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 $i = 1, \ldots, n_A$. 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 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