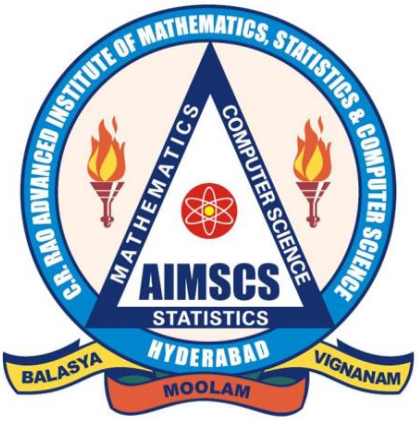
CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report



Author (s): **Daniel Ahfock, Saumyadipta Pyne, Sharon X. Lee & Geoffrey J. McLachlan**

Title of the Report: **Partial identification in the statistical matching problem**

Research Report No.: **RR2015-12**

Date: **November 16, 2015**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

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*

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. We propose a Gibbs sampler to draw from the set of possible completions. 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.

Keywords: Statistical matching, data integration, missing data, positive-definite matrix completion

2010 MSC: 00-01, 99-00

*Corresponding author
**Principal corresponding author
Email address: daniel.ahfock@uqconnect.edu.au (Daniel Ahfock)

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 distribution 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 distribution function belongs to some parametric family $\{f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$. 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 data set, under an identifiable model we can recover the true value of the parameters. In a partially identified model, given an infinite data set, 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.

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 $\boldsymbol{Z}$ and 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 $i = 1, \ldots, n_A$. Similarly, let $\boldsymbol{s}_j^B$ be a row vector representing the jth observation in file B for $j = 1, \ldots, n_B$. 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 identfied set for parameter θ_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, so that the distribution function is

$$f(x, y, z) = \phi_3 \left(x, y, z; \boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{bmatrix} \right),$$

and the standard stitching 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. We conclude by investigating extensions

to skew normal data.

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

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

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

$$\boldsymbol{\Sigma}_{YZ} = \begin{bmatrix} \sigma_{Y_1 Z_1} & \sigma_{Y_1 Z_2} & \dots & \sigma_{Y_1 Z_{d_Z}} \\ \sigma_{Y_2 Z_1} & \sigma_{Y_2 Z_2} & \dots & \sigma_{Y_2 Z_{d_Z}} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{Y_{d_Y} Z_1} & \sigma_{Y_{d_Y} Z_2} & \dots & \sigma_{Y_{d_Y} Z_{d_Z}} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{YX} = \begin{bmatrix} \sigma_{Y_1 X_1} & \sigma_{Y_1 X_2} & \dots & \sigma_{Y_1 X_{d_X}} \\ \sigma_{Y_2 X_1} & \sigma_{Y_2 X_2} & \dots & \sigma_{Y_2 X_{d_X}} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{Y_{d_Y} X_1} & \sigma_{Y_{d_Y} X_2} & \dots & \sigma_{Y_{d_Y} X_{d_X}} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{XZ} = \begin{bmatrix} \sigma_{X_1 Z_1} & \sigma_{X_1 Z_2} & \dots & \sigma_{X_1 Z_{d_Z}} \\ \sigma_{X_2 Z_1} & \sigma_{X_2 Z_2} & \dots & \sigma_{X_2 Z_{d_Z}} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{X_{d_X} Z_1} & \sigma_{X_{d_X} Z_2} & \dots & \sigma_{X_{d_X} Z_{d_Z}} \end{bmatrix}.$$

Here $\sigma_{X_i Z_j}$ denotes the covariance between X_i and Z_j. Finding an explicit expression for values of the matrix $\mathbf{\Sigma}_{YZ}$ that satisfy the condition in (1) is an open problem (Rässler & 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 $\mathbf{\Sigma}$ is a correlation matrix, we wish to find the identified set of correlations. For univariate Z, $\mathbf{\Sigma}_{YZ}$ is a column vector and the ellipsoid is governed by the equation

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

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

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

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

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 $\mathbf{\Sigma}_{YZ}$. 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 & Scheuren, 2001; D'Orazio et al., 2006).

We propose 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 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 & 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 capture 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 formulae 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 $\mathbf{\Sigma}$ is a correlation matrix.

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

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

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 $\mathbf{\Sigma}_{\widetilde{X}\widetilde{X}}$ denote the correlation matrix of $\widetilde{\boldsymbol{X}}$. We also let $\mathbf{\Sigma}_{\widetilde{Y}\widetilde{X}}$ denote the row vector containing the correlations between $\widetilde{Y}$ and $\widetilde{\boldsymbol{X}}$. Finally let $\mathbf{\Sigma}_{\widetilde{Z}\widetilde{X}}$ denote the row vector containing the correlations between $\widetilde{\boldsymbol{X}}$ and $\widetilde{Z}$.

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

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

where

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

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

While the full conditionals are easy to specify and sample from, the Hammersly-Clifford positivity condition does not apply so it is not guaranteed that the Gibbs sampler will converge to the correct stationary distribution (Hammersly & Clifford, 1971). We have to establish that the Markov chain defined by the Gibbs sampler is irreducible. Laurent & Varvitsiotis (2014) show that the identified set as defined in (1) will always be a convex set. Given fixed values for the other parameters, the identified set for $\mathbf{\Sigma}_{YZ}$ 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.

The Gibbs sampler requires an initial positive-definite completion of the covariance matrix. We recommend setting $\mathbf{\Sigma}_{YZ} = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XZ}$, 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 & Carlin, 1996). Roberts & 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.

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 $\mathbf{\Sigma}$ was specified to have a compound symmetry structure with correlation 0.75, thus having the form

$$\mathbf{\Sigma} = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \\ Z_1 \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & Y_1 & Y_2 & Z_1 \end{array} \\ \begin{bmatrix} 1.00 & 0.75 & 0.75 & 0.75 & 0.75 \\ 0.75 & 1.00 & 0.75 & 0.75 & 0.75 \\ 0.75 & 0.75 & 1.00 & 0.75 & - \\ 0.75 & 0.75 & 0.75 & 1.00 & - \\ 0.75 & 0.75 & - & - & 1.00 \end{bmatrix} \end{array}.$$

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. Figure 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 a shaded polygon determined by the convex hull of the samples and see that we have identified the boundaries of the space. Figure 2 shows trace plots and running mean plots to assess convergence of the Markov chain (Cowles & Carlin, 1996). There is no evidence of poor mixing.

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

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

The covariance matrix resembles an AR(0.9) correlation structure. We generated $n_A = 10000$ samples for file A and $n_B = 10000$ 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. Table 2 reports the range of the samples for each partially identified parameter. 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. Table 3 reports the range of the samples for each partially identified parameter. Comparing these intervals to those in Table 2 we do not obtain significantly different results due to the use of maximum likelihood estimates.

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 & Azzalini, 2006).

Let $\boldsymbol{S} = (\boldsymbol{X}^\mathsf{T}, \boldsymbol{Y}^\mathsf{T}, \boldsymbol{Z}^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ have dimension 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

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

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

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

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

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

We generated $n_A = 10000$ samples for file A and $n_B = 10000$ samples for file B. The data are plotted in Figures 3 and 4.

We estimated the identifiable parameters using maximum likelihood and then applied the Gibbs sampler to explore the identified set for $\boldsymbol{\Sigma}_{YZ}$. We use five thousand burn in iterations and drew fifty thousand samples. Table 4 summarises the output of the Gibbs sampler and Figure 5 shows a pairs plot of the samples. We report the observed range for each parameter. 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. Table 5 reports the results from this secondary run of the sampler. The obtained intervals are very similar to those in Table 4.

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.

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.

Acknowledgments

SP is supported by Ramalingaswami Fellowship of DBT, India.

References

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

Arellano-Valle, R. B., & Genton, M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, *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. *Journal of Multivariate Analysis*, *79*, 99–113.

Conti, P. L., Marella, D., & Scanu, M. (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics & Data Analysis*, *68*, 311–325.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*, 883–904.

D'Orazio, M. (2015). Integration and imputation of survey data in R: the StatMatch package. *Romanian Statistical Review*, *63*, 57–68.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley Series in Survey Methodology. New York: Wiley.

Gelfand, A. E., Smith, A. F., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.

Grone, R., Johnson, C. R., Sá, E. M., & Wolkowicz, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra and its Applications*, *58*, 109–124.

Hammersly, J., & Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished*, .

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

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

Laurent, M., & Varvitsiotis, A. (2014). Positive semidefinite matrix completion, universal rigidity and the strong Arnold property. *Linear Algebra and its Applications*, *452*, 292–317.

Lee, S. X., & McLachlan, G. J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, *7*, 241–266.

Moriarity, C., & Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, *17*, 407–422.

Moriarity, C., & 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*.

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. *Proceedings of the National Academy of Sciences*, *106*, 8519–8524.

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

Rässler, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica*, *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. *Annals of Applied Probability*, (pp. 1291–1302).

Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, *2*, 91–102.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*, 87–94.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*, 167–195.

7. Appendix

7.1. The SUN Distribution

Arellano-Valle & Genton (2005) introduced the fundamental skew normal distribution (FUSN). A random variable $\boldsymbol{S}$ is said to have the $\text{FUSN}_{p,q}$ distribution if $\boldsymbol{S} \stackrel{d}{=} [\boldsymbol{V}|\boldsymbol{U} > \boldsymbol{0}]$, 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 & 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 $\boldsymbol{S} \stackrel{d}{=} [\boldsymbol{V}|\boldsymbol{U} > \boldsymbol{0}]$ 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 & Dalla Valle (1996), Branco & Dey (2001), and Lachos et al. (2010); see Lee & McLachlan (2013). This corresponds to a highly specialised form of the SUN distribution, where $q = 1$, $\tau = 0$, and $\Gamma = 1$.

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

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

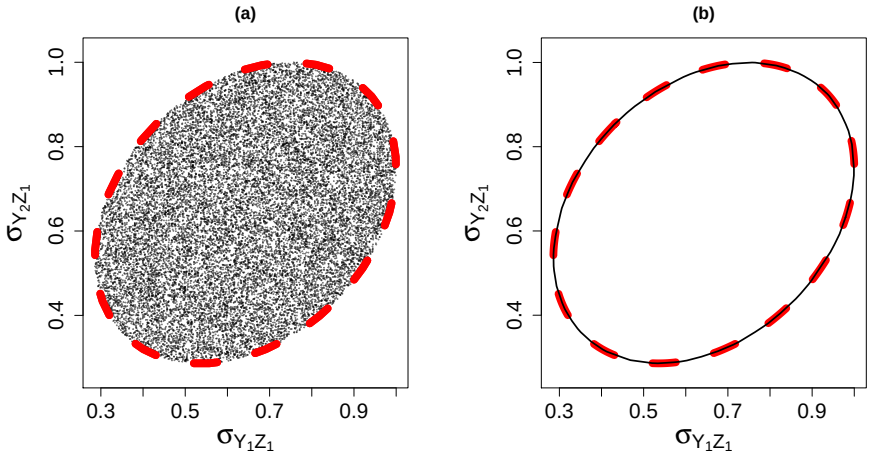


Figure 1: (a) Draws from the Gibbs sampler as black points. (b) The solid line denotes the convex hull of the Gibbs samples. The dashed ellipse shows the border of the true identified set in both (a) and (b).

	True Value	Lower Bound	Upper Bound
$\sigma_{Y_1 Z_1}$	0.810	0.354	0.950
$\sigma_{Y_2 Z_1}$	0.729	0.188	0.980
$\sigma_{Y_2 Z_1}$	0.900	0.264	0.925
$\sigma_{Y_2 Z_2}$	0.810	0.093	0.972

Table 2: Estimates of the identified range for each parameter using maximum likelihood estimates for the identifiable parameters.

Parameter	True Value	Lower Bound	Upper Bound
$\sigma_{Y_1 Z_1}$	0.810	0.359	0.952
$\sigma_{Y_1 Z_2}$	0.729	0.192	0.991
$\sigma_{Y_2 Z_1}$	0.900	0.263	0.918
$\sigma_{Y_2 Z_2}$	0.810	0.091	0.972

Table 3: Estimates of the identified range for each parameter using true values for the identifiable parameters.

Parameter	True Value	Lower Bound	Upper Bound
$\sigma_{Y_1 Z_1}$	6	5.094	7.050
$\sigma_{Y_1 Z_2}$	-6	-6.940	-4.962
$\sigma_{Y_2 Z_1}$	-6	-7.060	-5.056
$\sigma_{Y_2 Z_2}$	6	4.922	6.950

Table 4: Estimates of the identified range for each parameter using maximum likelihood estimates for the identifiable parameters.

Parameter	True Value	Lower Bound	Upper Bound
$\sigma_{Y_1 Z_1}$	6	5.003	6.996
$\sigma_{Y_1 Z_2}$	-6	-6.997	-5.004
$\sigma_{Y_2 Z_1}$	-6	-6.998	-5.006
$\sigma_{Y_2 Z_2}$	6	5.001	6.998

Table 5: Estimates of the identified range using true values for the identifiable parameters.

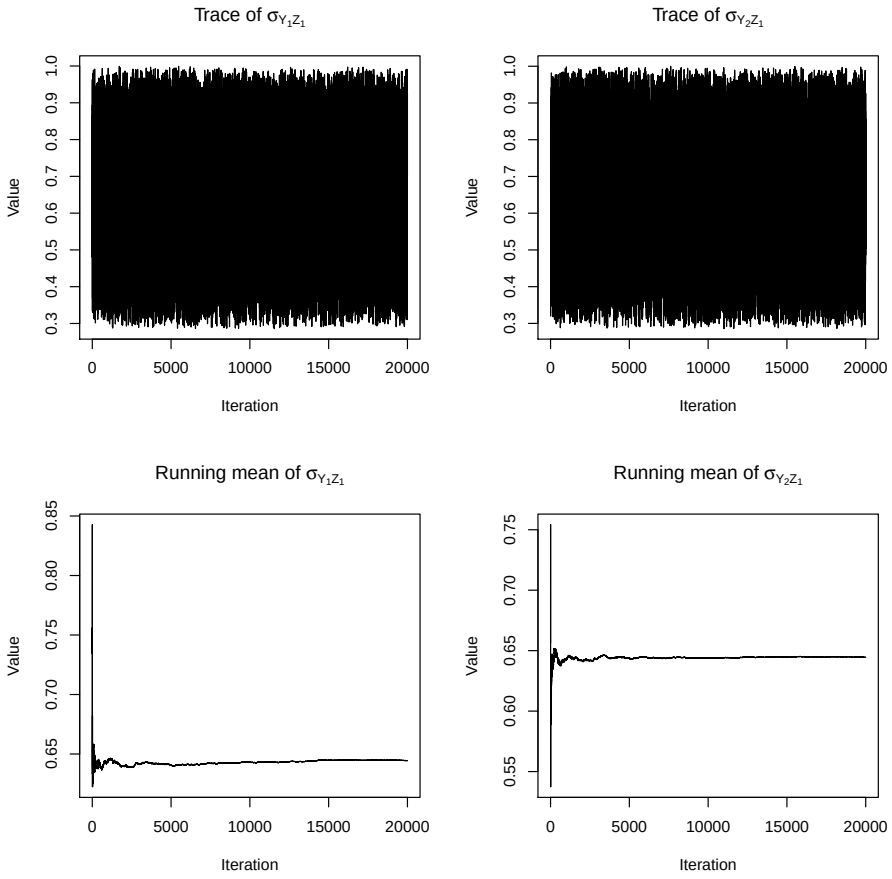


Figure 2: Convergence diagnostics

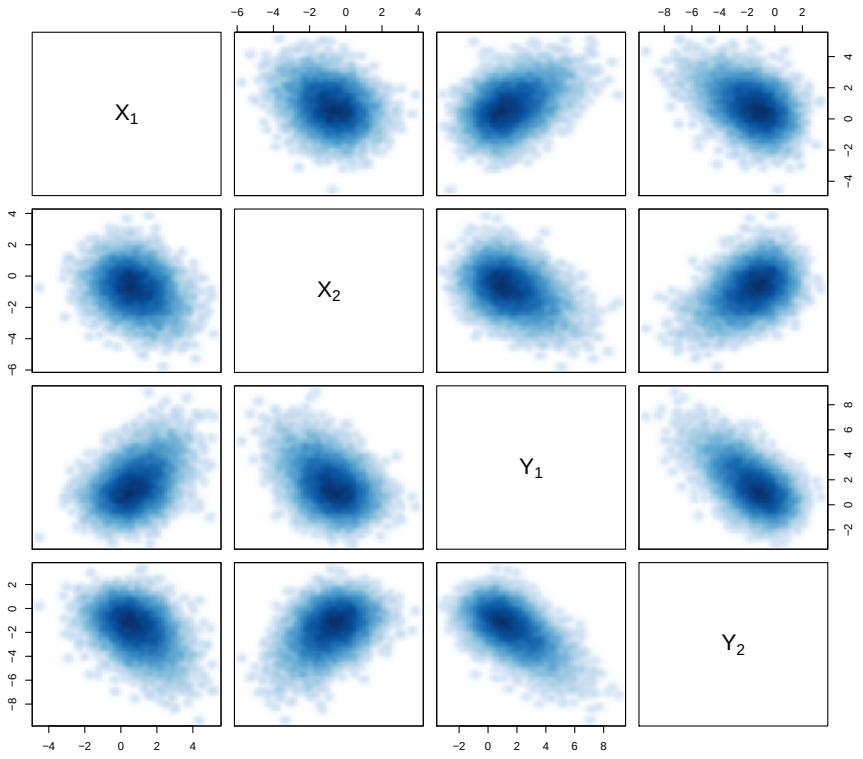


Figure 3: Samples in file A of skew normal example.

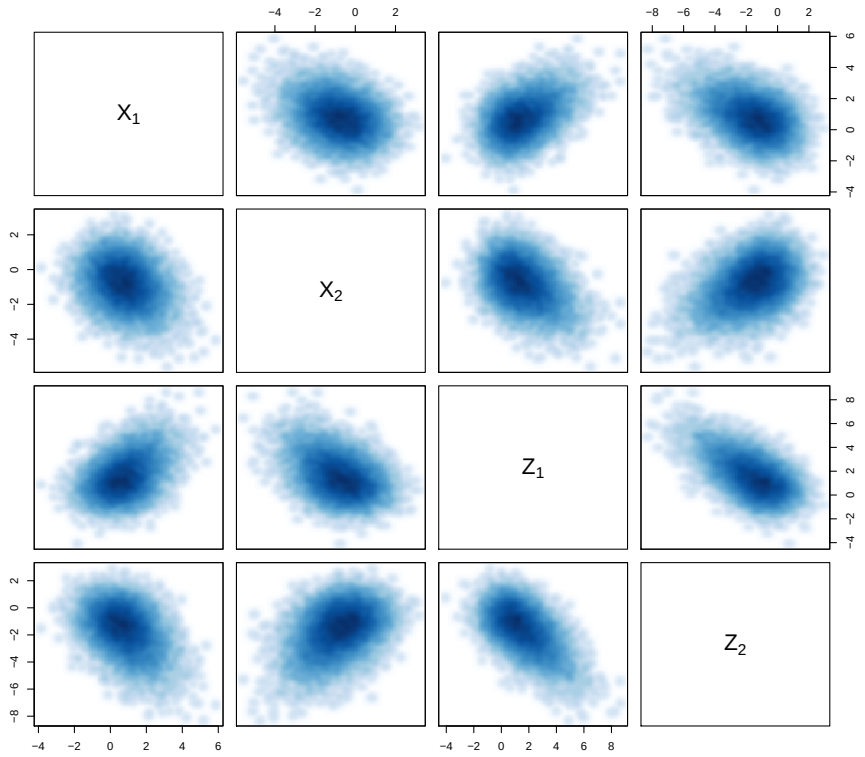


Figure 4: Samples in file B of skew normal example.

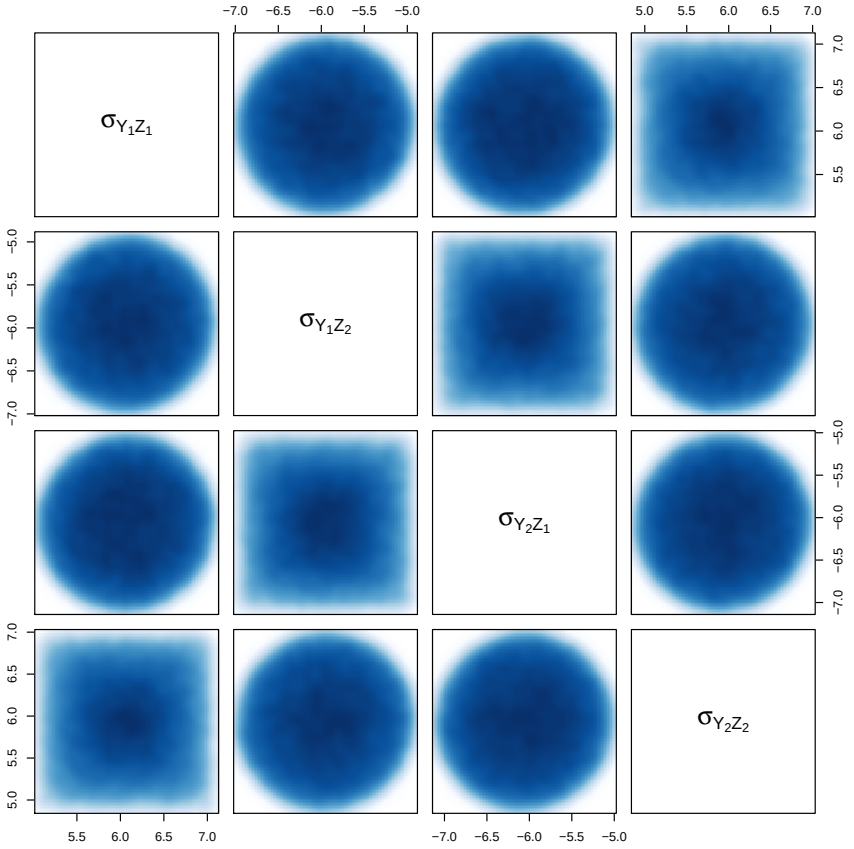


Figure 5: Output of the Gibbs sampler from the skew normal example