Computational Statistics and Data Analysis 104 (2016) 79–90

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Partial identification in the statistical matching problem

Daniel Ahfock [a,*], Saumyadipta Pyne [b,c], Sharon X. Lee [a], Geoffrey J. McLachlan [a]

[a] Department of Mathematics, University of Queensland, Australia
[b] Public Health Foundation of India, IIPH Hyderabad, India
[c] CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India

ARTICLE INFO

Article history:
Received 20 November 2015
Received in revised form 15 June 2016
Accepted 15 June 2016
Available online 22 June 2016

Keywords:
Data integration
Missing data
Positive-definite matrix completion
Statistical matching

ABSTRACT

The statistical matching problem involves the integration of multiple datasets where some variables are not observed jointly. This missing data pattern leaves most statistical models unidentifiable. Statistical inference is still possible when operating under the framework of partially identified models, where the goal is to bound the parameters rather than to estimate them precisely. In many matching problems, developing feasible bounds on the parameters is equivalent to finding the set of positive-definite completions of a partially specified covariance matrix. Existing methods for characterising the set of possible completions do not extend to high-dimensional problems. A Gibbs sampler to draw from the set of possible completions is proposed. The variation in the observed samples gives an estimate of the feasible region of the parameters. The Gibbs sampler extends easily to high-dimensional statistical matching problems.

1. Introduction

The statistical matching problem involves the integration of multiple datasets where we have a set of variables common to all datasets, and other variables which only appear in some datasets. In the simplest terms, we have two samples A and B of n_A and n_B independent observations, respectively, from the same population. In sample A we have measurements on sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, and in sample B we have observations on variables $\boldsymbol{X}$ and $\boldsymbol{Z}$. Our objective is to recover the joint density function $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ from the lower dimensional datasets. The statistical matching problem is a special class of a missing data problem, where the defining characteristic is that we have no joint observations of $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

We often assume that the joint density function belongs to some parametric family $\{f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$, where Ω denotes some parameter space. The objective is to perform statistical inference on the parameter $\boldsymbol{\theta}$. Because of the missing data structure in the statistical matching scenario some of the parameters may be unidentifiable. Statistical inference is still possible if the model is viewed as a partially identified model. The concept of partially identified models stems from the belief that identification is not a simple binary issue. In a partially identified model, the range of values that the parameter $\boldsymbol{\theta}$ can take while leaving the observed data likelihood function unchanged is some non-trivial set. Informally, given an infinite dataset, under an identifiable model we can recover the true value of the parameters. In a partially identified model, given an infinite dataset, we are limited to being able to restrict the parameters to some feasible set. In a partially identified model, some elements of $\boldsymbol{\theta}$ may be point-wise identifiable while others are only partially identifiable.

Standard estimation approaches for missing data problems can exhibit pathological behaviour when the model is only partially identified. Use of the EM algorithm (Dempster et al., 1977) is complicated by the fact that the observed data

[*] Correspondence to: MRC Biostatistics Unit, Cambridge, UK.
E-mail address: daniel.ahfock@uqconnect.edu.au (D. Ahfock).

http://dx.doi.org/10.1016/j.csda.2016.06.005

Table 1
Missing data structure in the canonical statistical matching problem. Observed dimensions for each observation have been shaded.

	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Z}$
$\boldsymbol{s}_1^A$	$\boldsymbol{s}_{1X}^A$	$\boldsymbol{s}_{1Y}^A$	-
$\boldsymbol{s}_2^A$	$\boldsymbol{s}_{2X}^A$	$\boldsymbol{s}_{2Y}^A$	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{s}_{n_A}^A$	$\boldsymbol{s}_{n_AX}^A$	$\boldsymbol{s}_{n_AY}^A$	-
$\boldsymbol{s}_1^B$	$\boldsymbol{s}_{1X}^B$	-	$\boldsymbol{s}_{1Z}^B$
$\boldsymbol{s}_2^B$	$\boldsymbol{s}_{2X}^B$	-	$\boldsymbol{s}_{2Z}^B$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{s}_{n_B}^B$	$\boldsymbol{s}_{n_BX}^B$	-	$\boldsymbol{s}_{n_BZ}^B$

likelihood will not have a unique maximiser, the likelihood will have a ridge over the allowable range of the partially identified parameters. It can be shown that the EM parameter estimates will converge to values that are located on likelihood ridge, and that the limiting estimates depend on the choice of initial values (Schafer, 1997, p. 53). This phenomenon is illustrated for the statistical matching problem in Section 4.2.2 of Rässler (2002). The high sensitivity to the initial values complicates the interpretation of the single point estimate returned by the EM algorithm. Bayesian approaches easily extend to partially identified models, however the posterior distribution can be highly sensitive to the prior, even in large samples (Gustafson, 2015). Credible intervals can also have very poor frequentist coverage (Moon and Schorfheide, 2012). We will pursue a frequentist strategy for estimating the partially identified parameters.

In the statistical matching problem, the partially identified parameters are often elements of a covariance matrix. It is typical to have all elements of the covariance matrix identifiable, other than the values that require joint observations on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. In this setting, estimating the identified set corresponds to determining the set of positive-definite completions of a partially specified covariance matrix. Existing methods for doing so are not applicable when both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are multivariate (D'Orazio, 2015). We take a new sampling based approach to characterising the identified set which is easily applicable to high-dimensional problems. We propose a Gibbs sampler to draw values uniformly from the identified set of covariance parameters. The range of the sampled values gives a direct measure of the uncertainty attached to the partially identified parameters. The Gibbs sampler extends the range of datasets that can be analysed using the statistical matching methodology.

2. The statistical matching problem

A standard mathematical description of the statistical matching problem is as follows (Rässler, 2002). Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be multivariate random variables with joint density function $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$. Assume we have a sample of n_A i.i.d. observations distributed according to $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$, which we will call file A, and another independent sample of size n_B from $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$, which we will call file B. Let $\boldsymbol{s}_i^A$ be a row vector representing the ith observation in file A for $i = 1, \dots, n_A$. Similarly, let $\boldsymbol{s}_j^B$ be a row vector representing the jth observation in file B for $j = 1, \dots, n_B$. The ith observation in file A can be written as $\boldsymbol{s}_i^A = (\boldsymbol{s}_{iX}^A, \boldsymbol{s}_{iY}^A, \boldsymbol{s}_{iZ}^A)$, where $\boldsymbol{s}_{iX}^A$ is a row vector representing the value of $\boldsymbol{X}$ and $\boldsymbol{s}_{iY}^A$, $\boldsymbol{s}_{iZ}^A$ are row vectors representing the values of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively. We can also form an identical partition $\boldsymbol{s}_j^B = (\boldsymbol{s}_{jX}^B, \boldsymbol{s}_{jY}^B, \boldsymbol{s}_{jZ}^B)$ for observation j in file B. Let the observations in file A have the $\boldsymbol{Z}$ values missing and the observations in file B have the $\boldsymbol{Y}$ values missing. Table 1 represents the data matrix in the statistical matching problem. We can consider inference in the statistical matching problem to be inference under a partially identified model. We call a model partially identified if the observed data likelihood is flat for a range of the parameters (Tamer, 2010). The identified set for a parameter is the range of values it can take without altering the observed data likelihood function. We use the notation $\boldsymbol{\Theta}(\boldsymbol{\theta}_j)$ to denote the identified set for parameter $\boldsymbol{\theta}_j$. When analysing a partially identified model we are interested in forming a set of plausible values for the non point-identified parameters. For example, assume we have observations $(X, Y, Z)^{\mathsf{T}}$ from a trivariate normal distribution,

$$N_3\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{bmatrix}\right),$$

and the standard statistical matching problem applies. The likelihood function formed from the observed data will not depend on σ_{YZ}, and so σ_{YZ} can be considered to be a partially identified parameter. All the parameters are point-wise identifiable other than σ_{YZ}. Even though we do not have any data to estimate σ_{YZ} from, as we do not observe Y and Z jointly, our modelling assumptions induce non-trivial bounds on the parameter. Given the other parameters, the possible values that σ_{YZ} can take are limited to those which result in a positive-definite covariance matrix for the underlying trivariate normal distribution. The identified set for the parameter σ_{YZ} is given by

$$\Theta(\sigma_{YZ}) = \left\{ \sigma_{YZ} : \begin{bmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{bmatrix} \text{ is positive-definite} \right\}.$$

In this example we will obtain an interval for σ_{YZ} which is a function of the other covariance parameters (Rässler, 2002). When estimating parameters from data, consistent estimators for the identified parameters allow us to construct a consistent estimator of the identified set. In the trivariate normal example above, we could use the maximum likelihood estimates for the identifiable parameters $\sigma_{XX}, \sigma_{XY}, \sigma_{YY}, \sigma_{ZZ}$ and σ_{XZ}. This estimated identified set can be used to gauge the amount of uncertainty surrounding the partially identified parameters (Conti et al., 2013).

Characterising the identified set has been a difficult problem in statistical matching. Most investigations only consider multivariate normal data and that will be our focus for now. We first discuss previous work on statistical matching before presenting our Gibbs sampler.

3. Matching in the multivariate normal case

We now assume that $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The joint distribution can be represented as

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \end{bmatrix} \sim N_d \left(\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZX} & \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix} \right),$$

where we have applied an obvious partition of the parameters. Given the other parameters in the model, the identified set for $\boldsymbol{\Sigma}_{YZ}$ is

$$\Theta(\boldsymbol{\Sigma}_{YZ}) = \left\{ \boldsymbol{\Sigma}_{YZ} : \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZX} & \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix} \text{ is positive-definite} \right\}. \tag{1}$$

We can expand the matrices $\boldsymbol{\Sigma}_{YZ}$, $\boldsymbol{\Sigma}_{YX}$ and $\boldsymbol{\Sigma}_{XZ}$ as follows

$$\boldsymbol{\Sigma}_{YZ} = \begin{bmatrix} \sigma_{Y_1Z_1} & \sigma_{Y_1Z_2} & \cdots & \sigma_{Y_1Z_{d_Z}} \\ \sigma_{Y_2Z_1} & \sigma_{Y_2Z_2} & \cdots & \sigma_{Y_2Z_{d_Z}} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{Y_{d_Y}Z_1} & \sigma_{Y_{d_Y}Z_2} & \cdots & \sigma_{Y_{d_Y}Z_{d_Z}} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{YX} = \begin{bmatrix} \sigma_{Y_1X_1} & \sigma_{Y_1X_2} & \cdots & \sigma_{Y_1X_{d_X}} \\ \sigma_{Y_2X_1} & \sigma_{Y_2X_2} & \cdots & \sigma_{Y_2X_{d_X}} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{Y_{d_Y}X_1} & \sigma_{Y_{d_Y}X_2} & \cdots & \sigma_{Y_{d_Y}X_{d_X}} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{XZ} = \begin{bmatrix} \sigma_{X_1Z_1} & \sigma_{X_1Z_2} & \cdots & \sigma_{X_1Z_{d_Z}} \\ \sigma_{X_2Z_1} & \sigma_{X_2Z_2} & \cdots & \sigma_{X_2Z_{d_Z}} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{X_{d_X}Z_1} & \sigma_{X_{d_X}Z_2} & \cdots & \sigma_{X_{d_X}Z_{d_Z}} \end{bmatrix}.$$

Here $\sigma_{X_iZ_j}$ denotes the covariance between X_i and Z_j. Finding an explicit expression for values of the matrix $\boldsymbol{\Sigma}_{YZ}$ that satisfy the condition in (1) is an open problem (Rässler and Kiels, 2009). For multivariate $\boldsymbol{X}$, $\boldsymbol{Y}$ and univariate Z, the identified set can be shown to be the interior of an ellipsoid.

Assuming without loss of generality $\boldsymbol{\Sigma}$ is a correlation matrix, we wish to find the identified set of correlations. For univariate Z, $\boldsymbol{\Sigma}_{YZ}$ is a column vector and the ellipsoid is governed by the equation

$$\left(\boldsymbol{\Sigma}_{YZ} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}\right)^{\mathsf{T}} \boldsymbol{A} \left(\boldsymbol{\Sigma}_{YZ} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}\right) = 1, \tag{2}$$

where $\boldsymbol{A} = (1 - \boldsymbol{\Sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ})^{-1} \cdot (\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY})^{-1}$ (Rässler and Kiels, 2009). In the case of univariate Y and Z, as considered by Moriarity and Scheuren (2001), this reduces to the interval $[C - \sqrt{W}, C + \sqrt{W}]$, where

$$C = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}, \tag{3}$$

$$W = \left(1 - \boldsymbol{\Sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}\right) \cdot \left(1 - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\right). \tag{4}$$

A simple rescaling of these formulas gives the identified set of the covariances.

The true values of the identifiable parameters can be substituted into these expressions to determine the allowable range for the partially identified parameter $\boldsymbol{\Sigma}_{YZ}$. This represents the absolute limit of the possible knowledge we can obtain about the joint relationship of $\boldsymbol{Y}$ and Z from our data (D'Orazio et al., 2006). When estimating parameters from data, we will assume that we have some consistent estimators of the model parameters that can be used to obtain a consistent estimate of the identified set. As there are no closed form expressions for the identified set when both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are multivariate, this direct approach to estimating the identified set cannot be used. When $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are both multivariate, numerical methods have been proposed for finding admissible completions of the covariance matrix. These methods involve grid search techniques which can be very inefficient in high dimensions (Moriarity and Scheuren, 2001; D'Orazio et al., 2006).

We suggest an alternative sampling based approach to characterising the identified set. We propose using a Gibbs sampler to draw values uniformly from the identified set. The range of observed values can then be used to infer the amount of uncertainty attached to each parameter. The general idea of sampling from the identified set when an analytical representation is intractable is also suggested in the unpublished work of Kline and Tamer (2015). Kline and Tamer (2015) do not consider the statistical matching problem, and give the broad recommendation to use slice sampling (Neal, 2003) to draw values from the identified set. We present a concrete Gibbs algorithm for the statistical matching problem.

The desire to fit a joint model in statistical matching is often to impute the missing data for downstream analyses. Multiple imputation is desirable to reflect the uncertainty introduced by the missing data. Initial work in this vein by Kadane (1978) and Rubin (1986) has been extended by Moriarity and Scheuren (2003) and Rässler (2003). These multiple imputation procedures often require the analyst to specify a range of values in the identified set. The multiple imputation procedure is thus somewhat ad-hoc, as there is no guarantee that the range of imputed datasets fully captures the uncertainty over the partially identified parameters.

For multivariate normal data, the statistical matching problem reduces to finding positive-definite completions of a partially specified covariance matrix. Finding the range of plausible values is important to accurately gauge the amount of uncertainty introduced into the statistical analysis by the missing data (Rodgers, 1984). Existing methods for characterising the identified set rely on mathematical formulas which have not been extended to problems with both multivariate $\boldsymbol{Y}$ and $\boldsymbol{Z}$. We will develop a Gibbs sampler that generalises easily to high-dimensional problems and simultaneously addresses the need for a principled method to generate values from the identified set.

4. Methods

Gibbs sampling is a powerful tool for sampling from constrained sets (Gelfand et al., 1992). Finding values that lie within a complex high-dimensional restricted set can be difficult. The full conditionals are often much easier to deal with as we only have to consider the feasible range of a single parameter. For the statistical matching problem, the full conditionals reduce to the issue of finding the identified interval for a single partially identified parameter. In other terms, the full conditionals are developed from a covariance matrix where only a single term is unspecified. If we wish to sample uniformly from the identified set, we will simply sample uniformly from the identified interval in each conditional distribution. Again without loss of generality, we assume that $\boldsymbol{\Sigma}$ is a correlation matrix.

Before stating the full conditional distributions we introduce some notations and definitions. Let $\sigma^{(-)}_{Y_rZ_s}$ denote all the elements of $\boldsymbol{\Sigma}_{YZ}$ other than $\sigma_{Y_rZ_s}$ for $r \in \{1, \dots, d_Y\}$ and $s \in \{1, \dots, d_Z\}$. Given r and s let

$$\widetilde{\boldsymbol{X}} = (X_1, X_2, \dots, X_{d_X}, Y_1, \dots, Y_{r-1}, Y_{r+1}, Y_{d_Y}, Z_1, \dots, Z_{s-1}, Z_{s+1}, Z_{d_Z}).$$

The dummy random variable $\widetilde{\boldsymbol{X}}$ represents all variables other than Y_r and Z_s. We also define $\widetilde{Y} = Y_r$ and $\widetilde{Z} = Z_s$. We let $\boldsymbol{\Sigma}_{\widetilde{X}\widetilde{X}}$ denote the correlation matrix of $\widetilde{\boldsymbol{X}}$. We also let $\boldsymbol{\Sigma}_{\widetilde{Y}\widetilde{X}}$ denote the row vector containing the correlations between $\widetilde{Y}$ and $\widetilde{\boldsymbol{X}}$. Finally let $\boldsymbol{\Sigma}_{\widetilde{Z}\widetilde{X}}$ denote the row vector containing the correlations between $\widetilde{\boldsymbol{X}}$ and $\widetilde{Z}$.

For all $r \in \{1, \dots, d_Y\}$ and $s \in \{1, \dots, d_Z\}$, the full conditional distribution is given by

$$p\left(\sigma_{Y_rZ_s} \,|\, \sigma^{(-)}_{Y_rZ_s}\right) \sim \operatorname{unif}\left(\widetilde{C} - \sqrt{\widetilde{W}}, \widetilde{C} + \sqrt{\widetilde{W}}\right), \tag{5}$$

where

$$\widetilde{C} = \boldsymbol{\Sigma}_{\widetilde{Y}\widetilde{X}} \boldsymbol{\Sigma}^{-1}_{\widetilde{X}\widetilde{X}} \boldsymbol{\Sigma}_{\widetilde{X}\widetilde{Z}}, \tag{6}$$

$$\widetilde{W} = \left(1 - \boldsymbol{\Sigma}_{\widetilde{Y}\widetilde{X}} \boldsymbol{\Sigma}^{-1}_{\widetilde{X}\widetilde{X}} \boldsymbol{\Sigma}_{\widetilde{X}\widetilde{Y}}\right)\left(1 - \boldsymbol{\Sigma}_{\widetilde{Z}\widetilde{X}} \boldsymbol{\Sigma}^{-1}_{\widetilde{X}\widetilde{X}} \boldsymbol{\Sigma}_{\widetilde{X}\widetilde{Z}}\right). \tag{7}$$

While the full conditionals are easy to specify and sample from, the gHammersley–Clifford positivity condition does not apply so it is not guaranteed that the Gibbs sampler will converge to the correct stationary distribution (Hammersley and Clifford, 1971). We have to establish that the Markov chain defined by the Gibbs sampler is irreducible. Laurent and Varvitsiotis (2014) show that the identified set as defined in (1) is a convex set. Given fixed values for the other parameters, the identified set for $\boldsymbol{\Sigma}_{YZ}$ can be considered the intersection of the cone of positive-definite matrices with a series of affine subspaces. As the intersection of convex sets is also convex, the identified set will be a convex set. As a consequence, the Markov chain is irreducible, and the Gibbs sampler will converge to the correct stationary distribution.

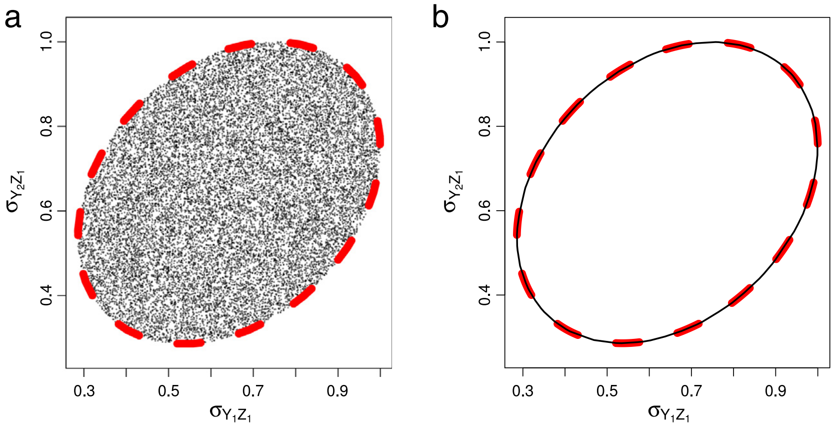

Fig. 1. (a) Draws from the Gibbs sampler as black points. (b) The solid line denotes the convex hull of the Gibbs samples. The dashed ellipse shows the border of the true identified set in both (a) and (b).

The Gibbs sampler requires an initial positive-definite completion of the covariance matrix. We recommend setting $\boldsymbol{\Sigma}_{YZ} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}$, which always provides a positive-definite completion provided that one exists (Grone et al., 1984). Determining an appropriate number of iterations to run the Gibbs sampler is a notoriously difficult problem (Cowles and Carlin, 1996). Roberts and Rosenthal (1998) consider the convergence properties of Gibbs samplers for uniform distributions on bounded regions. They establish that if the boundary satisfies a smoothness condition, the Gibbs sampler will be uniformly ergodic. If we are willing to assume a smoothness condition on the boundary of the identified set, the Gibbs sampler will not necessarily break down in high dimensions. As with general Markov Chain Monte Carlo methods, increasing the dimension of the target distribution can increase the amount of time it takes the sampler to reach the stationary distribution.

When using the proposed methodology to estimate the identified set from data there will be two sources of error, sampling error due to the finite number of observations and Monte Carlo error because the output from the Gibbs sampler is being used to characterise the identified set rather than a closed form representation. The input to the algorithm is a matrix of estimated covariances, that is subject to sampling error. Given the covariance estimates, the original dataset has no influence on the Gibbs algorithm and as such we expect the Monte Carlo error to be largely independent of the sample size. Similar reasoning can also be used to judge the sensitivity of the proposed method to outliers. If a robust estimator is used for the identifiable parameters, outliers should have little influence on the accuracy of the procedure.

A wide variety of models result in an unidentifiable covariance matrix in the statistical matching scenario. In the next section we show how the Gibbs sampler can be applied to matching problems involving multivariate normal, multivariate skew normal and normal mixture models. We also compare our Gibbs sampler to a conjugate Bayesian model on a real dataset.

5. Examples

5.1. Low-dimensional problem

To test the performance of the Gibbs sampling approach, we sampled from the identified set of a covariance matrix where $d_X = 2$, $d_Y = 2$ and $d_Z = 1$. In this scenario we can use the exact ellipsoid formula to calculate the identified set. The covariance matrix $\boldsymbol{\Sigma}$ was specified to have a compound symmetry structure with correlation 0.75, thus having the form

$$\boldsymbol{\Sigma} = \begin{array}{cccccc} & X_1 & X_2 & Y_1 & Y_2 & Z_1 \\ X_1 & 1.00 & 0.75 & 0.75 & 0.75 & 0.75 \\ X_2 & 0.75 & 1.00 & 0.75 & 0.75 & 0.75 \\ Y_1 & 0.75 & 0.75 & 1.00 & 0.75 & - \\ Y_2 & 0.75 & 0.75 & 0.75 & 1.00 & - \\ Z_1 & 0.75 & 0.75 & - & - & 1.00 \end{array}.$$

We applied the Gibbs sampler to explore the range of possible values for $\sigma_{Y_1Z_1}$ and $\sigma_{Y_2Z_1}$. We used five thousand burn in iterations, and took twenty thousand samples. Fig. 1 compares the output of the Gibbs sampler to the correct solution. This true identified set was calculated using the ellipsoid formula (2). The dashed ellipse in (a) and (b) represents the boundary of the true identified set. In (a) we plot the Gibbs samples and see that they cover the identified set uniformly. In (b) we plot the convex hull of the samples and see that we have identified the boundaries of the space. Trace plots and running mean plots showed no evidence of poor mixing of the chain (Cowles and Carlin, 1996).

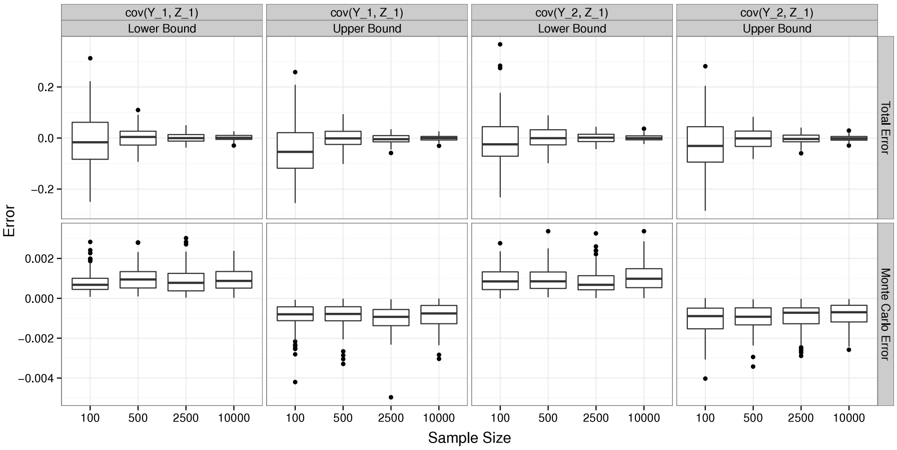

Fig. 2. Estimation error against sample size. The upper and lower bounds on each parameter were estimated using the Gibbs sampler. Total error is based on a comparison to the true identified set. Monte Carlo error is based on a comparison to the closed form maximum likelihood estimate of the bounds.

This low-dimensional problem is also useful to illustrate the impact of sample size on our methodology. Here it is possible to compute a maximum likelihood estimate of the identified set by substituting the maximum likelihood estimates of the identified parameters into the ellipsoid formula (2). The bounds from the Gibbs sampler output can be compared the closed form maximum likelihood bounds to gauge the loss in efficiency from using a computational approach.

We generated data from a multivariate normal model with the previous covariance matrix and zero mean vector for various sample sizes. We took equal number of observations in each file so $n_A = n_B$. We applied the missing data pattern for the matching problem with bivariate $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ as labelled previously. At each sample size we generated one hundred datasets. On each dataset we estimate the identified set using the closed form solution and the Gibbs sampler. Fig. 2 gives boxplots of the total error and the Monte Carlo error at each sample size. The total error is calculated by subtracting the estimate from the Gibbs sampler from the true bounds of the identified set. The Monte Carlo error is calculated by subtracting the Gibbs estimate from the exact maximum likelihood estimate of the bounds. In the top row of Fig. 2 we see the total error decreasing with sample size. In contrast, in the bottom row we see that the Monte Carlo error has no relationship with the sample size.

5.2. Multivariate normal model

In this example we assess the Gibbs sampler on a statistical matching problem with bivariate $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. The generative model was a multivariate normal distribution with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \\ Z_1 \\ Z_2 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & Y_1 & Y_2 & Z_1 & Z_2 \end{array} \\ \begin{bmatrix} 1.00 & 0.90 & 0.81 & 0.73 & 0.66 & 0.59 \\ 0.90 & 1.00 & 0.90 & 0.81 & 0.73 & 0.66 \\ 0.81 & 0.90 & 1.00 & 0.90 & 0.81 & 0.73 \\ 0.73 & 0.81 & 0.90 & 1.00 & 0.90 & 0.81 \\ 0.66 & 0.73 & 0.81 & 0.90 & 1.00 & 0.90 \\ 0.59 & 0.66 & 0.73 & 0.81 & 0.90 & 1.00 \end{bmatrix} \end{array}.$$

The covariance matrix resembles an AR(0.9) correlation structure. We generated $n_A = 10\,000$ samples for file A and $n_B = 10\,000$ samples for file B. We estimated the identifiable parameters using maximum likelihood and then applied the Gibbs sampler. We used five thousand burn in iterations and drew fifty thousand samples. The range of the samples for each partially identified parameter is reported in the maximum likelihood segment of Table 2. Looking at the interval widths, we see we have different levels of information about each parameter. The upper bound for all parameters is close to one, but the lower bounds range from 0.09 to 0.35. We are sure of at least moderate correlation between Y_1 and Z_1 but cannot conclude the same for Y_2 and Z_2. This is interesting as the true correlation between Y_1 and Z_1 is the same as the true correlation between Y_2 and Z_2. We can at least rule out negative correlations for the partially identified parameters. To gauge the quality of our estimate of the identified set we repeated the Gibbs sampling process using the correct values for the other covariance parameters instead of maximum likelihood estimates. These bounds are reported in the exact values segment of Table 2. The estimated bounds change very little when using the exact values for the identified parameters.

This simulation scenario is also used to investigate how the dimension of the identified set affects the performance of the Gibbs sampler. As mentioned earlier we expect dimension to primarily affect the required number of Gibbs iterations to reach the stationary distribution. We assessed convergence to the stationary distribution using the interval ratio statistic proposed by Brooks and Gelman (1998). The convergence diagnostic is calculated using m independent chains started from

Table 2
Estimated bounds in the multivariate normal example. The header row indicates what values of the identified parameters were used in the partially completed covariance matrix.

Parameter	True value	Maximum likelihood estimates		Exact values	
		Lower bound	Upper bound	Lower bound	Upper bound
$\sigma_{Y_1Z_1}$	0.810	0.354	0.950	0.359	0.952
$\sigma_{Y_1Z_2}$	0.730	0.188	0.980	0.192	0.991
$\sigma_{Y_2Z_1}$	0.900	0.264	0.925	0.263	0.918
$\sigma_{Y_2Z_2}$	0.810	0.093	0.972	0.091	0.972

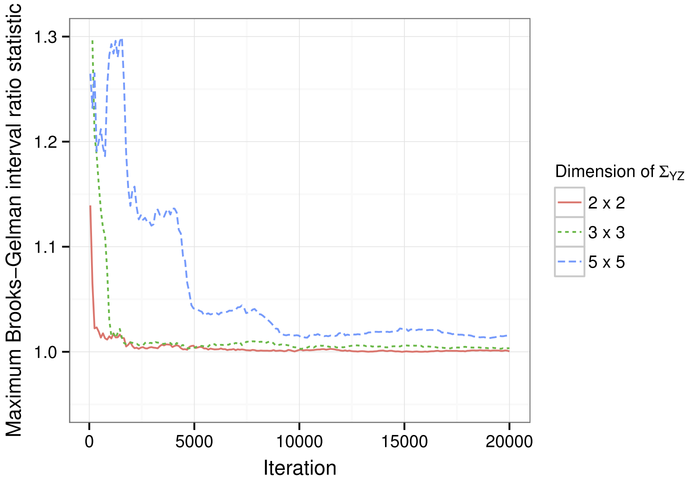

Fig. 3. Maximum interval ratio statistic against iteration.

overdispersed initial values. At a given iteration we take the ratio of the $(1-\alpha)\%$ interval from the pooled chains to the average width of $(1-\alpha)\%$ intervals from within each chain. Interval ratios are calculated for each variable in the joint target distribution. As the m chains approach the stationary distribution the ratio statistic should approach one. This convergence diagnostic was chosen as it is not based on approximate normality of the target distribution, and can detect if chains are not fully exploring the support of the target distribution.

We generated data from a zero mean p-dimensional normal distribution for $p = 6, 9, 15$. We again used a similar autoregressive structure for each covariance matrix. Element Σ_{ij} was set to $0.9^{|i-j|}$ for $i, j = 1, \ldots, p$. The first $p/3$ random variables were designated as the $\boldsymbol{X}$ variables, the second $p/3$ as $\boldsymbol{Y}$ and the third $p/3$ as $\boldsymbol{Z}$. Ten thousand samples were taken in both file A and file B for all values of p. We then used the Gibbs sampler to estimate the identified set for the $(p/3)^2$ partially identified parameters on each dataset. Twenty thousand iterations were run on each dataset.

To assess the speed of convergence we ran $m = 3$ independent chains and calculated the interval ratio statistic with $\alpha = 0.05$ at a grid of iteration numbers. The convergence diagnostic is plotted against iteration in Fig. 3. For ease of comparison we only plot the maximum statistic over the $(p/3)^2$ partially identified parameters at each iteration. Increasing the dimension of the identified set increases the number of iterations before apparent convergence. This suggests that the number of burn in iterations should increase with the dimension of the identified set.

5.3. Skew-normal model

We now consider characterising the identified set when the observations come from a skew normal model. We take the general definition of the skew normal distribution to be the unified skew normal (SUN) distribution (details in Appendix). The Gibbs technique is effective for SUN models as the SUN distribution can be expressed as the conditional distribution of a regular multivariate normal model (Arellano-Valle and Azzalini, 2006).

Let $\boldsymbol{S} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top, \boldsymbol{Z}^\top)^\top$, where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ have dimensions d_X, d_Y and d_Z respectively. Suppose that our observations come from the restricted skew normal distribution (Pyne et al., 2009), $\boldsymbol{S} \sim SUN_{p,1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, 1, 0)$, which is a special case of the SUN distribution (see details in Appendix). We will also form a corresponding partition of the parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZX} & \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix}, \qquad \boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}_X \\ \boldsymbol{\Delta}_Y \\ \boldsymbol{\Delta}_Z \end{bmatrix}.$$

Under the standard statistical matching problem, the only unidentifiable parameter is again $\boldsymbol{\Sigma}_{YZ}$. Due to the underlying conditioning representation of the SUN distribution, we face the problem of finding values of $\boldsymbol{\Sigma}_{YZ}$ such that the covariance

Table 3
Estimated bounds in the skew normal example. The header row indicates what values of the identified parameters were used in the partially completed covariance matrix.

Parameter	True value	Maximum likelihood estimates		Exact values	
		Lower bound	Upper bound	Lower bound	Upper bound
$\sigma_{Y_1Z_1}$	6	5.094	7.050	5.003	6.996
$\sigma_{Y_1Z_2}$	−6	−6.940	−4.962	−6.997	−5.004
$\sigma_{Y_2Z_1}$	−6	−7.060	−5.056	−6.998	−5.006
$\sigma_{Y_2Z_2}$	6	4.922	6.950	5.005	6.998

matrix of the latent multivariate normal distribution (9) is positive-definite. The identified set is

$$\Theta(\mathbf{\Sigma}_{YZ}) = \left\{ \mathbf{\Sigma}_{YZ} : \begin{bmatrix} 1 & \mathbf{\Delta}_X^{\mathsf{T}} & \mathbf{\Delta}_Y^{\mathsf{T}} & \mathbf{\Delta}_Z^{\mathsf{T}} \\ \mathbf{\Delta}_X & \mathbf{\Sigma}_{XX} & \mathbf{\Sigma}_{XY} & \mathbf{\Sigma}_{XZ} \\ \mathbf{\Delta}_Y & \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} & \mathbf{\Sigma}_{YZ} \\ \mathbf{\Delta}_Z & \mathbf{\Sigma}_{ZX} & \mathbf{\Sigma}_{ZY} & \mathbf{\Sigma}_{ZZ} \end{bmatrix} \text{ is positive-definite} \right\}.$$

We used the Gibbs sampler to estimate the identified set in a statistical matching problem with bivariate $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ from a joint restricted skew normal model. The parameters of the generative model were set to

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{array}{c|cccccc} & X_1 & X_2 & Y_1 & Y_2 & Z_1 & Z_2 \\ \hline X_1 & 2 & -1 & 2 & -2 & 3 & -3 \\ X_2 & -1 & 2 & -2 & 2 & -3 & 3 \\ Y_1 & 2 & -2 & 5 & -4 & 6 & -6 \\ Y_2 & -2 & 2 & -4 & 5 & -6 & 6 \\ Z_1 & 3 & -3 & 6 & -6 & 10 & -9 \\ Z_2 & -3 & 3 & -6 & 6 & -9 & 10 \end{array}, \quad \mathbf{\Delta} = \begin{bmatrix} 1 \\ -1 \\ 2 \\ -2 \\ 3 \\ -3 \end{bmatrix}.$$

We generated $n_A = 10\,000$ samples for file A and $n_B = 10\,000$ samples for file B. The data are plotted in the supplementary material (see Appendix B).

We estimated the identifiable parameters using maximum likelihood and then applied the Gibbs sampler to explore the identified set for $\mathbf{\Sigma}_{YZ}$. We use five thousand burn in iterations and drew fifty thousand samples. The estimated bounds on each partially identified parameter are reported in the maximum likelihood segment of Table 3. A pair's plot of the samples is in the supplementary material (see Appendix B). Despite the fact that we have no joint observations of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ we are able to bound the unidentified parameters roughly within the intervals $[5, 7]$ and $[-7, -5]$. The surprising tightness of the bounds is due to the strong impact of the skewness parameters on the covariance matrix of the underlying latent multivariate normal distribution. We again obtained an alternative estimate of the identified set using the true values of the identifiable parameters instead of maximum likelihood estimates. The bounds from this secondary run are shown in the exact values segment of Table 3. Use of exact values only changes the estimated bounds slightly.

5.4. Mixture model

The proposed methodology can also be applied to Gaussian mixture models, as the only partially identified parameters are the covariance parameters between the $\boldsymbol{Y}$ and $\boldsymbol{Z}$ random variables in each component distribution. We generated $n_A, n_B = 10\,000$ observations from a six dimensional, two component Gaussian mixture. Components were weighted equally. The first two variables are labelled as the $\boldsymbol{X}$ variables, the second two as the $\boldsymbol{Y}$ variables, and the final two as the $\boldsymbol{Z}$ variables. The mean of the first component was set as $(0, 0, 3, 0, 0, 3)^{\mathsf{T}}$ and the mean of the second component was set to $(3, 3, 0, 3, 3, 0)^{\mathsf{T}}$. The covariance matrices for the first and second components respectively were

$$\begin{array}{c|cccccc} & X_1 & X_2 & Y_1 & Y_2 & Z_1 & Z_2 \\ \hline X_1 & 1.00 & 0.50 & 0.25 & 0.12 & 0.06 & 0.03 \\ X_2 & 0.50 & 1.00 & 0.50 & 0.25 & 0.12 & 0.06 \\ Y_1 & 0.25 & 0.50 & 1.00 & 0.50 & 0.25 & 0.12 \\ Y_2 & 0.12 & 0.25 & 0.50 & 1.00 & 0.50 & 0.25 \\ Z_1 & 0.06 & 0.12 & 0.25 & 0.50 & 1.00 & 0.50 \\ Z_2 & 0.03 & 0.06 & 0.12 & 0.25 & 0.50 & 1.00 \end{array}, \quad \begin{array}{c|cccccc} & X_1 & X_2 & Y_1 & Y_2 & Z_1 & Z_2 \\ \hline X_1 & 1.00 & 0.90 & 0.81 & 0.73 & 0.66 & 0.59 \\ X_2 & 0.90 & 1.00 & 0.90 & 0.81 & 0.73 & 0.66 \\ Y_1 & 0.81 & 0.90 & 1.00 & 0.90 & 0.81 & 0.73 \\ Y_2 & 0.73 & 0.81 & 0.90 & 1.00 & 0.90 & 0.81 \\ Z_1 & 0.66 & 0.73 & 0.81 & 0.90 & 1.00 & 0.90 \\ Z_2 & 0.59 & 0.66 & 0.73 & 0.81 & 0.90 & 1.00 \end{array}.$$

The second component has a higher degree of collinearity across dimensions. The data are plotted in the supplementary material (see Appendix B). To estimate the identifiable parameters we used the EM algorithm, the specific case of Gaussian mixtures with missing data is described in Ghahramani and Jordan (1994). An initial clustering was obtained by first imputing the missing data using nearest neighbour matching, and then using k-means clustering. After estimating the identified parameters, we then applied the Gibbs sampler to estimate the identified set for each component. The output is summarised in histograms in Fig. 4. The true values of the covariance parameters are plotted as dashed lines. There is

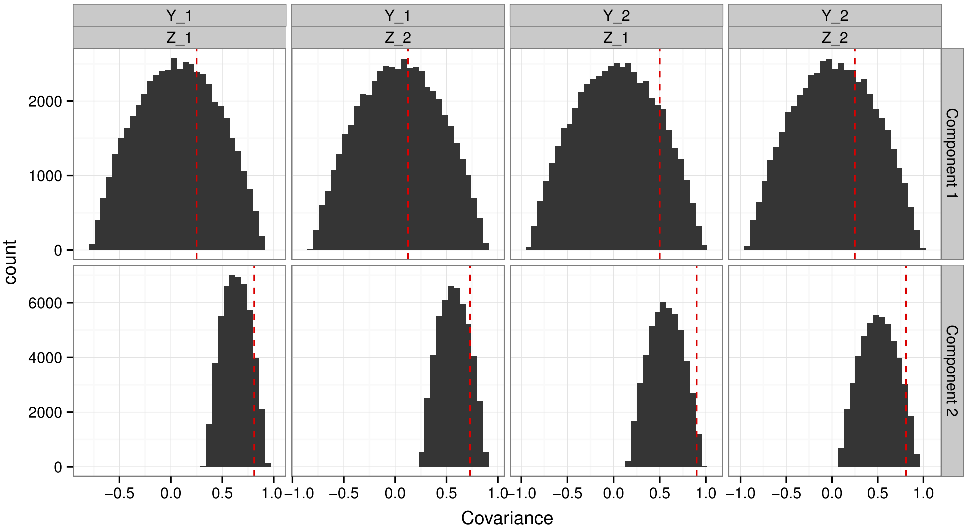

Fig. 4. Output of the Gibbs sampler from the mixture model example. Histograms for each partially identified parameter in both components are shown. Dashed lines give the true covariances.

Table 4
Different matching scenarios considered with the abalone dataset.

	X	***Y***	***Z***
Scenario 1	Length, diameter	Height	Shell weight
Scenario 2	Length, diameter	Height, whole weight	Shell weight

more uncertainty over the partially identified parameters in component 1 than in component 2, even though there are roughly equal numbers of observations from each component. This is because the higher correlations in component 2 result in the positive-definite constraint being more restrictive than in component 1.

Our methodology could also be used with mixtures of skew normal distributions. Identifiable parameters could be estimated by extending the algorithm in Lin et al. (2009) for skew normal data with missing values to mixture models.

5.5. Data application

We also compare our method to an existing Bayesian approach on a real dataset. We used the abalone dataset available from the UCI machine learning repository (Lichman, 2013). The dataset consists of various physical measurements on abalone. We took the 1307 observations on female abalone. The data was log transformed to make a multivariate normal model appropriate. The data are plotted in the supplementary material (see Appendix B). The complete dataset was split into two files, with $n_A = 653$ observations in the first and $n_B = 654$ in the second.

We compared our method to a Bayesian approach for multivariate normal data with missing values proposed by Schafer (1997). A Normal Inverse-Wishart prior is used on the parameters. An Inverse-Wishart prior is placed on the covariance matrix, $\Sigma \sim \text{InvWishart}(\mathbf{\Lambda}_0^{-1}, \nu_0)$ where $\mathbf{\Lambda}_0^{-1}$ is a $p \times p$ scale matrix, and ν_0 is a scalar giving the degrees of freedom. The prior mean is normally distributed conditional on $\mathbf{\Sigma}$, $\boldsymbol{\mu}|\mathbf{\Sigma} \sim N(\boldsymbol{m}_0, \Sigma/\kappa_0)$, where $\boldsymbol{m}_0$ is a p-length vector, and κ_0 is a scalar. Gibbs sampling can be used to draw from the posterior using data augmentation. When the data is weakly informative it is recommended to use a so called ridge prior, which improves numerical stability of posterior sampling and shrinks posterior correlations towards zero (Schafer, 1997, section 6.3.4). The ridge prior consists of setting $\nu_0 = \epsilon$, $\Lambda_0^{-1} = \epsilon D$, where D is a diagonal matrix containing estimated variances for each of the variables and ϵ is some constant greater than zero. A flat prior is induced on $\boldsymbol{\mu}$ by setting $\kappa_0 = 0$ leaving the choice of $\boldsymbol{m}_0$ arbitrary. The value of ϵ can be interpreted as the extra degrees of freedom being introduced by the prior. In the statistical matching scenario the prior on Σ_{YZ} conditional on the other parameters must be proper in order for the posterior to be proper (Gelfand and Sahu, 1999). The smallest value of ϵ which gives a proper Inverse-Wishart prior on Σ is $\epsilon = p$.

We compare our sampling based method to the Bayesian posterior in two different matching scenarios. These are described in Table 4. Low-dimensional problems are used as it is much easier to visually compare the distributions on the partially identified parameters. Maximum likelihood was used to estimate the identified parameters. We set $\epsilon = p$ in the ridge prior, to minimise the influence of the prior distribution on Σ while still keeping it proper. The matrix $\boldsymbol{D}$ was formed from the maximum likelihood estimates. We used five thousand burn in iterations with twenty thousand subsequent samples for both approaches.

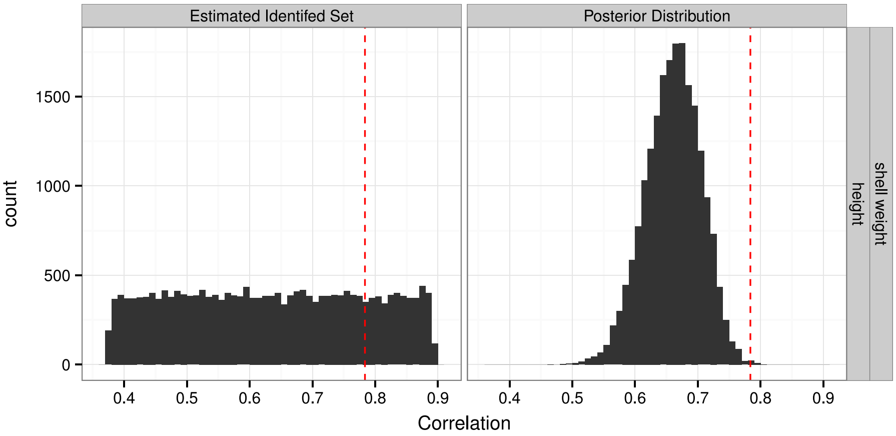

Fig. 5. Samples of the partially identified correlation in scenario one on the abalone dataset. Dashed lines represent the sample correlation.

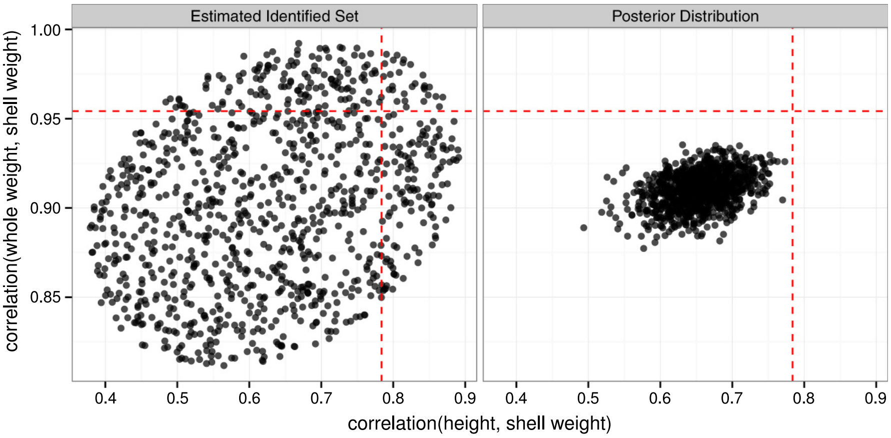

Fig. 6. Samples of the partially identified correlations in scenario two on the abalone dataset. The twenty thousand original draws have been thinned to two thousand for plotting. Dashed lines represent the sample correlations.

Fig. 5 compares uniform samples from the estimated identified set to samples from the posterior distribution in scenario one. For ease of interpretation we report the identified set of correlations. The sample correlation between height and shell weight on the complete dataset was 0.78, and is plotted as the dashed vertical line. Our approach yields samples drawn uniformly from the interval $(0.38, 0.9)$. Interestingly, the posterior appears to concentrate in a region in the middle of the estimated identified set. The posterior probability of correlations greater than 0.75 is very small, which is undesirable as the sample correlation coefficient is in this tail region. Our frequentist sampling strategy gives samples in the neighbourhood of the true observed correlation.

Fig. 6 compares uniform samples from the estimated identified set to samples from the posterior distribution in scenario two. The sample correlations from the complete dataset are plotted as dashed lines. The posterior again concentrates in a small portion of the estimated identified set. There is almost no posterior mass in the joint region where the two sample correlations lie, indicated by the intersection of the two dashed lines. Samples from the estimated identified set are much more dispersed, and cover the region where the true observed correlations lie. The Bayesian approach appears to underestimate the uncertainty surrounding the partially identified parameters, which can lead to misleading inference.

6. Conclusion

The statistical file matching problem is a data integration problem where missing data leaves some parameters unidentifiable. When trying to fit a parametric model, the goal is often to characterise the identified set of the parameters rather than to deliver a point estimate. Principled methods to establish uncertainty bounds are crucial in statistical matching problems to accurately represent the limitations of the observed data. The objective in statistical matching often reduces to finding positive-definite completions of a partially specified covariance matrix. Existing techniques for finding the set of possible completions are not applicable to high-dimensional datasets. We propose a Gibbs sampler that provides a simple and computationally efficient method to explore the identified set in high-dimensional statistical matching problems.

The proposed methodology involves sampling plausible values of the partially identified parameters. It was surprising to see that a Bayesian model with a weak conjugate prior did not give similar results to our method. It would be interesting to see if our algorithm resembles posterior sampling under a particular prior. Objective Bayesian inference on covariance matrices can be difficult, and this issue appears to be compounded in the partially identified setting (Sun and Berger, 2007).

Statistical matching can be challenging when standard parametric families are inappropriate models for the data. Parametric copulas are a very flexible tool for multivariate modelling (Smith et al., 2012; Kosmidis and Karlis, 2015), that could be useful for statistical matching problems. Parametric copulas are also subject to partial identification in the statistical matching problem (Ding and Song, 2016). Estimation of partially identified copula parameters using our Gibbs sampler could make statistical matching possible on a wider range of datasets.

Multiple imputation is frequently used in statistical matching to supply complete datasets for downstream analyses. Existing multiple imputation procedures require the user to specify a range of completed covariance matrices, introducing subjectivity into the process. The Gibbs sampler is an automatic method which should facilitate more objective multiple imputation procedures.

Acknowledgements

SP is supported by Ramalingaswami Fellowship of DBT, India (BT/RLF/Re-entry/12/2011), and GJM by the Australian Research Council.

Appendix A

A.1. The SUN Distribution

Arellano-Valle and Genton (2005) introduced the fundamental skew normal distribution (FUSN). A random variable $\boldsymbol{S}$ is said to have the FUSN$_{p,q}$ distribution if $\boldsymbol{S} \overset{d}{=} [\boldsymbol{V}|\boldsymbol{U} > \boldsymbol{0}]$, where $\boldsymbol{V}$ is a p dimensional multivariate normal random vector, and $\boldsymbol{U}$ is a q dimensional random vector defined on the same probability space. The probability density function of $\boldsymbol{S}$ can be expressed as

$$f(\boldsymbol{s}; \boldsymbol{\mu}; \boldsymbol{\Sigma}) = K_q^{-1}\phi_p(\boldsymbol{s}; \boldsymbol{\mu}, \boldsymbol{\Sigma})Q_q(\boldsymbol{s}), \tag{8}$$

where $K_q = \mathbb{E}\left(Q_q(\boldsymbol{V})\right) = P(\boldsymbol{U} > \boldsymbol{0})$ is a normalising constant and $Q_q(\boldsymbol{s}) = P(\boldsymbol{U} > \boldsymbol{0}|\boldsymbol{V} = \boldsymbol{s})$. The term $Q_q(\boldsymbol{s})$ can be interpreted as a skewing function. This is a very general formulation which encompasses the vast majority of skew normal distributions in the literature. An important special case of the FUSN family is the unified skew normal (SUN) distribution, which is also known as the closed skew normal (CSN) distribution or the hierarchical skew normal (HSN) distribution; see Arellano-Valle and Azzalini (2006). In the SUN family, we assume that $\boldsymbol{U}$ and $\boldsymbol{V}$ have a joint multivariate normal distribution. We say that $\boldsymbol{S} \sim \text{SUN}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\tau})$ if $\boldsymbol{S} \overset{d}{=} [\boldsymbol{V}|\boldsymbol{U} > \boldsymbol{0}]$ for q-dimensional $\boldsymbol{U}$ and p-dimensional $\boldsymbol{V}$, where

$$\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{\Delta}^T \\ \boldsymbol{\Delta} & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{9}$$

For the simulation in Section 5.3, we focus on one of the most commonly used formulations of the skew normal distribution – the restricted skew normal distribution – as adopted by Pyne et al. (2009) and equivalent to Azzalini and Dalla Valle (1996) and Branco and Dey (2001), and Lachos et al. (2010); see Lee and McLachlan (2013). This corresponds to a highly specialised form of the SUN distribution, where $q = 1$, $\tau = 0$, and $\Gamma = 1$.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.06.005.

References

Arellano-Valle, R.B., Azzalini, A., 2006. On the unification of families of skew-normal distributions. Scand. J. Statist. 561–574.
Arellano-Valle, R.B., Genton, M.G., 2005. On fundamental skew distributions. J. Multivariate Anal. 96, 93–116.
Azzalini, A., Dalla Valle, A., 1996. The multivariate skew-normal distribution. Biometrika 83, 715–726.
Branco, M.D., Dey, D.K., 2001. A general class of multivariate skew-elliptical distributions. J. Multivariate Anal. 79, 99–113.
Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7, 434–455.
Conti, P.L., Marella, D., Scanu, M., 2013. Uncertainty analysis for statistical matching of ordered categorical variables. Comput. Statist. Data Anal. 68, 311–325.
Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Statist. Assoc. 91, 883–904.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B (Methodolgical) 1–38.
Ding, W., Song, P.X.K., 2016. EM algorithm in Gaussian copula with missing data. Comput. Statist. Data Anal. 101, 1–11.
D'Orazio, M., 2015. Integration and imputation of survey data in R: the statmatch package. Rom. Stat. Rev. 63, 57–68.
D'Orazio, M., Di Zio, M., Scanu, M., 2006. Statistical Matching: Theory and Practice. In: Wiley Series in Survey Methodology, Wiley, New York.
Gelfand, A.E., Sahu, S.K., 1999. Identifiability, improper priors, and gibbs sampling for generalized linear models. J. Amer. Statist. Assoc. 94, 247–253.

Gelfand, A.E., Smith, A.F., Lee, T.M., 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J. Amer. Statist. Assoc. 87, 523–532.

Ghahramani, Z., Jordan, M.I., 1994. Supervised learning from incomplete data via an EM approach. In: Advances in Neural Information Processing Systems. p. 6.

Grone, R., Johnson, C.R., Sá, E.M., Wolkowicz, H., 1984. Positive definite completions of partial Hermitian matrices. Linear Algebra Appl. 58, 109–124.

Gustafson, P., 2015. Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.

Hammersley, J., Clifford, P., 1971. Markov fields on finite graphs and lattices.

Kadane, J., 1978. Some statistical problems in merging data files. 1978 Compendium of Tax Research, pp. 159–171.

Kline, B., Tamer, E., 2015. Bayesian inference in a class of partially identified models Technical Report 192431. Harvard University, Cambridge, Massachusetts.

Kosmidis, I., Karlis, D., 2015. Model-based clustering using copulas with applications. Stat. Comput. 1–21.

Lachos, V.H., Ghosh, P., Arellano-Valle, R.B., 2010. Likelihood based inference for skew-normal independent linear mixed models. Statist. Sinica 20, 303–322.

Laurent, M., Varvitsiotis, A., 2014. Positive semidefinite matrix completion, universal rigidity and the strong Arnold property. Linear Algebra Appl. 452, 292–317.

Lee, S.X., McLachlan, G.J., 2013. On mixtures of skew normal and skew t-distributions. Adv. Data Anal. Classif. 7, 241–266.

Lichman, M., 2013. UCI machine learning repository.

Lin, T.I., Ho, H.J., Chen, C.L., 2009. Analysis of multivariate skew normal models with incomplete data. J. Multivariate Anal. 100, 2337–2351.

Moon, H.R., Schorfheide, F., 2012. Bayesian and frequentist inference in partially identified models. Econometrica 80, 755–782.

Moriarity, C., Scheuren, F., 2001. Statistical matching: a paradigm for assessing the uncertainty in the procedure. J. Off. Stat. 17, 407–422.

Moriarity, C., Scheuren, F., 2003. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. J. Bus. Econom. Statist. 21, 65–73.

Neal, R.M., 2003. Slice sampling. Ann. Statist. 31, 705–767.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., et al., 2009. Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. 106, 8519–8524.

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

Rässler, S., 2003. A non-iterative Bayesian approach to statistical matching. Stat. Neerl. 57, 58–74.

Rässler, S., Kiels, H., 2009. How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model. In: Proceedings of the 57th Session of the International Statistical Institute.

Roberts, G.O., Rosenthal, J.S., 1998. On convergence rates of Gibbs samplers for uniform distributions. Ann. Appl. Probab. 1291–1302.

Rodgers, W.L., 1984. An evaluation of statistical matching. J. Bus. Econom. Statist. 2, 91–102.

Rubin, D.B., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. J. Bus. Econom. Statist. 4, 87–94.

Schafer, J., 1997. Analysis of Incomplete Multivariate Data. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.

Smith, M.S., Gan, Q., Kohn, R.J., 2012. Modelling dependence using skew *t* copulas: Bayesian inference and applications. J. Appl. Econometrics 27, 500–522.

Sun, D., Berger, J.O., 2007. Objective Bayesian analysis for the multivariate normal model. Bayesian Stat. 8, 525–562.

Tamer, E., 2010. Partial identification in econometrics. Annu. Rev. Econ. 2, 167–195.