighbour imputation does not produce any spurious clusters. Spurious clusters are not expected here as the addition of sepal length into the common $X$ group gives enough information to cluster the observations. The additional information should greatly reduce the number of mismatched species labels when finding nearest-
Figure 4: Iris dataset. Black squares, red triangles and blue circles denote observations from the species ‘setosa’, ‘virginica’ and ‘versicolor’ respectively.
neighbours across datasets. Parametric imputation gives very similar results to the nearest-neighbour method in Scenario 2. In both scenarios the imputed data shows weaker correlations than in the true dataset, this is because the component-wise identification restriction $\Sigma_Z^{(h)} = \Sigma_Y^{(h)} - \Sigma_X^{(h)} - 1\Sigma^{(h)}_X$ for $h = 1, \ldots, g$ is perhaps not appropriate for this dataset.

Figure 5: Actual and imputed data in Scenario 1 for the iris dataset.

Figure 6: Actual and imputed data in Scenario 2 for the iris dataset.

6.2 Flow cytometry data (single cell subpopulation)

We analysed a subset of flow cytometry data from a study on graft vs host disease [Brinkman et al. 2007]. See the supplementary material for more information on the dataset. The dataset is displayed in Figure 7. The data subset has $n = 1162$ observations on $p = 3$ markers. The dataset was split into two datasets of $n_A = n_B = 681$ observations. We took FL1.H, FL4.H and FL3.H as the $X$, $Y$ and $Z$ variables respectively. We applied nearest-neighbour matching and parametric imputation with a skew-normal model using the identification restriction described in Section 3.2. Figure 8 compares the imputed data to the actual data. Nearest-neighbour imputation gives different results than the parametric approach. The nearest-neighbour imputed data seems to exhibit the broad shape of the asymptotic imputation distribution that was studied in Section 4.1. If the skew-normal
model is appropriate we expect to see weaker joint associations between the $Y$ and $Z$ variables in the imputed dataset than in the true dataset. It appears that the parametric model does a better job of recovering the general shape of the $(Y, Z)$ joint distribution. The true sample correlation between the $(Y, Z)$ variables is $\rho_{YZ} = 0.39$. Parametric imputation gives a better estimate of the correlation ($\hat{\rho}_{YZ} = 0.44$) compared to nearest-neighbour imputation ($\hat{\rho}_{YZ} = 0.03$).

| Scenario 1 | Scenario 2 |
|------------|------------|
| $X$        | $Y$        | $Z$        |
| FL1.H      | FL4.H      | FL2.H      |
| FL3.H      | FL4.H      | FL2.H      |

Table 3: Matching scenarios using the two cell subpopulation flow cytometry dataset. Clustering using the $X$ variables is substantially easier in Scenario 2.

Figure 7: Flow cytometry dataset with a single cell subpopulation.

6.3 Flow cytometry data (two cell subpopulations)

We also consider another dataset from the graft vs host study. We extracted data on two cell subpopulations that were not considered in the previous data example. The data subset has $n = 1321$ observations on $p = 3$ markers. The dataset was split into two datasets of $n_A = 660$, $n_B = 661$ observations, and is plotted in Figure 9. We considered two different matching scenarios described in Table 6.3.

We applied nearest-neighbour matching and parametric imputation with a skew-normal mixture
model in each scenario. Figure 8 compares the imputed data to the actual and data in Scenario one. The matching variable FL3.H is not sufficiently informative for clustering in Scenario 1, and nearest-neighbour imputation introduces spurious clusters. Focusing on panel (b), we see that the majority of points from the second cell subpopulation (red diamonds) are no longer located in the top right corner of the plot. Nearest-neighbour matching shifts most observations in cell subpopulation 2 to either the top-left or bottom-right of the plot. This is a serious distortion of the joint relationship between the FL4.H and FL2.H variables that is present in the original dataset. Looking at panel (c) we see that parametric imputation does not suffer from this problem. We recover the correct general location of each cell cluster in the joint space of the FL4.H and FL2.H variables. Although the location of the second cell subpopulation is recovered well, we do not preserve the correct orientation of the second cell subpopulation.

Table 4 reports sample correlations for each cell subpopulation in Scenario one. Nearest-neighbour imputation gives poor estimates of the subpopulation correlations due to the mismatching of observations across clusters. In this scenario the parametric imputation strategy underestimates the \( Y, Z \) correlation for both groups.

Figure 9 compares the imputed data to the actual and data in Scenario two. Groupwise correlation estimates are also reported in Table 4. In Scenario 2, the matching variable FL3.H is able to separate the two cell subpopulations. As such, nearest-neighbour matching produces the correct number of clusters, however the correlation estimate in component two (\( \hat{\rho}_{YZ} = 0.13 \)) is much lower than the true correlation in the source dataset (\( \rho_{YZ} = 0.74 \)). Parametric imputation also produces two clusters, but seems to preserve more of the distributional shape than the nearest-neighbour approach. Nearest-neighbour matching gives very small correlation estimates for each component. Parametric imputation gives a good estimate of the correlation in subpopulation 2, but underestimates the correlation in subpopulation 1. It is interesting to compare the results in the two scenarios. Both parametric imputation and nearest-neighbour matching produces different correlation estimates in each scenario. The quality of results appears to be sensitive to the choice of matching variable. The matching variable in Scenario 1 is not informative for clustering and this appears to impact the faithfulness of the imputed data for both methods.
Figure 9: Two cell subpopulation data. Black circles and red triangles denote observations from cell subpopulations 1 and 2 respectively.

| Correlation            | Scenario 1 | Scenario 2 |
|------------------------|------------|------------|
|                        | Component 1 | Component 2 | Component 1 | Component 2 |
| True sample            | 0.10       | 0.74       | 0.10        | 0.74        |
| Nearest-neighbour      | -0.04      | -0.69      | 0.03        | 0.13        |
| Parameteric Model      | 0.16       | 0.04       | 0.01        | 0.81        |

Table 4: Estimates of the $(Y,Z)$ correlation $\rho_{YZ}$ for the two cell subpopulation flow cytometry dataset.

7 Conclusion

In the statistical matching problem, the identifiability of $f(x, y, z; \theta)$ does not necessarily imply the conditional independence of $Y$ and $Z$ given $X$. As such, a parametric imputation strategy can give different results to nearest-neighbour imputation. Although nearest-neighbour imputation is a non-parametric method, it is not assumption free. The nearest-neighbour method makes a conditional independence assumption on the observed random variables that may not be appropriate if the generative model has some latent structure.

We showed that a model based approach to the statistical matching problem for skew-normal data and Gaussian mixture data can be implemented using data augmentation and the EM algorithm. By reparameterising the complete-data likelihood we can establish a relationship between the statistical matching problem and mixtures of regression models. The reparameterisation is important to give closed form E and M steps. Additionally, we can impose identification constraints on the model to
avoid introducing the missing $Y$ and $Z$ observations into the complete-data log likelihood. This is to minimise the computational expense of each EM iteration. The statistical matching of non-Gaussian data can be significantly more computationally demanding than non-Gaussian data, and this is an important consideration when working with large flow cytometry datasets.

The artefacts introduced by the nearest-neighbour method can lead to potential issues in downstream analysis, as the imputer’s model and the analyst’s model should be compatible [Rubin 1996, Meng 1994]. For example, if nearest-neighbour matching introduces spurious clusters, a downstream analyst using mixture models may overestimate the number of subpopulations in the dataset. The conditional independence assumption is not a necessity in order to perform statistical matching, and as such it may be worth exploring alternative non-parametric or semi-parametric imputation methods.

The statistical matching of mixture models is particularly interesting as the group separation on the common $X$ variables appears to strongly influence parameter estimation. In some situations, investigators may be able to engineer the data collection process such that informative variables are
assigned to the common $X$ group. This situation may arise in survey research where respondents are asked different sets of questions, with a smaller set of common questions given to every subject. The selection of common questions $X$ is interesting from an experimental design point of view.

The statistical matching problem is a challenging data integration scenario that can require specialised algorithms for missing data imputation. We have found that model based approaches can give different results to nonparameteric imputation schemes. However, the pathological nature of the statistical matching problem does limit the accuracy of any imputation scheme. The model based approach can introduce some systemic bias into the imputed data when there is model misspecification. The identification restrictions adopted in this work are all inspired by the conditional independence constraint for the Gaussian distribution $\Sigma_{YZ} = \Sigma_{Y|X} = \Sigma_{Y|X} \Sigma^{-1}_{XX} \Sigma_{XZ}$. This assumption is quite strong, and is not testable. We have seen in the data applications that there is no guarantee that it will hold on real data. As such, an important future research direction is the identification of alternative parameter constraints that are testable.

References

Aluja-Banet, T., Daunis-i Estadella, J., and Pellicer, D. (2007). GRAFT, a complete system for data fusion. *Computational Statistics & Data Analysis* **52**, 635–649.

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society* **59**, 2–5.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association* **52**, 200–203.

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.

Arellano-Valle, R. B., Branco, M. D., and Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics* **34**, 581–601.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)* **61**, 579–602.

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R., and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation* **13**, 691–700.

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

D’Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley Series in Survey Methodology. Wiley, New York.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Jones, P. and McLachlan, G. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics* **34**, 233–240.

Lee, G., Finn, W., and Scott, C. (2011). Statistical file matching of flow cytometry data. *Journal of Biomedical Informatics* **44**, 663–676.

Lee, S. X. and McLachlan, G. J. (2016). Finite mixtures of canonical fundamental skew t-distributions. *Statistics and Computing* **26**, 573–589.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley series in probability and statistics. Wiley, Hoboken, 2nd edition.

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association* **50**, 870–876.

Marella, D., Scanu, M., and Luigi Conti, P. (2008). On the matching noise of some nonparametric imputation procedures. *Statistics & Probability Letters* **78**, 1593–1600.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–558.

Moriarity, C. and Scheuren, F. (2003). A note on Rubin’s statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **21**, 65–73.

O’Neill, K., Aghaeepour, N., Parker, J., Hogge, D., Karsan, A., Dalal, B., and Brinkman, R. R. (2015). Deep profiling of multitube flow cytometry data. *Bioinformatics* **31**, 1623–1631.

Pedreira, C. E., Costa, E. S., Barrena, S., Lecrevisse, Q., Almeida, J., van Dongen, J. J. M., and Orfao, A. (2008). Generation of flow cytometry data files with a potentially infinite number of dimensions. *Cytometry Part A* **73**, 834–846.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafer, D. A., et al. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences* **106**, 8519–8524.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association* **67**, 306–310.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics Series. Springer-Verlag.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of multivariate analysis*. Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J., 3rd edition.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.
A Appendix

A.1 Conditional skew-normal

The conditional distributions \( f(y_i|x_i; \theta) \) and \( f(z_i|x_i; \theta) \) belong to a generalised family of skew-normal distributions that also have a latent variable representation (Arellano-Valle and Azzalini, 2006; Arellano-Valle et al., 2006). Let \( TN(\mu, \sigma^2, a) \) denote a lower truncated normal distribution where \( \mu \) and \( \sigma^2 \) give the mean and variance of the underlying normal distribution and \( a \) gives the lower truncation bound. We say \( U \sim TN(\mu, \sigma^2, a) \) if \( U = [W | W > a] \) where \( W \sim N(\mu, \sigma^2) \). The conditional distribution of \( Y \) and \( Z \) given \( X \) can be represented as

\[
\begin{bmatrix}
  Y \\
  Z 
\end{bmatrix} = \begin{bmatrix}
  \mu_{Y|X} \\
  \mu_{Z|X}
\end{bmatrix} + \begin{bmatrix}
  \delta_{Y|X} \\
  \delta_{Z|X}
\end{bmatrix} U_X + \begin{bmatrix}
  V_{Y|X} \\
  V_{Z|X}
\end{bmatrix},
\]

where \( U_X \) has a truncated normal distribution \( TN(\tau_X, \gamma_X, 0) \), and \( V_{Y|X} \) and \( V_{Z|X} \) are jointly normally distributed. The location parameters are given by

\[
\begin{bmatrix}
  \mu_{Y|X} \\
  \mu_{Z|X}
\end{bmatrix} = \begin{bmatrix}
  \mu_Y \\
  \mu_Z
\end{bmatrix} - \begin{bmatrix}
  \Sigma_{YX} \\
  \Sigma_{ZX}
\end{bmatrix}^{-1} \Sigma_{XX}^{-1} (x - \mu_X).
\]

The latent \( U_X \) is distributed as a truncated normal \( TN(\tau_X, \gamma_X, 0) \) random variable, where

\[
\tau_X = \delta_X^T (\Sigma_{XX} + \delta_X \delta_X^T)^{-1} (x - \mu_X),
\]

\[
\gamma_X = 1 - \delta_X^T (\Sigma_{XX} + \delta_X \delta_X^T)^{-1} \delta_X.
\]

The latent \( V_{Y|X} \) and \( V_{Z|X} \) have a normal distribution,

\[
\begin{bmatrix}
  V_{Y|X} \\
  V_{Z|X}
\end{bmatrix} \sim N \left( \begin{bmatrix}
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  \Sigma_{YY|X} & \Sigma_{YZ|X} \\
  \Sigma_{ZY|X} & \Sigma_{ZZ|X}
\end{bmatrix} \right).
\]
The conditional covariance matrix is
\[
\begin{bmatrix}
\Sigma_{YY|X} & \Sigma_{YZ|X} \\
\Sigma_{YX|X} & \Sigma_{ZZ|X}
\end{bmatrix} = \begin{bmatrix}
\Sigma_{YX} & \Sigma_{YZ}
\end{bmatrix} - \begin{bmatrix}
\Sigma_{XX} + \delta_X \delta_X^T
\end{bmatrix}^{-1} \begin{bmatrix}
\Sigma_{XY} & \Sigma_{XZ}
\end{bmatrix} - \frac{1}{\gamma_X} \delta_X \delta_X^T.
\]

The conditional skewness parameters are
\[
\begin{bmatrix}
\delta_{Y|X} \\
\delta_{Z|X}
\end{bmatrix} = \frac{1}{\gamma_X} \begin{bmatrix}
\delta_Y \\
\delta_Z
\end{bmatrix} - \begin{bmatrix}
\Sigma_{XX} + \delta_X \delta_X^T
\end{bmatrix}^{-1} \begin{bmatrix}
\Sigma_{XY} & \Sigma_{XZ}
\end{bmatrix} \begin{bmatrix}
\delta_X \\
\delta_X
\end{bmatrix}.
\]

A.2 Asymptotic analysis of skew-normal

The hierarchical representation of a skew-normal random variable is useful to analyse the asymptotic behaviour of the nearest-neighbour method for skew-normal data. The asymptotic form of the nearest-neighbour imputation distribution (37) will involve the true conditional distributions \(f(y_i|x_i; \theta)\) and \(f(z_i|x_i; \theta)\). The true marginal \((Y_i, Z_i)\) distribution can be expressed through a hierarchical model
\[
Y_i, Z_i | X_i = x_i, U_X = u_X \sim N \left( \begin{bmatrix} \mu_{Y|X} \\ \mu_{Z|X} \end{bmatrix} + \begin{bmatrix} \delta_{Y|X} \\ \delta_{Z|X} \end{bmatrix} u_X, \begin{bmatrix} \Sigma_{YY|X} & \Sigma_{YZ|X} \\ \Sigma_{YX|X} & \Sigma_{ZZ|X} \end{bmatrix} \right),
\]
\[
U_X | X_i = x_i \sim TN(\tau_X, \gamma_X, 0),
\]
\[
X_i \sim f(x_i; \mu_X, \Sigma_{XX}, \delta_X).
\]

We can also characterise the asymptotic nearest-neighbour imputation distribution through a hierarchical model. Asymptotically, the effective model used by nearest-neighbour matching is that conditional on \(X_i = x_i\),
\[
Z_i | X_i = x_i, U_X^{(1)} = u_X^{(1)} \sim N \left( \mu_{Z|X} + \delta_{Z|X} u_X^{(1)}, \Sigma_{ZZ|X} \right),
\]
\[
Y_i | X_i = x_i, U_X^{(2)} = u_X^{(2)} \sim N \left( \mu_{Y|X} + \delta_{Y|X} u_X^{(2)}, \Sigma_{YY|X} \right),
\]
\[
U_X^{(1)}, U_X^{(2)} | X_i = x_i \sim^{i.i.d} TN(\tau_X, \gamma_X, 0),
\]
\[
X_i \sim f(x_i; \mu_X, \Sigma_{XX}, \delta_X).
\]

Comparing to (A.6) to (A.2), we see that an important difference between the nearest-neighbour imputation distribution and the true marginal distribution is the use of two independent latent scaling variables \(U_X^{(1)}\) and \(U_X^{(2)}\) instead of the single latent variable \(U_X\) that appears in the true model. A second difference is that \(Z_i\) and \(Y_i\) are sampled independently conditional on \(X_i\). Levels (A.4) and (A.5) are independent conditional on the lower layers. The conditional covariance \(\Sigma_{YZ|X}\) that appears in the true generative process (A.2) does not influence the nearest-neighbour imputation distribution. With these systematic differences, it is reasonable to expect that nearest-neighbour imputation will underestimate the level of dependence between the \(Y\) and \(Z\) variables. If we believe that the skew-normal model (19) is appropriate, it is difficult to justify the use of nearest-neighbour matching as the asymptotic behaviour is not consistent with our initial assumptions.
A.3 Asymptotic analysis of Gaussian mixture

The true joint distribution on the \((Y_i, Z_i)\) variables has the hierarchical representation:

\[
Y_i, Z_i | X_i = x_i, S_i = h \sim f(y_i, z_i | x_i; \theta_h), \tag{A.8}
\]

\[
S_i | X_i = x_i \sim \text{Categorical}(\tau_1, \ldots, \tau_g), \tag{A.9}
\]

\[
X_i \sim f(x_i; \Psi), \tag{A.10}
\]

where for \(j = 1, \ldots, g:\)

\[
\tau_j = \frac{\pi_j f(x_i; \theta_j)}{\sum_{k=1}^{g} \pi_k f(x_i; \theta_k)}. \tag{A.11}
\]

Although conventional to include the cluster indicator \(S_i\) at the bottom of the hierarchy, for the purposes of our analysis it is beneficial to sample it conditional on \(X_i = x_i\). We again analyse the asymptotic behaviour nearest-neighbour matching by determining the form of the asymptotic imputation distribution \(g(x_i, y_i, z_i)\) \(^{[37]}\). The true conditional distributions \(f(y_i | x_i; \theta)\) and \(f(z_i | x_i; \theta)\) can also be expressed as finite mixture models. In full,

\[
f(y_i | x_i; \Psi) = \sum_{j=1}^{g} \tau_j f(y_i | x_i; \theta_j)
\]

\[
f(z_i | x_i; \Psi) = \sum_{j=1}^{g} \tau_j f(z_i | x_i; \theta_j)
\]

where the weights \(\tau_1, \ldots, \tau_g\) are given by \([A.11]\). Asymptotically speaking, the nearest-neighbour imputation distribution \(g(x_i, y_i, z_i)\) has the hierarchical representation:

\[
Z_i | X_i = x_i, T_i = j \sim f(z_i | x_i; \theta_j) \tag{A.12}
\]

\[
Y_i | X_i = x_i, S_i = h \sim f(y_i | x_i; \theta_h), \tag{A.13}
\]

\[
S_i, T_i | X_i = x_i \sim \text{i.i.d. Categorical}(\tau_1, \ldots, \tau_g) \tag{A.14}
\]

\[
X_i \sim f(x_i; \Psi), \tag{A.15}
\]

An important difference between the nearest-neighbour imputation distribution and the true conditional model is the introduction of a second latent variable \(T_i\) \([A.14]\). If the sampled \(T_i\) is not equal to the sampled \(S_i\) at level \([A.14]\), the \(Y_i\) and \(Z_i\) random variables will not be drawn from the same component distribution in levels \([A.13]\) and \([A.12]\). The probability of a mismatch is related to how reliably the \(X\) variables can be used to classify observations, entering through the \(\tau_j\) terms \([A.11]\). The imputed joint distribution \(g(y, z)\) can introduce spurious clusters if the \(X\) variables are not useful for clustering the observations.
Web Appendix: Introduction

Here we give the E-step and M-step for the EM algorithms for the statistical matching of non-Gaussian data. Recall the hierarchical model for the skew normal distribution:

\[
\begin{bmatrix}
X_i \\
Y_i \\
Z_i
\end{bmatrix} = \begin{bmatrix}
\mu_X \\
\mu_Y \\
\mu_Z
\end{bmatrix} + \begin{bmatrix}
\delta_X \\
\delta_Y \\
\delta_Z
\end{bmatrix} U_i + V_i
\]  \hspace{1cm} (A.16)

where \(U_i \sim TN(0, 1, 0)\) and

\[
V_i \sim N \left( \mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix} \right). \]

The combined data matrix from dataset \(A\) and dataset \(B\) is partitioned as

\[
\begin{bmatrix}
x_1 & y_1 & z_1 \\
\vdots & \vdots & \vdots \\
x_{n_A} & y_{n_A} & z_{n_A} \\
x_{n_A+1} & y_{n_A+1} & z_{n_A+1} \\
\vdots & \vdots & \vdots \\
x_{n_A+n_B} & y_{n_A+n_B} & z_{n_A+n_B}
\end{bmatrix},
\]

where shaded cells are observed and blank cells are missing.

Web Appendix A: Skew-normal distribution

E-step

Using the hierarchical model for the skew normal distribution \([19]\), define the following conditional expectations given the current parameter estimates:

\[
e_{1,i} = \begin{cases} 
E[U_i | x_i, y_i] & \text{if } i = 1, \ldots, n_A, \\
E[U_i | x_i, z_i] & \text{if } i = n_A + 1, \ldots, n
\end{cases}
\]

\[
e_{2,i} = \begin{cases} 
E[U_i^2 | x_i, y_i] & \text{if } i = 1, \ldots, n_A, \\
E[U_i^2 | x_i, z_i] & \text{if } i = n_A + 1, \ldots, n
\end{cases}
\]
Let
\[ m_i = \begin{cases} 
\begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} + \begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \right)^{-1} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} 
& \text{if } i = 1, \ldots, n_A \\
\begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix} + \begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \right)^{-1} \begin{bmatrix} x_i \\ z_i \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix} 
& \text{if } i = n_A + 1, \ldots, n_A + n_B 
\end{cases} \]

Furthermore define the constants
\[ c_1 = \begin{cases} 
1 - \begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} + \begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \begin{bmatrix} \delta_X^T & \delta_Y^T \end{bmatrix} \right)^{-1} \begin{bmatrix} \delta_X \\ \delta_Y \end{bmatrix} 
& \text{if } i = 1, \ldots, n_A \\
1 - \begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix} + \begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \begin{bmatrix} \delta_X^T & \delta_Z^T \end{bmatrix} \right)^{-1} \begin{bmatrix} \delta_X \\ \delta_Z \end{bmatrix} 
& \text{if } i = n_A + 1, \ldots, n_A + n_B 
\end{cases} \]

Let \( \phi(\cdot) \) denote the standard normal density function and let \( \Phi(\cdot) \) denote the cumulative distribution function of the standard normal distribution. We have the closed form expressions for the expectations (Pyne et al., 2009).

\[ e_{1,i} = \begin{cases} 
m_i + c_1 [\phi(m_i/c_1)/\Phi(m_i/c_1)] & \text{if } i = 1, \ldots, n_A \\
m_i + c_2 [\phi(m_i/c_2)/\Phi(m_i/c_2)] & \text{if } i = n_A + 1, \ldots, n_A 
\end{cases} \]

\[ e_{2,i} = \begin{cases} 
m_i^2 + c_1^2 + c_1 m_i [\phi(m_i/c_1)/\Phi(m_i/c_1)] & \text{if } i = 1, \ldots, n_A \\
m_i^2 + c_2^2 + c_2 m_i [\phi(m_i/c_2)/\Phi(m_i/c_2)] & \text{if } i = n_A + 1, \ldots, n_A 
\end{cases} \]

M-step

For \( i = 1, \ldots, n \) let
\[ P_i = (x_i - \mu_X)^T (x_i - \mu_X) - e_{1,i} (\delta_X (x_i - \mu_X)^T) - ((x_i - \mu_X) \delta_X^T) e_{1,i} + e_{2,i} \delta_X \delta_X^T \]
\[ D_i = e_{1,i} (x_i - \mu_X) \]
\[ L_i = (x_i - e_{1,i} \delta_X) \]
The maximisation of the $X$ parameters follows from existing results on the skew normal distribution (Pyne et al., 2009):

$$
\hat{\Sigma}_{XX} = \left( \sum_{i=1}^{n} P_{i} \right),
\hat{\mu}_{X} = \left( \sum_{i=1}^{n} L_{i} \right),
\hat{\delta}_{X} = \left( \sum_{i=1}^{n} D_{i} \right) \left\{ \sum_{i=1}^{n} e_{2,i} \right\}^{-1}.
$$

Recall that $\Gamma_{A}$ and $\Gamma_{B}$ contain the regression parameters for the conditional regression models:

$\Gamma_{A} = [\alpha_{Y} \; \lambda_{Y} \; \beta_{Y}]^{T}$ and $\Gamma_{B} = [\alpha_{Z} \; \lambda_{Z} \; \beta_{Z}]^{T}$.

The regressions for dataset A and dataset B can be written as

$$
Y_{A} = B_{A} \Gamma_{A} + \epsilon_{A},
Z_{B} = B_{B} \Gamma_{B} + \epsilon_{B},
$$

where $\epsilon_{A} \sim MN(I_{n_{A}}, \Omega_{Y})$ and $\epsilon_{B} \sim MN(I_{n_{B}}, \Omega_{Z})$. The design matrices for the regressions now include the latent $U_{i}$ terms. The complete-data design matrices are given by

$$
B_{A} = \begin{bmatrix}
1 & u_{1} & x_{1}^{T} \\
1 & u_{2} & x_{2}^{T} \\
\vdots \\
1 & u_{n_{A}} & x_{n_{A}}^{T}
\end{bmatrix},
B_{B} = \begin{bmatrix}
1 & u_{n_{A}+1} & x_{n_{A}+1}^{T} \\
1 & u_{n_{A}+2} & x_{n_{A}+2}^{T} \\
\vdots \\
1 & u_{n_{A}+n_{B}} & x_{n_{A}+n_{B}}^{T}
\end{bmatrix}.
$$

(A.19)

We need to calculate the conditional expectations $E[B_{A}^{T}B_{A}]$ and $E[B_{B}^{T}B_{B}]$ given the current parameter estimates and the observed data. For $i = 1, \ldots, n$ let

$$
G_{i} = \begin{bmatrix}
1 & e_{1,i} \\
x_{i}
\end{bmatrix} \begin{bmatrix}
1 & e_{1,i} \\
x_{i}^{T}
\end{bmatrix} + \begin{bmatrix}
0 & 0 & 0 \\
0 & e_{2,i} - e_{1,i} & 0 \\
0 & 0 & 0
\end{bmatrix}.
$$

From standard results on multiple outcome regression models (Rencher and Christensen, 2012), we have a closed form M-step for the regression parameters

$$
\hat{\Gamma}_{Y} = \left( \sum_{i=1}^{n_{A}} G_{i} \right)^{-1} \left( \sum_{i=1}^{n_{A}} \begin{bmatrix}
1 \\
x_{i}
\end{bmatrix} y_{i}^{T} \right),
\hat{\Gamma}_{X} = \left( \sum_{i=n_{A}+1}^{n} G_{i} \right)^{-1} \left( \sum_{i=n_{A}+1}^{n} \begin{bmatrix}
1 \\
x_{i}
\end{bmatrix} z_{i}^{T} \right).
$$
For $i = 1, \ldots, n$ let

$$R_i = \begin{cases} y_i y_i^T - y_i \begin{bmatrix} 1 & e_{1,i} & x_i^T \end{bmatrix} \hat{\Gamma}_Y - \hat{\Gamma}_Y^T \begin{bmatrix} 1 & e_{1,i} & x_i^T \end{bmatrix} & y_i^T + \hat{\Gamma}_Y^T G_i \hat{\Gamma}_Y, \\ z_i z_i^T - z_i \begin{bmatrix} 1 & e_{1,i} & x_i^T \end{bmatrix} \hat{\Gamma}_Z - \hat{\Gamma}_Z^T \begin{bmatrix} 1 & e_{1,i} & x_i^T \end{bmatrix} & z_i^T + \hat{\Gamma}_Z^T G_i \hat{\Gamma}_Z. \end{cases}$$

The M-step for the error covariance matrices in the regressions (A.17) and (A.18) is as follows:

$$\hat{\Omega}_Y = n_A^{-1} \left( \sum_{i=1}^{n_A} R_i \right)$$

$$\hat{\Omega}_Z = n_B^{-1} \left( \sum_{i=n_A+1}^{n} R_i \right).$$

The estimates in terms of the original parameters of the skew-normal density can be obtained by substituting into the following equations:

$$\mu_Y = \alpha_Y + \beta_Y \mu_X$$

$$\mu_Z = \alpha_Z + \beta_Z \mu_X$$

$$\delta_Y = \lambda_Y + \beta_Z \delta_X$$

$$\delta_Z = \lambda_Z + \beta_Z \delta_X$$

$$\Sigma_{YX} = \beta_Y \Sigma_{XX}$$

$$\Sigma_{ZX} = \beta_Z \Sigma_{XX}$$

$$\Sigma_{YY} = \Omega_Y + \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{YX}^T$$

$$\Sigma_{ZZ} = \Omega_Z + \Sigma_{ZX} \Sigma_{XX}^{-1} \Sigma_{ZX}^T$$

$$\Sigma_{YZ} = \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{ZX}.$$
Web Appendix B: Mixtures of Gaussians

E-step

For $h = 1, \ldots, g$, define the posterior class probabilities as

$$
\tau_{i,h} = \frac{N \left( \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} ; \mu^{(h)}_X, \mu^{(h)}_Y, \mu^{(h)}_Z \right) \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix} : \Sigma^{(h)}_{XX}, \Sigma^{(h)}_{XY}, \Sigma^{(h)}_{YY} \right) \cdot \begin{bmatrix} 1 \\ y_i \end{bmatrix} : \Sigma^{(h)}_{YY}, \Sigma^{(h)}_{YZ}, \Sigma^{(h)}_{ZZ} \right)}{\sum_{r=1}^{g} N \left( \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} ; \mu^{(r)}_X, \mu^{(r)}_Y, \mu^{(r)}_Z \right) \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix} : \Sigma^{(r)}_{XX}, \Sigma^{(r)}_{XY}, \Sigma^{(r)}_{YY} \right) \cdot \begin{bmatrix} 1 \\ y_i \end{bmatrix} : \Sigma^{(r)}_{YY}, \Sigma^{(r)}_{YZ}, \Sigma^{(r)}_{ZZ} \right)}^{-1}
$$

if $i = 1, \ldots, n_A$.

M-step

For the maximisation of the $X$ parameters we follow standard results for Gaussian mixture models.

$$
\hat{\mu}^{(h)}_X = \left( \sum_{i=1}^{n} \tau_{i,h} x_i \right) \left( \sum_{i=1}^{n} \tau_{i,h} \right)^{-1}
$$

$$
\hat{\Sigma}^{(h)}_{XX} = \left( \sum_{i=1}^{n} \tau_{i,h} (x_i - \hat{\mu}^{(h)}_X) (x_i - \hat{\mu}^{(h)}_X)^T \right) \left( \sum_{i=1}^{n} \tau_{i,h} \right).
$$

Recall the conditional regression specification for dataset A,

$$
Y_i | X = x_i, S_i = h \sim N(\alpha^{(h)}_Y + \beta^{(h)}_Y X_i, \Omega^{(h)}_Y), \quad (A.20)
$$

where $\beta^{(h)}_Y = \Sigma^{(h)}_{YX} [\Sigma^{(h)}_{XX}]^{-1}$, $\alpha^{(h)}_Y = \mu^{(h)}_Y - \beta^{(h)}_Y \mu^{(h)}_X$ and $\Omega^{(h)}_Y = \Sigma^{(h)}_{YY} - \Sigma^{(h)}_{YX} [\Sigma^{(h)}_{XX}]^{-1} \Sigma^{(h)}_{XY}$. A similar model applies for dataset B,

$$
Z_i | X = x_i, S_i = h \sim N(\alpha^{(h)}_Z + \beta^{(h)}_Z X_i, \Omega^{(h)}_Z), \quad (A.21)
$$

where $\beta^{(h)}_Z = \Sigma^{(h)}_{ZX} [\Sigma^{(h)}_{XX}]^{-1}$, $\alpha^{(h)}_Z = \mu^{(h)}_Z - \beta^{(h)}_Z \mu^{(h)}_X$ and $\Omega^{(h)}_Z = \Sigma^{(h)}_{ZZ} - \Sigma^{(h)}_{ZX} [\Sigma^{(h)}_{XX}]^{-1} \Sigma^{(h)}_{XZ}$. The maximisation for the conditional regression parameters then follows from existing results on mixtures of regression models (Jones and McLachlan [1992]). For $i = 1, \ldots, n$ define:

$$
G_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix} [1 \ x_i^T],
$$

The $M$-step for the regression coefficients is given by:

$$
\hat{\Gamma}^{(h)}_Y = \left( \sum_{i=1}^{n_A} \tau_{i,h} G_i \right)^{-1} \left( \sum_{i=1}^{n_A} \tau_{i,h} \begin{bmatrix} 1 \\ x_i \end{bmatrix} y_i^T \right), \quad h = 1, \ldots, g.
$$

$$
\hat{\Gamma}^{(h)}_Z = \left( \sum_{i=n_A+1}^{n} \tau_{i,h} G_i \right)^{-1} \left( \sum_{i=n_A+1}^{n} \tau_{i,h} \begin{bmatrix} 1 \\ x_i \end{bmatrix} z_i^T \right), \quad h = 1, \ldots, g.
$$

31
For the residual error variances in (A.20) and (A.21) we have

\[ \hat{\Omega}^{(h)}_Y = \frac{1}{\sum_{i=1}^{n_A} \tau_{i,h}} \sum_{i=1}^{n_A} \tau_{i,h} \left( y_i - \Gamma^{(h)T}_Y \begin{bmatrix} 1 \\ x_i \end{bmatrix} \right) \left( y_i - \hat{\Gamma}^{(h)T}_Y \begin{bmatrix} 1 \\ x_i \end{bmatrix} \right)^T, \quad h = 1, \ldots, g. \]

\[ \hat{\Omega}^{(h)}_Z = \frac{1}{\sum_{i=n_A+1}^{n} \tau_{i,h}} \sum_{i=n_A+1}^{n} \tau_{i,h} \left( z_i - \Gamma^{(h)T}_Z \begin{bmatrix} 1 \\ x_i \end{bmatrix} \right) \left( z_i - \hat{\Gamma}^{(h)T}_Z \begin{bmatrix} 1 \\ x_i \end{bmatrix} \right)^T. \quad h = 1, \ldots, g. \]

For \( h = 1, \ldots, g \), the maximum likelihood estimates of the original parameters can then be obtained by substituting into the following formulae:

\[ \mu^{(h)}_Y = \alpha^{(h)}_Y + \beta^{(h)}_Y \mu^{(h)}_X \]
\[ \mu^{(h)}_Z = \alpha^{(h)}_Z + \beta^{(h)}_Z \mu^{(h)}_X \]
\[ \Sigma^{(h)}_Y = \beta^{(h)} Y \Sigma^{(h)}_X \]
\[ \Sigma^{(h)}_Z = \beta^{(h)} Z \Sigma^{(h)}_X \]
\[ \Sigma^{(h)}_{YY} = \Omega^{(h)}_Y + \Sigma^{(h)}_Y [\Sigma^{(h)}_{XX}]^{-1} \Sigma^{(h)}_{YX} \]
\[ \Sigma^{(h)}_{ZZ} = \Omega^{(h)}_Z + \Sigma^{(h)}_Z [\Sigma^{(h)}_{XX}]^{-1} \Sigma^{(h)}_{ZX} \]
\[ \Sigma^{(h)}_{YZ} = \Sigma^{(h)}_{YX} [\Sigma^{(h)}_{XX}]^{-1} \Sigma^{(h)}_{ZX}. \]

Web Appendix C: Mixtures of Skew-normal distributions

E-step

For \( h = 1, \ldots, g \), define the posterior class probabilities as

\[ \tau_{i,h} = \begin{cases} f \left( \begin{bmatrix} x_i \\ y_i \end{bmatrix} : \mu^{(h)}_X, \Sigma^{(h)}_{XX}, \Sigma^{(h)}_{YX}, \Sigma^{(h)}_{YY}, \delta^{(h)}_Y \right) \left( \Sigma^{(h)}_{YX} \right)^{-1} \left( f \left( \begin{bmatrix} x_i \\ y_i \end{bmatrix} : \mu^{(h)}_X, \Sigma^{(h)}_{XX}, \Sigma^{(h)}_{YX}, \Sigma^{(h)}_{YY}, \delta^{(h)}_Y \right) \right)^{-1} & \text{if } i = 1, \ldots, n_A \\
\end{cases} \]

\[ \begin{cases} f \left( \begin{bmatrix} x_i \\ z_i \end{bmatrix} : \mu^{(h)}_Z, \Sigma^{(h)}_{ZZ}, \Sigma^{(h)}_{ZX}, \delta^{(h)}_Z \right) \left( \Sigma^{(h)}_{ZX} \right)^{-1} \left( f \left( \begin{bmatrix} x_i \\ z_i \end{bmatrix} : \mu^{(h)}_Z, \Sigma^{(h)}_{ZZ}, \Sigma^{(h)}_{ZX}, \delta^{(h)}_Z \right) \right)^{-1} & \text{if } i = n_A + 1, \ldots, n \\
\end{cases} \]

Where \( f(v; \mu, \Sigma, \delta) \) denotes the skew normal density with parameters \( \mu, \Sigma \) and \( \delta \) evaluated at the point \( v \). Using the hierarchical model for the skew normal distribution (A.16), define the following conditional expectations given:

\[ e^{1,i,h} = \begin{cases} E[U_i | x_i, y_i, S_i = h] & \text{if } i = 1, \ldots, n_A, \\
E[U_i | x_i, z_i, S_i = h] & \text{if } i = n_A + 1, \ldots, n \\
\end{cases} \]

\[ e^{2,i,h} = \begin{cases} E[U_i^2 | x_i, y_i, S_i = h] & \text{if } i = 1, \ldots, n_A, \\
E[U_i^2 | x_i, z_i, S_i = h] & \text{if } i = n_A + 1, \ldots, n \\
\end{cases} \]
For $h = 1, \ldots, g$, define
\[
\begin{align*}
m_{i,h} &= \begin{cases} 
\begin{bmatrix} \delta_X^{(h)T} & \delta_Y^{(h)T} \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX}^{(h)} & \Sigma_{XY}^{(h)} \\ \Sigma_{YX}^{(h)} & \Sigma_{YY}^{(h)} \end{bmatrix} + \begin{bmatrix} \delta_X^{(h)} \\ \delta_Y^{(h)} \end{bmatrix} \right)^{-1} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \mu_X^{(h)} \\ \mu_Y^{(h)} \end{bmatrix} & \text{if } i = 1, \ldots, n_A \\
\begin{bmatrix} \delta_X^{(h)T} & \delta_Z^{(h)T} \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX}^{(h)} & \Sigma_{XZ}^{(h)} \\ \Sigma_{ZX}^{(h)} & \Sigma_{ZZ}^{(h)} \end{bmatrix} + \begin{bmatrix} \delta_X^{(h)} \\ \delta_Z^{(h)} \end{bmatrix} \right)^{-1} \begin{bmatrix} x_i \\ z_i \end{bmatrix} - \begin{bmatrix} \mu_X^{(h)} \\ \mu_Z^{(h)} \end{bmatrix} & \text{if } i = n_A + 1, \ldots, n_A + n_B
\end{cases}
\end{align*}
\]

Furthermore define the constants
\[
\begin{align*}
c_{1,h} &= 1 - \begin{bmatrix} \delta_X^{(h)T} & \delta_Y^{(h)T} \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX}^{(h)} & \Sigma_{XY}^{(h)} \\ \Sigma_{YX}^{(h)} & \Sigma_{YY}^{(h)} \end{bmatrix} + \begin{bmatrix} \delta_X^{(h)} \\ \delta_Y^{(h)} \end{bmatrix} \right)^{-1} \begin{bmatrix} \delta_X^{(h)} \\ \delta_Y^{(h)} \end{bmatrix} \\
c_{2,h} &= 1 - \begin{bmatrix} \delta_X^{(h)T} & \delta_Z^{(h)T} \end{bmatrix} \left( \begin{bmatrix} \Sigma_{XX}^{(h)} & \Sigma_{XZ}^{(h)} \\ \Sigma_{ZX}^{(h)} & \Sigma_{ZZ}^{(h)} \end{bmatrix} + \begin{bmatrix} \delta_X^{(h)} \\ \delta_Z^{(h)} \end{bmatrix} \right)^{-1} \begin{bmatrix} \delta_X^{(h)} \\ \delta_Z^{(h)} \end{bmatrix}
\end{align*}
\]

Let $\phi(\cdot)$ denote the standard normal density function and let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution. We have the closed form expressions for the expectations.
\[
\begin{align*}
e_{1,i,h} &= \begin{cases} 
(m_{i,h} + c_{1,h} [\phi(m_{i,h}/c_{1,h})]/\Phi(m_{i,h}/c_{1,h})) & \text{if } i = 1, \ldots, n_A \\
(m_{i,h} + c_{2,h} [\phi(m_{i,h}/c_{2,h})]/\Phi(m_{i,h}/c_{2,h})) & \text{if } i = n_A + 1, \ldots, n_A
\end{cases}
\end{align*}
\]
\[
\begin{align*}
e_{2,i,h} &= \begin{cases} 
(m_{i,h}^2 + c_{1,h}^2 + c_{1,h} m_{i,h} [\phi(m_{i,h}/c_{1,h})]/\Phi(m_{i,h}/c_{1,h})) & \text{if } i = 1, \ldots, n_A \\
(m_{i,h}^2 + c_{2,h}^2 + c_{2,h} m_{i,h} [\phi(m_{i,h}/c_{2,h})]/\Phi(m_{i,h}/c_{2,h})) & \text{if } i = n_A + 1, \ldots, n_A
\end{cases}
\end{align*}
\]

**M-step**

For $i = 1, \ldots, n$ let
\[
P_{i,h} = (x_i - \mu_X^{(h)}) (x_i - \mu_X^{(h)})^T - e_{1,i,h} (\delta_X^{(h)} (x_i - \mu_X^{(h)})^T) - ((x_i - \mu_X^{(h)}) \delta_X^{(h)T}) e_{1,i,h} + e_{2,i,h} \delta_X^{(h)T} \delta_X^{(h)T}
\]
\[
D_{i,h} = e_{1,i,h} (x_i - \mu_X^{(h)})
\]
\[
L_{i,h} = (x_i - e_{1,i,h} \delta_X^{(h)})
\]
The maximisation of the $X$ parameters follows from existing results on mixtures of skew normal distributions [Pyne et al., 2009]. For $h = 1, \ldots, g$:

$$
\hat{\Sigma}^{(h)}_{X X} = \left( \sum_{i=1}^{n} P_{i,h} \right) \left( \sum_{i=1}^{n} \tau_{i,h} \right)^{-1}
$$

$$
\hat{\mu}_X = \left( \sum_{i=1}^{n} L_{i,h} \right) \left( \sum_{i=1}^{n} \tau_{i,h} \right)^{-1}
$$

$$
\hat{\delta}_X = \left( \sum_{i=1}^{n} D_{i,h} \right) \left\{ \sum_{i=1}^{n} \tau_{i,h} e_{2,i,h} \right\}^{-1}
$$

The complete-data model can be expressed in terms of conditional regression specifications. Let

$$
\beta^{(h)}_Y = \Sigma_{Y X} [\Sigma_{X X}]^{-1}, \alpha^{(h)}_Y = \mu_Y^{(h)} - \beta_Y^{(h)} \mu_X^{(h)}, \Omega_Y^{(h)} = \Sigma_{Y Y} - \Sigma_{Y X} [\Sigma_{X X}]^{-1} \Sigma_{X Y}^{(h)} \text{ and } \lambda_Y^{(h)} = \delta_Y^{(h)} - \beta_Y^{(h)} \delta_X^{(h)}.
$$

$$
Y_i | X_i = x_i, U_i = u_i, S_i = h \sim N(\alpha_Y^{(h)} + \lambda_Y^{(h)} u_i + \beta_Y^{(h)} x_i, \Omega_Y^{(h)}) \quad (A.22)
$$

Let

$$
\beta^{(h)}_Z = \Sigma_{Z X} [\Sigma_{X X}]^{-1}, \alpha^{(h)}_Z = \mu_Z^{(h)} - \beta_Z^{(h)} \mu_X^{(h)}, \Omega_Z^{(h)} = \Sigma_{Z Z} - \Sigma_{Z X} [\Sigma_{X X}]^{-1} \Sigma_{X Z}^{(h)} \text{ and } \lambda_Z^{(h)} = \delta_Z^{(h)} - \beta_Z^{(h)} \delta_X^{(h)}.
$$

Similarly the conditional distribution of $Z_i$ given $X_i$ and the latent scaling variable $U_i$ can be represented as a regression model

$$
Z_i | X_i = x_i, U_i = u_i, S_i = h \sim N(\alpha_Z^{(h)} + \lambda_Z^{(h)} u_i + \beta_Z^{(h)} x_i, \Omega_Z^{(h)}). \quad (A.23)
$$

We need to calculate the conditional expectations of the predictors in the regressions given the current parameter estimates and the observed data. For $i = 1, \ldots, n$ and $h = 1, \ldots, g$ let

$$
G_{i,h} = \begin{bmatrix}
1 & e_{1,i,h} & x_i^T \\
1 & e_{1,i,h} & x_i^T \\
0 & 0 & 0 \\
0 & e_{2,i,h} - e_{1,i,h}^2 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

Recall that $\Gamma_A^{(h)}$ and $\Gamma_B^{(h)}$ contain the regression parameters for component $h$, so $\Gamma_A^{(h)} = [\alpha_Y^{(h)} \lambda_Y^{(h)} \beta_Y^{(h)}]^T$ and $\Gamma_B^{(h)} = [\alpha_Z^{(h)} \lambda_Z^{(h)} \beta_Z^{(h)}]^T$ for $h = 1, \ldots, g$. We have a closed form M-step for the $Y$ regression parameters in (A.22) and the $Z$ regression parameters parameters in (A.23):

$$
\hat{\Gamma}_Y = \left( \sum_{i=1}^{n_A} \tau_{i,h} G_{i,h} \right)^{-1} \left( \sum_{i=1}^{n_A} \tau_{i,h} \begin{bmatrix} 1 \\ e_{1,i,h} \end{bmatrix} y_i^T \right)
$$

$$
\hat{\Gamma}_X = \left( \sum_{i=n_A+1}^{n} \tau_{i,h} G_{i,h} \right)^{-1} \left( \sum_{i=n_A+1}^{n} \tau_{i,h} \begin{bmatrix} 1 \\ e_{1,i,h} \end{bmatrix} z_i^T \right)
$$
For $h = 1, \ldots, g$ define:

$$R_{i,h} = \begin{cases} 
    y_i^T y_i - y_i \begin{bmatrix} 1 & e_{1,i,h} & x_{i,h}^T \end{bmatrix} \hat{\Gamma}_Y - \hat{\Gamma}_Y^T & \text{if } i = 1, \ldots, n_A \\
    z_i^T z_i - z_i \begin{bmatrix} 1 & e_{1,i,h} & x_{i,h}^T \end{bmatrix} \hat{\Gamma}_Z - \hat{\Gamma}_Z^T & \text{if } i = n_A + 1, \ldots, n_A + n_B.
\end{cases}$$

The M-step for the error covariance matrices in the regressions (34) and (35) is as follows:

$$\hat{\Omega}_Y^{(h)} = \frac{1}{\sum_{i=1}^{n_A} \tau_{i,h}} \left( \sum_{i=1}^{n_A} \tau_{i,h} R_{i,h} \right) \quad h = 1, \ldots, g.$$\[\hat{\Omega}_Z^{(h)} = \frac{1}{\sum_{i=n_A+1}^{n} \tau_{i,h}} \left( \sum_{i=n_A+1}^{n} \tau_{i,h} R_{i,h} \right) \quad h = 1, \ldots, g.\]

The estimates in the original parameterisation for the skew-normal model can be obtained using the following results. For $h = 1, \ldots, g$ we have the relationships:

$$\begin{align*}
    \mu_Y^{(h)} &= \alpha_Y^{(h)} + \beta_Y^{(h)} \mu_X^{(h)} \\
    \mu_Z^{(h)} &= \alpha_Z^{(h)} + \beta_Z^{(h)} \mu_X^{(h)} \\
    \delta_Y^{(h)} &= \lambda_Y^{(h)} + \beta_Y^{(h)} \delta_X^{(h)} \\
    \delta_Z^{(h)} &= \lambda_Z^{(h)} + \beta_Z^{(h)} \delta_X^{(h)} \\
    \Sigma_{YX}^{(h)} &= \beta_Y^{(h)} \Sigma_{XX}^{(h)} \\
    \Sigma_{ZX}^{(h)} &= \beta_Z^{(h)} \Sigma_{XX}^{(h)} \\
    \Sigma_{YX}^{(h)} &= \Omega_Y^{(h)} + \Sigma_{YX}^{(h)} [\Sigma_{XX}^{(h)}]^{-1} \Sigma_{YX}^{(h)}^T \\
    \Sigma_{ZZ}^{(h)} &= \Omega_Z^{(h)} + \Sigma_{Z}^{(h)} [\Sigma_{XX}^{(h)}]^{-1} \Sigma_{Z}^{(h)}^T \\
    \Sigma_{YZ}^{(h)} &= \Sigma_{YX}^{(h)} [\Sigma_{XX}^{(h)}]^{-1} \Sigma_{Z}^{(h)}^T .
\end{align*}$$