UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 2011

A SEMI-PARAMETRIC BAYESIAN GENERALIZED LEAST SQUARE ESTIMATOR

Ruochen Wu Melvyn Weeks

20 February 2020

Generalized Least Square (GLS) estimators have been vastly applied in empirical studies to improve the efficiency of estimation. However, parametric GLS still imposes certain assumptions on the form of the covariance matrix of the unobservable, and the efficiency gain of GLS in fact depends on these assumptions being correct. In this paper we propose a semi-parametric Bayesian GLS estimator to cope with such heterogeneity. A Dirichlet process prior is put on the distribution of the covariance matrices of the unobservables, leading to a model that could be interpreted as the mixture of a variable number of normal distributions. Our methods let the number of heterogeneous groups be data driven, and so is the group membership of each observation. The semi-parametric Bayesian Seemingly Unrelated Regression (SUR) for equation systems, as well as Random Effects Model (REM) and Correlated Random Effects Model (CREM) for panel data are then described as special cases of the GLS estimators. A series of simulation experiments is designed to explore the performance of our methods, and demonstrates that they provide more reliable inference than the parametric Bayesian GLS. We then apply our semi-parametric Bayesian SUR and REM/CREM methods to empirical examples.

A Semi-Parametric Bayesian Generalized Least Square Estimator

Ruochen Wu
School of Economics
Fudan University

Melvyn Weeks*
Faculty of Economics and Clare College,
University of Cambridge

February 28, 2020

Abstract

Generalized Least Square (GLS) estimators have been vastly applied in empirical studies to improve the efficiency of estimation. However, parametric GLS still imposes certain assumptions on the form of the covariance matrix of the unobservable, and the efficiency gain of GLS in fact depends on these assumptions being correct. In this paper we propose a semi-parametric Bayesian GLS estimator to cope with such heterogeneity. A Dirichlet process prior is put on the distribution of the covariance matrices of the unobservables, leading to a model that could be interpreted as the mixture of a variable number of normal distributions. Our methods let the number of heterogeneous groups be data driven, and so is the group membership of each observation. The semi-parametric Bayesian Seemingly Unrelated Regression (SUR) for equation systems, as well as Random Effects Model (REM) and Correlated Random Effects Model (CREM) for panel data are then described as special cases of the GLS estimators. A series of simulation experiments is designed to explore the performance of our methods, and demonstrates that they provide more reliable inference than the parametric Bayesian GLS. We then apply our semi-parametric Bayesian SUR and REM/CREM methods to empirical examples.

JEL Classification Code: C3

Keywords: Bayesian semi-parametric, generalized lease square estimator, Dirichlet process, equation system, seemingly unrelated regression, panel data, random effects model, correlated random effects model.

*Contact Author: Dr. M. Weeks, Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK. Email: mw217@econ.cam.ac.uk. We have benefited from comments provided by Oliver Linton, ...

Contents

List of Tables

List of Figures

1 Introduction

The Generalized Least Square (GLS) estimator is a family of econometric methods that have seen numerous applications in empirical economics. As pointed out by Wooldridge (2003), GLS type estimators accommodate a deviation from the assumption that the covariance matrix of the errors, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, reflects i.i.d errors. However, the efficiency gains of GLS estimators are conditional on the set of implied restrictions being correct. Such restrictions exist with the most popular GLS type estimators, including the Seemingly Unrelated Regression (SUR) for equation systems and the random effects (RE) estimator for panel data.

In the analysis of panel data the RE assumes that individual specific, time-invariant features are uncorrelated with the explanatory variables. A useful extension to RE is the correlated random effects (CRE) estimator (Chamberlain, 1980; Wooldridge, 2005; Murtazashvili and Wooldridge, 2008), which allows the individual effects to be correlated with the explanatory variables usually as a linear function of the means of the regressors.

A GLS estimator can be applied to CRE model, with the efficiency gain conditional on the assumption of identically distributed errors. In reality heterogeneity is a major concern in empirical analysis, given that observations on individuals or households reflect variation in demographics such as size of the household, and the level of the education. It is a challenge for the analyst who seeks reliable inference with the data to capture the form of the heterogeneity in observations.

The standard Bayesian approach assumes that the error distribution is multivariate normal. Recent developments in Bayesian methods allow the use of prior information to relax this assumption. For example, the Dirichlet prior has been introduced to accommodate heterogeneity both in errors (see Chigira and Shiba, 2015 for an example) and in regression parameters (Allenby et al., 1998). A notable drawback is that the dimension of the mixing distribution is usually unknown.

Bayesian semi-parametric methods introduce flexibility in the sense of letting the data and the prior determine the structure of heterogeneity jointly. The Dirichlet Process ($\mathcal{DP}$) prior[1] can be used to allocate observations into groups, with those in the same group following a common distribution. In this sense, relative to a hierarchical model of heterogeneity that mixes a predetermined, fixed number of normals, the use of $\mathcal{DP}$ priors represents a more flexible approach.

In this paper we propose a semi-parametric Bayesian GLS estimator that incorporates the $\mathcal{DP}$ prior. The motivation is to maintain the efficiency gains of GLS estimators, whilst accounting for heterogeneity in the unobservables by allowing hyperparameters to differ among observations. The resulting distribution of the error terms involve a mixture of normal distributions where the number of the normal components is influenced by both the prior and the data. Our procedure embodies much of the flexibility of a finite mixture of normals without requiring additional computations/procedures to determine the number of components and impose penalties for over-fitting. The aforementioned SUR in combination with the random and correlated random effects model, are introduced as special cases of the semi-parametric Bayesian GLS estimators for equation systems and panel data, respectively.

A useful point of departure for our approach follows from the classical treatment of a number of relatively standard econometric problems. For example, classical instrumental variables estimators such as two stage least squares do not make any specific assumptions regarding the distribution of the error terms beyond independence and identically distributed. However, the Bayesian treatment of this model has traditionally relied on the assumption that the error terms are bivariate normal (cf. Chao and Phillips, 1998; Geweke, 1996; Kleibergen and van Dijk, 1998;

[1] See Escobar and West, 1995 and 1998 and MacEachern, 1998 for a reference of the Dirichlet Process prior.

Rossi et al., 2005.)

As an example, a Bayesian semi-parametric approach to the instrumental variable problem has been adopted by Conley et al. (2008). Instead of assuming a bivariate normal distribution for the structural and reduced form equations, the authors introduced a Dirichlet process prior for the hyperparameters. This provides a semi-parametric version of the two stage least square estimator, where the errors of the two stages jointly follow a non-parametric mixture of normal distributions.

The rest of the paper is organized as follows. In Section 2 we introduce the generic form of the Dirichlet process, and demonstrate its use as a prior for semi-parametric Bayesian GLS. Then two special cases of the GLS are described. The DP-SUR method is introduced in Section 3. Sections 3.3 and 3.4 present the simulation design and results, respectively, for the DP-SUR. Two empirical examples are given in Section 3.5. Section 4 motivates and introduces our semi-parametric Bayesian GLS methods for panel data, the DP-REM and DP-CREM. Simulation designs for the panel setting are in Section **??**, and the results are in Section 4.2. The DP-REM and DP-CREM methods are then applied to two empirical examples in Section 4.3. Section 5 concludes the paper.

2 Bayesian GLS with Dirichlet Process Prior

We introduce parametric and semi-parametric Bayesian methods dealing with heterogeneity in the errors. In Section 2.1 we start with parametric Bayesian GLS assuming that the errors follow the t-distribution. In Section 2.2 we allow the errors to follow a finite mixture of normal distributions. In Section 2.3 we describe the non-parametric $\mathcal{DP}$ mixture model for an infinite number of normal distributions. Section 2.4 then introduces the generic Bayesian GLS with $\mathcal{DP}$ mixture.

2.1 t Distributed Errors

Consider the basic linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector, $\boldsymbol{X}$ is a $N \times K$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $K \times 1$ vector of coefficients, and $\boldsymbol{\varepsilon}$ is the $N \times 1$ error vector. A common assumption underpinning a parametric Bayesian approach is that the errors ε_i are $i.i.d$ normally distributed with zero mean and constant variance σ^2. When micro data are used individuals or firms may have different unobservable characteristics such that the errors are likely to be heteroskedastic, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2\psi_i)$. The covariance matrix of the error vector can then be written as

$$\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Psi} = \sigma^2 \begin{bmatrix} \psi_1^2 & 0 & \cdots & 0 \\ 0 & \psi_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_N^2 \end{bmatrix}. \tag{2}$$

As $\boldsymbol{\Psi}$ is positive definite, there exists a $N \times N$matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P'} = \boldsymbol{I_N}$. A GLS type estimator could be employed to account for the heteroskedasticity. Following Koop (2003), the standardized regression is then written as

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}, \tag{3}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{X}$, and $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{P}\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The likelihood of the Bayesian GLS may then be written as

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}|\tilde{\boldsymbol{y}}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})\right]. \tag{4}$$

As the form of heteroskedasticity is generally unknown, some structure must be introduced so that the Bayesian GLS can be operationalised. Geweke (1993) introduced a Bayesian GLS estimator where a prior is put on ψ_i to capture the heteroskedasticity in the errors, namely

$$\psi_i|\nu \overset{iid}{\sim} \mathcal{IG}(1,\nu), \tag{5}$$

where $\mathcal{IG}(1,\nu)$ stands for an inverse gamma distribution with mean 1 and ν degrees of freedom. Given (5), the errors have a distribution that is a scale mixture of normal distributions[2], with $\varepsilon_i \sim \mathcal{N}(0,\sigma^2\psi_i)$. For ψ_i distributed inverse gamma, this particular scale mixture is equivalent to the t-distribution with ν degrees of freedom. Although the model with the t-distributed errors is flexible, this approach depends upon the assumption that the normal distributions are mixed with inverse gamma distributed variances.

As noted out by Koop (2003), relaxing this assumption results in more flexible models. In Section 2.2 we describe a parametric method with a Dirichlet prior to mix a finite number of normal distributions, so that the errors are no longer restricted to having a t-distribution. In Section 2.3, the non-parametric infinite mixture of normal distributions with the Dirichlet process prior is introduced.

2.2 Parametric Mixture Model

To accommodate more general cases of heterogeneity in the distribution of y_i, a natural point of departure is a hierarchical parametric model for heterogeneity. For example, we might assume that the data y_i given parameters $\boldsymbol{\theta}_i$ is normally distributed, letting parameters $\theta_i \sim F(\boldsymbol{\varphi})$, with hyper-parameters $\boldsymbol{\varphi}$. The degree of the heterogeneity is reflected in the number of unique values of $\boldsymbol{\theta}_i$, i.e. the number of normal distributions that form the distribution of y_i.

To represent the heterogeneity in the distribution of $\boldsymbol{y_i}$, a natural specification is a mixture model which groups observations by hyper-parameters, with the observations in the same group sharing the same hyper-parameter. We begin with a mixture of a pre-determined finite (say K) number of groups. Observation i will be assigned to group $c_i = k \in \{1,2,\ldots,K\}$ with probability p_k. The distribution of each group is governed by a group-specific hyper-parameter, say $\theta^*_{c_i}$, where $\theta^*_{c_i}$ denotes the unique value of the hyperparameters of group c_i[3].

We can think of the hierarchical data generation process as follows. We start with the group identifier of each observation being drawn from K-dimension multinomial distribution with probabilities $p = \{p_1,p_2,\ldots,p_K\}$. The hyperparameters of group c_i ($\theta^*_{c_i}$) are then generated from their distribution G_0, with the final step being the generation of observations from the normal distributions with their hyperparameters.

The conjugate prior for the multinomial distribution is the Dirichlet distribution given by

$$p = \{p_1,p_2,\ldots,p_K\} \sim \text{Dir}\left(\alpha_1,\alpha_2,\ldots,\alpha_K\right)$$

with probability density function

$$f\left(p_1,p_2,\ldots,p_K\right) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}. \tag{6}$$

$\Gamma(\cdot)$ denotes the gamma function, and $\alpha_1,\ldots,\alpha_K$ are the concentration parameters of the Dirichlet distribution.

[2] That is, the mean of the normal distributions (0 in this case) are all the same, while the variances (scales) are different.

[3] As an example, if observations i and j are in the same group, such that $c_i = c_j$ and $\theta^*_{c_i} = \theta^*_{c_j}$.

The finite normal mixture model has emerged as a widely applied methodology for capturing heterogeneity in both linear and non-linear models, including Allenby et al. (1998), Li and Tobias (2011) and Chigira and Shiba (2015). However, the main limitation is that it takes a fairly difficult test procedure to determine the "correct" number of mixing components. Below we introduce the infinite mixture model which does not require the number of mixing components to be determined a priori.

2.3 Non-parametric Mixture Model

The main restriction of the finite mixture model is that the number of groups is fixed. It is more reasonable to let the data and the prior determine the number of groups jointly with a non-parametric model that mixes an infinite number of normal distributions.

2.3.1 Dirichlet Process

The extension of the Dirichlet distribution to the infinite dimension case is referred to as a Dirichlet process ($\mathcal{DP}$). Introduced by by Ferguson (1973),[4] $\mathcal{DP}$ is the conjugate prior for an infinite dimension, non-parametric multinomial distribution.[5] The generic form of the $\mathcal{DP}$ can be written as

$$F \sim \mathcal{DP}\left(\alpha, F_0\right), \tag{7}$$

where $\alpha > 0$ is the concentration parameter, and F_0 is the base distribution. Following Escobar and West (1995 and 1998) and MacEachern (1998), $\mathcal{DP}$ may be interpreted as a *distribution of distributions*, that is, a draw from a $\mathcal{DP}$ is a probability distribution itself. F is a random distribution that is discrete with probability one (even given a continuous base distribution) with expectation the base distribution F_0.[6] The level of discreteness is adjusted by α, the concentration parameter.

To illustrate, in Figure 1[7] we present draws from the Dirichlet process $\mathcal{DP}(\alpha, F_0)$. Note that draws from a Dirichlet process are discrete distributions and they become less concentrated, i.e. there are more atoms, with increasing α. Each row contains 3 sets of draws from $F_0 \equiv \mathcal{N}(0,1)$ for fixed α. The four rows use different values for $\alpha \in \{1, 10, 100, 1000\}$. One can see that smaller α will lead to (on average) more discrete[8] F.

2.3.2 Chinese Restaurant Process

Given that the $\mathcal{DP}$ is a non-parametric distribution (i.e. not defined with a finite number of parameters) it is not possible to directly draw from it. Below we describe the Chinese restaurant process (CRP), which provides a way to construct the $\mathcal{DP}$.

Below we can think of a "customer" in the CRP as a metaphor for the realizations of the hyper-parameters θ when they are drawn from a distribution $F \sim \mathcal{DP}\left(\alpha, F_0\right)$. Since F is discrete with probability one, draws from it will form groups, for which "tables" represent a metaphor for discrete probability mass points, in the form of a group of customers.

At the moment a new customer arrives, $n-1$ customers (i.e. $n-1$ existing realizations $\{\theta_1, \theta_2, \ldots, \theta_{n-1}\}$)

are sitting at K tables numbered $\{1, 2, \ldots, K\}$. A defining feature of the $\mathcal{DP}$ is a new customer (the n^{th}) will sit at an existing table with a probability proportional to the number

[4] See Teh (2011) and Gershman and Blei (2012).

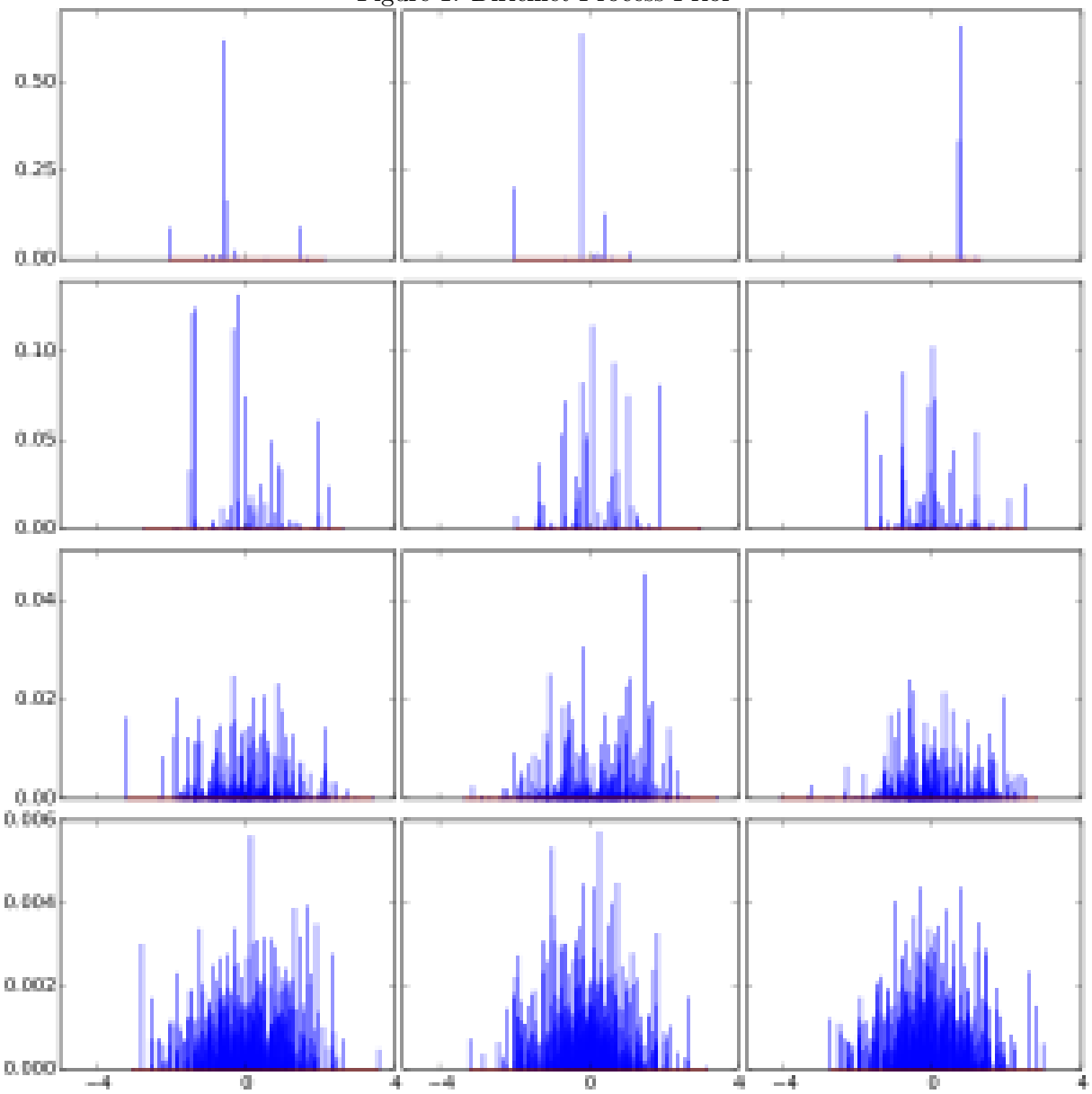
[5] An infinite mixture of normal distributions is an infinite dimension, non-parametric multinomial distribution.

[6] F_0 could be either discrete or continuous. See Ferguson (1973).

[7] The figure is reproduced from Wikipedia: https://en.wikipedia.org/wiki/Dirichlet_process

[8] In fact, in the extreme case where $\alpha = 0$, the $\mathcal{DP}$ will have one single atom only, while in the other extreme case where $\alpha \to \infty$, $F \to F_0$ weakly as a draw from the $\mathcal{DP}$.

Figure 1: Dirichlet Process Prior



of people sitting at the table; and will sit at an unoccupied table, i.e. the $(K+1)^{th}$, with a probability proportional to the concentration parameter α. A customer (a realisation θ_i) sitting at table k will have group id $c_i = k$. The location of the table (i.e. the unique value of θ's in group k) in the restaurant is the same for all the customers sitting there (i.e. $\theta_i = \theta^*_{c_i} = \theta^*_k$).

The predictive probability of θ_n taking an existing value or a new value may be written as

$$\begin{aligned}&\Pr\left\{\theta_n = \theta^*_k | \theta_1, \theta_2, \ldots, \theta_{n-1}\right\} \\ &= \begin{cases} \frac{\sum_{i=1}^{n-1} \delta_{\theta^*_k}(\theta_i)}{n-1+\alpha} = \frac{\sum_{i=1}^{n-1} \delta_k(c_i)}{n-1+\alpha} & \text{if } 1 \le k \le K \\ \frac{\alpha}{n-1+\alpha} & \text{if } k = K+1 \text{ (i.e. } \theta_n = \theta^*_{K+1} \sim F_0) \end{cases},\end{aligned} \tag{8}$$

where $\delta_{\theta^*_k}(\theta_i)$ ($\delta_k(c_i)$) is an indicator taking the value 1 when $\theta_i = \theta^*_k$ ($c_i = k$), and 0 otherwise.

The construction of the $\mathcal{DP}$ based on the predictive probability leads to what we observed in Figure 1 i.e. the $\mathcal{DP}$ is more concentrated with small α. Such features make the DP a logical basis for constructing a prior distribution with no other knowledge on which to group the customer or indeed the hyper-parameters.

As shown by Aldous (1985), F is a draw from the $\mathcal{DP}$ if $\theta_1, \theta_2, \ldots, \theta_n$[9] are generated according to the CRP, i.e.

$$\begin{aligned} F|\alpha, F_0 &\sim \mathcal{DP}(\alpha, F_0) \\ \theta_i|F &\overset{iid}{\sim} F. \end{aligned} \tag{9}$$

The CRP provides a mechanism to draw from F, which is a draw from the $\mathcal{DP}$ distribution.

2.3.3 Dirichlet Process Mixture Model

A model with a $\mathcal{DP}$ prior on the *distribution* of parameters is called a $\mathcal{DP}$ mixture model. The $\mathcal{DP}$ mixture can represent general forms of heterogeneity in the distributions of the observations. Unlike models based upon the t-distribution where the form of the mixture is fixed a priori, in the $\mathcal{DP}$ mixture model observations that could reasonably be assumed to originate from the same distribution are grouped together a posteriori.

The $\mathcal{DP}$ normal mixture model can be represented as

$$\begin{aligned} F|\alpha, F_0 &\sim \mathcal{DP}(\alpha, F_0) \\ \theta_i|F &\overset{iid}{\sim} F \\ \boldsymbol{y_i}|\theta_i &\sim \mathcal{N}(\theta_i). \end{aligned} \tag{10}$$

The posterior probability of θ_i having the same value as one of the existing θ_{-i} may be written as

$$\Pr\left\{\theta_i = \theta^*_k | \theta_{-i}, \boldsymbol{y_i}, \alpha\right\} \propto \frac{n_k}{n-1+\alpha} \mathcal{N}\left(\boldsymbol{y_i}|\theta^*_k\right), \tag{11}$$

where θ^*_k and n_k denote, respectively, the unique value of group k and the number of observations already in group k. The posterior probability of θ_i assuming a new value from the base distribution, i.e. $\theta_i = \theta^*_{new} \sim F_0$, may be written as

$$\Pr\left\{\theta_i = \theta^*_{new} | \theta_{-i}, \boldsymbol{y_i}, \alpha, F_0\right\} \propto \frac{\alpha}{n-1+\alpha} \int \mathcal{N}\left(\boldsymbol{y_i}|\theta^*_{new}\right) p\left(\theta^*_{new}|F_0\right) \mathrm{d}\theta^*_{new}, \tag{12}$$

where $p\left(\theta^*_{new}|F_0\right)$ is the probability density of the new value θ^*_{new} given F_0, the base distribution of the $\mathcal{DP}$ prior.

[9] The realisations $\theta_1, \theta_2, \ldots, \theta_n$ generated according to (8) are not independent given that the n^{th} realisation is generated conditioned on the $n-1$ realizations before. However, these realisations are exchangeable, and therefore independent conditional on a distribution F.

From (11) and (12) one can see that the probability of θ_i taking an existing or a new value is proportional to the normal probability density of the observation y_i conditioned on the value of the parameter (θ_k^* or θ_{new}^*), Note that a larger concentration parameter α of the $\mathcal{DP}$ prior makes the probability that $\theta_i = \theta_k^*$ ($\theta_i = \theta_{new}^*$) smaller (larger). In addition, as in (10), the probability of $\theta_i = \theta_k^*$ is proportional to the number of observations in group k. Third, when θ_i takes a new unique value, it is drawn from the base distribution (F_0) of the $\mathcal{DP}$ prior, leading to the probability of $\theta_i = \theta_{new}^*$ in (12)

It should also be noted that in the $\mathcal{DP}$ mixture model, the prior assigns the parameters θ_i's into groups. Observations $\boldsymbol{y_i}$ in the same group k will share the same unique value θ_k^*. This feature of the $\mathcal{DP}$ mixture model, namely assigning observations into groups sharing the same distribution[10] can be exploited to relax the identical distribution assumptions made by the parametric Bayesian GLS. In addition, it is not necessity to specify the number of groups beforehand. We now introduce how the $\mathcal{DP}$ prior is applied to introduce a semi-parametric Bayesian GLS.

2.4 Semi-parametric Bayesian GLS

In the context of semi-parametric Bayesian GLS, a $\mathcal{DP}$ prior is introduced on the distribution of the hyperparameters of the errors.

Consider a general linear regression

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}, \tag{13}$$

where i indexes the observation, $\boldsymbol{y_i}$ is a $Q \times 1$ vector of dependent variables, $\boldsymbol{X_i}$ is a $Q \times K$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $K \times 1$ vector of coefficients, and $\boldsymbol{\varepsilon_i}$ is a $Q \times 1$ vector of errors. Our semi-parametric GLS estimator introduces a $\mathcal{DP}$ prior on the distribution of the error covariance matrix which will be used to weight the observations. Given the usual assumption of zero means for the errors then $\theta_i = \boldsymbol{\Sigma_i}$. The hierarchical prior can then be written as

$$\begin{aligned} F|\alpha, F_0 &\sim \mathcal{DP}(\alpha, F_0) \\ \boldsymbol{\Sigma_i}|F &\overset{iid}{\sim} F \\ \boldsymbol{\varepsilon_i}|\boldsymbol{\Sigma_i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_i}). \end{aligned} \tag{14}$$

Due to the discreteness of F under the $\mathcal{DP}$ prior, the value of some covariance matrices $\boldsymbol{\Sigma_i}$ will the same, thus putting $\boldsymbol{\Sigma_i}$ into groups denoted by c_i. This "grouping" characteristic can help to reveal the structure of the unobserved heterogeneity in the data.

The GLS estimator weights the observations according to their covariance matrix. The likelihood of $\boldsymbol{\beta}$ is then

$$p\left(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{c_i}^*\right) = \frac{1}{(2\pi)^{Q/2}}|\boldsymbol{\Sigma}_{c_i}^*|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})'\boldsymbol{\Sigma}_{c_i}^{*-1}(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})\right]. \tag{15}$$

where $\boldsymbol{\Sigma}_{c_i}^*$ is the unique covariance matrix of group c_i. Given the choice of prior for $\boldsymbol{\beta}$, one could generate draws from the posterior of the parameters with MCMC methods. Consider the conjugate normal prior for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{b_0}, \boldsymbol{V_0}), \tag{16}$$

where $\boldsymbol{b_0}$ and $\boldsymbol{V_0}$ are, respectively, the prior mean and covariance matrix of $\boldsymbol{\beta}$. The posterior of $\boldsymbol{\beta}$ may then be written as

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\Sigma}_{c_i}^* \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{V}), \tag{17}$$

[10] See (Gershman and Blei, 2012).

where

$$\boldsymbol{V} = \left(\boldsymbol{V_0}^{-1} + \sum_{i=1}^{N} \boldsymbol{X_i}' \boldsymbol{\Sigma}_{c_i}^{*-1} \boldsymbol{X_i} \right)^{-1}, \tag{18}$$

and

$$\boldsymbol{b} = \boldsymbol{V} \left(\boldsymbol{V_0}^{-1} \boldsymbol{b_0} + \sum_{i=1}^{N} \boldsymbol{X_i}' \boldsymbol{\Sigma}_{c_i}^{*-1} \boldsymbol{y_i} \right). \tag{19}$$

Note that (17), (18) and (19) have the same form as the posterior of the parametric Bayesian GLS estimator assuming *i.i.d.* normal errors. In the case of semi-parametric Bayesian GLS, the errors are associated with different hyperparameters, such that each observation i is weighted by $\boldsymbol{\Sigma}_{c_i}^{*}$. DP-GLS is generic, without making any assumptions on the form of the covariance matrix other than all $\boldsymbol{\Sigma}_{c_i}^{*}$ being positive definite and symmetric.

Note that each observation $\boldsymbol{y_i}$ is assumed to be a $Q \times 1$ vector, in order to incorporate cases such as the equation systems and panel data. With an equation system, Q will be the number of equations in the system, while with panel data, Q is the number of time series in the panel. We now proceed to explain how the semi-parametric Bayesian GLS estimators work with these structures.

3 Semi-parametric Seemingly Unrelated Regression

Below we introduce the SUR equation system and demonstrate how the $\mathcal{DP}$ prior is incorporated. Without loss of generality we consider a system of two equations

$$\begin{aligned} y_{1i} &= \beta_{10} + x_{11,i}\beta_{11} + x_{12,i}\beta_{12} + \varepsilon_{1i} \\ y_{2i} &= \beta_{20} + x_{21,i}\beta_{21} + x_{22,i}\beta_{22} + x_{23,i}\beta_{23} + \varepsilon_{2i}, \end{aligned} \tag{20}$$

where y_{mi} denotes observation i for equation m ($m = 1, 2$) and $x_{mk,i}$ ($k = 1, 2, 3$) are the explanatory variables. β_{ml} ($l = 0, 1, 2, 3$) denote the coefficients, and ε_{mi} are the errors. The model can be written in matrix form

$$\begin{aligned} \boldsymbol{y_1} &= \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\varepsilon_1} \\ \boldsymbol{y_2} &= \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon_2}, \end{aligned} \tag{21}$$

where $\boldsymbol{y_m} = \{y_{mi}\}$, $\boldsymbol{\varepsilon_m} = \{\varepsilon_{mi}\}$ are $N \times 1$ vectors. $\boldsymbol{X_1} = [\boldsymbol{\iota}, \boldsymbol{x_{11}}, \boldsymbol{x_{12}}]$ and $\boldsymbol{X_2} = [\boldsymbol{\iota}, \boldsymbol{x_{21}}, \boldsymbol{x_{22}}, \boldsymbol{x_{23}}]$ are $N \times 3$ and $N \times 4$ matrices, respectively, where $\boldsymbol{\iota}$ is an $N \times 1$ vector of ones. $\boldsymbol{\beta_1} = \{\beta_{1l}\}$ and $\boldsymbol{\beta_2} = \{\beta_{2l}\}$ are 3×1 and 4×1 vectors, respectively.

In the presence of correlated errors there exists an efficiency gain by utilising a system estimator. The Seemingly Unrelated Regression (SUR, Zellner, 1962) was introduced for this task. Instead of $\varepsilon_{1i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_1^2)$ and $\varepsilon_{2i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_2^2)$, as in the OLS case, we let $\boldsymbol{\varepsilon_i} = (\varepsilon_{1i}\ \varepsilon_{2i})' \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. We follow the benchmark Bayesian non-parametric mixture model, i.e. the mixture of normal distributions, assuming normality for each $\boldsymbol{\varepsilon_i}$. The covariance matrix of $\boldsymbol{\varepsilon}$ is then

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{I} = \begin{bmatrix} \sigma_{11}^2 \boldsymbol{I_N} & \sigma_{12}^2 \boldsymbol{I_N} \\ \sigma_{21}^2 \boldsymbol{I_N} & \sigma_{22}^2 \boldsymbol{I_N} \end{bmatrix}, \text{ s.t. } \sigma_{12}^2 = \sigma_{21}^2, \tag{22}$$

where "$\otimes$" stands for the Kronecker product. One could transform the observations with this covariance matrix, so that the errors follow the standard normal distribution $\mathcal{N}(0, 1)$, with the likelihood, prior and posterior of the parameters defined similarly as in equations (15) to (19)[11].

[11] However, the covariance matrix $\boldsymbol{\Omega}$ has a specific form of the SUR in (22), instead of the general, positive definite symmetric form of a covariance matrix.

3.1 DP prior for SUR

Although the SUR model accounts for the cross-equation correlation of errors, as Wooldridge (2003) has noted, the errors are assumed to be identically distributed. Moreover, unlike the classical GLS estimator, this distribution is usually assumed to be normal. In this section we propose a new DP-SUR method that makes no a priori assumptions on the family of distribution of the errors. Given (21), the covariance matrix of the error for observation i is given by

$$\boldsymbol{\Sigma_i} = \begin{bmatrix} \sigma^2_{11,i} & \sigma^2_{12,i} \\ \sigma^2_{21,i} & \sigma^2_{22,i} \end{bmatrix}, \; s.t. \; \sigma^2_{12,i} = \sigma^2_{21,i}. \tag{23}$$

If we allow each observation i have its own covariance matrix, flexibility of the error distribution comes with a curse of dimensionality, which increase proportionally with the number of observations. Assigning the observations into groups represents a compromise.

The $2N \times 2N$ covariance matrix of the error vector is

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma^{*2}_{11,c_1} & 0 & \cdots & 0 & \sigma^{*2}_{12,c_1} & 0 & \cdots & 0 \\ 0 & \sigma^{*2}_{11,c_2} & \cdots & 0 & 0 & \sigma^{*2}_{12,c_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^{*2}_{11,c_N} & 0 & 0 & \cdots & \sigma^{*2}_{12,c_N} \\ \sigma^{*2}_{21,c_1} & 0 & \cdots & 0 & \sigma^{*2}_{22,c_1} & 0 & \cdots & 0 \\ 0 & \sigma^{*2}_{21,c_2} & \cdots & 0 & 0 & \sigma^{*2}_{22,c_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^{*2}_{21,c_N} & 0 & 0 & \cdots & \sigma^{*2}_{22,c_N} \end{bmatrix}. \tag{24}$$

where c_i denotes the group id of observation i, and the superscript $*$ denotes the group-specific hyperparameter. For $c_i = c_j$, $i, j \in \{1, 2, \ldots, N\}$ observations i and j share the same group ID and hyperparameters, such that $\sigma^{*2}_{pq,c_i} = \sigma^{*2}_{pq,c_j}$, where $p, q \in \{1, 2\}$ index the equations in the system.

Assuming that the number of groups are known, the Dirichlet prior may be used to perform the mixing. A less restrictive approach utilises a non-parametric approach by introducing a $\mathcal{DP}$ prior for the distribution of $\boldsymbol{\Sigma_i}$ as in (14). A natural choice of the base distribution F_0 is the conjugate prior for the covariance matrix of a multivariate normal distribution, the inverse Wishart distribution, i.e.

$$F_0 \equiv \mathcal{IW}(\nu, \boldsymbol{W}) \textbf{??} \tag{25}$$

where ν, $\boldsymbol{W}$ are hyperparameters of the inverse Wishart distribution. Given (**??**) the posterior distribution of the covariance matrices are also distributed inverse Wishart, which is easy to draw from using the Gibbs sampler. The main difference from the parametric Bayesian SUR is that the covariance matrix $\boldsymbol{\Omega}$ is now given in (24), which allows each group of observations to have its own unique values for the parameters.

3.2 MCMC algorithm

A Gibbs sampler (see Geweke, 1996) is available for the SUR model. The Gibbs sampler draws two sets of parameters from their posteriors: the covariance matrix of the errors $\boldsymbol{\Sigma}$ and the regression parameters $\boldsymbol{\beta}$, namely

$$\boldsymbol{\Sigma} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta} \tag{26}$$

$$\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\Sigma}. \tag{27}$$

When introducing the hierarchical structure which includes the $\mathcal{DP}$ prior, a number of extra parameters are included in the MCMC algorithm. These are the covariance matrices of the errors,

$\Theta = \{\boldsymbol{\Sigma_i}\}$, and α, the concentration parameter of the $\mathcal{DP}$ prior. The Gibbs sampler now consists of

$$\Theta|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \alpha \tag{28}$$
$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \Theta, \alpha \tag{29}$$
$$\alpha|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \Theta. \tag{30}$$

The major difference between the two Gibbs sampler lies in (26) and (28). In (26) the errors have the same covariance matrix $\boldsymbol{\Sigma}$. In contrast, there will be $K \leq N$ unique values in Θ in equation (28) due to the discreteness of F under the $\mathcal{DP}$ prior; observations with the same value of $\boldsymbol{\Sigma_i}$ are assigned to the same group. With the last draw of $\boldsymbol{\beta}$, the residuals can be obtained, which are used as the data to take a draw for Θ.

In making draws of the concentration parameter α using (30), we adopt the $\mathcal{DP}$ prior introduced by Conley et al. (2008), namely

$$p(\alpha) \propto \left(1 - \frac{\alpha - \alpha_{min}}{\alpha_{max} - \alpha_{min}}\right)^{\tau}, \tag{31}$$

where α_{min} and α_{max} are the pre-set lower and upper bound of α. Larger α lead to more groups being generated on average, i.e. the $\mathcal{DP}$ being less discrete. According to the distribution of the number of groups K conditioned on α in Antoniak (1974), we could determine α_{min} and α_{max} by setting the mode of number of groups to K_{min} and K_{max}. In this paper we let K_{min} be 1 and K_{max} be 5% of the number of observations. Following the suggestion of Conley et al. (2008), we set τ to 0.8. The hyper-parameters α_{max} has been adjusted according to K_{max} being 10% and 50% of the sample size. In our experiments the results are insensitive to these changes in the hyper-parameters in the prior of the concentration parameter α.

3.3 A Simulation Experiment

In this section we conduct a simulation experiments designed to compare our method to the Bayesian SUR described in Section 3. As the main focus of this paper is the potential efficiency gains over GLS type estimators, we evaluate the performance of the DP-SUR and normal Bayesian SUR focusing upon the posterior standard deviations of the parameters estimated with the two methods. All simulation experiments are based upon the two equation system in (20).

The experiments are designed to highlight the performance of the estimators along the following dimensions:

(i) heterogeneity in the errors;

(ii) the tail of the error distribution;

(iii) sample size.

For (i) we check the performance of our DP-SUR approach against a model where the errors are distributed *i.i.d.* multivariate normal. In the heterogeneous case, the most direct way is to generate the errors from a mixture of multivariate normal distributions. Exploiting the fact that a scale mixture of normal distributions is Student t-distributed given am inverse gamma the mixing distribution[12], we use the multivariate t-distribution to represent the case where the errors are heterogeneous, with each observation following a different normal distribution.[13]

[12] See Andrews and Mallows (1974)

[13] Simulating data from a mixture of multivariate normal distributions can be problematic given the influence of the number of components, the covariance matrices[14] of the normal components and the weights assigned to each component.

To accommodate (ii), we vary the degrees of freedom (df) of the multivariate t-distribution. Smaller degrees of freedom leads to heavier tails, which indicates that a larger proportion of observations follow normal distributions that are "flatter", i.e. less concentrated around the mean.

To determine the robustness of our method, we include a set of simulations where the errors follow a log-normal distribution. For if an $m \times 1$ random vector variable $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being the $m \times m$ covariance matrix, then $\boldsymbol{Z} = \exp(\boldsymbol{W})$ is multivariate log-normally distributed with mean and covariance matrix

$$\mathbb{E}[\boldsymbol{Z}]_k = e^{\frac{1}{2}\boldsymbol{\Sigma}_{kk}} \tag{32}$$

$$Cov(\boldsymbol{Z})_{jk} = e^{\frac{1}{2}(\boldsymbol{\Sigma}_{jj}+\boldsymbol{\Sigma}_{kk})}\left(e^{\boldsymbol{\Sigma}_{jk}} - 1\right), \tag{33}$$

where $k, j = 1, 2, \ldots, m$ are the row and column subscripts.

The log-normal distribution has seen a wide range of applications in empirical studies. For example, with perhaps the exception of the top 1-3 percent of the population, income has been shown to follow a log-normal distribution (Clementi and Gallegati, 2005). In addition, extreme realizations are more likely to be generated from the multivariate log-normal distribution, as it is fat-tailed.

Using (21), the explanatory variables are drawn from normal distributions with parameters

$$x_{11,i} \overset{iid}{\sim} \mathcal{N}(1,1),\ x_{12,i} \overset{iid}{\sim} \mathcal{N}(3,1),$$

and

$$x_{21,i} \overset{iid}{\sim} \mathcal{N}(-2,1),\ x_{22,i} \overset{iid}{\sim} \mathcal{N}(4,1),\ x_{23,i} \overset{iid}{\sim} \mathcal{N}(-1,1). \tag{34}$$

We set $\boldsymbol{\beta}_1 = (1. - 0.51.6)'$ and $\boldsymbol{\beta}_2 = (1.5 - 1.2 - 0.72)'$. We generate errors from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \tag{35}$$

Without loss of generality, we let the variances be identical, and fix the correlation between the errors in the two equations at 0.5[15]. We set three samples sizes for the simulation experiment: 100, 250, and 500.

When generating errors from a multivariate t-distribution, we set the location parameter to $\boldsymbol{\mu} = \mathbf{0}$, and shape parameter to $\boldsymbol{\Sigma}$ as in the case of the multivariate normal errors. As noted, the parameter that controls the tail behaviour of multivariate t-distributions is the df. We set the df to 2[16], 3 and 4. For df=2 the tails of the corresponding multivariate t-distribution are much heavier than that of the multivariate normal with the same location and shape parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. When the df is 4, the tails of the multivariate t-distribution are only slightly heavier than the multivariate normal. Our DP-SUR should demonstrate relative efficiency in all three situations, as the multivariate t-distributions has heavier tails than the multivariate normal. Gains in efficiency will be decreasing in df, given that the tails are less heavy.

The errors in the multivariate log-normal scenario are constructed by first drawing from multivariate normal distributions, and then taking the natural exponent of these draws. Because the log-normal distribution has a positive mean, it is necessary to demean[17], so that the errors will have zero means. Our DP-SUR is expected to have efficiency gains in the log-normal case given it is asymmetric and heavy tailed.

[15] There should be no loss of generality, given we are comparing a semi-parametric and a parametric SUR estimator. One would be more interested in the correlation if she were comparing the SUR to the OLS, which does not take the cross equation correlation into consideration.

[16] We do not use df 1 as the t distribution does not even have a mean in this case.

[17] see (32) for the expression of the mean of the multivariate log-normal distribution

3.4 DP-SUR Simulation Results

Below we present the simulation results.[18] We present the posterior standard deviations (S.D.) estimated with both our DP-SUR method and the Bayesian SUR assuming multivariate normal errors[19]. Each table contains 9 columns, presenting the posterior S.D. obtained by the two methods percentage difference between them for the three sample sizes. In the multivariate t-distribution case, the results are presented with the df being 2, 3, 4, and infinity (i.e. the multivariate normal case), respectively.

Multivariate t-distributed Errors

Table 1 presents the posterior S.D.s and the percentage difference[20] between the S.D. estimated with the DP-SUR and the normal SUR, both averaged over the samples. We observe that the DP-SUR gives smaller posterior S.D when df is 2, 3 and 4. The percentage differences when df is 2 are above 40%, above 20% when df is 3, and around 15% when df is 4 as shown in the upper three panels of the table. Efficiency gains increase with sample size as more extreme values of the errors are realised. While the parametric SUR standardize all the realizations with the same $\mathbf{\Sigma}$, these extreme realizations will be assigned to distributions with larger $\mathbf{\Sigma_i}$'s by the DP-SUR, thus given smaller weights, and more efficiency gains is achieved.

Our results are consistent with expectations. The efficiency gains of the semi-parametric DP-SUR are the largest when the df is 2 (with the heaviest tails). Efficiency gains fall with the df increasing, given less heavy tails of the distribution of the errors. In fact, the lowest panel in Table 2 where the df is infinity, we observe that the posterior S.D. estimated with the two methods are very close. The S.D. for DP-SUR is slightly larger than their SUR counterparts. This is not surprising since when the distribution of the errors is multivariate normal, the parametric method is more parsimonious, using the correct structure for the covariance matrix of the errors. Among the three sample sizes, the differences between the S.D. are the largest[21] when the the sample size is 100. This is expected as the DP-SUR requires more information to allow the observation to have different parameters, and it has a larger impact when the sample size is small.

Multivariate Log-normal Errors

The posterior S.D.s are presented in Table 2. We observe that the DP-SUR posterior S.D. are more than 55% smaller than those calculated using the Bayesian SUR assuming *i.i.d.* normal errors. The efficiency gains increase with sample size, which reach more than 65% in the case of 500 observations. As with the case of t distributed errors, this is due to the fact that more extreme realizations of errors are present in larger samples, and lead to more efficiency gains by grouping them.

3.5 DP-SUR Empirical Examples

Below we apply our DP-SUR method to an economic model of the demand for factors of production with generalized Leontief cost function (Diewert, 1971), which is an equation system with

[18] We carry out 100 simulations for each sample size, which proved sufficient to achieve stable results even with the smallest sample size.

[19] The tables containing the posterior means can be found in the Appendices. For the tables of posterior means, there are 6 columns presenting the means estimated by the two methods for the 3 correlations.

[20] $\Delta\% = (\text{S.D.}_{\cdot\text{SUR}} - \text{S.D.}_{\cdot\text{DP}})/\text{S.D.}_{\cdot\text{SUR}} \times 100\%$

[21] Nevertheless, the differences are still small in magnitudes, less than 2.5% for all coefficients

Table 1: Posterior S.D., multivariate t errors

df = 2									
Sample size	100			250			500		
Parameters	DP	SUR	Δ%	DP	SUR	Δ%	DP	SUR	Δ%
β_{10}	0.4184	0.8103	42.59%	0.2328	0.5262	50.09%	0.1644	0.4175	55.49%
β_{11}	0.1149	0.2155	41.57%	0.0732	0.1596	49.56%	0.0458	0.1126	54.22%
β_{12}	0.1254	0.2358	41.74%	0.0696	0.1515	49.30%	0.0498	0.1223	54.12%
β_{20}	0.5085	0.9791	42.67%	0.3261	0.7255	48.60%	0.2366	0.5649	53.30%
β_{21}	0.1107	0.2080	41.49%	0.0649	0.1359	48.48%	0.0508	0.1198	52.78%
β_{22}	0.1136	0.2141	41.63%	0.0653	0.1366	48.28%	0.0487	0.1139	52.34%
β_{23}	0.1271	0.2391	41.63%	0.0631	0.1321	48.38%	0.0467	0.1098	52.55%
df = 3									
Sample size	100			250			500		
Parameters	DP	SUR	Δ%	DP	SUR	Δ%	DP	SUR	Δ%
β_{10}	0.4121	0.5327	20.63%	0.2250	0.3098	26.25%	0.1619	0.2289	28.09%
β_{11}	0.1114	0.1452	21.20%	0.0708	0.0973	26.05%	0.0446	0.0634	28.53%
β_{12}	0.1231	0.1586	20.40%	0.0670	0.0923	26.20%	0.0486	0.0687	28.01%
β_{20}	0.4980	0.6497	21.15%	0.3174	0.4344	25.73%	0.2260	0.3220	28.82%
β_{21}	0.1079	0.1397	20.63%	0.0630	0.0866	25.91%	0.0488	0.0694	28.70%
β_{22}	0.1114	0.1442	20.51%	0.0636	0.0869	25.63%	0.0462	0.0661	29.11%
β_{23}	0.1230	0.1609	21.35%	0.0617	0.0844	25.61%	0.0449	0.0638	28.70%
df = 4									
Sample size	100			250			500		
Parameters	DP	SUR	Δ%	DP	SUR	Δ%	DP	SUR	Δ%
β_{10}	0.3931	0.4570	13.98%	0.2211	0.2625	15.19%	0.1578	0.1913	17.09%
β_{11}	0.1079	0.1248	13.52%	0.0696	0.0826	15.22%	0.0440	0.0529	16.58%
β_{12}	0.1179	0.1363	13.49%	0.0659	0.0783	15.32%	0.0476	0.0574	16.81%
β_{20}	0.4856	0.5568	12.78%	0.3091	0.3686	15.49%	0.2210	0.2702	17.82%
β_{21}	0.1054	0.1200	12.15%	0.0618	0.0733	15.08%	0.0473	0.0582	18.41%
β_{22}	0.1086	0.1236	12.17%	0.0617	0.0739	15.88%	0.0456	0.0554	17.30%
β_{23}	0.1204	0.1376	12.56%	0.0603	0.0717	15.19%	0.0438	0.0534	17.77%
df = ∞									
Sample size	100			250			500		
Parameters	DP	SUR	Δ%	DP	SUR	Δ%	DP	SUR	Δ%
β_{10}	0.3076	0.3016	-2.04%	0.2055	0.2031	-1.20%	0.1275	0.1270	-0.39%
β_{11}	0.0854	0.0837	-2.09%	0.0569	0.0565	-0.85%	0.0372	0.0371	-0.56%
β_{12}	0.0909	0.0891	-1.97%	0.0611	0.0605	-1.13%	0.0366	0.0365	-0.42%
β_{20}	0.4796	0.4696	-2.19%	0.2754	0.2730	-0.93%	0.1826	0.1810	-0.96%
β_{21}	0.0879	0.0860	-2.25%	0.0575	0.0572	-0.58%	0.0388	0.0384	-1.14%
β_{22}	0.0974	0.0955	-2.03%	0.0568	0.0564	-0.83%	0.0395	0.0395	-0.25%
β_{23}	0.0888	0.0868	-2.31%	0.0557	0.0552	-0.93%	0.0355	0.0355	-0.23%

Table 2: Posterior S.D., multivariate log-normal errors

Log-normal									
Sample size	100			250			500		
Parameters	DP	SUR	Δ%	DP	SUR	Δ%	DP	SUR	Δ%
β_{10}	0.2221	0.5537	56.72%	0.1536	0.4525	63.95%	0.1001	0.2927	64.94%
β_{11}	0.0720	0.1831	57.64%	0.0415	0.1266	65.06%	0.0272	0.0834	66.61%
β_{12}	0.0672	0.1741	58.37%	0.0438	0.1338	65.20%	0.0280	0.0867	66.95%
β_{20}	0.3363	0.8722	57.82%	0.2150	0.6088	63.42%	0.1382	0.4064	65.06%
β_{21}	0.0800	0.2119	58.78%	0.0439	0.1277	64.38%	0.0270	0.0826	66.39%
β_{22}	0.0716	0.1890	58.64%	0.0438	0.1277	64.44%	0.0297	0.0896	65.87%
β_{23}	0.0662	0.1749	58.58%	0.0398	0.1147	64.05%	0.0270	0.0806	65.62%

the number of equations as that of factors.[22] To make our empirical demonstration as general as possible, we do not impose symmetry or homogeneity restrictions on the model.

The dataset, taken from Malikov et al. (2016), contains 2397 observations on 285 large U.S. banks between 2001 and 2010. The data includes quantities and prices of the inputs, i.e. labour, physical assets and borrowed funding, and the quantity of output, which is the loans made by a bank. Given the relatively large sample size, it is possible for us to explore the performance of the DP-SUR with different sample sizes.

The demand for factors equation system may be written as

$$a_L = \frac{L}{Y} = \beta_{LL} + \beta_{LA}\frac{P_A}{P_L} + \beta_{LF}\frac{P_F}{P_L} + \beta_{LT}T + \varepsilon_L \tag{36}$$

$$a_A = \frac{A}{Y} = \beta_{AA} + \beta_{AL}\frac{P_L}{P_A} + \beta_{AF}\frac{P_F}{P_A} + \beta_{AT}T + \varepsilon_A \tag{37}$$

$$a_F = \frac{F}{Y} = \beta_{FF} + \beta_{FL}\frac{P_L}{P_F} + \beta_{FA}\frac{P_A}{P_F} + \beta_{FT}T + \varepsilon_F, \tag{38}$$

where L, A and F denote the quantity of labour, physical assets and borrowed funds, respectively; T denotes the trend variable; Y denotes output, and P_k is the price of factor k, with $k \in \{L, A, F\}$. For the errors we assume $(\varepsilon_{Li}, \varepsilon_{Ai}, \varepsilon_{Fi}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_i})$ where $\boldsymbol{\Sigma_i}$ is the covariance matrix of observation i. We allow the errors to be correlated across the three equations in the system, i.e. $\operatorname{cov}(\varepsilon_{ki}, \varepsilon_{si}) \neq 0$, with $k, s \in \{L, A, F\}$ indexing equations.

Given that the main objects of interest are the price elasticities, we report the posterior means and S.D. of the price elasticities of the three factors. With the generalized Leontief cost function, the cross price elasticities of the factors are given by

$$e_{ks} = \frac{1}{2}\frac{\beta_{ks}\left(P_k/P_s\right)^{-1/2}}{a_k}, \quad \forall k \neq s, \tag{39}$$

The own price elasticities are

$$e_{kk} = -\frac{1}{2}\frac{\sum_{s\neq k}\beta_{ks}\left(P_k/P_s\right)^{-1/2}}{a_k}. \tag{40}$$

Table 3 contains the posterior means of the price elasticities of the demand for factors. One can see that with both the 800 observation sub-sample and the full sample, the posterior means of all the elasticities are relatively small indicating that the demand for factors (labour, physical assets and borrowed fundings) of U.S. banks are relatively price inelastic.

Note that the own price elasticities of labour and physical assets are negative in both samples. In contrast, the own price elasticity of borrowed fundings is positive, although we note that the absolute values are extremely small[23] compared to those of the labour and physical assets. This shows that the demand for borrowed fundings is inelastic in the production of the U.S. banking industry. This is not surprising as the output of the banks is loans made to their customers, and funding is the source of loans.

There are some differences between the posterior means estimated with the DP-SUR and the Bayesian SUR assuming normal errors. Such differences are not observed in the simulation studies. However, it should be noted that the data was generated exactly with the model, i.e. the regression equation was correctly specified, which is not guaranteed with the empirical data.

[22] Note that the SUR and OLS estimators are exactly the same when all the equations in the system share the same explanatory variables.

[23] The posterior S.D. are also relatively large for this elasticity as shown in Table 4.

Table 3: Elasticities, U.S. banking industry: posterior means

800-observation sub-sample						
Input	Labour		Assets		Fundings	
Parameters	DP	SUR	DP	SUR	DP	SUR
Wage	-0.189	-0.422	0.375	0.254	-0.016	-0.007
Asset Price	-0.009	0.131	-0.690	-0.585	0.012	0.002
Funding Price	0.198	0.291	0.315	0.331	0.004	0.004
Full sample, 2397 observations						
Input	Labour		Assets		Fundings	
Parameters	DP	SUR	DP	SUR	DP	SUR
Wage	-0.149	-0.201	0.406	0.385	0.004	-0.001
Asset Price	-0.094	-0.068	-0.686	-0.668	-0.016	-0.005
Funding Price	0.243	0.270	0.280	0.283	0.012	0.006

Table 4 presents the posterior S.D. of the price elasticities estimated with the two samples. We observe that the posterior S.D estimated with the DP-SUR are smaller than those estimated with the Bayesian SUR assuming normality for all the price elasticities. This is not unexpected, as the elasticities are functions of the regression parameters in the equation system, which are estimated with smaller posterior S.D. with the semi-parametric DP-SUR than the parametric Bayesian SUR. The greatest percentage difference ($\Delta\%$) with the 800-observation sub-sample takes place with the cross price elasticity of the demand for labour with respect to the price of physical assets, for which the DP-SUR posterior S.D is 38.27% smaller than the SUR counterpart. With the full sample, the largest percentage difference is observed with the cross price elasticity of the demand for borrowed fundings with respect to the price of physical assets, which reached 39.15%.

Table 4: Elasticities, U.S. banking industry: posterior S.D.

800-observation sub-sample									
Input	Labour			Assets			Fundings		
Price	DP	SUR	$\Delta\%$	DP	SUR	$\Delta\%$	DP	SUR	$\Delta\%$
Wage	0.0341	0.0541	36.99%	0.0438	0.0484	9.60%	0.0276	0.0305	9.52%
Asset Price	0.0228	0.0369	38.27%	0.0304	0.0359	15.22%	0.0160	0.0230	30.59%
Funding Price	0.0236	0.0381	38.02%	0.0267	0.0317	15.82%	0.0148	0.0160	7.16%
Full sample, 2397 observations									
Input	Labour			Assets			Fundings		
Price	DP	SUR	$\Delta\%$	DP	SUR	$\Delta\%$	DP	SUR	$\Delta\%$
Wage	0.0189	0.0222	14.97%	0.0198	0.0210	6.01%	0.0102	0.0147	30.74%
Asset Price	0.0106	0.0131	19.38%	0.0143	0.0151	5.21%	0.0067	0.0110	39.15%
Funding Price	0.0154	0.0178	13.68%	0.0157	0.0161	2.48%	0.0081	0.0098	17.43%

4 Semi-parametric Approach to Random Effects Model

In this section we propose a semi-parametric Bayesian approach by introducing $\mathcal{DP}$ priors on the variances of the random effects and the errors. We follow the same approach as in the DP-SUR method in terms of applying $\mathcal{DP}$ prior on the hyperparameters.

Consider the following panel data model

$$y_{it} = \beta_1 x_{1it} + \cdots + \beta_K x_{Kit} + u_i + \eta_{it} = \beta_1 x_{1it} + \cdots + \beta_K x_{Kit} + \varepsilon_{it}, \tag{41}$$

where i and t index the cross section and time series dimensions of the data, respectively, y_{it} is the dependent variable, x_{kit} denote the explanatory variables, and the β_k, $k = 1, \ldots, K$ are the coefficients. u_i is the time-invariant unobservable of individual i, and η_{it} the error term.

In Bayesian methods the difference between the fixed and random effects lies in the choice of prior for the individual effects u_i. Fixed effects Bayesian methods assume a non-hierarchical prior for u_i, while for the random effects a hierarchical prior is assumed. The prior for u_i may be written as

$$u_i|d^2 \overset{iid}{\sim} \mathcal{N}\left(0, d^2\right), \tag{42}$$

where d^2 is the variance[24] of u_i. Assuming $\eta_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, the posterior under such a prior is then

$$u_i|\boldsymbol{y_i}, \boldsymbol{\beta}, d^2, \sigma^2 \sim \mathcal{N}\left(\mu_i, s^2\right), \tag{43}$$

where $\mu_i = s^2\sigma^{-2}\iota_T'(\boldsymbol{y_i} - \boldsymbol{X_i\beta})$, $s^2 = (d^{-2} + \sigma^{-2}\iota_T'\iota_T)^{-1}$, with ι_T being a $T \times 1$ vector of 1s. $\boldsymbol{X_i} = [x_{1it}, \ldots, x_{Kit}]$ is a $T \times K$ matrix of explanatory variables, and $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]'$ is a $T \times 1$ vector.

The likelihood of $\boldsymbol{\beta}$ marginalized over u_i in the Bayesian REM is

$$p\left(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{Q/2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})'\boldsymbol{\Sigma}(\boldsymbol{y_i} - \boldsymbol{X_i\beta})\right], \tag{44}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the $T \times 1$ vector $\boldsymbol{\varepsilon_i} = [\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}]'$. Assuming that $\mathbb{E}\left[\eta_{it}u_j|\boldsymbol{X}\right] = 0,\ \forall\ i, t, j$ (Greene, 2012), the covariance matrix of the composite error $\boldsymbol{\varepsilon_i}$ is

$$\mathrm{Cov}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\Sigma}_{T\times T} = \begin{bmatrix} \sigma^2 + s^2 & s^2 & \cdots & s^2 \\ s^2 & \sigma^2 + s^2 & \cdots & s^2 \\ \vdots & \vdots & \ddots & \vdots \\ s^2 & s^2 & \cdots & \sigma^2 + s^2 \end{bmatrix}, \tag{45}$$

where σ^2 is the variance of η_{it}, and s^2 is the variance of u_i.

We relax the identically distributed assumptions for η_{it} and u_i, respectively by introducing $\mathcal{DP}$ priors on the variances. This will have the effect of grouping errors over both the cross section dimension i and the time series dimension t, with those in the same group sharing the same hyperparameter. The $\mathcal{DP}$ prior is written as

$$\begin{aligned} G &\sim \mathcal{DP}\left(\alpha_\eta, G_0\right) \\ \sigma_{it}^2|G &\sim G, \end{aligned} \tag{46}$$

where α_η and G_0 are the concentration parameter and base distribution of the $\mathcal{DP}$ prior, respectively.

For c_{it} denoting the group id of η_{it}, we write the covariance matrix of the $T \times 1$ vector $\boldsymbol{\eta_i}$ as

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta_i}} = \begin{bmatrix} \sigma_{c_{i1}}^{*2} & 0 & \cdots & 0 \\ 0 & \sigma_{c_{i2}}^{*2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{c_{iT}}^{*2} \end{bmatrix}_{T\times T}. \tag{47}$$

Here if $c_{it} \neq c_{is}$, then η_{it} and η_{is} are not in the same group, and $\sigma_{c_{it}}^{*2} \neq \sigma_{c_{is}}^{*2}$. Therefore the grouping of the hyper-parameters of η_{it} is neither restricted to the cross section i nor the time

[24] Note that the variances of u_i and η_{it} is often assumed to be random, and have their own priors. For the moment we leave them fixed for the sake of simplicity.

series t, accommodating heteroscedasticity across t for a given i, and across i for a given t. To our knowledge, the restriction that $c_{it} = c_{jt}, \forall i, j, t$ has remained in the literature on semi-parametric REM (see Kleinman and Ibrahim, 1998 and Kyung et al., 2010).

The prior for the individual effects u_i in our DP-REM can be written using the following hierarchical structure:

$$\begin{aligned} F &\sim \mathcal{DP}\left(\alpha_u, F_0\right) \\ d_i^2|F &\sim F \\ u_i|d_i^2, F &\sim \mathcal{N}\left(0, d_i^2\right). \end{aligned} \tag{48}$$

d_i^2 is the prior variance of the random effects u_i, α_u is the concentration parameter, and F_0 the base distribution of the $\mathcal{DP}$ prior. The use of an independent $\mathcal{DP}$ prior on the hyper-parameters of individual effects u_i, generates groupings over the N individuals in the cross section. Random effects u_i's that belong to the same group are generated from a distribution with the same hyper-parameter, thereby relaxing the REM assumption that the individual effects are identically distributed. For u_i and u_j in the same group, then $d_i^2 = d_j^2$.

We maintain the assumption that the idiosyncratic errors η_{it} and the individual effects u_i satisfy $\mathbb{E}\left[\eta_{it} u_j | \boldsymbol{X}\right] = 0, \ \forall\, i, t, j$, and introduce independent $\mathcal{DP}$ priors[25] for their hyper-parameters. Then the correlation between ε_{it} and ε_{is} $(t \neq s)$ is solely caused by u_i, i.e., the form of the covariance matrix of the $T \times 1$ vector of compound errors $\boldsymbol{\varepsilon_i}$ has the same structure as in (45).

Although the u_i are no longer identically distributed, for each particular u_i a conjugate normal prior[26] can be introduced. The posterior of each u_i is then a normal distribution, the means and variances of which are different across the cross section i, i.e.

$$u_i|\boldsymbol{y_i}, \boldsymbol{\beta}, d_{c_i}^{*2}, \sigma_{c_{it}}^{*2} \sim \mathcal{N}\left(\mu_i, s_i^2\right), \tag{49}$$

where

$$\mu_i = s_i^2 \iota_T' \boldsymbol{\Sigma}_{\boldsymbol{\eta_i}}^{-1} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta}) \tag{50}$$

is the posterior mean of u_i. The posterior variance is

$$s_i^2 = \left(d_{c_i}^{*-2} + \iota_T' \boldsymbol{\Sigma}_{\boldsymbol{\eta_i}}^{-1} \iota_T\right)^{-1}. \tag{51}$$

From (51) we observe that the posterior variance of the random effects u_i is the sum of $d_{c_i}^{*-2}$ (the inverse of the unique value of d_i^2) the prior variance of u_i, and all the elements in $\boldsymbol{\Sigma}_{\boldsymbol{\eta_i}}$. Given that the posterior variance s_i^2 is a function of $\boldsymbol{\Sigma}_{\boldsymbol{\eta_i}} \neq \boldsymbol{\Sigma}_{\boldsymbol{\eta_j}}$, each random effect u_i will have a unique posterior variance s_i^2. This feature of our DP-REM takes advantage of panel data to allow each $T \times 1$ compound error vector $\boldsymbol{\varepsilon_i}$ to have its own unique distribution, which is infeasible with cross sectional data (as in the DP-SUR case before).

It is useful to see the two $\mathcal{DP}$ priors for the REM from the perspective of the CRP. In this case, instead of individual customers entering one Chinese restaurant, we might think of teachers and students going for lunch. In this school, there are N classes, indexed $i \in \{1, \ldots, N\}$. Each class is headed by a teacher with students in each i^{th} class , indexed $t \in \{1, \ldots, T\}$.

The N teachers enter the same restaurant. Each of the $N \times T$ students enter another restaurant, and decide whether to sit at a new table or an occupied table with some students[27] who are not necessarily from her own class. This allows the grouping of students over both

[25] It is also possible to introduce, e.g., a Hierarchical Dirichlet Process prior (Teh et al., 2005) for both the hyper-parameters of u_i and η_{it}. However, as here we are considering the relaxation of the identical distribution assumptions on u_i and η_{it} in the REM framework, this falls out of the purview of this paper.

[26] Here we adopt a prior whose mean is 0, and variance is 1000.

[27] The probabilities for both teachers and students to sit at a table with some person(s) already, or at a new table are the same as in Section 2.3.2.

the cross section and time series dimensions. Such a way of grouping is the most unrestricted, and it is in accordance with the fact that generally how students decide to sit together is not determined by their class or student numbers.

In this example we can think of teacher i as representing d_i^2, the variance of the individual effect u_i. A $\mathcal{DP}$ prior is introduced so they are grouped over the cross section dimension. Student it represents σ_{it}^2, the variances of the idiosyncratic error η_{it}. We allow $\sigma_{it}^2 \neq \sigma_{is}^2 \quad \forall s$, heteroscedasticity over time for given i, and $\sigma_{it}^2 \neq \sigma_{jt}^2 \quad \forall s$, heteroscedasticity over i for given t. $\boldsymbol{\Sigma_{\eta_i}}$ in (47) is then different for each i.

In our semi-parametric model we relax the assumption that u_i and η_{it} are both identically distributed by assigning them to groups with a $\mathcal{DP}$ prior for their respective hyper-parameters. The $\mathcal{DP}$ priors are independent, as reflected in the CRP by the teachers and students dining at two separate restaurants. This reflects the maintained assumption that the individual effects u_i and η_{it} are mean independent, i.e. $\mathbb{E}\left[\eta_{it} u_j | \boldsymbol{X}\right] = 0, \ \forall \ i, t, j$. Under this assumption, individual specific characteristics are captured by the individual effects u_i, which are grouped with a $\mathcal{DP}$ prior by their hyper-parameters. Meanwhile, though the idiosyncratic errors η_{it} are indexed with both an individual subscript i and a time subscript t, there is nothing individual or time specific in them. Thus, the grouping of η_{it} with no regard of the individual and time dimensions is viable with an independent $\mathcal{DP}$ prior for their hyper-parameters.

The semi-parametric DP-REM differs from the parametric REM in that the form of the error covariance matrix for the i^{th} individual is written as

$$\text{Cov}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\Sigma_i} = \begin{bmatrix} \sigma_{c_{i1}}^{*2} + s_i^2 & s_i^2 & \cdots & s_i^2 \\ s_i^2 & \sigma_{c_{i2}}^{*2} + s_i^2 & \cdots & s_i^2 \\ \vdots & \vdots & \ddots & \vdots \\ s_i^2 & s_i^2 & \cdots & \sigma_{c_{iT}}^{*2} + s_i^2 \end{bmatrix}, \tag{52}$$

where each observation i is associated with different hyperparameters. For the choice of base distributions, we use the inverse gamma distribution, the conjugate prior for the variance of a normal distribution, i.e.

$$\begin{aligned} F_0 &\equiv \mathcal{IG}\left(a_u, b_u\right) \\ G_0 &\equiv \mathcal{IG}\left(a_\eta, b_\eta\right), \end{aligned} \tag{53}$$

where a_u and a_η are the shape hyper-parameters, and b_u and b_η denote, respectively, the rate hyper-parameters of F_0 and G_0. The likelihood of $\boldsymbol{\beta}$ marginalized over u_i is given by

$$p\left(\boldsymbol{y_i} | \boldsymbol{\beta}, \boldsymbol{\Sigma_i^*}\right) = \frac{1}{(2\pi)^{Q/2}} |\boldsymbol{\Sigma_i^*}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})' \boldsymbol{\Sigma_i^{*-1}} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})\right]. \tag{54}$$

Compared with the marginal likelihood in (44), the covariance matrix of the composite error vector $\boldsymbol{\varepsilon_i}$ is allowed to be different for each individual i in the panel. Given a conjugate normal prior for $\boldsymbol{\beta}$, i.e.

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{b_0}, \boldsymbol{V_0}),$$

where $\boldsymbol{b_0}$ and $\boldsymbol{V_0}$ denote respectively, prior mean and covariance matrix of $\boldsymbol{\beta}$, the posterior of $\boldsymbol{\beta}$ marginalized over u_i is

$$\boldsymbol{\beta} | \boldsymbol{y}, d_{c_i}^{*2}, \sigma_{c_{it}}^{*2} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{V}). \tag{55}$$

$$\boldsymbol{V} = \left(\boldsymbol{V_0^{-1}} + \sum_{i=1}^{N} \boldsymbol{X_i'} \boldsymbol{\Sigma_i^{-1}} \boldsymbol{X_i}\right)^{-1}, \tag{56}$$

denotes the posterior covariance matrix, and $\boldsymbol{b}$ is the posterior mean vector, which we write as

$$\boldsymbol{b} = \boldsymbol{V}\left(\boldsymbol{V_0^{-1}} \boldsymbol{b_0} + \sum_{i=1}^{N} \boldsymbol{X_i'} \boldsymbol{\Sigma_i^{-1}} \boldsymbol{y_i}\right). \tag{57}$$

For $\Theta_\eta = \{\sigma_{it}^2\}$, $U = \{u_i\}$, and $\Theta_u = \{d_i^2\}$, a Gibbs sampler for this DP-REM can be written as:

$$\begin{aligned}
\Theta_\eta|\boldsymbol{y}, \boldsymbol{X}, U, \boldsymbol{\beta}, \Theta_u, \alpha_u, \alpha_\eta \\
\Theta_u|\boldsymbol{y}, \boldsymbol{X}, U, \boldsymbol{\beta}, \Theta_\eta, \alpha_u, \alpha_\eta \\
U|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \Theta_u, \Theta_\eta, \alpha_u, \alpha_\eta \\
\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, U, \Theta_u, \Theta_\eta, \alpha_u, \alpha_\eta \\
\alpha_u|\boldsymbol{y}, \boldsymbol{X}, U, \boldsymbol{\beta}, \Theta_u, \Theta_\eta, \alpha_\eta \\
\alpha_\eta|\boldsymbol{y}, \boldsymbol{X}, U, \boldsymbol{\beta}, \Theta_u, \Theta_\eta, \alpha_u.
\end{aligned} \tag{58}$$

The Gibbs sampler for the regression parameters, hyperparameters and concentration parameters of the two $\mathcal{DP}$ are similar to those for the DP-SUR in 3.2. In the DP-REM, the random effects have a mixture of normal distributions. The posterior mean and variance of each particular u_i are in (50) and (51), respectively. For each i a u_i is drawn from $\mathcal{N}\left(\mu_i, s_i^2\right)$ with the Gibbs sampler.

4.1 Correlated Random Effects Model

We also consider the Correlated Random Effects Model (CREM) introduced by Mundlak (1978) and further discussed by Chamberlain (1980). A natural extension of the REM, CREM offers a middle ground between the fixed and random effects. Consider the following model for the panel data

$$y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + v_i + \eta_{it}, \tag{59}$$

where v_i is the random effects. While maintaining the GLS structure of the REM, CREM allows the individual effects to be correlated with $\boldsymbol{X}_i$, representing the correlation using a linear function of the means of $\boldsymbol{X}_i$, i.e.

$$v_i = \beta_3 \bar{x}_{1i} + \beta_4 \bar{x}_{2i} + u_i, \tag{60}$$

The CREM model is then

$$y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 \bar{x}_{1i} + \beta_4 \bar{x}_{2i} + u_i + \eta_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 \bar{x}_{1i} + \beta_4 \bar{x}_{2i} + \varepsilon_{it}. \tag{61}$$

Then the $\mathcal{DP}$ prior could be introduced on u_i and η_{it} as in the REM case.

4.2 DP-REM/CREM Results

We carry out a series of simulation experiments to demonstrate the performance of our DP-REM and DP-CREM methods relative to the standard Bayesian REM and CREM. The simulation experiments have been designed for the same purpose as those for the DP-SUR in Section 3.3.

For the REM model we assume

$$y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + u_i + \eta_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \varepsilon_{it} \tag{62}$$

where the explanatory variables are generated from the following normal distributions

$$x_{1,it} \overset{iid}{\sim} \mathcal{N}(1,1), \;\; x_{2,it} \overset{iid}{\sim} \mathcal{N}(3,1).$$

We set the coefficients in (62) to

$$\beta_1 = 5, \;\; \beta_2 = 10.$$

The coefficients in the CREM model

$$y_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 \bar{x}_{1i} + \beta_4 \bar{x}_{2i} + u_i + \eta_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 \bar{x}_{1i} + \beta_4 \bar{x}_{2i} + \varepsilon_{it} \tag{63}$$

are set to

$$\beta_1 = 5, \;\; \beta_2 = 10, \;\; \beta_3 = -2, \;\; \beta_4 = 2. \tag{64}$$

Below we present the simulation results. We first report the results where the errors, u_i and η_{it} are assumed to follow t-distributions and then those with log-normal distributions.

t-Distributed Random Effects and Errors

Table 5 reports the averages of the posterior S.D.s of the REM coefficients estimated with both methods, and the average of the percentage differences between the DP-REM and REM posterior S.D.s. The largest differences between the two estimators with respect to the posterior S.D. are observed when df = 2, where the t-distributions of the random effects and the errors have the heaviest tails. As expected, these differences decrease as the df increase, where the tails of the t distributions become less 'heavy'. In the lowest panel where the errors have normal distributions (equivalent to df being infinity), the DP-REM and normal REM posterior S.D. are almost the same, as the t-distribution is the normal distribution in this case.

We also note that the percentage differences increase slightly when the sample size becomes larger for all three finite df. This is expected given that there are more extreme realizations in larger samples, and our DP-REM method detects such heterogeneity and assign them into the same group. In contrast, the Bayesian REM method assuming normality flattens the normal posterior distribution for the extreme values, leading to larger posterior S.D.

Table 5: Posterior S.D., REM with t distributed unobservables

df = 2									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0751	0.1419	43.18%	0.0484	0.0956	47.88%	0.0340	0.0671	47.92%
β_2	0.0499	0.0883	41.45%	0.0316	0.0575	47.04%	0.0213	0.0422	48.50%
df = 3									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0627	0.0803	20.61%	0.0366	0.0475	22.36%	0.0290	0.0379	22.94%
β_2	0.0419	0.0530	20.05%	0.0237	0.0309	22.65%	0.0183	0.0242	24.12%
df = 4									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0588	0.0667	11.51%	0.0346	0.0393	11.54%	0.0270	0.0310	12.59%
β_2	0.0394	0.0442	10.66%	0.0225	0.0257	12.15%	0.0171	0.0198	13.67%
df = ∞									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0484	0.0484	-0.02%	0.0274	0.0274	-0.20%	0.0211	0.0210	-0.19%
β_2	0.0301	0.0301	0.03%	0.0180	0.0179	-0.45%	0.0139	0.0139	0.01%

Table 6 reports the averages of posterior S.D. of the CREM coefficients, and the averages of the percentage differences between the two estimators. β_1 and β_2 denote the two original explanatory variables, whereas β_3 and β_4 capture the effect of the respective sample means for each individual in the panel. The findings are similar to the REM case in that the percentage differences between the posterior S.D. estimated with our DP-CREM and the parametric Bayesian CREM are the largest with df equal to 2, and decrease with the increase in the df. Such differences between the two methods regarding the posterior S.D. are almost zero when the df is infinity, when the t-distribution becomes normal distribution. The percentage differences also increase slightly in the three finite df cases when the sample size becomes large due to more extreme values in the unobservables.

Table 6: Posterior S.D., CREM with t distributed unobservables

df = 2									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0860	0.1482	41.82%	0.0446	0.0935	48.33%	0.0394	0.0765	49.06%
β_2	0.0793	0.1409	41.60%	0.0471	0.0978	48.84%	0.0371	0.0790	50.10%
β_3	0.3665	0.6575	37.62%	0.2743	0.4854	42.63%	0.1595	0.3255	48.31%
β_4	0.1545	0.2768	40.26%	0.1197	0.2001	43.61%	0.0671	0.1412	49.43%
df = 3									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0689	0.0887	21.09%	0.0368	0.0472	21.51%	0.0294	0.0387	23.24%
β_2	0.0659	0.0845	20.99%	0.0383	0.0493	21.96%	0.0304	0.0398	22.92%
β_3	0.3077	0.3884	18.53%	0.1833	0.2434	23.10%	0.1373	0.1825	24.09%
β_4	0.1279	0.1619	19.54%	0.0749	0.0989	23.23%	0.0575	0.0763	24.06%
df = 4									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0639	0.0723	11.29%	0.0345	0.0399	12.97%	0.0256	0.0299	14.09%
β_2	0.0607	0.0689	11.60%	0.0358	0.0414	13.09%	0.0263	0.0306	14.04%
β_3	0.2857	0.3175	9.35%	0.1740	0.1982	11.91%	0.2547	0.2944	12.80%
β_4	0.1177	0.1316	10.09%	0.0707	0.0808	12.25%	0.0915	0.1056	12.78%
df = ∞									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0518	0.0517	-0.31%	0.0295	0.0295	-0.14%	0.0228	0.0228	-0.02%
β_2	0.0503	0.0502	-0.25%	0.0298	0.0296	-0.59%	0.0225	0.0224	-0.59%
β_3	0.2427	0.2416	-0.51%	0.1359	0.1365	0.34%	0.1076	0.1075	-0.10%
β_4	0.0952	0.0951	-0.13%	0.0586	0.0586	0.03%	0.0431	0.0431	-0.26%

Log-normal Distributed Random Effects and Errors

Table 7 contains the average of posterior S.D. estimated with the DP-REM and normal REM, and the percentage differences between them. It can be seen that our DP-REM posterior S.D. are smaller than those estimated by Bayesian REM assuming normality in all cases. Due to the fact that the log-normal distribution is heavy tailed, the percentage differences are more than 70% in all cases, which increase slightly when the sample size gets larger, as more extreme realizations are present in larger samples.

Table 7: Posterior S.D., REM with log-normal distributed unobservables

Log-normal									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0733	0.3656	79.56%	0.0384	0.2228	82.43%	0.0305	0.1788	82.86%
β_2	0.0608	0.2350	73.37%	0.0311	0.1471	78.75%	0.0240	0.1163	79.02%

The posterior S.D. of the DP-CREM and CREM averaged over the simulated samples are reported in Table 8, along with the average of the percentage difference between the two S.D. It can be seen that the posterior S.D. estimated with our DP-CREM are more than 70% smaller than those estimated with the normal Bayesian CREM for all coefficients. The percentage differences also increase when the sample size becomes larger leading to the presence of more extreme realizations.

Table 8: Posterior S.D., CREM with log-normal distributed unobservables

Log-normal									
Sample size	100			300			500		
Parameters	DP	REM	Δ%	DP	REM	Δ%	DP	REM	Δ%
β_1	0.0754	0.3920	81.45%	0.0381	0.2479	85.24%	0.026	0.199	86.67%
β_2	0.0769	0.3785	81.37%	0.0390	0.2578	85.15%	0.024	0.194	87.17%
β_3	0.7722	1.7634	67.37%	0.3002	1.1358	74.96%	0.157	0.898	80.77%
β_4	0.3342	0.7583	70.91%	0.1080	0.4828	78.79%	0.063	0.378	82.24%

4.3 DP-REM/CREM Empirical Examples

In this section we present the results based upon two empirical examples. In the first we estimate the cost function of U.S. banks, and in the second we estimate the wages of U.S. workers.

Bank Cost Function

We first apply our DP-REM and DP-CREM methods to the dataset in Feng and Serletis (2009) on the costs of 218 U.S. banks whose assets are between 1 and 3 billion dollars (2000 value), covering a period of 8 years from 1998 to 2005. There are three inputs, labour, borrowed funds and physical capital; and three outputs, consumer loans, non-consumer loans and securities. The functional form is the simple translog cost function (Christensen and Greene, 1976). For industry i with n inputs and m outputs we write

$$\begin{aligned} \ln C_{it} = & \sum_{j=1}^{m} \alpha_j \ln q_{j,it} + \frac{1}{2} \sum_{j=1}^{m} \sum_{k=1}^{m} \delta_{jk} \ln q_{j,it} \cdot \ln q_{k,it} + \sum_{r=1}^{n} \beta_r \ln p_{r,it} \\ & + \frac{1}{2} \sum_{r=1}^{n} \sum_{s=1}^{n} \phi_{rs} \ln p_{r,it} \cdot \ln p_{s,it} + \sum_{r=1}^{n} \sum_{j=1}^{m} \gamma_{rj} \ln p_{r,it} \cdot \ln q_{j,it} + u_i + \eta_{it}. \end{aligned} \tag{65}$$

where C is cost, q_j is output quantity j, and p_r is input price r. We impose the linear homogeneity in input prices on the cost function, which in the translog case can be expressed as

$$\begin{aligned}
&\sum_{r=1}^{n} \beta_r = 1, \\
&\sum_{s=1}^{n} \phi_{sr} = 0, \ r = 1, 2, \ldots, n, \\
&\sum_{r=1}^{n} \gamma_{rj} = 0, \ j = 1, 2, \ldots, m.
\end{aligned} \tag{66}$$

Table 9 contains the posterior means and S.D. of the free coefficients in the REM. To differentiate the inputs from outputs, we index the three outputs with numbers, and index the inputs by letters, with l and f standing for labour and borrowed funds, respectively. The posterior mode of the number of clusters in the random effects is 2, and that in the errors is 3, which is the evidence of the existence of heterogeneity. The degree of heterogeneity is not very high according to this result. Thus, as will be discussed later, the reduction in the posterior S.D. is not as large as when the degree of heterogeneity is high, such as when the errors are log-normally distributed in the simulation studies.

The posterior means of all the coefficients for first order terms (the β's and α's) are of the same magnitude with the two methods with the exception of α_1 for customer loans, but it is insignificant with the REM. Although a number of the coefficients for the crossproduct terms have different signs across the two methods, the coefficients are not significant. Consistent with the detection of heterogeneity in the random effects and the errors, the posterior S.D. of our DP-SUR method are smaller than those estimated by parametric Bayesian REM for all coefficients. Most of the percentage differences presented in the last column are more than 10%, with the largest being 24.51% for δ_{33}.

Table 9: U.S. Bank Cost Function REM

	Mean		S.D.		
Coefficients	DP-REM	REM	DP-REM	REM	Δ%
β_l	-0.6583	-0.6099	0.1480	0.1795	17.56%
β_f	1.4475	1.0758	0.1068	0.1091	2.12%
α_1	-0.1908	-0.0240	0.0799	0.0944	15.35%
α_2	0.6870	0.5777	0.1405	0.1562	10.05%
α_3	0.9486	0.8439	0.1284	0.1558	17.55%
ϕ_{ll}	0.1087	0.0405	0.0170	0.0214	20.44%
ϕ_{ff}	0.1670	0.1266	0.0055	0.0071	22.37%
ϕ_{lf}	-0.1664	-0.0891	0.0079	0.0097	18.57%
δ_{11}	0.0141	0.0145	0.0033	0.0041	20.02%
δ_{22}	0.1267	0.1098	0.0192	0.0218	12.13%
δ_{33}	0.0934	0.1030	0.0134	0.0178	24.51%
δ_{12}	0.0033	-0.0113	0.0067	0.0082	17.41%
δ_{13}	0.0011	-0.0032	0.0048	0.0063	23.39%
δ_{23}	-0.1441	-0.1318	0.0117	0.0152	22.98%
γ_{l1}	0.0049	0.0181	0.0057	0.0063	9.82%
γ_{l2}	-0.0030	0.0370	0.0131	0.0151	13.60%
γ_{l3}	0.0086	-0.0030	0.0112	0.0135	17.36%
γ_{f1}	-0.0157	-0.0220	0.0035	0.0043	19.51%
γ_{f2}	0.0172	-0.0241	0.0088	0.0100	12.21%
γ_{f3}	0.0217	0.0570	0.0064	0.0080	19.82%

We also estimate the model with the DP-CREM.[28] The posterior mode of the number of clusters in the random effects and the errors are 2 and 3, respectively, indicating the existence of heterogeneity. The coefficients of the explanatory variables (not the sample means of them), have DP-CREM posterior S.D. smaller than their CREM counterparts, and to similar magnitudes as in Table 9. However, the regression parameters are all highly insignificant for the sample means of the explanatory variables, as their posterior S.D. are very large compared with their posterior means. This indicates that the data do not support the CREM specification, i.e. the correlation between v_i and $\boldsymbol{X_i}$ is a linear function of the means of explanatory variables.

4.4 U.S. Individual Wage

In this section we present the results of a wage model for U.S. workers using the data in Cornwell and Rupert (1988). The data covers 595 individuals over a period of 7 years, from 1976 to 1982. This sample size allows us to demonstrate our method with sub-samples of 100 and 250 individuals.

The model is given by

$$\ln Wage_{it} = \beta_1 E_{it} + \beta_2 M_{it} + \beta_3 F_{it} + \beta_4 Ed_{it} + v_i + \varepsilon_{it},$$

where the dependent variable is the logged wage, and the explanatory variables are experience in years (E), dummies for marriage status (M) and the individual being female (F), as well as the years of education (Ed). As there is a strong reason to suspect that the unobserved individual effect v_i to be endogenous due to omitted variables such as personal capability and motivation, we apply our DP-CREM model, and write v_i as

$$v_i = \tilde{\beta}_1 \overline{E}_i + \tilde{\beta}_2 \overline{M}_i + u_i. \tag{67}$$

The means of experience and marriage status of individual i are included as they are the two time variant variables in the original model.

Table 10: U.S. Individual Wage CREM

	Mean				S.D.					
Sample size	100		250		100			250		
Parameters	DP	REM	DP	REM	DP	REM	Δ%	DP	REM	Δ%
β_1	0.100	0.102	0.094	0.099	0.0025	0.0032	24.18%	0.0015	0.0020	25.49%
β_2	0.027	0.038	-0.043	-0.070	0.0375	0.0524	28.43%	0.0231	0.0308	25.07%
β_3	5.184	1.945	5.077	1.825	0.2646	0.4256	37.83%	0.1576	0.2602	39.42%
β_4	0.085	0.334	0.075	0.288	0.0189	0.0213	11.22%	0.0091	0.0131	30.74%
$\tilde{\beta}_1$	-0.091	-0.058	-0.085	-0.058	0.0046	0.0077	40.21%	0.0028	0.0056	50.17%
$\tilde{\beta}_2$	5.450	1.603	5.668	2.261	0.2968	0.3381	12.23%	0.1493	0.2091	28.60%

Table 10 contains the posterior means, S.D. and the percentage difference between the S.D. estimated with our DP-CREM and Bayesian CREM assuming normality for both the 100 and 250 individual sub-sample. The two coefficients for the means of time variant explanatory variables, $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are both significant with both sub-samples, indicating that v_i actually is correlated to the explanatory variables, confirming our suspicion. As for the coefficients for the explanatory variables themselves (β_1 to β_4), the posterior means are all of the same signs with the DP-CREM and DP-REM. As expected, the experience and education are both positively correlated with the wage of the workers. The coefficient for the gender dummy is also positive,

[28] The results are not presented here for the sake of clearness, but are available on request.

demonstrating that within the two datasets, there is no evidence of gender discrimination in the wages. Heterogeneity is detected in both sub-samples, as the posterior modes of the numbers of clusters in the random effects and errors are 2 and 3 with the 100 individual sub-sample, and 3 and 4 with the 250 individual one. Our semi-parametric DP-CREM provides smaller posterior S.D. for all coefficients with both sub-samples.

5 Conclusion

In this paper we address the potential violation of the assumptions made by parametric Bayesian GLS estimators that the unobservables are homogeneous and normally distributed. Such assumptions are likely to be problematic in two ways. One is that the observations are likely to be heterogeneous, particularly in situations where micro data are used. The other is that making assumptions on the distribution of the unobservable is always risky, for one could never truly know it in empirical researches. We introduce a Dirichlet process ($\mathcal{DP}$) prior on the distribution of the parameters of the errors, so that the number of groups is decided jointly by the data and the prior. The distribution of the errors is then a non-parametric mixture of a variable number of normal distributions, which is able to model a large variety of distributions. Two special cases of the GLS type estimators are then introduced, which are the SUR for equation systems and the REM/CREM for panel data models.

Our DP-SUR and DP-REM/DP-CREM methods are demonstrated with a series of simulation experiments consisting of three scenarios, where the unobservables follow normal distributions, t-distributions which are one type of scale mixtures of normals, and log-normal distributions, respectively. The results show that in the homogeneous normal case, our DP-SUR and DP-REM/DP-CREM methods give similar posterior means and S.D. to those estimated by their parametric counterparts assuming normality. When the errors follow t-distributions, the degrees of freedom of the t-distribution control how heavy the tails are, which reflects the strength of heterogeneity in the unobservables. Our simulation results show that the posterior S.D. estimated with the DP-SUR and DP-REM/DP-CREM are smaller than those estimated with the normal Bayesian methods. Such efficiency gains are the largest when df is 2 that represents the strongest heterogeneity. The efficiency gains become smaller when the df increases, for the tails are less heavy, i.e. the heterogeneity is less strong. The simulations with log-normal unobservables are used to demonstrate the robustness of our method with asymmetric, fat tailed distributions. The results demonstrated that the posterior S.D. estimated with our DP-SUR and DP-REM/DP-CREM method are more than 50% smaller than those estimated with the parametric Bayesian estimators assuming normality. Moreover, the efficiency gains increase slightly with larger sample sizes when the distribution of the unobservables are non-normal, which is the result of more extreme realizations in large samples, and our semi-parametric methods effectively group these extreme observations.

We apply our DP-SUR method to estimate the demands for production factors with the generalized Leontief cost function using a dataset of the U.S. banking industry. We estimated the model with an 800-observation sub-sample and the full sample. Heterogeneity is detected in both the sub-sample and the full sample. The DP-SUR posterior S.D. are smaller than the normal Bayesian SUR ones for all the demand elasticities, which shows that it is more preferable to use a semi-parametric method such as our DP-SUR.

Our DP-REM/DP-CREM are applied to two datasets as well. The first is a U.S. bank cost functions data. The REM seems to fit the datasets better. Heterogeneity is detected in the U.S. bank data, and our DP-REM achieved smaller posterior S.D. than the parametric Bayesian REM. The second application is a U.S. individual wage model, where there is a strong reason to suspect that the unobserved individual effects are correlated to the explanatory variables like education due to unobserved individual features such as personal capability. Such suspicion is

confirmed by the CREM. In addition, our DP-CREM detects heterogeneity in this dataset as well, and obtains smaller posterior S.D. than the parametric Bayesian CREM.

A Posterior Means for DP-SUR Simulations

A.1 Multivariate t-distributed Errors

Table 11 gives the posterior means averaged over the samples that are estimated with the DP-SUR and SUR assuming normality. One can see that the posterior means estimated with both our semi-parametric DP-SUR and the Bayesian SUR assuming normality are similar to each other. In addition, they are both close to the true values of the coefficients in all cases.

Table 11: Posterior means, multivariate t errors

df = 2							
Sample size	Truth	100		250		500	
Parameters		DP	SUR	DP	SUR	DP	SUR
β_{10}	1	1.035	0.942	1.017	1.044	0.986	1.051
β_{11}	-0.5	-0.500	-0.484	-0.491	-0.488	-0.500	-0.519
β_{12}	1.6	1.590	1.606	1.591	1.585	1.606	1.590
β_{20}	1.5	1.524	1.568	1.482	1.540	1.484	1.387
β_{21}	-1.2	-1.195	-1.194	-1.201	-1.194	-1.206	-1.230
β_{22}	-0.7	-0.702	-0.719	-0.697	-0.700	-0.696	-0.684
β_{23}	2	2.003	2.006	1.998	1.980	2.010	2.004
df = 3							
Sample size	Truth	100		250		500	
Parameters		DP	SUR	DP	SUR	DP	SUR
β_{10}	1	1.008	1.002	1.015	1.030	0.999	0.989
β_{11}	-0.5	-0.497	-0.493	-0.502	-0.499	-0.498	-0.499
β_{12}	1.6	1.600	1.600	1.596	1.591	1.598	1.601
β_{20}	1.5	1.507	1.507	1.493	1.502	1.503	1.504
β_{21}	-1.2	-1.194	-1.183	-1.198	-1.199	-1.202	-1.199
β_{22}	-0.7	-0.699	-0.696	-0.698	-0.699	-0.703	-0.702
β_{23}	2	1.995	1.989	1.999	2.001	1.995	1.996
df = 4							
Sample size	Truth	100		250		500	
Parameters		DP	SUR	DP	SUR	DP	SUR
β_{10}	1	0.973	0.979	1.000	1.000	0.989	0.992
β_{11}	-0.5	-0.500	-0.499	-0.516	-0.519	-0.495	-0.495
β_{12}	1.6	1.610	1.609	1.606	1.608	1.603	1.600
β_{20}	1.5	1.516	1.536	1.519	1.522	1.509	1.495
β_{21}	-1.2	-1.205	-1.208	-1.192	-1.192	-1.202	-1.202
β_{22}	-0.7	-0.708	-0.713	-0.700	-0.701	-0.704	-0.701
β_{23}	2	1.996	1.997	2.000	1.995	1.998	1.999
df = ∞							
Sample size	Truth	100		250		500	
Parameters		DP	SUR	DP	SUR	DP	SUR
β_{10}	1	0.9895	0.9864	0.9853	0.9846	0.9813	0.9835
β_{11}	-0.5	-0.4969	-0.4974	-0.4959	-0.4957	-0.4904	-0.4905
β_{12}	1.6	1.6030	1.6040	1.6040	1.6044	1.6005	1.5999
β_{20}	1.5	1.4706	1.4706	1.5122	1.5122	1.4675	1.4677
β_{21}	-1.2	-1.2115	-1.2118	-1.1959	-1.1965	-1.1976	-1.1977
β_{22}	-0.7	-0.6984	-0.6985	-0.7004	-0.7003	-0.6911	-0.6912
β_{23}	2	1.9950	1.9953	2.0000	2.0004	1.9973	1.9979

A.2 Multivariate Log-normal Errors

Table 12 contains the posterior means estimated with the two methods with multivariate log-normal errors. In all three samples the two posterior means of all the slope parameters are similar, and are close to the truth. The intercepts β_{10} and β_{20} estimated with our DP-SUR, however, are farther away from the true values. The fact that the log-normal distribution is skewed influences the posterior means of intercepts, when the $\mathcal{DP}$ mixture model mixes normal distributions to model the log-normal distribution.

Table 12: Posterior means, multivariate log-normal errors

Log-normal							
Sample size	Truth	100		250		500	
Parameters		DP	SUR	DP	SUR	DP	SUR
β_{10}	1	0.199	0.996	0.130	0.973	0.116	0.987
β_{11}	-0.5	-0.501	-0.492	-0.499	-0.488	-0.496	-0.504
β_{12}	1.6	1.600	1.598	1.606	1.605	1.602	1.603
β_{20}	1.5	0.697	1.553	0.665	1.557	0.654	1.568
β_{21}	-1.2	-1.201	-1.180	-1.202	-1.205	-1.199	-1.198
β_{22}	-0.7	-0.701	-0.696	-0.703	-0.719	-0.705	-0.720
β_{23}	2	2.004	2.012	2.002	1.993	1.995	1.988

B Posterior Means for DP-REM/CREM

B.1 t-distributed Errors

Table 13 contains the average of the posterior means over the samples of the coefficients in REM with t-distributed random effects and errors. It can be seen that the averaged posterior means estimated with our DP-REM and parametric Bayesian REM are all almost identical, and they are all close to the true value of the coefficients for all four cases, i.e. df being 2, 3, 4, and infinity, where the t-distribution becomes normal distribution.

The average of posterior means of the CREM coefficients are presented in Table 14. Similar to the REM case, the average of the posterior means estimated with both our DP-CREM and parametric Bayesian CREM are similar to each other, and close to the pre-set true values of the coefficients in all cases.

B.2 Log-normal Errors

Table 15 gives the average of the posterior means of the REM with log-normal distributed random effects and errors. One could see that the the DP-REM and Bayesian REM assuming normality obtain relatively close means to the true values of the coefficients, with the DP-REM posterior means being slightly closer to the truths.

Table 16 presents the posterior means averaged over the simulation samples of the DP-CREM and normal Bayesian CREM with log-normal distributed random effects and errors. The posterior means of the coefficients for the explanatory variables are very close to the truth with both our DP-CREM and Bayesian CREM assuming normality. The posterior means of the coefficients for the means of the explanatory variables are slightly farther away from the truth. The skewness of the log-normal distribution influences the posterior means of the intercepts, which are time invariant for each individual i in the random effects. In the CREM case, the sample means of each individual's explanatory variables, $\bar{x}_{1i}$ and $\bar{x}_{2i}$, are also time invariant like the intercept. As a result, the posterior means of their coefficients, β_3 and β_4, are more different from the truth compared with β_1 and β_2, as the log-normal distribution is skewed.

Table 13: Posterior means, REM with t distributed unobservables

df = 2							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	5.003	5.005	5.002	5.000	5.008	5.013
β_2	10	9.999	10.002	10.000	9.997	9.996	9.990
df = 3							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	4.997	4.996	5.004	5.005	4.999	4.999
β_2	10	10.001	10.000	9.998	9.997	9.999	9.999
df = 4							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	4.999	4.999	4.998	4.998	4.997	4.998
β_2	10	10.001	10.001	10.001	10.003	9.998	9.998
df = ∞							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	5.003	5.003	5.000	5.000	4.998	4.998
β_2	10	9.998	9.999	9.999	9.999	10.000	10.000

References

[1] Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*. Springer, Berlin, 1-198.

[2] Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, 384-389.

[3] Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 99-102.

[4] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 1152-1174.

[5] de Carvalho, V. I., Jara, A., Hanson, T. E., & de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Analysis*, 8(3), 623-646.

[6] Chao, J. C., & Phillips, P. C. (1998). Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior. *Journal of Econometrics*, 87(1), 49-86.

[7] Chigira, H., & Shiba, T. (2015). Dirichlet Prior for Estimating Unknown Regression Error Heteroskedasticity. *TERG Discussion Papers*, 341, 1-17.

[8] Clementi, F., & Gallegati, M. (2005). Pareto's law of income distribution: Evidence for Germany, the United Kingdom, and the United States. In *Econophysics of Wealth Distributions* (pp. 3-14). Springer, Milano.

Table 14: Posterior means, CREM with t distributed unobservables

df = 2							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	4.999	5.003	4.996	4.988	5.001	4.997
β_2	10	10.002	9.996	10.001	10.005	10.000	9.999
β_3	-2	-1.973	-2.020	-1.988	-1.961	-2.005	-2.003
β_4	2	1.991	2.023	1.993	1.990	2.006	2.011
df = 3							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	4.996	4.993	5.002	5.005	4.994	4.995
β_2	10	10.000	9.997	10.000	10.003	9.997	9.995
β_3	-2	-1.987	-1.970	-1.983	-1.968	-1.986	-1.996
β_4	2	1.995	1.991	1.997	1.987	1.999	2.004
df = 4							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	4.998	4.999	5.002	5.001	4.999	4.997
β_2	10	10.003	10.003	10.002	10.001	9.998	9.993
β_3	-2	-1.999	-1.996	-1.999	-1.996	-2.005	-2.004
β_4	2	2.000	1.999	1.996	1.996	2.008	2.014
df = ∞							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	5.000	5.000	4.998	4.998	5.003	5.003
β_2	10	10.001	10.001	9.998	9.998	10.001	10.001
β_3	-2	-1.999	-2.000	-2.004	-2.005	-1.998	-1.998
β_4	2	1.998	1.998	2.004	2.005	1.999	1.999

Table 15: Posterior means, REM with log-normal distributed unobservables

Log-normal							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	5.147	5.491	5.124	5.414	5.113	5.371
β_2	10	10.357	11.163	10.369	11.232	10.359	11.222

Table 16: Posterior means, CREM with log-normal distributed unobservables

Log-normal							
Sample size	Truth	100		300		500	
Parameters		DP	REM	DP	REM	DP	REM
β_1	5	5.002	5.023	4.999	4.974	5.000	5.040
β_2	10	9.999	10.020	10.007	10.015	10.000	9.995
β_3	-2	-1.908	-1.269	-1.851	-1.657	-1.832	-1.432
β_4	2	2.669	3.734	2.590	3.896	2.577	3.818

[9] Conley, T. G., Hansen, C. B., McCulloch, R. E., & Rossi, P. E. (2008). A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1), 276-305.

[10] Cornwell, C., & Rupert, P. (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, 3(2), 149-155.

[11] Christensen, L. R., & Greene, W. H. (1976). Economies of scale in US electric power generation. *Journal of political Economy*, 84(4, Part 1), 655-676.

[12] Diewert, W. E. (1971). An application of the Shephard duality theorem: a generalized Leontief production function. *Journal of Political Economy*, 79(3), 481-507.

[13] Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430), 577-588.

[14] Escobar, M. D., & West, M. (1998). Computing nonparametric hierarchical models. In: Dey, D., MüIler, P., & Sinha, D. *Practical nonparametric and semiparametric Bayesian statistics*. Springer, New York, 1-22.

[15] Feng, G., & Serletis, A. (2009). Efficiency and productivity of the US banking industry, $1998-2005$: Evidence from the Fourier cost function satisfying global regularity conditions. *Journal of Applied Econometrics*, 24(1), 105-138.

[16] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2), 209-230.

[17] Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1), 1-12.

[18] Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1).

[19] Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75(1), 121-146.

[20] Greene, William H. (2012). Econometric Analysis, Seventh Edition. Upper Saddle River, New Jersey: Prentice Hall.

[21] Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F., & Thiébaut, R. (2019). Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *The Annals of Applied Statistics*, 13(1), 638-660.

[22] Kleinman, K. P., & Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, 921-938.

[23] Kleibergen, F., & van Dijk, H. K. (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14(6), 701-743.

[24] Kloek, T., & Van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica*, 46, 1-19.

[25] Koop, G. (2003). Bayesian Econometrics. John Wiley & Sons.

[26] Kumbhakar, S. C., & Tsionas, E. G. (2011). Stochastic error specification in primal and dual production systems. *Journal of Applied Econometrics*, 26(2), 270-297.

[27] Kyung, M., Gill, J., & Casella, G. (2010). Estimation in Dirichlet random effects models. *The Annals of Statistics*, 38(2), 979-1009.

[28] Li, M., & Tobias, J. L. (2011). Bayesian inference in a correlated random coefficients model: Modeling causal effect heterogeneity with an application to heterogeneous returns to schooling. *Journal of econometrics*, 162(2), 345-361.

[29] Li, C., Casella, G., & Ghosh, M. (2018). Estimation of regression vectors in linear mixed models with Dirichlet process random effects. *Communications in Statistics-Theory and Methods*, 47(16), 3935-3954.

[30] MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In: Dey, D., Müller, P., & Sinha, D. *Practical nonparametric and semiparametric Bayesian statistics*. Springer, New York, 23-43.

[31] Malikov, E., Kumbhakar, S. C., & Tsionas, M. G. (2016). A cost system approach to the stochastic directional technology distance function with undesirable outputs: the case of US banks in 2001-2010. *Journal of Applied Econometrics*, 31(7), 1407-1429.

[32] Murtazashvili, I., & Wooldridge, J. M. (2008). Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *Journal of Econometrics*, 142(1), 539-552.

[33] Rossi, P. E., Allenby, G. M., & McCulloch, R. (2012). Bayesian statistics and marketing. John Wiley & Sons.

[34] Strutz, T. (2010). *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Vieweg and Teubner.

[35] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (pp. 1385-1392).

[36] Teh, Y. W. (2011). Dirichlet Process. In *Encyclopedia of machine learning*, pp. 280-287. Springer US.

[37] Wiesenfarth, M., Hisgen, C. M., Kneib, T., & Cadarso-Suarez, C. (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. *Journal of Business & Economic Statistics*, 32(3), 468-482.

[38] White, H. (2014). *Asymptotic theory for econometricians*. Academic press.

[39] Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2), 133-138.

[40] Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics*, 87(2), 385-390.

[41] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298), 348-368.

[42] Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. Wiley, New York.