Università
Ca'Foscari
Venezia

Scuola Dottorale di Ateneo
Graduate School

Dottorato di ricerca
in Economia
Ciclo XXIX
Anno di discussione 2017

Contributions to Bayesian Nonparametric and Objective Bayes Literature

Settore scientifico disciplinare di afferenza: SECS-P/05

Tesi di Dottorato di Luca Rossini, Matricola: 956080

Coordinatore del Dottorato
Prof. Giacomo Pasini

Tutore del Dottorando
Prof.ssa Monica Billio

Co-tutore del Dottorando
Prof. Roberto Casarin

Contributions to Bayesian Nonparametric and Objective Bayes Literature

by

Luca Rossini

Matricola 956080

A thesis submitted in partial fulfillment
for the degree of

Doctor of Philosophy in Economics

Department of Economics

Ca' Foscari University of Venice

Supervisors :

Prof.ssa Monica Billio
Prof. Roberto Casarin

January, 2017

Declaration

I, Rossini Luca, hereby declare that this Ph.D thesis titled "Contributions to Bayesian Non-parametric and Objective Bayes Literature" is a presentation of my original research work for the degree of Doctor of Philosophy in Economics under the guidance and supervision of Prof. Monica Billio and Prof. Roberto Casarin. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

I affirm that this work has not been submitted previously in the same or similar form to another examination committee at another department, university or country.

Signed:

_____

Date:

_____

Studente: Rossini, Luca Matricola: 956080

Dottorato di ricerca in Economia
Ciclo XXIX
Titolo della tesi: Contributions to Bayesian Nonparametric and Objective Bayes Literature

Abstract

The contribution of this dissertation are discussed in three self-contained chapters.

Chapter 1 and **Chapter 2** contribute to the literature on Bayesian nonparametrics by proposing two approaches, the first one to multiple time series and the second one to conditional copula models.

Chapter 1 sets up a novel Bayesian nonparametric prior for SUR models, which allows shrinkage of SUR coefficients toward multiple locations and identification of group of coefficients. Our two-stage hierarchical distribution consists in a hierarchical Dirichlet process on the parameters of a Normal-Gamma. We use this new model for extracting contagion networks with linkage clustering effects. The proposed method has been applied both to simulated results and to a macroeconomic dataset.

Chapter 2 analyses a conditional copula model from a Bayesian nonparametric perspective. The conditional copula models allow us to model the effect of a covariate driving the strength of dependence between the main variables. The previous methodology has been applied to the twins data and socioeconomic variables of the parents.

In **Chapter 3**, we focus on the Yule–Simon distribution, which has been introduced for the analysis of frequency of data (stock options frequency, frequency of surnames, etc). For this distribution, we derive two objective Bayes prior, the Jeffreys and the loss-based prior, proving some theoretical results. We apply both the priors to simulated examples of different sample sizes and we study the effectiveness of these priors to real data examples (e.g. finance, surnames and music hit charts). In the same chapter we propose a Gibbs sampling algorithm for the analysis of the posterior distribution of a Yule–Simon

distribution when a Gamma prior is chosen for the shape parameter. The effectiveness of the data augmented algorithm has been proved through simulated examples, including count data regression, and an application to text analysis.

Chapter 1 is a joint work with Monica Billio and Roberto Casarin and is currently under revision for submission. Chapter 2 is a joint work with Luciana Dalla Valle and Fabrizio Leisen and has received revision from Journal of the Royal Statistical Society (Series C). Chapter 3 is made by two different papers both joint works with Fabrizio Leisen and Cristiano Villa. The first paper has been submitted and is under review, while the second one will appear in the *"Journal of Statistical Computation and Simulation"*.

Keywords: Bayesian Nonparametrics, Multivariate Time Series, Sparsity, Shrinkage, Conditional Copulas, Objective Bayesian, Loss-Based Prior, Jeffreys' Prior, Data Augmentation.

Contents

List of Tables

List of Figures

Acknowledgements

I acknowledge my two supervisors Monica Billio and Roberto Casarin from the University Ca' Foscari of Venice for their guidance and their support during the development of my thesis. I am thankful to them for the opportunity that they gave me to collaborate with them and with professors from other institutions during my Ph.D.

I acknowledge Fabrizio Leisen and Cristiano Villa (University of Kent, U.K.) for their support, for accepting my visit to their institutions and for the future collaborations built during my period in Canterbury. I acknowledge Luciana Dalla Valle (Plymouth University, U.K.) for the collaborations and all the partecipants to the conferences for the helpful suggestions. Many thanks goes to the editor and referees of *Journal of Statistical Computation and Simulation*.
I would like to thank Stefano Tonellato (Ca' Foscari University of Venice), Dimitris Korobilis (University of Essex, U.K.) and Concepcion Ausin (Universidad Carlos III de Madrid, Spain) for their valuable comments to the manuscript and their constructive suggestions.
I gratefully acknowledge the PhD scholarship award from the University Ca' Foscari of Venice by the Italian Ministry of Education, University and Research (MIUR).

Special thanks to my fiancee, Alice, and my parents, Marilena and Gianluigi, for helping me during these three years of Ph.D.

Chapter 1

Bayesian Nonparametric Sparse Seemingly Unrelated Regression Model (SUR)

Abstract. Seemingly unrelated regression (SUR) models are used in studying the interactions among economic variables of interest. In a high dimensional setting and when applied to large panel of time series, these models have a large number of parameters to be estimated and suffer of inferential problems. In order to avoid overparametrization and overfitting issues, shrinkage priors have been introduced, which usually shrink some parameters to zero.

We propose a novel Bayesian nonparametric prior for SUR models, which allows shrinkage of SUR coefficients toward multiple locations and identification of group of coefficients. Our two-stage hierarchical distribution consists in a hierarchical Dirichlet process on the parameters of a Normal-Gamma distribution.

This new multiple shrinkage prior model allows us to extract network structures from panel data and to cluster the network edges between panel units. Applications both to

This chapter is based on: Billio, M., Casarin, R. and Rossini, L. (2016). *"Bayesian Nonparametric Sparse Seemingly Unrelated Regression Model (SUR)"*. Working papers N. 20/WP/2016, Dept of Economics, Ca' Foscari University of Venice. Working paper available at http://arxiv.org/abs/1608.02740.

simulated data and to macroeconomic contagion show important gains from our prior compared to existing priors in terms of parameter estimation and predictive abilities.

Keywords: Bayesian nonparametrics; Bayesian model selection; shrinkage; Large vector autoregression; Network representation.

1.1 Introduction

In the last decade, high dimensional models and large datasets have increased their importance in economics (e.g., see Scott and Varian (2013)). The use of large dataset has been proved to improve the forecasts in large macroeconomic and financial models (see, Banbura et al. (2010), Carriero et al. (2013), Koop (2013), Stock and Watson (2012)). For analyzing and better forecasting them, seemingly unrelated regression (SUR) models have been introduced (Zellner, 1962, 1971), where the error terms are independent across time, but may have cross-equation contemporaneous correlations. SUR models require estimation of large number of parameters with few observations. In order to avoid overparametrization, overfitting and dimensionality issues, Bayesian inference and suitable classes of prior distributions have been proposed.

In vector autoregressive (VAR) modeling (see Sims (1980, 1992)) Bayesian inference and related prior on the VAR parameters should be introduced to solve these problems (see Litterman (1980)). Litterman (1986), Doan et al. (1984) and Sims and Zha (1998) specify particular priors constraint on the VAR parameters for Bayesian VAR and Canova and Ciccarelli (2004) discuss prior choice for panel VAR models.

Unfortunately these classes of priors may be not effective in dealing with overfitting in very large SUR models. Thus, new priors have been proposed. George et al. (2008) introduced Stochastic Search Variable Selection (SSVS) and spike-and-slab prior distribution. Wang (2010) develops a sparse SUR model with Gaussian errors, where the coefficients shrink near zero in both the regression coefficients and the error precision matrix. Korobilis (2013) extend the use of SSVS to restricted VARs and particularly to select variables in linear and nonlinear VAR using MCMC methods (see Koop and Korobilis (2010) for an introduction). Ahelgebey et al. (2015, 2016) propose Bayesian graphical VAR (BGVAR)

and sparse BGVAR. Both SSVS and BGVAR use two separate sets of restrictions for the contemporaneous and lagged interactions, where the SSVS uses the reduced-form model, while in the BGVAR the restrictions are directly used in the structural model and help to solve the identification problem of the SVAR using the graph structures. Furthermore, the two models differ in the computational part, where George et al. (2008) use a single-move Gibbs sampler, while Ahelgebey et al. (2015) focus on a collapsed and multi-move Gibbs sampler. Koop and Korobilis (2015) build on SSVS prior of George et al. (2008) a new parametric prior, which takes into a ccount the panel descriptions and Korobilis (2016) proposed in the same way new parametric and semi-parametric priors for panel VAR.

In this paper, a novel Bayesian nonparametric hierarchical prior for multivariate time series is proposed, which allows shrinkage of the SUR coefficients to multiple locations using a Normal-Gamma distribution with location, scale and shape parameters unknown. In our sparse SUR (sSUR), some SUR coefficients shrink to zero, due to the shrinking properties of the lasso-type distribution at the first stage of our hierarchical prior, thus improving efficiency of parameters estimation, prediction accuracy and interpretation of the temporal dependence structure in the time series. We use a Bayesian Lasso prior, which allows us to reformulate the SUR model as a penalized regression problem, in order to determine which SUR coefficients shrink to zero (see Tibshirani (1996) and Park and Casella (2008)).

For alternative shrinkage procedures, see also Zou and Hastie (2005) (elastic-net), Zou and Zhang (2009) (Adaptive elastic-net Lasso), Gefang (2014) (Doubly adaptive elastic-net Lasso).

As regards to the second stage of the hierarchy, a mixture of hyperprior distributions for the Normal-Gamma hyperparameters, which allows for shrinkage of different locations has been used. This mixture consists of two different components, where we assigned a Dirichlet process hyperpriors to achieve parameters parsimony due to clustering of the

SUR coefficients. We build on Bassetti et al. (2014), which propose a vector of dependent Dirichlet process prior to capture similarities in clustering effects across time series and on MacLehose and Dunson (2010), which propose a Bayesian semiparametric approach that allows shrinkage to multiple locations using a mixture of double exponential priors with location and scale parameters assigned through a Dirichlet process hyperpriors to allow groups of coefficients to be shrunk toward the same mean.

Hence, after the seminal papers of Ferguson (1973), Lo (1984) and Sethuraman (1994), Dirichlet process priors and their multivariate extensions (e.g., see Müller et al. (2004), Griffin and Steel (2006), Hatjispyros et al. (2011), Hjort et al. (2010) for a review of Bayesian nonparametrics), are now widely used due to the availability of efficient algorithms for posterior computations (Escobar and West, 1995; MacEachern and Müller, 1998; Papaspiliopoulos and Roberts, 2008; Walker, 2007; Kalli et al., 2011), including but not limited to applications in time series settings (Hirano, 2002; Chib and Hamilton, 2002; Rodriguez and ter Horst, 2008; Jensen and Maheu, 2010; Griffin, 2011; Griffin and Steel, 2011; Bassetti et al., 2014; Jochmann, 2015).

As regards to the posterior approximation, we develop a MCMC algorithm. We rely on slice sampling by Kalli et al. (2011), which is an improved version of the algorithm of Walker (2007) and on the paper of Hatjispyros et al. (2011), where they present an approach to modeling dependent nonparametric random density functions through mixture of DP model.

Another contribution of this paper relates to the extraction of network structures from panel data. As see, through the macroeconomic contagion application, we contribute to the literature of financial and macroeconomic connectedness (Demirer et al., 2015; Diebold and Yilmaz, 2014). The network connectedness has a central role in the financial, systemic and credit risk measurement and helps us to understand fundamental macroeconomic risks (see

Acharya et al. (2012), Billio et al. (2012) and Bianchi et al. (2015)). Our sparse Bayesian nonparametric prior allows us to catch the most relevant linkages between different units of the panel at different lags and for estimating the exact number of cluster in the network. We show, through the definition of an adjacency matrix based on the pairwise probability and the co-clustering matrix, the transmission of shocks from and to specific countries at different lags (Diebold and Yilmaz (2015), Barigozzi and Brownlees (2016), Brownlees and Engle (2016) and Diebold and Yilmaz (2016)). Applications both to simulated data and to macroeconomic contagion show important gain from our prior compared to existing priors in terms of parameter estimation and predictive abilities.

The paper is organized as follows. Section 1.2 introduces our sparse Bayesian SUR model and the prior assumptions on the hyperparameters. In Section 1.3 we explain the computational details of the model and the Gibbs sampling, while Section 1.4 through simulated results illustrates the performance of the methodology compared to existing popular prior for VAR and SUR models. Finally, in Section 1.5 an empirical macroeconomic exercise on contagion shows clear advantages of the proposed prior.

1.2 A sparse Bayesian SUR model

In this section, we review preliminary notions on seemingly unrelated regression (SUR) models and vector autoregressive (VAR) models. Furthermore we focus on the prior specifications for our specific sparse SUR.

1.2.1 SUR and VAR models

Zellner (1962) introduces the seemingly unrelated regression (SUR) model and analyzes individual relationships that are linked by the fact that their disturbances are correlated. Hence, SUR models have many applications in different fields, for example demand func-

tions can be estimated for different households for a given commodity or for different commodities.

In a SUR model with N units (or groups of cross-section observations) we consider a sequence of m_i-dimensional vectors of dependent variables, $\mathbf{y}_{i,t}$, that follow individual regressions:

$$\mathbf{y}_{i,t} = X_{i,t}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_{i,t}, \quad t = 1, \ldots, T \quad i = 1, \ldots, N, \tag{1.1}$$

where $X_{i,t}$ is the $(m_i \times n_i)-$ matrix of observations on n_i explanatory variables with a possible constant term for individual i at time t, $\boldsymbol{\beta}_i = (\beta_{i,1}, \ldots, \beta_{i,n_i})$ is a n_i-vector of unknown coefficients, and $\varepsilon_{i,t}$ is a random error. We write (1.1) in a stacked regression form:

$$\mathbf{y}_t = X_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \quad t = 1, \ldots, T, \tag{1.2}$$

where $\mathbf{y}_t = (\mathbf{y}'_{1,t}, \ldots, \mathbf{y}'_{N,t})'$ is the $m \times 1$ vector of observations, with $m = \sum_{i=1}^{N} m_i$, $X_t = \text{diag}(X_{1,t}, \ldots X_{N,t})$ is the $m \times n$ matrix of observations on the explanatory variables at time t with $n = \sum_{i=1}^{N} n_i$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_N)'$, the $n-$vector of coefficients and $\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}'_{1,t}, \ldots, \boldsymbol{\varepsilon}'_{m,t})'$ is the vector of errors distributed as $\mathcal{N}_m(\mathbf{0}, \Sigma)$, where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_s$ are independent for $t \neq s$.

The use of SUR models is important to gain efficiency in estimation by combining different equations and to impose or test restrictions that involve parameters in different equations.

An important special case of the SUR model is vector autoregressive (VAR) model. Due to the work of Sims (1980), VAR models have acquired a permanent place in the toolkit of applied macroeconomics to study the impact of a policy decision on the variables of interest. A VAR model of order p (VAR(p)) is defined as

$$\mathbf{y}_t = \mathbf{b} + \sum_{i=1}^{p} B_i\mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \tag{1.3}$$

for $t = 1, \ldots, T$, where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{m,t})'$, $\mathbf{b} = (b_1, \ldots, b_m)'$ and B_i is a $(m \times m)$ matrix of coefficients. We assume that $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{m,t})'$ follows a Gaussian distribution $\mathcal{N}_m(\mathbf{0}, \Sigma)$ with mean $\mathbf{0}$ and covariance matrix Σ.

The VAR(p) can be obtained as a special case of (1.2) by setting $N = 1$, $m = m_1$ and writing (1.3) in a stacked regression form:

$$\mathbf{y}_t = (I_m \otimes \mathbf{x}_t')\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \tag{1.4}$$

where $\mathbf{x}_t = (1, y_{t-1}', \ldots, y_{t-p}')'$ is the vector of predetermined variables, $\boldsymbol{\beta} = \text{vec}(B)$, where $B = (\mathbf{b}, B_1, \ldots, B_p)$, $\otimes$ is the Kronecker product and vec the column-wise vectorization operator that stacks the columns of a matrix in a column vector.

1.2.2 Prior assumption

The number of parameters to estimate in (1.2) is $q = r + (m+1)m/2$, with $r = \sum_{i=1}^{N} r_i$, $r_i = n_i$. For large value of m, q can be large and add some problems during the estimation, such as overfitting, or unstable predictions and difficult-to-interpret descriptions of the temporal dependence. In order to avoid overparameterization issues and the overfitting problems, a hierarchical strategy in prior specification has been suggested in the Bayesian dynamic panel modelling literature (e.g., Canova and Ciccarelli (2004), Kaufmann (2010), and Bassetti et al. (2014)). The hierarchical prior can be used to incorporate cross-equation interdependences and various degrees of information pooling across units (e.g., see Chib and Greenberg (1995) and Min and Zellner (1993)), while a different stream of literature is using instead a prior model which induces sparsity (e.g., MacLehose and Dunson (2010), Wang (2010)).

In this paper we combine the two strategies and define a hierarchical prior distribution which induces sparsity on the vector of coefficients $\boldsymbol{\beta}$. In order to regularize (1.2), we incorporate a penalty using a lasso prior $f(\boldsymbol{\beta}) = \prod_{j=1}^{r} \mathcal{NG}(\beta_j|0, \gamma, \tau)$, where $\mathcal{NG}(\beta|\mu, \gamma, \tau)$

denotes the normal-gamma distribution with location parameter μ, shape parameter $\gamma > 0$ and scale parameter $\tau > 0$. The normal-gamma distribution has density function

$$f(\beta|\mu,\gamma,\tau) = \frac{\tau^{\frac{2\gamma+1}{4}}|\beta-\mu|^{\gamma-\frac{1}{2}}}{2^{\gamma-\frac{1}{2}}\sqrt{\pi}\Gamma(\gamma)}K_{\gamma-\frac{1}{2}}(\sqrt{\tau}|\beta-\mu|),$$

where $K_\gamma(\cdot)$ represents the modified Bessel function of the second kind with the index γ (see Abramowitz and Stegun (1972)). The normal-gamma distribution has the double exponential distribution as a special case for $\gamma = 1$ and can be represented as a scale mixture of normals (see Andrews and Mallows (1974) and Griffin and Brown (2006)):

$$\mathcal{NG}(\beta|\mu,\gamma,\tau) = \int_0^{+\infty} \mathcal{N}(\beta|\mu,\lambda)\mathcal{G}a(\lambda|\gamma,\tau/2)d\lambda, \tag{1.5}$$

where $\mathcal{G}a(\cdot|a,b)$ denotes a gamma distribution[1].

The normal-gamma distribution in (1.5) induces shrinkage toward the prior mean of μ, but we can extend the lasso model specification by introducing a mixture prior with separate location parameter μ_j^*, separate shape parameter γ_j^* and separate scale parameter τ_j^* such that: $f(\boldsymbol{\beta}) = \prod_{j=1}^r \mathcal{NG}(\beta_j|\mu_j^*,\gamma_j^*,\tau_j^*)$. In our paper, we favor the sparsity of the parameters through the use of carefully tailored hyperprior and we use a nonparametric Dirichlet process prior (DPP), which reduces the overfitting problem and the curse of dimensionality by allowing for parameters clustering due to the concentration parameter and the base measure choice.

Also, following Bassetti et al. (2014), we assume that N blocks of parameters can be exogenously defined. The blocks correspond to series from different countries which share

[1] The gamma distribution of τ ($\tau \sim \mathcal{G}a(a,b)$) used in this paper is parametrized as:

$$f(\tau|a,b) = \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau)\mathbb{I}_{(0,+\infty)}(\tau)$$

a sparse component but have possibly different clustering features. Our framework can be extended to include dependence in the clustering features (Bassetti et al., 2014; Taddy, 2010; Griffin and Steel, 2011).

In our case we define $\boldsymbol{\theta}^* = (\boldsymbol{\mu}^*, \boldsymbol{\gamma}^*, \boldsymbol{\tau}^*)$ as the parameters of the Normal-Gamma distribution, and assume a prior $\mathbb{Q}_l$ for $\boldsymbol{\theta}^*_{lj}$, that is

$$\beta_j \overset{ind}{\sim} \mathcal{NG}(\beta_j|\mu^*_j, \gamma^*_j, \tau^*_j), \tag{1.6}$$

$$\boldsymbol{\theta}^*_{lj}|\mathbb{Q}_l \overset{i.i.d.}{\sim} \mathbb{Q}_l, \tag{1.7}$$

for $j = 1, \ldots, r_l$ and $l = 1, \ldots, N$.

Following a construction of the hierarchical prior similar to the one proposed in Hatjispyros et al. (2011) we define the vector of random measures

$$\begin{aligned} \mathbb{Q}_1(d\boldsymbol{\theta}_1) &= \pi_1\mathbb{P}_0(d\boldsymbol{\theta}_1) + (1-\pi_1)\mathbb{P}_1(d\boldsymbol{\theta}_1), \\ &\vdots \\ \mathbb{Q}_N(d\boldsymbol{\theta}_N) &= \pi_N\mathbb{P}_0(d\boldsymbol{\theta}_N) + (1-\pi_N)\mathbb{P}_N(d\boldsymbol{\theta}_N), \end{aligned} \tag{1.8}$$

with the same sparse component $\mathbb{P}_0$ in each equation and with the following hierarchical construction as previously explained,

$$\begin{aligned} &\mathbb{P}_0(d\boldsymbol{\theta}) \sim \delta_{\{(0,\gamma_0,\tau_0)\}}(d(\mu,\gamma,\tau)), \\ &\mathbb{P}_l(d\boldsymbol{\theta}) \overset{i.i.d.}{\sim} \mathrm{DP}(\tilde{\alpha}, G_0), \quad l = 1, \ldots, N, \\ &\pi_l \overset{i.i.d.}{\sim} \mathcal{B}e(\pi_l|1, \alpha_l), \quad l = 1, \ldots, N, \\ &(\gamma_0, \tau_0) \sim g(\gamma_0, \tau_0|\nu_0, p_0, s_0, n_0), \\ &G_0 \sim \mathcal{N}(\mu|c, d) \times g(\gamma, \tau|\nu_1, p_1, s_1, n_1) \end{aligned} \tag{1.9}$$

where $\delta_{\{\boldsymbol{\psi}_0\}}(\boldsymbol{\psi})$ denotes the Dirac measure indicating that the random vector $\boldsymbol{\psi}$ has a degenerate distribution with mass at the location $\boldsymbol{\psi}_0$, and $g(\gamma_0, \tau_0)$ is the conjugate joint

prior distribution (see Miller (1980)) with density

$$g(\gamma_0, \tau_0|\nu_0, p_0, s_0, n_0) \propto \tau_0^{\nu_0\gamma_0 - 1} p_0^{\gamma_0 - 1} \exp\{-s_0\tau_0\} \frac{1}{\Gamma(\gamma_0)^{n_0}}, \tag{1.10}$$

and hyperparameters fixed such that $\nu_0 > 0$, $p_0 > 0$, $s_0 > 0$ and $n_0 > 0$. From Miller (1980), we construct the gamma two-parameters $g(\gamma, \tau) = g(\tau|\gamma)g(\gamma)$, where $g(\tau|\gamma) \sim \mathcal{G}a(\nu_0\gamma, s_0)$ and we marginalize out such that:

$$g(\gamma) = \int_0^\infty g(\gamma, \tau) d\tau = C \frac{\Gamma(\nu_0\gamma)}{\Gamma(\gamma)^{n_0}} \frac{p_0^{\gamma-1}}{s_0^{\nu_0\gamma}}, \tag{1.11}$$

$$g(\tau|\gamma) = \frac{g(\gamma, \tau)}{g(\gamma)} = \frac{\tau^{\nu_0\gamma - 1} e^{-s_0\tau}}{\Gamma(\nu_0\gamma)} s_0^{\nu_0\gamma}, \tag{1.12}$$

with a normalizing constant C such that $1 = \int_0^\infty g(\gamma) d\gamma$. Based on MacLehose and Dunson (2010) and on our computational experiments, we assume the following parameter setting for the sparse and nonsparse component in the gamma two parameters distribution, $g(\gamma, \tau)$,

$$v_0 = 30 \quad s_0 = 1/30 \quad p_0 = 0.5 \quad n_0 = 18,$$

$$v_1 = 3 \quad s_1 = 1/3 \quad p_1 = 0.5 \quad n_1 = 10.$$

As described in the hierarchical prior representations in (1.8) and in (1.9), with probability π (distributed as a beta[2]) a coefficient β_j is shrunk toward zero as in standard lasso, while with probability $(1 - \pi)$ the coefficient is distributed as a $DP(\tilde{\alpha}, G_0)$. The amount of shrinkage is determined by the shape and scale parameter (γ, τ), which moves as a two-parameters gamma (Miller (1980)).

[2] The beta distribution for x ($x \sim \mathcal{B}e(\alpha, \beta)$) used in this paper is parametrized as follows:

$$f(x|\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1} \mathbb{I}_{[0,1]}(x)$$

where $\mathrm{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ and $\alpha, \beta > 0$

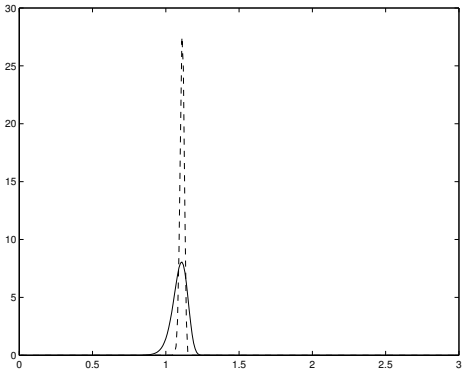


FIGURE1.1: Probability density function $f(\gamma)$ for sparse ($v_0 = 30, s_0 = 1/30, p_0 = 0.5, n_0 = 18$, dashed line) and nonsparse ($v_1 = 3, s_1 = 1/3, p_1 = 0.5, n_1 = 10$, solid line) case.

The Dirichlet Process, $\mathrm{DP}(\tilde{\alpha}, G_0)$, can be defined by using the stick-breaking representation (Sethuraman (1994)) given by:

$$\mathbb{P}_l(\cdot) = \sum_{j=1}^{\infty} w_{lj} \delta_{\{\boldsymbol{\theta}_{lj}\}}(\cdot) \quad l = 1, \ldots, N. \tag{1.13}$$

Following the definition of the dependent stick-breaking processes, proposed by MacEachern (1999) and MacEachern (2001) the atoms θ_{lj} and the weights w_{lj} (for $l = 1, \ldots, N$) are stochastically independent and satisfy the following hypothesis:

- θ_{lj} is an independent and identically distributed (i.i.d.) sequence of random elements with common probability distribution G_0 ($\theta_{lj} \overset{\text{iid}}{\sim} G_0$);
- the weights (w_{lj}) are determined through the stick-breaking construction for $j > 1$, while for $j = 1$ $w_{l1} = v_{l1}$:

$$w_{lj} = v_{lj} \prod_{k=1}^{j-1} (1 - v_{lk}) \quad l = 1, \ldots, N$$

with $v_j = (v_{1j}, \ldots, v_{Nj})$ independent random variables taking values in $[0,1]^N$ distributed as a $\mathcal{B}e(1, \tilde{\alpha})$ such that $\sum_{j \geqslant 1} w_{lj} = 1$ almost surely for every $l = 1, \ldots, N$.

After this definition, we are able to construct a random density function $f(\beta|\mathbb{P})$ based on an infinite mixture representation similar to the well known Dirichlet process mixture model (Lo (1984)):

$$f_l(\boldsymbol{\beta}|\tilde{\mathbb{P}}_l) = \int K(\boldsymbol{\beta}|\boldsymbol{\theta})\tilde{\mathbb{P}}_l(d\boldsymbol{\theta}), \tag{1.14}$$

where $K(\boldsymbol{\beta}|\boldsymbol{\theta})$ is a density for each $\boldsymbol{\theta} \in \Theta$, the so called density kernel and $\tilde{\mathbb{P}}_l$ is a random measure. In our paper, the density kernel is defined as $K(\boldsymbol{\beta}|\boldsymbol{\theta}) = \mathcal{NG}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau})$. Following the definition of the density kernel and using the representation as infinite mixture, we have that, for each $l = 1, \ldots, N$, the equation (1.14) has the following representation

$$\begin{aligned} f_l(\boldsymbol{\beta}|\mathbb{Q}_l) &= \pi_l f(\boldsymbol{\beta}|\mathbb{P}_0) + (1-\pi_l) f(\boldsymbol{\beta}|\mathbb{P}_l) = \pi_l \int \mathcal{NG}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \mathbb{P}_0(d(\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau})) \\ &+ (1-\pi_l) \int \mathcal{NG}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \mathbb{P}_l(d(\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau})) \\ &= \pi_l \mathcal{NG}(\boldsymbol{\beta}|0, \gamma_0, \tau_0) + (1-\pi_l) \sum_{k=1}^{\infty} w_{lk} \mathcal{NG}(\boldsymbol{\beta}|\mu_{lk}, \gamma_{lk}, \tau_{lk}) \\ &= \sum_{k=0}^{\infty} \breve{w}_{lk} \mathcal{NG}(\boldsymbol{\beta}|\breve{\theta}_{lk}), \end{aligned}$$

where

$$\breve{w}_{lk} = \begin{cases} \pi_l, & k = 0 \\ (1-\pi_l) w_{lk}, & k > 0 \end{cases} \qquad \breve{\theta}_{lk} = \begin{cases} (0, \gamma_0, \tau_0), & k = 0 \\ (\mu_{lk}, \gamma_{lk}, \tau_{lk}), & k > 0. \end{cases}$$

As regards to the choice of the prior for Σ, we model it by considering its restrictions induced by a graphical model structuring. A graph G is defined by the pair (L, E), where L is the vertex set and E is the edge-set, or the set of linkages. In our case the prior over

the graph structure is defined as a Bernoulli distribution with parameter ψ, which is the probability of having an edge. That is, a m node graph $G = (L, E)$, with $|L|$ the cardinality of the set of nodes and with $|E|$ edges has a prior probability:

$$p(G) \propto \prod_{i,j} \psi^{e_{ij}} (1-\psi)^{(1-e_{ij})} = \psi^{|E|} (1-\psi)^{T-|E|}, \tag{1.15}$$

with $e_{ij} = 1$ if $(i, j) \in E$ and $|T| = \binom{|L|}{2}$ is the maximum number of edges, while to induce sparsity we choose $\psi = 2/(p-1)$ which would provide a prior mode at p edges. Conditional on a specified graph G we assume a Hyper Inverse Wishart prior distribution for Σ that is:

$$\Sigma \sim \mathcal{HIW}_G(b, \tilde{L}), \tag{1.16}$$

where b means the degrees of freedom and $\tilde{L}$ is the scale hyperparameters. The density function of the $\mathcal{HIW}$ is represented in the Appendix A.

1.3 Computational details

In this section we develop a Gibbs sampler algorithm in order to approximate the posterior distribution. For simplicity of notations we focus on the bivariate case, $N = 2$ and consequently $l = 1, 2$, and, without loss of generality, we can extend the following representation to the multivariate case.

First of all, we focus on the slice latent variables for $l = 1, 2$ through the introduction of the latent variables, $u_{lj}, j = 1, \ldots, r_1$, for f_l, which allows us to write the infinite mixture model in an easy way. Hence we represent the full conditional of β_{1j} as follows,

$$f_1(\beta_{1j}, u_{1j} | (\mu_1, \gamma_1, \tau_1), w_1) = \pi_1 \sum_{k=0}^{\infty} \mathbb{I}(u_{1j} < \tilde{w}_{1k}) \mathcal{NG}(\beta_{1j} | (0, \gamma_{1k}, \tau_{1k})) +$$

$$+ (1-\pi_1) \sum_{k=1}^{\infty} \mathbb{I}(u_{1j} < w_{1k}) \mathcal{NG}(\beta_{1j} | \mu_{1k}, \gamma_{1k}, \tau_{1k})$$

$$= \pi_1 \mathbb{I}(u_{1j} < \tilde{w}_0)\mathcal{NG}(\beta_{1j}|(0, \gamma_0, \tau_0))+$$

$$+ (1-\pi_1)\sum_{k=1}^{\infty} \mathbb{I}(u_{1j} < w_{1k})\mathcal{NG}(\beta_{1j}|\mu_{1k}, \gamma_{1k}, \tau_{1k}),$$

where we introduce a variable $\tilde{w}_{1k}$ such that we can apply the slice sampler and then we assume $\tilde{w}_{1k} = \tilde{w}_0 = 1$ if $k = 0$ and $\tilde{w}_{1k} = 0$ for $k > 0$ and, for simplicity of notations, we denote $(0, \gamma_{1,0}, \tau_{1,0}) = (0, \gamma_0, \tau_0)$.

Moving to the density function f_2, we introduce the latent variables $u_{2j}, j = 1, \ldots, r_2,$, which allows us to write the following density:

$$f_2(\beta_{2j}, u_{2j}|(\mu_2, \gamma_2, \tau_2), w_2) = \pi_2 \mathbb{I}(u_{2j} < \tilde{w}_0)\mathcal{NG}(\beta_{2j}|(0, \gamma_0, \tau_0))+$$

$$+ (1-\pi_2)\sum_{k=1}^{\infty} \mathbb{I}(u_{2j} < w_{2k})\mathcal{NG}(\beta_{2j}|\mu_{2k}, \gamma_{2k}, \tau_{2k}).$$

The introduction of the slice variables (u_{1j}, u_{2j}) allows us to reduce the dimensionality of the problem from a mixture with an infinite number of components to a similar finite mixture model. In particular, letting

$$\mathcal{A}_{w_1}(u_{1j}) = \{k : u_{1j} < w_{1k}\}, \qquad j = 1, \ldots, r_1,$$

$$\mathcal{A}_{w_2}(u_{2j}) = \{k : u_{2j} < w_{2k}\}, \qquad j = 1, \ldots, r_2,$$

then it can be proved that the cardinality of the sets $(\mathcal{A}_{w_1}, \mathcal{A}_{w_2})$ is almost surely finite.

Therefore, we express f_1 and f_2 as an augmented random joint probability density function for β_{1j}, β_{2j} and u_{1j}, u_{2j}

$$f_l(\beta_{lj}, u_{lj}|(\mu_l, \gamma_l, \tau_l), w_l) = \pi_l \mathbb{I}(u_{lj} < \tilde{w}_0)\mathcal{NG}(\beta_{lj}|0, \gamma_0, \tau_0)$$

$$+ (1-\pi_l)\sum_{k \in \mathcal{A}_{w_l}(u_{lj})} \mathcal{NG}(\beta_{lj}|\mu_{lk}, \gamma_{lk}, \tau_{lk}).$$

We iterate the data augmentation principle for each f_l (with $l = 1, 2$) through the introduction of two auxiliary variables, the latent variables δ_{lj} $(j = 1, \ldots, r_l)$ and the allocation

variables d_{lj} $(j = 1, \ldots, r_l)$. The first variable described above selects one of the two random measures $\mathbb{P}_0$ and $\mathbb{P}_l$, hence, when δ_{lj} is equal to zero, we choose the sparse component $\mathbb{P}_0$, while if it is one, we choose the nonsparse component $\mathbb{P}_l$ and we need to introduce the allocation variables. The second variable of interest, d_{lj}, selects the components of the Dirichlet mixture $\mathbb{P}_l$ to which each single coefficient β_{lj} is allocated to. Then the density function can be expressed as

$$f_l(\beta_{lj}, u_{lj}, d_{lj}, \delta_{lj}) = \Big(\mathbb{I}(u_{lj} < \tilde{w}_{d_{lj}})\mathcal{NG}(\beta_{lj}|0, \gamma_0, \tau_0)\Big)^{1-\delta_{lj}} \times$$
$$\Big(\mathbb{I}(u_{lj} < w_{ld_{lj}})\mathcal{NG}(\beta_{lj}|\mu_{ld_{lj}}, \gamma_{ld_{lj}}, \tau_{ld_{lj}})\Big)^{\delta_{lj}} \pi_l^{1-\delta_{lj}}(1-\pi_l)^{\delta_{lj}}.$$

From (1.5), we demarginalize the Normal-Gamma distribution by introducing a latent variable λ_{lj} for each β_{lj} such that the joint distribution has the following representation:

$$f_l(\beta_{lj}, \lambda_{lj}, u_{lj}, d_{lj}, \delta_{lj}) =$$
$$= \Big(\mathbb{I}(u_{lj} < \tilde{w}_{d_{lj}})\mathcal{N}(\beta_{lj}|0, \lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_0, \tau_0/2)\Big)^{1-\delta_{lj}} \times$$
$$\Big(\mathbb{I}(u_{lj} < w_{ld_{lj}})\mathcal{N}(\beta_{lj}|\mu_{ld_{lj}}, \lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_{ld_{lj}}, \tau_{ld_{lj}}/2)\Big)^{\delta_{lj}} \pi_l^{1-\delta_{lj}}(1-\pi_l)^{\delta_{lj}}.$$

Hence, we describe the joint posterior distribution based on the distribution previously defined as follows

$$f(\Theta, \Sigma, \Lambda, U, D, V, \Delta|Y) \propto$$
$$\prod_{t=1}^{T}(2\pi|\Sigma|)^{-1/2}\exp\left(-\frac{1}{2}\big(y_t - X_t'\beta\big)'\Sigma^{-1}\big(y_t - X_t'\beta\big)\right) \times$$
$$\prod_{j=1}^{r_1} f_1(\beta_{1j}, \lambda_{1j}, u_{1j}, d_{1j}, \delta_{1j}) \prod_{j=1}^{r_2} f_2(\beta_{2j}, \lambda_{2j}, u_{2j}, d_{2j}, \delta_{2j}) \times \tag{1.17}$$
$$\prod_{k>1}\mathcal{B}e(v_{1k}|1, \alpha)\mathcal{B}e(v_{2k}|1, \alpha)\mathcal{HIW}_G(b, L) \times g(\gamma_0, \tau_0|\nu_0, p_0, s_0, n_0) \times$$

$$\prod_{k>1} \mathcal{N}(\mu_{1k}|c,d)g(\gamma_{1k},\tau_{1k}|\nu_1,p_1,s_1,n_1)\mathcal{N}(\mu_{2k}|c,d)g(\gamma_{2k},\tau_{2k}|\nu_1,p_1,s_1,n_1).$$

The distribution defined in (1.17) is not tractable thus we apply Gibbs sampling to draw random numbers from it. We use the notation $U = \{u_{lj} : j = 1,2,\ldots,r_l \text{ and } l = 1,2,\ldots,N\}$, $V = \{v_{lj} : j = 1,2,\ldots \text{ and } l = 1,2,\ldots,N\}$ to describe the latent variables and the stick-breaking components; $D = \{d_{lj} j = 1,2,\ldots,r_l \text{ and } l = 1,2,\ldots,N\}$ and $\Delta = \{\delta_{lj} j = 1,2,\ldots,r_l \text{ and } l = 1,2,\ldots,N\}$ to describe the new variables that we have introduced in this section. The Gibbs sampler iterates over the following steps using the conditional independence between the different variables as seen in the appendix:

1. The stick-breaking and the latent variables U, V are updated given $[\Theta, \beta, \Sigma, G, \Lambda, D, \Delta, \pi, Y]$;
2. The latent variable Λ is updated given $[\Theta, \beta, \Sigma, G, U, V, D, \Delta, \pi, Y]$;
3. The parameters of the Normal-Gamma distribution Θ are updated given $[\beta, \Sigma, G, \Lambda, U, V, D, \Delta, \pi, Y]$;
4. The coefficients β of the SUR model are updated given $[\Theta, \Sigma, G, \Lambda, U, V, D, \Delta, \pi, Y]$;
5. The matrix of variance-covariance Σ is updated given $[\Theta, \beta, G, \Lambda, U, V, D, \Delta, \pi, Y]$;
6. The Graph G is updated given $[\Theta, \beta, \Sigma, \Lambda, U, V, D, \Delta, \pi, Y]$;
7. The allocation and the latent variables D, Δ are updated given $[\Theta, \beta, \Sigma, G, \Lambda, U, V, \pi, Y]$;
8. The probability of being sparse π is updated given $[\Theta, \beta, \Sigma, G, \Lambda, U, V, D, \Delta, Y]$.

The full conditional distributions of the Gibbs sampler and the sampling methods are discussed in Appendix A.

1.4 Simulation experiments

This section illustrates the performance of our Bayesian nonparametric sparse model with simulated data. We generate different datasets sample size $T = 100$ from a VAR model with lag $p = 1$:

$$\mathbf{y}_t = B\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \quad \text{for } t = 1, \ldots, 100,$$

where the dimension of y_t and of the square matrix of coefficients B takes different values, $m = 20$ (small dimension), $m = 40$ (medium dimension), $m = 80$ (big dimension). Furthermore, the matrix of coefficients has different costruction, from a block-diagonal to a random form, as follows:

- if $m = 20$, the matrix of coefficients $B = \text{diag}\{B_1, \ldots, B_5\} \in \mathcal{M}_{(20,20)}$ is a block-diagonal matrix with blocks B_j $(j = 1, \ldots, 5)$ of (4×4) matrices on the main diagonal:

$$B_j = \begin{pmatrix} b_{11,j} & \ldots & b_{14,j} \\ \vdots & \vdots & \vdots \\ b_{41,j} & \ldots & b_{44,j} \end{pmatrix},$$

  where the elements are randomly taken from an uniform distribution $\mathcal{U}(-1.4, 1.4)$ and then checked for the stationarity conditions;

- if $m = 40$, the matrix of coefficients $B = \text{diag}(B_1, \ldots, B_{10})$ is a block-diagonal matrix with blocks B_j of (4×4) matrices on the main diagonal:

$$B_j = \begin{pmatrix} b_{11,j} & \ldots & b_{14,j} \\ \vdots & \vdots & \vdots \\ b_{41,j} & \ldots & b_{44,j} \end{pmatrix},$$

  where the elements are randomly taken from an uniform distribution $\mathcal{U}(-1.4, 1.4)$ and then checked for the stationarity conditions;

	mean	mode
$m = 20$	9.48	9
$m = 40$	12.32	12
$m = 80$ (random)	11.49	11
$m = 80$ (blocks)	11.29	12

Table 1.1: Summary statistics of the number of clusters with different dimensions m.

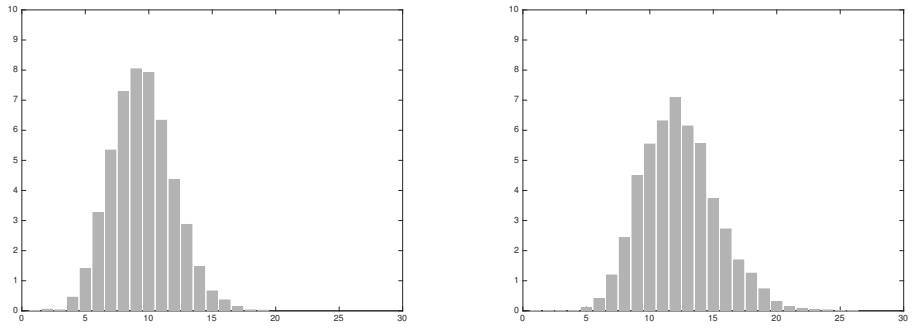
- if $m = 80$, we analyse two different situations, when
  - the matrix of coefficients $B = \text{diag}(B_1, \dots, B_{20})$ is a block-diagonal matrix with blocks B_j of (4×4) matrices on the main diagonal:

$$B_j = \begin{pmatrix} b_{11,j} & \dots & b_{14,j} \\ \vdots & \vdots & \vdots \\ b_{41,j} & \dots & b_{44,j} \end{pmatrix},$$

    where the elements are randomly taken from an uniform distribution $\mathcal{U}(-1.4, 1.4)$ and then checked for the stationarity conditions;
  - the (80×80) matrix of coefficients has 150 elements randomly chosen from an uniform distribution $\mathcal{U}(-1.4, 1.4)$ and then checked for the stationarity conditions.

For all the cases, we run the Gibbs sampler algorithm described in Section 1.3 and sample from the posterior distribution via Monte Carlo methods with $5,000$ iterations and a burn-in period of 500 iterations. Furthermore, we have chosen the hyperparameters for the sparse and non-sparse components as in Section 1.2.2 and the hyperparameters of the Hyper-inverse Wishart as in Section 1.2.2, where the degree of freedom is $b_0 = 3$ and the scale matrix $L = \mathbb{I}_n$. Figure 1.2 and A.1 show the histograms for the posterior distribution of the number of clusters for each sample sizes, the comparison between the construction of our simulated outputs and the posterior of the number of clusters highlights the good

fit of our Bayesian nonparametric hierarchical model, which is also confirmed by the mean and the mode of the number of cluster for every sample sizes (see Table 1.1).

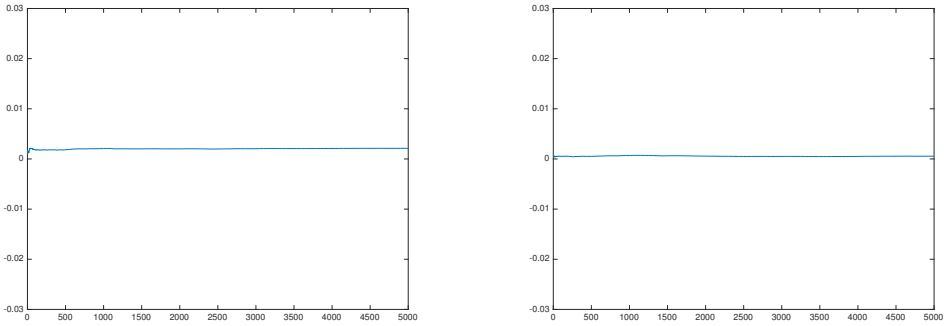


(a) m=20 (b) m=40

FIGURE1.2: Posterior distribution of the number of clusters for $m = 20$ (left) and for $m = 40$ (right).

Focusing on the posterior of the matrix of coefficients B, the proportion of elements whose true simulated values fall inside their 95% credible intervals is 0.96 (for $m = 20$), 0.983 (for $m = 40$), 0.9939 (for $m = 80$ in the block case) and 0.998 (for $m = 80$ in the random element case). We can compute the number of zeroes in the true simulated B, which are 325 ($m = 20$), 1452 ($m = 40$), 6105 and 6261 for $m = 80$ in the block and in the randomly case, respectively. If we compare these values with the posterior number of zeroes in the matrix B, which are 335 (for $m = 20$), 1461 (for $m = 40$), 6102 and 6192 for $m = 80$ in the block and in the randomly case, we have that the differences between them are small, which allow us to consider our approach feasible for the inference of sparse and nonsparse components.

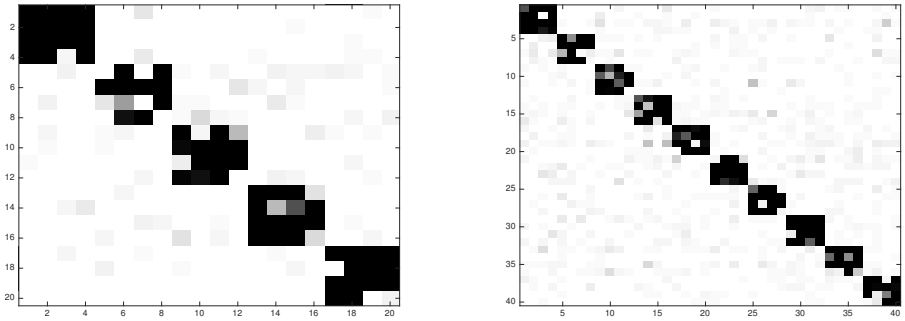
We evaluate the accurancy of our estimates by using the Hamming distance for the matrix of coefficients, which is the difference between the real values of the matrix of coefficients B and the posterior values of it. In definition, the Hamming distance is



(a) m=20 (b) m=40

FIGURE1.3: Hamming distance between B and its posteriors for $m = 20$ (left) and for $m = 40$ (right).

$|\mathbf{a} - \mathbf{b}|_H| = |\{i|a_i \neq b_i\}|$, where the difference $\mathbf{a} - \mathbf{b}$ contains negative values corresponding to points where $b_i > a_i$. Figure 1.3 and A.2 show this difference for different sample sizes and it converges to zero, which means that our posteriors for the matrix of coefficients are exactly what we were expecting.



(a) m=20 (b) m=40

FIGURE1.4: Posterior mean of the matrix of δ for $m = 20$ (left) and for $m = 40$ (right)

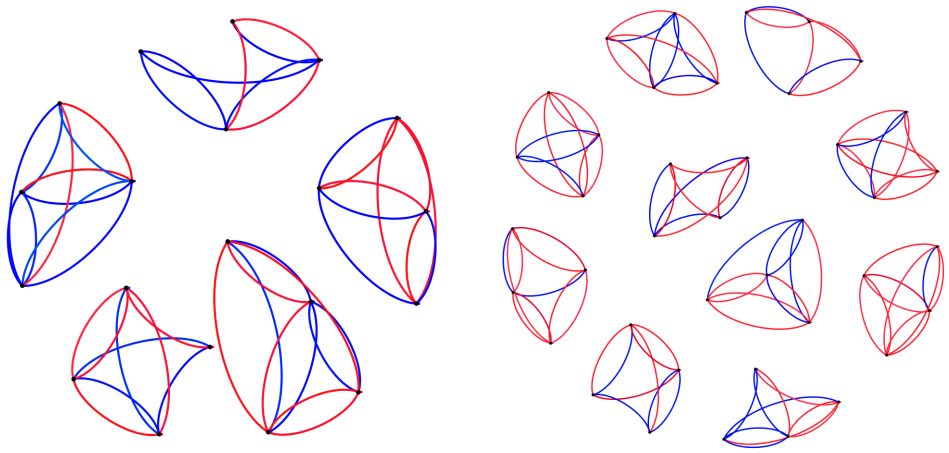
Figure 1.4 and A.3 explain the posterior mean of the matrix of δ, which shows us the choice of the components between the two random measures $\mathbb{P}_0$ and $\mathbb{P}_l$. In particular, we have that the white color explains if the coefficient δ is equal to zero (i.e. sparse com-

ponent), while the black one if the δ is equal to one, for nonsparse components. The representation in Figure 1.4 and A.3 correctly explain the sparsity in the matrix of coefficients through the definition of the matrix of the latent variable δ. In order to identificate the mixture components, we apply the least square clustering method proposed originally in Dahl (2006). The method is based on the choice of the nonsparse components and on the posterior pairwise probabilities of joint classification $P(D_i = D_j|Y, \delta = 1)$. To estimate this matrix, we use the following pairwise probability matrix:

$$P_{ij} = \frac{1}{M_{\text{i}}} \sum_{l=1}^{M_{\text{i}}} \delta_{D_i^l}(D_j^l)$$

where we use every pair of allocation variables D_i^l and D_j^l, with $i, j = 1, \ldots, T_{\text{nsp}}$, T_{nsp} is the number of nonsparse component and $l = 1, \ldots, M_{\text{i}}$, where M_{i} is the number of MCMC iterations. The definition of the pairwise posterior probabilities and of the co-clustering matrix for the atom locations μ allows us to built the weighted networks (see Figure 1.4 and Figure A.4), where the blue edges represent negative weights, while the red ones represent the positive weights. The curved edges follow a clockwise relations, which means that a node A is related to a node B if there is a clockwise curved edge between them and it allows us to explain the presence of different cliques in each simulated examples. As known, the representation with block matrices confirms the presence of different cliques, e.g. for $n = 20$ exactly 5 cliques, while increasing the dimensionality, augment the number of cliques.

We compare our prior with the Bayesian Lasso (Park and Casella (2008)), the Elastic-net (Zou and Hastie (2005)) and to a prior for imposing restrictions on the VAR based on Stochastic Search Variable Selection (SSVS) of George et al. (2008). For the SSVS, we use the default hyperparameters $\tau_1^2 = 0.0001$, $\tau_2^2 = 4$ and $\pi = 0.5$. We use the mean absolute



(a) m=20 (b) m=40

FIGURE1.5: Weighted network for $m = 20$ (left) and for $m = 40$ (right), where the blue edges mean negative weights and red ones represent positive weights.

deviation (MAD, Korobilis (2016)) for looking at the performance of the five different priors: our Bayesian nonparametric prior (BNP), Bayesian Lasso (B-Lasso), Elastic-net (E-Net), SSVS and OLS, unrestriced estimator, equivalent to diffuse prior.

If $\hat{\beta}$ is an estimate of B based on the five priors and $\tilde{\beta}$ is it true value from the DGP,

$$\text{MAD} = \frac{1}{n}\sum_{k=1}^{n}\left|Z_k\hat{\beta}_k - Z_k\tilde{\beta}_k\right|$$

where n denotes the number of VAR coefficients and Z_k is the k-th column of $Z = (I_m \otimes \mathbf{x}')$. Table 1.2 shows that the best perfomance is obtained from our prior for each dimension m and all the priors are performing well related to OLS.

1.5 Measuring contagion effects

We apply the proposed Bayesian nonparametric sparse model to a macroeconomic dataset and, following Diebold and Yilmaz (2015), we extract network structures to investigate the

	BNP	B-Lasso	E-net	SSVS	OLS
$m = 20$	0.228	0.2513	0.2582	0.2938	0.3382
$m = 40$	0.2663	0.3145	0.3143	0.401	0.4835
$m = 80$ (random)	0.2294	0.3011	0.2951	0.5413	0.7048
$m = 80$ (block)	0.2916	0.3773	0.3743	0.5633	0.7290

Table 1.2: Mean absolute deviation statistics for different m.

role of contagion effects between different cycles and the possible relations bewteen GDP of different countries. Furthermore, we study the transmission of shocks and contagion between different countries at different lags.

Following the literature on international business cycles in large models (Kose et al., 2003, 2010; Del Negro and Otrok, 2008) we use a multi-country macroeconomic dataset to study the role of contagion effects between different cycles in the panel, while Francis et al. (2012) and Kaufmann and Schumacher (2012) investigate the role of global business cycles for many different countries in large factor models.

For our analysis, we use a VAR(p), with quarterly lags of interest ($p = 4$) and focus on the GDP growth rate, which is the first difference of the logarithm of each GDP series. We consider a dataset of the most important OECD countries, which will be described below, from the first quarter of 1961 to the second quarter of 2015 for a total of $T = 215$ observations.

Due to missing values in the GDP time series of some countries, we choose a subset of all the OECD countries, which is formed by the most industrialised countries, and in particular we focus on two big macroareas, the European one and the rest of the world, where the latter is formed by the countries from Asia, Oceania, North and Central America and Africa. Hereafter, we describe more in details the two macroareas:

- Rest of the World - Australia, Canada, Japan, Mexico, South Africa, Turkey, United States;

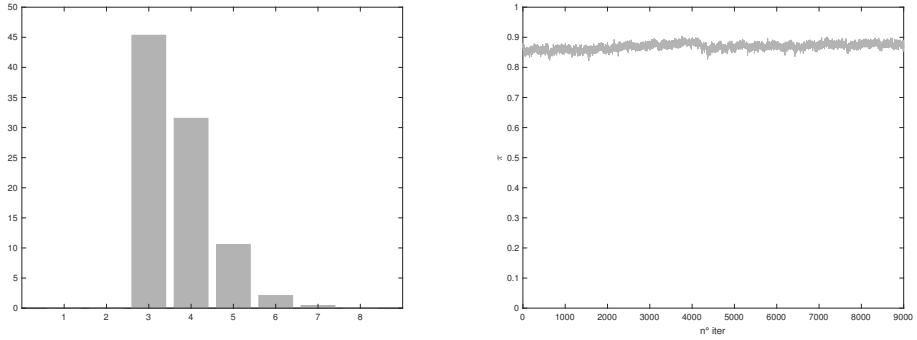


FIGURE1.6: Posterior distribution of the number of clusters for the macroeconomic application (left) and the posterior sample (right) for the probability of being sparse π.

- Europe - Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Iceland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom;

Based on our empirical and computational experiments (see Section 1.4), we run the Gibbs sampling algorithm described in Section 1.3 for $10,000$ iterations with a burn-in period of $1,000$ iterations adopting the same priors of the simulation studies. The location of the posterior mode (value equals to 3) of the histograms in Figure 1.6 allows us to conclude that following our approach there is evidence in favour of three type of macroeconomic contagion effects between the countries in our panel. Figure 1.6 shows the MCMC samples for the probability of being sparse, π, which has posterior mean 0.87 providing evidence of high sparsity in the model.

Figure 1.7 shows the pairwise posterior probabilities P_{ij} that two coefficients β_i and β_j belong to the same cluster. We can detect the presence of four different clusters as seen also from the co-clustering matrix based on the location atom (μ) generated at each iterations of the MCMC method, which has been build up from the least square marginal clustering. The procedure is the clustering D^l sampled at the l-th iteration which minimizes the sum

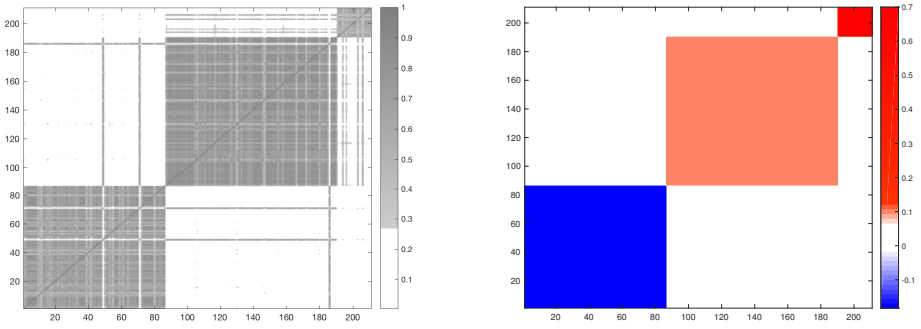


FIGURE1.7: Pairwise posterior probabilities for the clustering (left) and Co-clustering matrix for the atoms μ (right).

of squared deviations from the pairwise posterior probability:

$$l = \arg \min_{l \in \{1,\dots,M\}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\delta_{D_i^l}(D_j^l) - P_{ij} \right)^2$$

Figure 1.8 draws the weighted networks of the GDP connectivity between different countries with respect to different time lags (a) $t-1$, (b) $t-2$, (c) $t-3$ and (d) $t-4$. As seen from Figure 1.7, we have three types of relation: "negative", "positive" and "strong positive". Figure 1.8 shows the weighted networks at each lag, where blue edges represent negative weights and red ones positive weights, and nodes' size is based on the node degree, which is its number of links to other nodes. In terms of directional connectedness received from other countries (out-degree) or transmit to other countries (in-degree), we have:

- at lag $t-1$, Japan appears to be the country that received the highest percentage of shocks from other countries, followed by Spain and Australia.
- at lag $t-1$, Australia is the country that transmit the highest percentage of shocks to other countries, followed by France, Germany and United Kingdom.

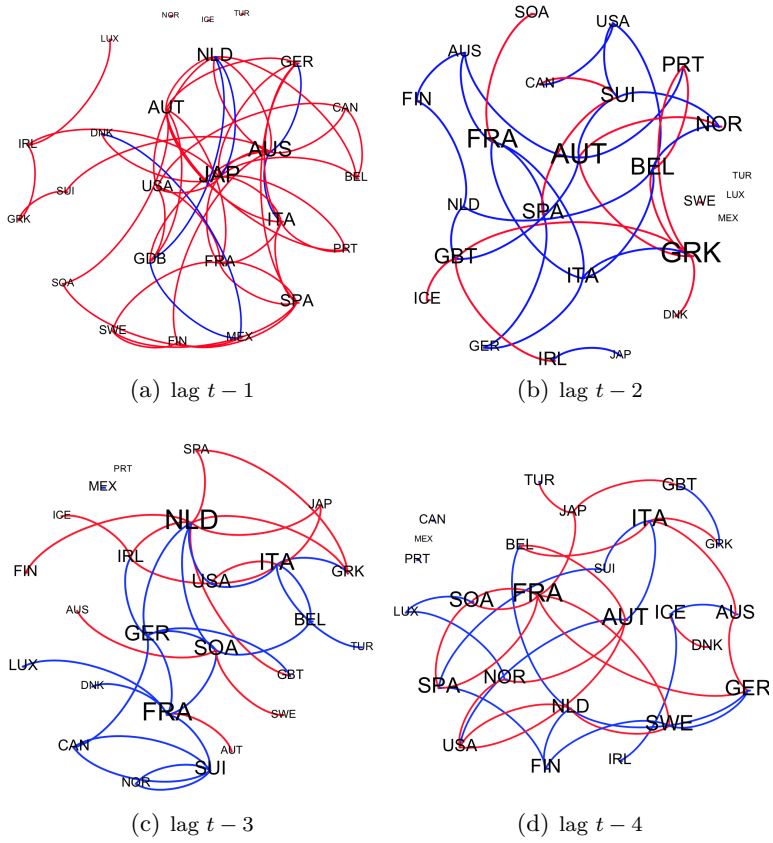
- at lag $t-2$, Greece, France and Austria are the countries that receive highest percentage of shocks from other countries.
- at lag $t-3$ and $t-4$, Germany and Italy receive highest percentage of shocks from other countries and Netherland transmits the highest percentage to other countries.

Table 1.3 shows the network statistics extracted from the four different graphs. Here, the average path length represents the average graph-distance between all pair of nodes, where connected nodes have graph distance 1. From Table 1.3 and Figure 1.8 the lag $t-1$ is more dense, from the average degree 2.92 and from the density of the graph 0.122, and has the highest number of links (73). Indeed, in the lag $t-3$ the average path length reaches its minimum value meaning a faster shock transmission.

	Links	Avg Degree	Density	Avg Path length
$t-1$	73	2.92	0.122	3.423
$t-2$	45	1.80	0.075	3.211
$t-3$	41	1.64	0.068	2.479
$t-4$	52	2.08	0.087	2.718

Table 1.3: The network statistics for the 4 different lags. The average path length represents the average graph-distance between all pairs of nodes. Connected nodes have graph distance 1.

Figure A.5 and Figure A.6 show the predictive distributions (solid lines) generated by the nonparametric approach conditioning on all values of Y_{it}, where $t = 1, \ldots, T$ and $i = 1, \ldots, 25$ (the number of the states) and the best normal fits (dashed lines) for the empirical distributions of all the series. From a comparison with the empirical distribution, we note that the nonparametric approach is able to capture skewness and excess of kurtosis in the data. Furthermore, we observe that for the majority of the countries of interest, the predictive densities (solid lines) generated with our nonparametric sparse approach have fatter tails than the tails of the best normal (dashed lines) and they have long left tails.



(a) lag $t-1$ (b) lag $t-2$

(c) lag $t-3$ (d) lag $t-4$

FIGURE1.8: Weighted Networks of GDP for OECD countries at lag: (a) $t-1$, (b) $t-2$, (c) $t-3$, (d) $t-4$, where blue edges represent negative weights and red ones positive weights. Nodes' size is based on the node degree.

Our Bayesian nonparametric sparse model is suitable for describing and predicting these data thanks to these features.

Figure A.7 and Figure A.8 show the one-step-ahead posterior predictive densities for Y_{it}, where $t = 50, \ldots, T$ and $i = 1, \ldots, 25$, evaluated at the current values of the explanatory variables $Y_{it-1}, \ldots, Y_{it-p}$. In the same plot, the grey area represents the heatmap sequence of the 95% high probability density region of the predictive densities (darker colors represent higher density values). These densities are used to predict the peaks and the troughs of

the cycles in the OECD countries. In particular we can see troughs near the 1980s and 2009s near the crisis in the majority of the European countries.

1.6 Conclusions

In this paper we have proposed a novel Bayesian nonparametric sparse model thourgh the introduction of multiple shrinkage priors. In order to capture the sparsity structure in the model, we introduce two stage of the hierarchy for the prior choice, where the first one consists in a Bayesian lasso conditionally independent Normal-Gamma prior and the second one is given by a random mixture distribution for the hyperparameters of the Normal-Gamma distribution with a particular base measure, based on the two-parameters gamma developed by Miller (1980).

The proposed hierarchical prior is used to proposed a Bayesian nonparametric model for VAR models. We provide an efficient Monte Carlo Markov Chain algorithm for the posterior computations and the effectiveness of this algorithm is assesed in simulation and real data exercises. These simulation studies illustrate the good performance of our model with different sample sizes and different constructions of the matrix of coefficients compared to existing priors in the literature.

Besides through simulation studies, the application to the GDP growth rates in different OECD countries shows the relevant linkages between different units of the panel at various lags and the estimation of the number of cluster in the network. Furthermore we found evidence of good predictive abilities of our Bayesian nonparametric model.

We conclude the paper with the indication of some future research lines. Our hierarchical prior and our nonparametric approach can be extended to the graphical models for the study of the financial contagion with the introduction of link functions (such as the probit or the logit function) or to the Factor autoregressive models (see Kaufmann and

Schumacher (2012)) for the analysis of the stochastic volatility processes.

Acknowledgements

We would like to thank all the conference partecipants for helpful discussion at: "9th Annual RCEA Bayesian Econometric Workshop" at Rimini Center for Economic Analysis; "Joint PhD Workshop Economics and Management" and "Internal Seminar" at Ca' Foscari University of Venice; "Statistics Seminar" at University of Kent; "9th Computational and Financial Econometrics Conference (CFE 2015)" in London; "ISBA 2016 World Meeting" in Sardinia; "3rd Bayesian Young Statistician Meeting" at University of Florence; "7th European Seminar on Bayesian Econometrics" at University of Venice; "10th International Conference on Computational and Financial Econometrics (CFE 2016)" at University of Seville.

Bibliography

Abramowitz, M. and Stegun, I. A. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York.

Acharya, V. V., Engle, R., and Richardson, M. (2012), "Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risks," *American Economic Review*, 102, 59–64.

Ahelgebey, D. F., Billio, M., and Casarin, R. (2015), "Bayesian graphical models for structural vector autoregressive processes," *Journal of Applied Econometrics*.

Ahelgebey, D. F., Billio, M., and Casarin, R. (2016), "Sparse Graphical Vector Autoregression: A Bayesian Approach," *Annals of Economics and Statistics*.

Andrews, D. and Mallows, C. (1974), "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society Series B*, 36, 99–102.

Banbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian vector autoregressions," *Journal of Applied Econometrics*, 25, 71–92.

Barigozzi, M. and Brownlees, C. (2016), "NETS: Network Estimation for Time Series," *Working Paper*.

Bassetti, F., Casarin, R., and Leisen, F. (2014), "Beta-product dependent Pitman-Yor processes for Bayesian inference," *Journal of Econometrics*, 180, 49–72.

Bianchi, D., Billio, M., Casarin, R., and Guidolin, M. (2015), "Modeling Contagion and Systemic Risk," *Working Paper*.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012), "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," *Journal of Financial Econometrics*, 104, 535–559.

Brownlees, C. and Engle, R. (2016), "SRISK: A Conditional Capital Shortfall Measure of Systemic Risk," *The Review of Financial Studies*, Forthcoming.

Canova, F. and Ciccarelli, M. (2004), "Forecasting and turning point prediction in a Bayesian panel VAR model," *Journal of Econometrics*, 120(2), 327–359.

Carriero, A., Clark, T., and Marcellino, M. (2013), "Bayesian VARs: specification choices and forecast accurancy," *Journal of Applied Econometrics*, 25, 400–417.

Chib, S. and Greenberg, E. (1995), "Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models," *Journal of Econometrics*, 68, 339–360.

Chib, S. and Hamilton, B. H. (2002), "Semiparametric Bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67–89.

Dagpunar, J. (1988), "Principles of Random Variate Generation," *Clarendon Oxford Science Pubblications*.

Dagpunar, J. (1989), "An easily implemented generalised inverse Gaussian generator," *Communications in Statistics - Simulation and Computation*, 18, 703–710.

Dahl, D. B. (2006), "Model-based clustering for expression data via a Dirichlet process

mixture model." in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. P. Müller, and M. Vannucci, pp. 201–218, Cambridge University Press.

Del Negro, M. and Otrok, C. (2008), "Dynamic factor models with time-varying parameters: measuring changes in international business cycles," *Fed New York*.

Demirer, M., Diebold, F. X., Liu, L., and Yilmaz, K. (2015), "Estimating Global Bank Network Connectedness," *Manuscript MIT, University of Pennsylvania and Koc University*.

Devroye, L. (2014), "Random variate generation for the generalized inverse Gaussian distribution," *Statistics and Computing*, 24, 239–246.

Diebold, F. X. and Yilmaz, K. (2014), "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms," *Journal of Econometrics*, 182, 119–134.

Diebold, F. X. and Yilmaz, K. (2015), *Measuring the Dynamics of Global Business Cycle Connectedness*, pp. 45–89, Oxford Univerisity Press.

Diebold, F. X. and Yilmaz, K. (2016), "Trans-Atlantic Equity Volatility Connectedness: U.S. and European Trans-Atlantic Equity Volatility Connectedness: U.S. and European Financial Institutions, 2004–2014," *Journal of Financial Econometrics*, 14, 81–127.

Doan, T., Litterman, R., and Sims, C. A. (1984), "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1–100.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Francis, N., Owyang, M., and Savascin, O. (2012), "An endogenously clustered factor approach to international business cycles," *Federal Reserve Bank of St. Louis Working Paper*.

Gefang, D. (2014), "Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage," *International Journal of Forecasting*, 30, 1–11.

George, E. I., Sun, D., and Ni, S. (2008), "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 142, 553–580.

Giudici, P. and Green, P. (1999), "Decomposable graphical Gaussian model determination," *Biometrika*, 86, 758–801.

Griffin, J. and Brown, P. (2006), "Alternative prior distributions for variable selection with very many more variables than observations," Tech. rep., University of Warwick.

Griffin, J. E. (2011), "Inference in infinite superpositions of non-Gaussian Ornstein-Uhlenbeck processes using Bayesian nonparametic methods," *Journal of Financial Econometrics*, 1, 1–31.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

Griffin, J. E. and Steel, M. F. J. (2011), "Stick-breaking autoregressive processes," *Journal of Econometrics*, 162, 383–396.

Halphen, E. (1941), "Sur un nouveau type de courbe de frequence," *Comptes Rendus des seances de l'Academie des Sciences*.

Hatjispyros, S. J., Nicoleris, T. N., and Walker, S. G. (2011), "Dependent mixtures of Dirichlet processes," *Computational Statistics & Data Analysis*, 55, 2011–2025.

Hirano, K. (2002), "Semiparametric Bayesian Inference in autoregressive panel data models," *Econometrica*, 70(2), 781–799.

Hjort, N. L., Homes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, Cambridge University Press.

Hoermann, W. and Leydold, J. (2013), "Generating Generalized Inverse Gaussian random variates," *Research Report Series*.

Jensen, J. M. and Maheu, M. J. (2010), "Bayesian semiparametric stochastic volatility modeling," *Journal of Econometrics*, 157, 306–316.

Jochmann, M. (2015), "Modeling U.S. inflation dynamics: A Bayesian nonparametric approach," *Econometric Reviews*, 34, 537–558.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in stochastic computation for high-dimensional graphical models," *Statistical Science*, pp. 388–400.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Kaufmann, S. (2010), "Dating and forecasting turning points by Bayesian clustering with dynamic structure: a suggestion with an application to Austrian data," *Journal of Applied Econometrics*, 25, 309–344.

Kaufmann, S. and Schumacher, C. (2012), "Finding relevant variables in sparse Bayesian

factor models: economic applications and simulation results," *Discussion Paper Deutsche Bundesbank*.

Koop, G. (2013), "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28, 177–203.

Koop, G. and Korobilis, D. (2010), "Bayesian multivariate time series methods for empirical macroeconomics," *Foundations and Trends in Econometrics*, 3, 267–358.

Koop, G. and Korobilis, D. (2015), "Model uncertainty in panel vector autoregressions," *European Economic Review*, 81, 115–131.

Korobilis, D. (2013), "VAR forecasting using Bayesian variable selection," *Journal of Applied Econometrics*, 28, 204–230.

Korobilis, D. (2016), "Prior selection for panel vector autoregressions," *Computational Statistics & Data Analysis*, 101, 110–120.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2003), "International Business Cycles: World, region and country specific factors," *American Economic Review*, 93, 1216–1239.

Kose, M. A., Otrok, C., and Whiteman, C. H. (2010), "Understanding the evolution of world business cycles," *Journal of International Economics*, 75, 110–130.

Litterman, R. (1980), "Techniques for forecasting with vector autoregressions," *University of Minnesota, Ph.D. Dissertation*.

Litterman, R. (1986), "Forecasting with Bayesian vector autoregressions-five years of experience," *Journal of Business and Economic Statistics*, 4, 25–38.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I. Density estimates," *The Annals of Statistics*, 12, 351–357.

MacEachern, S. N. (1999), "Dependent nonparametric processes," in *In ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA*, American Statistical Association.

MacEachern, S. N. (2001), "Decision theoretic aspects of dependent nonparametric processes." in *Bayesian Methods with Applications to Science, Policy and Official Statistics*, ed. E. George, pp. 551–560, Creta: ISBA.

MacEachern, S. N. and Müller, P. (1998), "Estimating mixtures of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

MacLehose, R. and Dunson, D. (2010), "Bayesian semiparametric multiple shrinkage," *Biometrics*, 66, 455–462.

Miller, R. (1980), "Bayesian analysis of the two-parameter Gamma distribution," *Technometrics*, 22.

Min, C. and Zellner, A. (1993), "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates," *Journal of Econometrics*, 56, 89–118.

Müller, P., Quintana, F., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society B*, 66, 735–749.

Papaspiliopoulos, O. and Roberts, G. (2008), "Retrospective Markov chain Monte Carlo for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Rodriguez, A. and ter Horst, E. (2008), "Bayesian dynamics density estimation," *Bayesian Analysis*, 3, 339–366.

Scott, S. L. and Varian, H. R. (2013), "Predicting the present with Bayesian structural time series," *International Journal of Mathematical Modelling and Numerical Optimisation*, 5.

Sethuraman, J. (1994), "A constructive definition of the Dirichlet process prior," *Statistica Sinica*, 2, 639–650.

Sims, C. A. (1980), "Macroeconomics and reality," *Econometrica*, 48, 1–48.

Sims, C. A. (1992), "Interpreting the macroeconomic time series facts: The effects of monetary policy," *European Economic Review*, 38, 975–1000.

Sims, C. A. and Zha, T. (1998), "Bayesian methods for dynamic multivariate models," *International Economic Review*, 39(4), 949–968.

Stock, J. H. and Watson, M. W. (2012), "Generalized shrinkage methods for forecasting using many predictors," *Journal of Business & Economic Statistics*, 30, 481–493.

Taddy, M. A. (2010), "An auto-regressive mixture model for dynamic spatial Poisson processes: Application to tracking the intensity of violent crime," *Journal of the American Statistical Association*, 105, 1403–1427.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, Series B (Methodological), 267–288.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

Wang, H. (2010), "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics," *Computational Statistics & Data Analysis*, 54, 2866–2877.

Zellner, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias," *Journal of the American Statistical Association*, 57, 500–509.

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

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, 67, 301–320.

Zou, H. and Zhang, H. H. (2009), "On the adaptive elastic-net with diverging number of parameters," *Annals of Statistics*, 37, 1733–1751.

Appendix A

Technical Details of Chapter 1

A.1 Gibbs sampling details

We introduce the following notations, for $k \geqslant 1$, and $l = 1, 2$,

$$\mathcal{D}_{lk} = \{j \in 1, \ldots, r_l : d_{lj} = k, \delta_{lj} = 1\},$$

$$\mathcal{D}^* = \{k | \mathcal{D}_{1k} \cup \mathcal{D}_{2k} \neq 0\}, \quad D^* = \max_{l=1,2} \max_{j \in \{1,\ldots,r_l\}} d_{lj},$$

where $\mathcal{D}_k$ denotes the set of indexes of the coefficients allocated to the k-th component of the mixture and $\mathcal{D}^*$ the set of indexes of the non-empty mixture components, while D^* is the number of stick-breaking components used in the mixture. As noted by Kalli et al. (2011), the sampling of infinitely many elements of Θ and V is not necessarily, since only the elements in the full conditional probability density functions of D, Δ are needed.

The maximum number of atoms and stick-breaking components to sample is $N^* = \max\{N_1^*, N_2^*\}$, where N_l^* is the smallest integer such that $\sum_{k=1}^{N_l^*} w_{lk} > 1 - u_l^*$, where $u_l^* = \min_{1 \leqslant j \leqslant n_l}\{u_{lj}\}$. In the following sections we explain in details all the steps of the Gibbs sampler, which is built on the slice sampler algorithm of Walker (2007) and of Kalli et al. (2011).

A.1.1 Update V,U

We treat V as three blocks of random length: $V = (V^*, V^{**}, V^{***})$, where

$$V^* = \{V_k : k \in \mathcal{D}^*\} = (v_{k1}, \ldots, v_{kD^*}),$$

$$V^{**} = (v_{kD^*+1}, \ldots v_{kN^*}), \quad V^{***} = \{V_k : k > N^*\}.$$

In order to sample from the conditional distribution of (U, V) a further blocking is used:

i) Sampling from the full conditional posterior distribution of V^*, is obtained by drawing v_{1k}, v_{2k}, with $k \leqslant D^*$ from the full conditionals

$$f(v_{1j}|\ldots) \propto \mathcal{B}e\left(1 + \sum_{j=1}^{r_1} \mathbb{I}(d_{1j} = d, \delta_{1j} = 1), \alpha + \sum_{j=1}^{r_1} \mathbb{I}(d_{1j} > d, \delta_{1j} = 1)\right),$$

$$f(v_{2j}|\ldots) \propto \mathcal{B}e\left(1 + \sum_{j=1}^{r_2} \mathbb{I}(d_{2j} = d, \delta_{2j} = 1), \alpha + \sum_{j=1}^{r_2} \mathbb{I}(d_{2j} > d, \delta_{2j} = 1)\right).$$

ii) Sampling form the full conditional posterior distribution of U is obtain by simulating from, for $1 \leqslant j \leqslant r_1$,

$$f(u_{1j}|\ldots) \propto \begin{cases} \mathbb{I}(u_{1j} < w_{1d_{1j}})^{\delta_{1j}} & \text{if } \delta_{1j} = 1, \\ \mathbb{I}(u_{1j} < 1)^{1-\delta_{1j}} & \text{if } \delta_{1j} = 0, \end{cases}$$

and, for $1 \leqslant j \leqslant r_2$,

$$f(u_{2j}|\ldots) \propto \begin{cases} \mathbb{I}(u_{2j} < w_{2d_{2j}})^{\delta_{2j}} & \text{if } \delta_{2j} = 1, \\ \mathbb{I}(u_{2j} < 1)^{1-\delta_{2j}} & \text{if } \delta_{2j} = 0. \end{cases}$$

iii) For (V^{**}, V^{***}) given $[\Theta, \Sigma, \Lambda, V^*, D, \Delta, Y]$, we need to sample only the elements of V^{**} from the prior distribution of the stick-breaking construction, that is, for each $l = 1, 2$,

$$f(v_{lj}|\ldots) \propto \mathcal{B}e(1, \alpha).$$

A.1.2 Update the mixing parameters $\boldsymbol{\lambda}$

We update the mixing parameters λ_{lj} $(l = 1, 2)$, where the full conditional posterior distribution of λ_{lj} is

$$f(\lambda_{lj}|\dots) \propto \lambda_{lj}^{-\frac{1}{2}(1-\delta_{lj})} \exp\left\{\left(-\frac{1}{2}\frac{1}{\lambda_{lj}}\beta_{lj}^2 - \frac{\tau_0}{2}\lambda_{lj}\right)(1-\delta_{lj})\right\}\lambda_{lj}^{(\gamma_0-1)(1-\delta_{lj})} \times$$

$$\times\, \lambda_{lj}^{-\frac{1}{2}\delta_{lj}} \exp\left\{-\frac{1}{2}\frac{1}{\lambda_{lj}}(\beta_{lj} - \mu_{ld_{lj}})^2\delta_{lj}\right\}\lambda_{lj}^{(\gamma_{ld_{lj}}-1)\delta_{lj}} \exp\left\{\left(-\frac{\tau_{ld_{lj}}}{2}\lambda_{lj}\right)\delta_{lj}\right\}$$

$$\propto \lambda_{lj}^{C_{lj}-1} \exp\left\{-\frac{1}{2}\left[A_{lj}\lambda_{lj} + \frac{B_{lj}}{\lambda_{lj}}\right]\right\} \propto \mathcal{G}i\mathcal{G}(A_{lj}, B_{lj}, C_{lj}),$$

where $\mathcal{G}i\mathcal{G}$ stays for Generalize Inverse Gaussian of parameters $A_{lj} > 0$, $B_{lj} > 0$ and C_{lj} a real parameter (see Halphen (1941), Hoermann and Leydold (2013), Devroye (2014), Dagpunar (1988) and Dagpunar (1989)), which, in our case, are defined as

$$A_{lj} = \left[(1-\delta_{lj})\tau_0 + \delta_{lj}\tau_{ld_{lj}}\right], \qquad B_{lj} = \left[(1-\delta_{lj})\beta_{lj}^2 + \delta_{lj}(\beta_{lj} - \mu_{ld_{lj}})^2\right],$$

$$C_{lj} = \left[(1-\delta_{lj})\gamma_0 + \gamma_{ld_{lj}}\delta_{lj} - \frac{1}{2}\right].$$

We use the λ_{lj} just drawn for construct the matrix $\Lambda_l = \text{diag}\{\boldsymbol{\lambda}_l\}$, where $\text{diag}\{\boldsymbol{\lambda}_l\}$ returns a diagonal matrix with the elements of $\boldsymbol{\lambda}_l = (\lambda_{l1}, \dots, \lambda_{lr_l})'$ on the main diagonal. In practice we have two different matrix, $\Lambda_1 = \text{diag}\{\lambda_{11}, \dots, \lambda_{1r_1}\}$ and $\Lambda_2 = \text{diag}\{\lambda_{21}, \dots, \lambda_{2r_2}\}$.

A.1.3 Update Θ

We consider two different cases: the sparse one, where the parameters are $(\mu_0, \gamma_0, \tau_0)$, and the nonsparse case, where the parameters are $(\mu_k, \gamma_k, \tau_k)$, with $k \geqslant 1$. Since the prior for μ_0 has unit probability mass at 0, the full conditional distribution of μ_0 is $f(\mu_0|\dots) = \delta_{\{0\}}(\mu_0)$. The full conditional distribution of the shape and scale parameters (γ_0, τ_0) is:

$$f((\gamma_0, \tau_0)|\dots) \propto g(\gamma_0, \tau_0|\nu_0, p_0, s_0, n_0) \prod_{j=1|\delta_{1j}=0}^{r_1} \left(\frac{(\tau_0/2)^{\gamma_0}}{\Gamma(\gamma_0)}\lambda_{1j}^{\gamma_0-1} \exp\left\{-\frac{\tau_0}{2}\lambda_{1j}\right\}\right)$$

$$\times \prod_{j=1|\delta_{2j}=0}^{r_2} \left(\frac{(\tau_0/2)^{\gamma_0}}{\Gamma(\gamma_0)} \lambda_{2j}^{\gamma_0 - 1} \exp\left\{ -\frac{\tau_0}{2} \lambda_{2j} \right\} \right), \tag{A.1}$$

where we assume that:

$$r_{1,0} = \sum_{j=1}^{r_1} (1 - \delta_{1j}) = r_1 - r_{1,1}, \qquad r_{1,1} = \sum_{j=1}^{r_1} \delta_{1j},$$

$$r_{2,0} = \sum_{j=1}^{r_2} (1 - \delta_{2j}) = r_2 - r_{2,1}, \qquad r_{2,1} = \sum_{j=1}^{r_2} \delta_{2j}.$$

The distribution in (A.1) has the same kernel of the prior distribution $g(\gamma_0, \tau_0 | \dots)$ given in (1.10), that is:

$$f((\gamma_0, \tau_0) | \dots) \propto \tau_0^{\nu_0 \gamma_0 - 1} p_0^{\gamma_0 - 1} \exp\{-s_0 \tau_0\} \frac{1}{\Gamma(\gamma_0)^{n_0}} \times$$

$$\times \frac{(\tau_0/2)^{r_{1,0}\gamma_0}}{\Gamma(\gamma_0)^{r_{1,0}}} \left(\prod_{j|\delta_{1j}=0} \lambda_{1j} \right)^{\gamma_0 - 1} \exp\left\{ -\frac{\tau_0}{2} \sum_{j|\delta_{1j}=0} \lambda_{1j} \right\}$$

$$\times \frac{(\tau_0/2)^{r_{2,0}\gamma_0}}{\Gamma(\gamma_0)^{r_{2,0}}} \left(\prod_{j|\delta_{2j}=0} \lambda_{2j} \right)^{\gamma_0 - 1} \exp\left\{ -\frac{\tau_0}{2} \sum_{j|\delta_{2j}=0} \lambda_{2j} \right\}$$

$$\propto g\Bigg(\gamma_0, \tau_0 | \nu_0 + r_{1,0} + r_{2,0}, p_0 \prod_{j|\delta_{1j}=0} \lambda_{1j} \prod_{j|\delta_{2j}=0} \lambda_{2j},$$

$$s_0 + \frac{1}{2} \sum_{j|\delta_{1j}=0} \lambda_{1j} + \frac{1}{2} \sum_{j|\delta_{2j}=0} \lambda_{2j}, n_0 + r_{1,0} + r_{2,0} \Bigg).$$

In order to draw samples from g we apply here a collapsed Gibbs sampler. Samples from $f(\gamma)$ are obtained by a Metropolis-Hastings (MH) algorithm with the prior as proposal, we start with a value of $\gamma^* \sim \mathcal{G}a(1/2, 2)$, we remind $q(\gamma)$ is the probability density function of γ and is distributed as a $\mathcal{G}a(1/2, 2)$. The acceptance probability of the MH step is:

$$\alpha(\gamma^*, \gamma_{\text{old}}) = \min\left\{ 1, \frac{f(\gamma^*) q(\gamma_{\text{old}})}{f(\gamma_{\text{old}}) q(\gamma^*)} \right\}. \tag{A.2}$$

The MH chain updates as follows:

$$\gamma_{\text{new}} = \begin{cases} \gamma_{\text{old}} & \text{if } u > \alpha(\gamma^*, \gamma_{\text{old}}), \\ \gamma^* & \text{if } u \leqslant \alpha(\gamma^*, \gamma_{\text{old}}), \end{cases}$$

where u is a random number from a standard uniform. Samples from the conditional $f(\tau|\gamma)$ are easily obtained since $f(\tau|\gamma)$ is a Gamma distribution.

In the nonsparse case, we generate samples $(\mu_{lk}, \gamma_{lk}, \tau_{lk})$, $k = 1, \ldots, N^*$, $l = 1, 2$, by applying a single move Gibbs sampler with full conditional distributions $f(\mu_{lk}|\ldots)$ and $f(\gamma_{lk}, \tau_{lk}|\ldots)$. The full conditional

$$\begin{aligned} f(\mu_{lk}|\ldots) &\propto \mathcal{N}(\mu_{lk}|c, d) \prod_{j|\delta_{lj}=1, d_{lj}=k} \mathcal{N}(\beta_{lj}|\mu_{lk}, \lambda_{lj}) \\ &\propto \frac{1}{\sqrt{2\pi d}} \exp\left\{-\frac{1}{2d}(\mu_{lk} - c)^2\right\} \prod_{j|\delta_{lj}=1, d_{lj}=k} \frac{1}{\sqrt{2\pi\lambda_{lj}}} \exp\left\{-\frac{1}{2\lambda_{lj}}(\beta_{lj} - \mu_{lk})^2\right\} \\ &\propto \exp\left\{-\frac{1}{2d}(\mu_{lk} - c)^2 - \sum_{j|\delta_{lj}=1, d_{lj}=k} \frac{1}{2\lambda_{lj}}(\beta_{lj} - \mu_{lk})^2\right\} \end{aligned}$$

is proportional to the normal $\mathcal{N}(\tilde{E}_k, \tilde{V}_k)$ with parameters $\tilde{E}_k = \tilde{V}_k \left(\frac{c}{d} + \sum_{j|\delta_{lj}=1, d_{lj}=k} \frac{\beta_{lj}}{\lambda_{lj}}\right)$ and $\tilde{V}_k = \left(\frac{1}{d} + \sum_{j|\delta_{lj}=1, d_{lj}=k} \frac{1}{\lambda_{lj}}\right)^{-1}$. On the other hand, the joint conditional posterior of (γ_{lk}, τ_{lk}) is:

$$f((\gamma_{lk}, \tau_{lk})|\ldots) \propto g(\gamma_{lk}, \tau_{lk}|\nu_1, p_1, s_1, n_1) \prod_{j|\delta_{lj}=1, d_{lj}=k} \left(\frac{(\tau_{lk}/2)^{\gamma_{lk}}}{\Gamma(\gamma_{lk})} \lambda_{lj}^{\gamma_{lk}-1} \exp\left\{-\frac{\tau_{lk}}{2}\lambda_{lj}\right\}\right), \tag{A.3}$$

where we have defined $r_{l,1k} = \sum_{j=1}^{r_l} \delta_{lj} \mathbb{I}(d_{lj} = k)$. Hence (A.3) can be reduced as

$$f((\gamma_{lk}, \tau_{lk})|\ldots) \propto \tau_{lk}^{\nu_1\gamma_{lk}-1} p_1^{\gamma_{lk}-1} \exp\{-s_1\tau_{lk}\} \frac{1}{\Gamma(\gamma_{lk})^{n_1}} \times$$

$$\times \frac{(\tau_{lk}/2)^{r_{l,1k}\gamma_{lk}}}{\Gamma(\gamma_{lk})^{r_{l,1k}}} \left(\prod_{j|\delta_{lj}=1,d_{lj}=k} \lambda_{lj} \right)^{\gamma_{lk}-1} \exp\left\{ -\frac{\tau_{lk}}{2} \sum_{j|\delta_{lj}=1,d_{lj}=k} \lambda_{lj} \right\}$$

$$\propto g\left(\gamma_{lk}, \tau_{lk} | \nu_1 + r_{l,1k}, p_1 \prod_{j|\delta_{lj}=1,d_{lj}=k} \lambda_{lj}, s_1 + \frac{1}{2} \sum_{j|\delta_{lj}=1,d_{lj}=k} \lambda_{lj}, n_1 + r_{l,1k} \right),$$

for $k \in \mathcal{D}^*$ and from the prior G_0 for $k \notin \mathcal{D}^*$. As in the sparse case, we apply a MH algorithm, with the acceptance probability as described in (A.2).

A.1.4 Update $\boldsymbol{\beta}$

The full conditional posterior distribution of $\boldsymbol{\beta}$ is:

$$f(\boldsymbol{\beta}_l | \dots) \propto \exp\left\{ -\frac{1}{2} \left(\sum_t \boldsymbol{\beta}_l' X_t' \Sigma^{-1} X_t \boldsymbol{\beta}_l + \right.\right.$$

$$\left.\left. - 2\boldsymbol{\beta}_l' \sum_t X_t' \Sigma^{-1} \mathbf{y}_t \right) \right\} - \prod_{j=1}^{n} \exp\left\{ -\frac{1}{2} \frac{\boldsymbol{\beta}_l^2}{\lambda_{lj}} (1 - \delta_{lj}) - \frac{1}{2\lambda_{lj}} (\boldsymbol{\beta}_l - \mu_{d_{lj}})^2 \delta_{lj} \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} \left(\sum_t \boldsymbol{\beta}_l' X_t' \Sigma^{-1} X_t \boldsymbol{\beta}_l + \right.\right.$$

$$\left.\left. - 2\boldsymbol{\beta}_l' \sum_t X_t' \Sigma^{-1} \mathbf{y}_t \right) - \frac{1}{2} \left(\boldsymbol{\beta}_l' \Lambda_l^{-1} \boldsymbol{\beta}_l - 2\boldsymbol{\beta}_l' \Lambda_l^{-1} (\boldsymbol{\mu}_l^* \odot \boldsymbol{\delta}_l) \right) \right\}$$

$$\sim \mathcal{N}_{r_l}(\tilde{\mathbf{v}}_\mathbf{l}, M_l),$$

where

$$M_l = \left(\sum_t X_t' \Sigma^{-1} X_t + \Lambda_l^{-1} \right)^{-1},$$

$$\tilde{\mathbf{v}}_\mathbf{l} = M_l \left(\sum_t X_t' \Sigma^{-1} \mathbf{y}_t + \Lambda_l^{-1} (\boldsymbol{\mu}_l^* \odot \boldsymbol{\delta}_l) \right),$$

and $\boldsymbol{\mu}_l^* = (\mu_{ld_{l1}}, \dots, \mu_{ld_{lr_l}})'$, $\boldsymbol{\delta}_l = (\delta_{l1}, \dots, \delta_{lr_l})'$.

A.1.5 Update Σ

Let $\mathcal{S} = \{S_1, \ldots, S_{n_S}\}$ and $\mathcal{P} = \{P_1, \ldots, P_{n_P}\}$ be the set of separators and of prime components, respectively, of the graph G. So the density of the hyper-inverse Wishart for Σ conditional on the graph G is:

$$p(\Sigma) = \prod_{P \in \mathcal{P}} p(\Sigma_P) \left(\prod_{S \in \mathcal{S}} p(\Sigma_S) \right)^{-1}, \tag{A.4}$$

where

$$p(\Sigma_P) \propto |\Sigma_P|^{-(b+2\text{Card}(P))/2} \exp\left\{ -\frac{1}{2} \text{tr}(\Sigma_P^{-1} L_P) \right\}, \tag{A.5}$$

with L_P is the positive-definite symmetric diagonal block of $\tilde{L}$ corresponding to Σ_P.

By using the sets $\mathcal{S}$ and $\mathcal{P}$ and since we are working with the decomposable graph, we know that the likelihood of the graphical gaussian model can be approximated as the ratio between the likelihood in the prime components and the likelihood in the separator components. So the posterior for Σ factorizes as follows:

$$\begin{aligned}
p(\Sigma|\ldots) &\propto \prod_{t=1}^{T} (2\pi)^{n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} \left(y_t - X_t'\beta\right)' \Sigma^{-1} \left(y_t - X_t'\beta\right) \right) p(\Sigma) \\
&\propto |\Sigma|^{T/2} \exp\left(-\frac{1}{2} \text{tr}\left(\sum_t \left(y_t - X_t'\beta\right)' \Sigma^{-1} \left(y_t - X_t'\beta\right) \right) \right) p(\Sigma) \\
&\propto \frac{\prod_{P \in \mathcal{P}} |\Sigma_P|^{-T/2} \exp\left(-\frac{1}{2} \text{tr}\left(\sum_t \left(y_t - X_t'\beta\right)' \Sigma_P^{-1} \left(y_t - X_t'\beta\right) \right) \right)}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-T/2} \exp\left(-\frac{1}{2} \text{tr}\left(\sum_t \left(y_t - X_t'\beta\right)' \Sigma_S^{-1} \left(y_t - X_t'\beta\right) \right) \right)} \times \\
&\frac{\prod_{P \in \mathcal{P}} |\Sigma_P|^{-(b+2\text{Card}(P))/2} \exp\left\{ -\frac{1}{2} \text{tr}(\Sigma_P^{-1} L_P) \right\}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-(b+2\text{Card}(S))/2} \exp\left\{ -\frac{1}{2} \text{tr}(\Sigma_S^{-1} L_S) \right\}} \\
&\propto \frac{\prod_{P \in \mathcal{P}} |\Sigma_P|^{-(b+2\text{Card}(P)+T)/2}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-(b+2\text{Card}(S)+T)/2}}
\end{aligned}$$

$$\frac{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma_P^{-1}\left(\sum_t (y_t - X_t'\beta)'(y_t - X_t'\beta) + L_P\right)\right)\right)}{\exp\left(-\frac{1}{2}\text{tr}\left(\Sigma_S^{-1}\left(\sum_t (y_t - X_t'\beta)'(y_t - X_t'\beta) + L_S\right)\right)\right)}.$$

So we have that the posterior distribution for Σ is drawn from:

$$p(\Sigma|\dots) \propto \mathcal{HIW}_G\left(b + T, \tilde{L} + \sum_{t=1}^{T}(y_t - X_t'\beta)'(y_t - X_t'\beta)\right).$$

A.1.6 Update Graph G

We apply a Markov chain Monte Carlo for multivariate graphical models for learning the graph structure G (see Giudici and Green (1999) and Jones et al. (2005)). We see due to the prior independence assumption of the parameters that:

$$p(\mathbf{y}|G) = \iint \prod_{t=1}^{T}(2\pi)^{-n/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}(y_t - X_t'\beta)\Sigma^{-1}(y_t - X_t'\beta)\right)$$
$$p(\beta)p(\Sigma|G)d\beta d\Sigma.$$

This integral is difficult to compute and evaluate analytically and we apply a Candidate's formula along the line of Chib and Greenberg (1995) and Wang (2010). Following Jones et al. (2005) we apply a local-move Metropolis-Hastings based on the conditional posterior $p(G|\dots)$. A candidate G' is sampled from a proposal distribution $q(G'|G)$ and accepted with probability

$$\alpha = \min\left\{1, \frac{p(G'|\mathbf{y})q(G|G')}{p(G|\mathbf{y})q(G'|G)}\right\}.$$

We use the add/delete edge move proposal of Jones et al. (2005).

A.1.7 Update D and Δ

The full conditionals of D are obtain by sampling from the two different cases, when $\delta_{lj} = 1$ and $\delta_{lj} = 0$ $(l = 1, 2)$. Starting for $\delta_{lj} = 1$, we have

$$P(d_{lj} = d, \delta_{lj} = 1|\dots) \propto (1 - \pi_l)\mathcal{N}(\beta_{lj}|\mu_{ld}, \lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_{ld}, \tau_{ld}/2)\mathbb{I}(u_{lj} < w_{ld})$$

$$\propto \frac{(1-\pi_l)\mathcal{N}(\beta_{lj}|\mu_{ld},\lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_{ld},\tau_{ld}/2)}{\sum_{k\in A_{w_l}(u_{lj})}\mathcal{N}(\beta_{lj}|\mu_{lk},\lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_{lk},\tau_{lk}/2)} \quad \forall d \in A_{w_l}(u_{lj}),$$

for $\delta_{lj}=1$, while we have

$$P(d_{lj}=d,\delta_{lj}=0|\dots)\propto \pi_l \mathbb{I}(u_{lj}<\tilde{w}_{ld}),$$

with $d \in A_{\tilde{w}}(u_{lj})$, where $A_{\tilde{w}}(u_{lj})=\{k: u_{lj}<\tilde{w}_k\}$ which is equal to $\{0\}$, because $\tilde{w}_k=0$, $\forall k>0$,

$$P(d_{lj}=d,\delta_{lj}=0|\dots)\propto \begin{cases} \pi_l\mathbb{I}(u_{lj}<1)\mathcal{N}(\beta_{lj}|0,\lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_0,\tau_0/2) & \text{if } d=0,\\ 0 & \text{if } d>0. \end{cases}$$

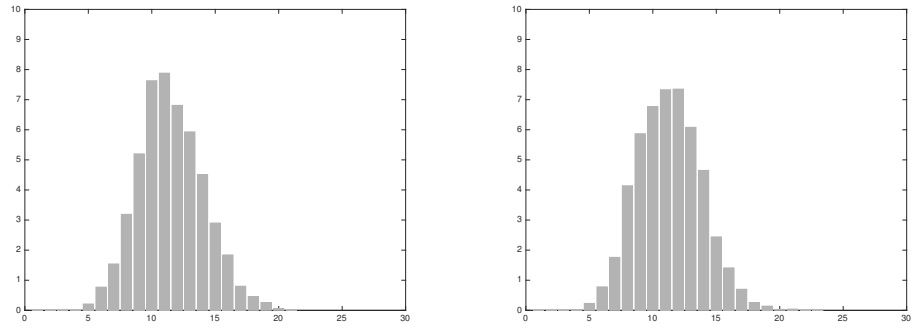
$$\propto \pi_l\mathcal{N}(\beta_{lj}|0,\lambda_{lj})\mathcal{G}a(\lambda_{lj}|\gamma_0,\tau_0/2) \quad \text{if } d=0.$$

A.1.8 Update $\pi=(\pi_1,\pi_2)$

We assume that the prior for π_l is $\mathcal{B}e(1,\alpha_l)$, so we have that the full conditional for π_l is,

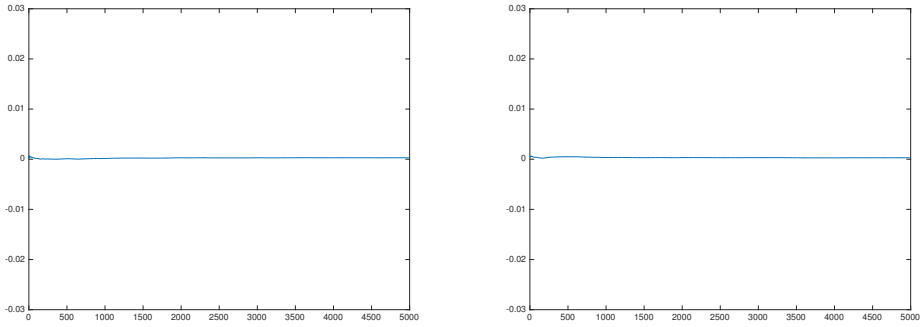
$$f(\pi_l|\dots)\propto \mathcal{B}e\left(r_l+1-\sum_{i=1}^{r_l}\mathbb{I}(\delta_{li}=1),\alpha_l+\sum_{i=1}^{r_l}\mathbb{I}(\delta_{li}=1)\right).$$

A.2 Simulated and Real Data Results



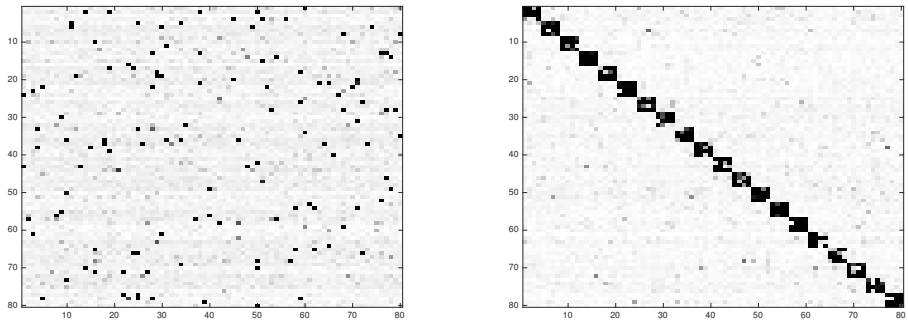
(a) m=80 with random elements (b) m=80 with block matrices

FIGUREA.1: Posterior distribution of the number of clusters for $m = 80$, with random elements in the B matrix (left) and with block matrix (right).



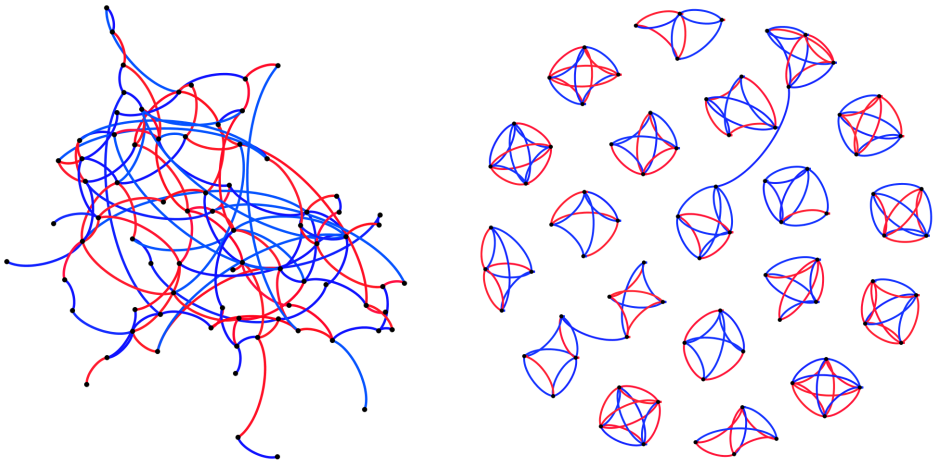
(a) m=80 with random numbers (b) m=80 with blocks

FIGUREA.2: Hamming distance between B and its posteriors for $m = 80$, with random elements (left) and with block matrix (right).



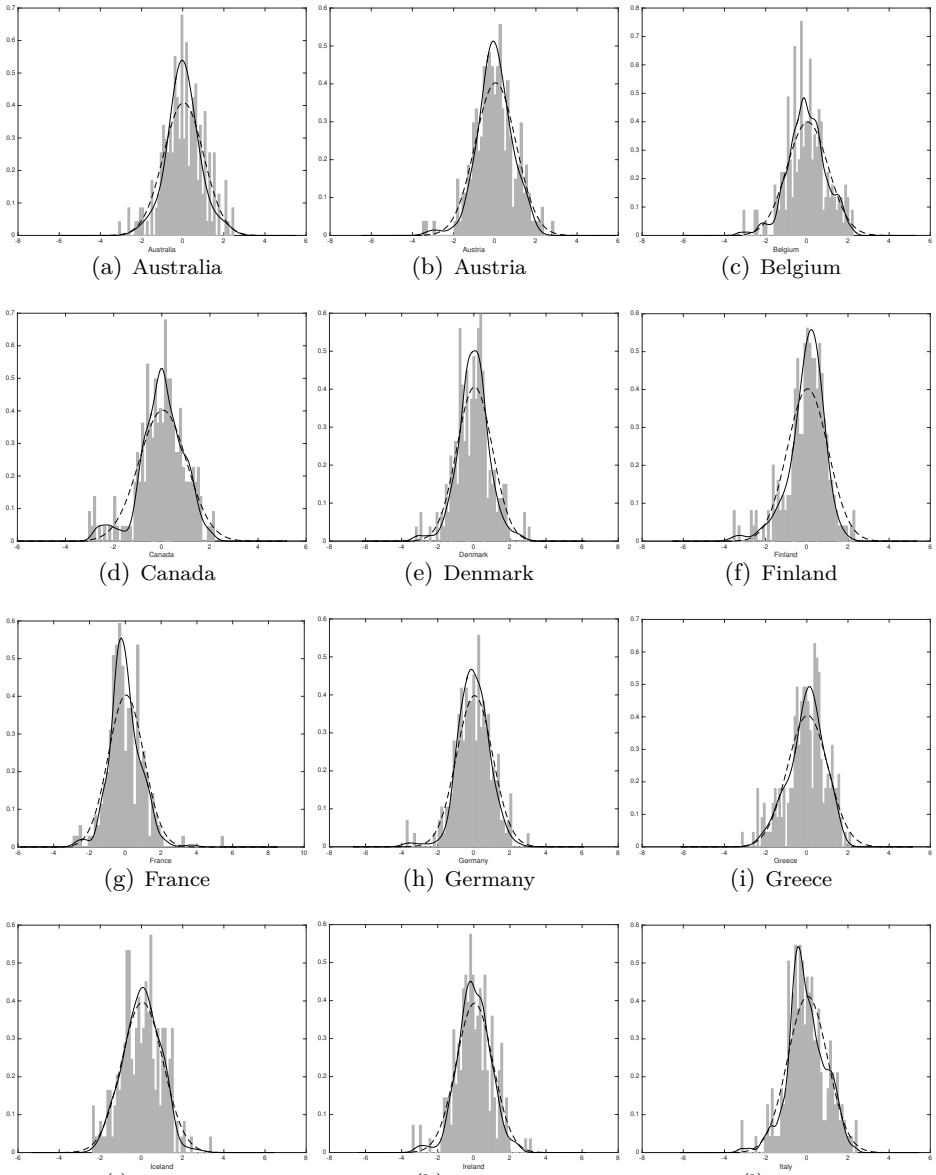
(a) m=80 with random numbers (b) m=80 with block matrices

FIGUREA.3: Posterior mean of the matrix of δ for $m = 80$ with random element (left) and with block matrix (right).



(a) m=80 with random numbers (b) m=80 with block matrices

FIGUREA.4: Weighted network for $m = 80$ with random elements in the B matrix (left) and with block matrix (right), where the blue edges mean negative weights and red ones represent positive weights.



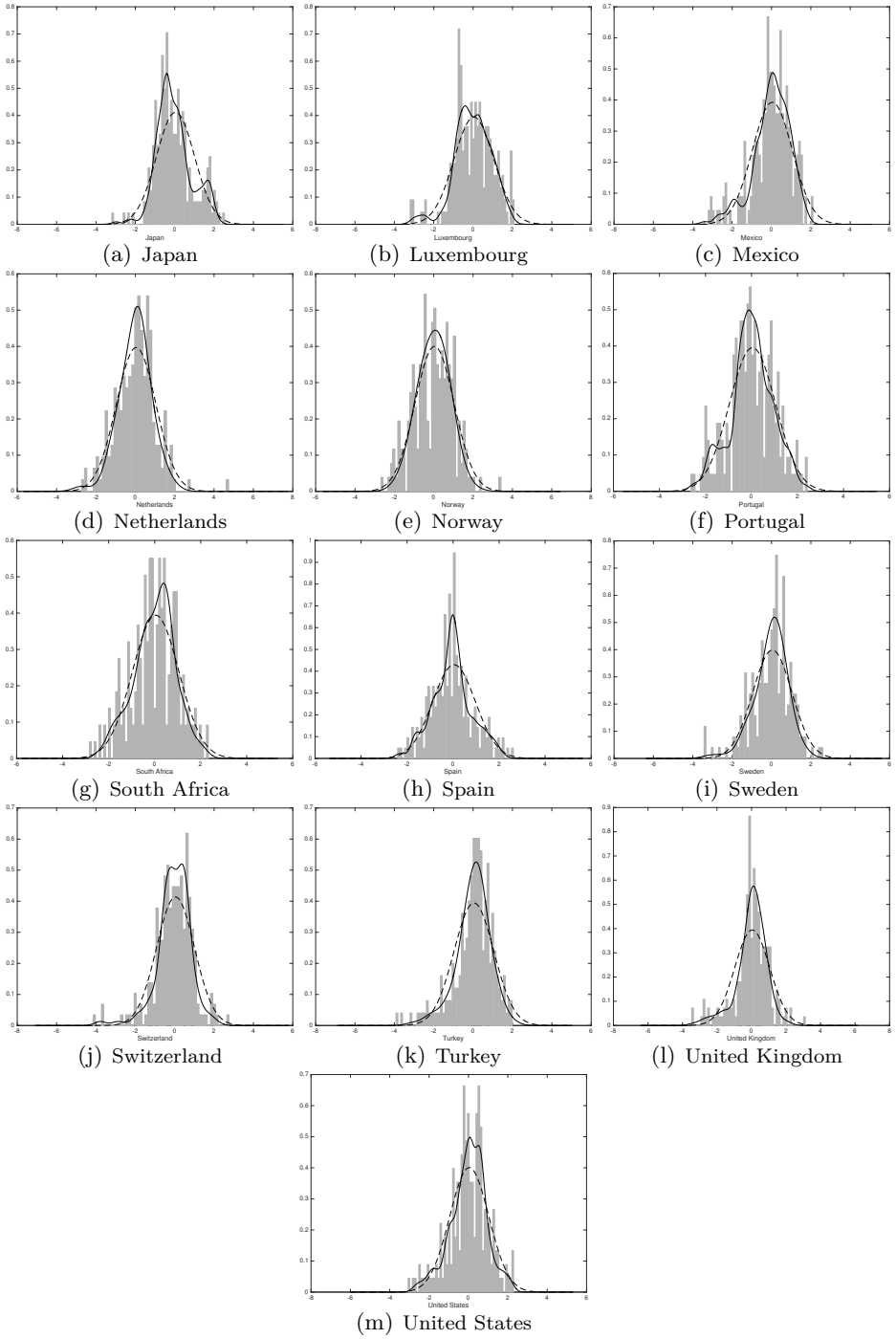
(a) Australia (b) Austria (c) Belgium

(d) Canada (e) Denmark (f) Finland

(g) France (h) Germany (i) Greece

(j) Iceland (k) Ireland (l) Italy

FIGUREA.5: GDP growth rates Y_{it} (histogram), predictive distribution (solid line) and best normal (dashed line) for all the countries of the panel.



(a) Japan (b) Luxembourg (c) Mexico

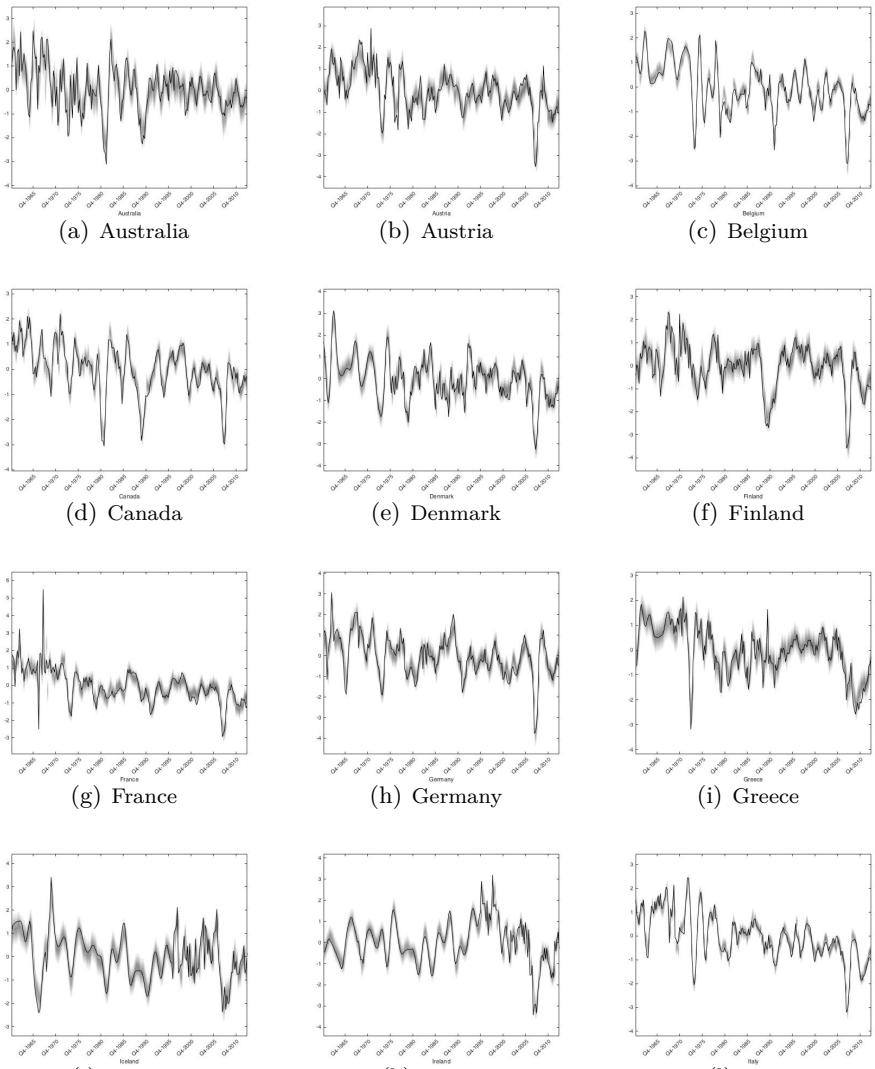
(d) Netherlands (e) Norway (f) Portugal

(g) South Africa (h) Spain (i) Sweden

(j) Switzerland (k) Turkey (l) United Kingdom

(m) United States

FIGUREA.6: GDP growth rates Y_{it} (histogram), predictive distribution (solid line) and best normal (dashed line) for all the countries of the panel.



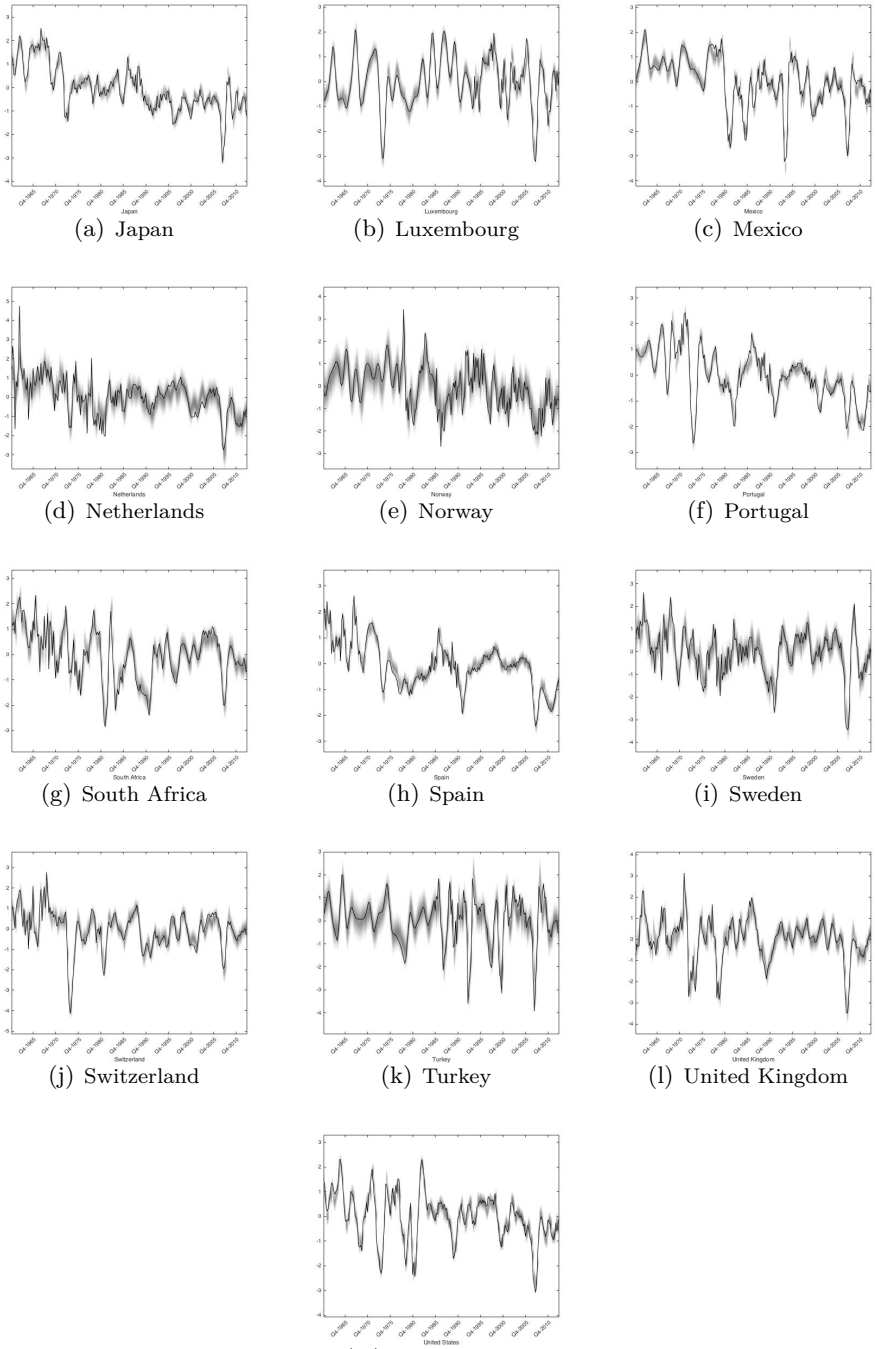
(a) Australia (b) Austria (c) Belgium

(d) Canada (e) Denmark (f) Finland

(g) France (h) Germany (i) Greece

(j) Iceland (k) Ireland (l) Italy

FIGUREA.7: Predictive results for all countries. In each plot: GDP growth rates Y_{it} (black lines); heatmap (grey areas) of the 95% high probability density region of the predictive density functions (darker colors represent higher density values) evaluated at each time point, for $t = 1, \ldots, T$ at the value of the predictors $Y_{it-1}, \ldots, Y_{it-p}$ for $i = 1, \ldots, 25$.



(a) Japan (b) Luxembourg (c) Mexico

(d) Netherlands (e) Norway (f) Portugal

(g) South Africa (h) Spain (i) Sweden

(j) Switzerland (k) Turkey (l) United Kingdom

(m) United States

FIGUREA.8: Predictive results for all countries. In each plot: GDP growth rates Y_{it} (black lines); heatmap (grey areas) of the 95% high probability density region of the predictive density functions (darker colors represent higher density values) evaluated at each time point, for $t = 1, \ldots, T$ at the value of the predictors $Y_{it-1}, \ldots, Y_{it-p}$ for $i = 1, \ldots, 25$.

Chapter 2

Bayesian Nonparametric Conditional Copula Estimation of Twin Data

Abstract. Several studies on heritability in twins aim at understanding the different contribution of environmental and genetic factors to specific traits. Considering the National Merit Twin Study, our purpose is to correctly analyse the influence of the socioeconomic status on the relationship between twins' cognitive abilities. Our methodology is based on conditional copulas, which allow us to model the effect of a covariate driving the strength of dependence between the main variables. We propose a flexible Bayesian nonparametric approach for the estimation of conditional copulas, which can model any conditional copula density. Our methodology extends the work of Wu, Wang, and Walker (2015) by introducing dependence from a covariate in an infinite mixture model. Our results suggest that environmental factors are more influential in families with lower socio-economic position.

Keywords: Bayesian nonparametrics, Conditional Copula models, Slice sampling.

This chapter is based on: Dalla Valle, L., Leisen, F. and Rossini, L. (2016). *"Bayesian Nonparametric Conditional Copula Estimation of Twin Data"*, Working Papers N. 08/WP/2016, Dept. of Economics, Ca' Foscari University of Venice. Working paper available at http://arxiv.org/abs/1603.03484.

2.1 Introduction

The literature on heritability of traits in children often focusses on twins, due to the shared environmental factors and the association of genetical characteristics. Among studies on the heritability of diseases, Wang et al. (2011) applied an efficient estimation method to mixed-effect models to analyze disease inheritance in twins.

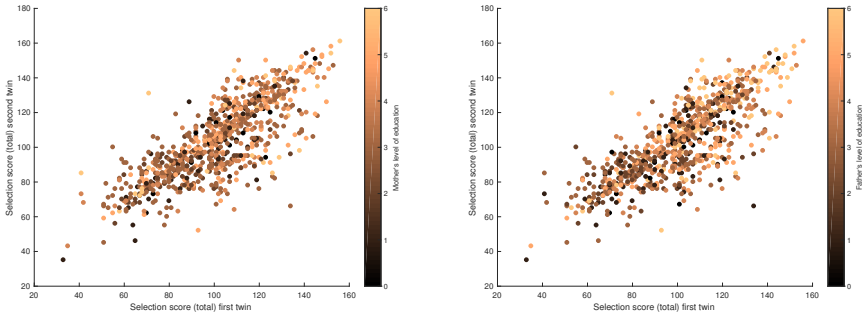
One of the main purposes of studies on heritability is to estimate the different contribution of genetic and environmental factors to traits or outcomes (see, for example, the latent class twin method of Baker (2016)). Bates et al. (2013) studied the interactions between environmental and genetic effects to intelligence in twins, showing that higher socioeconomic status is associated with higher intelligence scores. Bioecological theory states that environmental factors may significantly influence the heritability of certain characteristics, such as cognitive ability, which is the readiness for future intellectual or educational pursuits. Several studies have found that cognitive ability is more pronounced and evident among children raised in higher socioeconomic status families. Such families can offer greater opportunities to children, due to their socioeconomic wealth status, and represent stimulating environments where children's inherited capabilities may become more manifest.

The aim of this paper is to correctly analyse the effect of socioeconomic factors on the relationship between twins' cognitive abilities. From a sample of 839 US adolescent twin pairs who completed the National Merit Scholarship Qualifying Test, we consider each twin's overall school performance (measured by a total score including English, Mathematics, Social Science, Natural Science and Word Usage), the mother's and father's education level and the family income. The data are plotted in Figure 2.1, which shows the scatterplots of the twins' school performances, on each axis, against the socioeconomic variables,

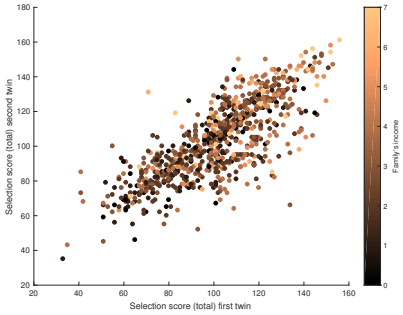
whose values are in different colours (dark brown denotes low values, while light rose denotes high values). Figure 2.1 indicates that the twins' school performances are positively correlated and their dependence is influenced by the values of the socioeconomic variables (the mother's (panel (a)), the father's level of education (panel (b)) and the family income (panel (c))). Indeed, most of the light rose dots (denoting high values of the covariates) are grouped in the upper right corner, while the dark brown dots (denoting low values of the covariates) lie in the bottom left corner of each plot. Hence, the higher the parents' education or family income, the higher the twins' school performance. This means that the association between the twins' performance scores is a function of each covariate and it varies according to the values of the covariates.

In Figure 2.2 we selected only data corresponding to the minimum and maximum value of each covariate and we produced the scatterplots of the twins' school performance scores. The left plots correspond to the minimum value of each covariate, while the right plots correspond to the maximum value of each covariate. The top plots refer to the mother's level of education, the central plots refer to the father's level of education and the bottom plots refer to the family income. In all three cases we notice that, as already pointed out, low values of covariates correspond to low performance scores, while high values of covariates correspond to high performance scores. In addition, the scatterplot points corresponding to low covariate values (left plots) tend to lie closely to the diagonal, while the points corresponding to high covariate values (right plots) tend to be more spread around the diagonal. This suggests that the school outcomes of children belonging to less affluent families are generally more similar to each other, while the outcomes of children belonging to privileged families tend to be more different to each other.

In order to model the dependence structure between the twins' school performances, we used copulas, which are popular modeling approaches in multivariate statistics allowing the



(a) Mother's education (b) Father's education



(c) Family income

FIGURE2.1: Scatterplots of the twins overall scores with respect to the mother's (panel (a)) and father's level of education (panel (b)) and family income (panel (c)).

separation of the marginal components of a joint distribution from its dependence structure. More precisely, Sklar (1959) proved that a d-dimensional distribution H of the random variables $Y_1, \ldots, Y_d$ can be fully described by its marginal distributions and a function $C : [0,1]^d \to [0,1]$, called copula, through the relation $H(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d))$, where H is the joint cumulative density function. In the literature, copulas have been applied to model the dependence between variables in a wide variety of fields (see Kolev, dos Anjos, and Vaz de Mendes (2006) and Cherubini, Luciano, and Vecchiato (2004)). In particular, applications of copula models involved lifetime data analysis (Andersen (2005)), survival analysis of Atlantic halibut (Braekers and Veraverbeke (2005)) and transfusion-

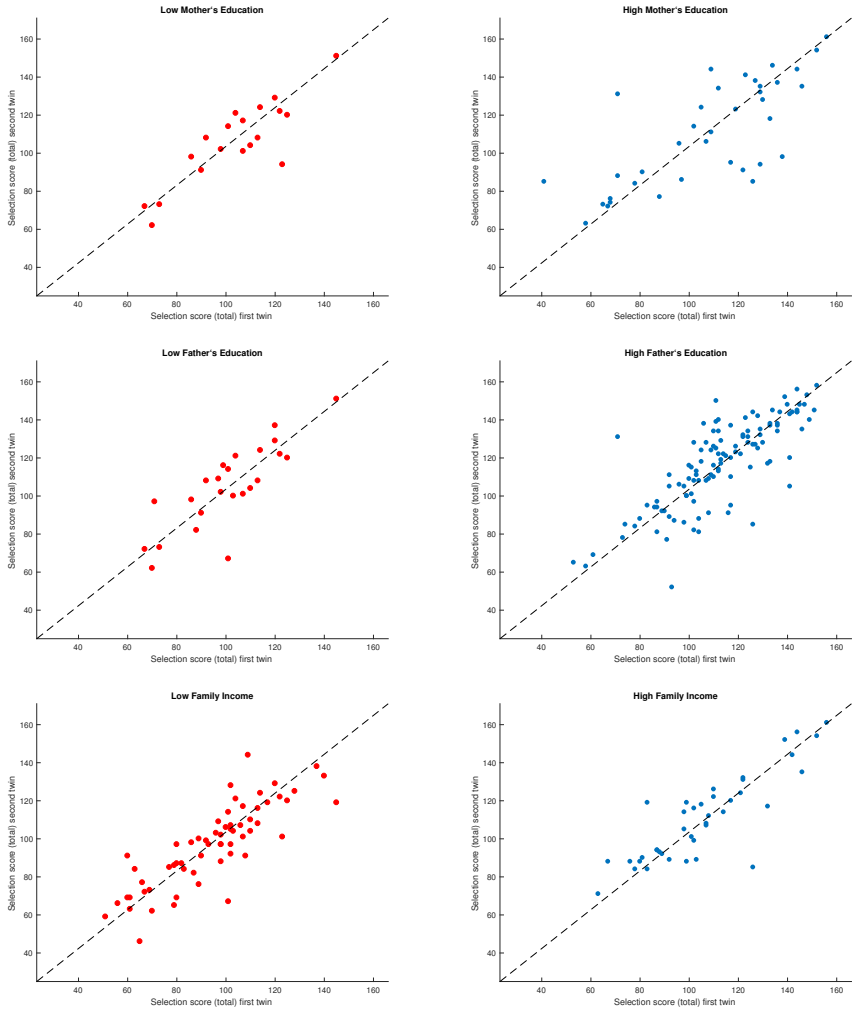


FIGURE2.2: Scatterplots of the twins overall scores with respect to the mother's (top panel) and father's level of education (middle panel) and family income (bottom panel). Each panel shows on the left (right) the points corresponding to the minimum (maximum) value of the covariate. The black line corresponds to the 45 degrees diagonal.

related AIDS and cancer analysis (Emura and Wang (2012), Huang and Zhang (2008) and Owzar, Jung, and Sen (2007)).

The introduction of covariate adjustments to copulas has attracted an increased interest in recent years, since it allows the dependence structure to be explained by a specific covari-

ate. Craiu and Sabeti (2012) propose a conditional copula approach in regression settings where the bivariate outcome can be mixed or continuous. Patton (2006) introduce time-variation in the dependence structure of ARMA models (see also Jondeau and Rockinger (2006) and Bartram, Taylor, and Wang (2007) for other applications of time-series analysis to dependence modelling). The paper of Acar, Craiu, and Yao (2010) provides a nonparametric procedure to estimate the functional relationship between copula parameters and covariates, showing that the gestational age drives the strength of dependence between the birth weights of twins. Abegaz, Gijbels, and Veraverbeke (2012) and Gijbels, Omelka, and Veraverbeke (2012) propose semiparametric and nonparametric methodologies for the estimation of conditional copulas, establishing consistency and asymptotic normality results for the estimators. The methodology is then applied to examine the influence of the gross domestic product (GDP), in USD per capita, on the life expectancies of males and females at birth. Following this literature, we adopt a conditional copula approach to model the effect of a covariate, such as the parents' education or the family income, on the strength of dependence between twins' school performances.

The literature offers a rich range of copula families, such as elliptical copulas (e.g. Gaussian and Student's t) and archimedean copulas (e.g. Frank, Gumbel, Clayton and Joe copulas) to accommodate various dependence structures. Nonetheless, the choice of the copula family may be controversial and it is still an open problem (see Joe (2014)). To overcome this issues, Wu et al. (2015) propose a Bayesian nonparametric procedure to estimate any unconditional copula density function. The authors combine the well-known Gaussian copula density with the modeling flexibility of the Bayesian nonparametric approach, proposing to use an infinite mixture of Gaussian copulas. Our paper extends the work of Wu et al. (2015) to the conditional copula setting, by proposing a novel methodology which combines the advantages of a conditional copula approach with the modeling

flexibility of Bayesian nonparametrics. In particular, we included a conditional covariate component to explain the variables dependence structure, allowing us further flexibility to the copula density modelling. Up to our knowledge, this is the first Bayesian nonparametric proposal in the conditional copulas literature.

The outline of the chapter is the following. In Section 2.2 we briefly review the literature about conditional copulas and Bayesian nonparametric copula estimation. In Section 2.3 we introduce our novel Bayesian nonparametric conditional copula setting. Section 2.4 provides an algorithm for estimating the posterior parameters and Section 2.5 illustrates the performance of the methodology. Section 2.6 is devoted to the application of our methodology to the analysis of the National Merit Twin Study. Concluding remarks are given in section 2.7.

2.2 Preliminaries

In this Section, we review some preliminary notions about conditional copulas and illustrate the Bayesian nonparametric copula density estimation introduced in Wu et al. (2015). In what follows, we focus on the bivariate case for simplicity, however the arguments can be easily extended to more than two dimensions.

2.2.1 Copula and Sklar's Theorem

Copulas are particular functions which account for dependence between multivariate data. Sklar (1959) introduces the idea of copula for separating the joint distribution function $H(y_1, y_2)$ into two parts respectively. The first part describes the dependence structure of the distribution, while the second one describes the marginal distribution functions F_i, for $i = 1, 2$.

Definition 2.2.1. *Let $Y = (Y_1, Y_2)$ be a random vector with distribution function H and*

with marginal distribution functions F_i, $Y_i \sim F_i$, $i = 1, 2$. *A distribution function* C *with uniform marginals on* $[0, 1]$ *is called "copula" of* Y *if:*

$$H = C(F_1, F_2).$$

If the marginal distributions are continuous and $F_i(Y_i) \sim \mathcal{U}(0, 1)$ then C is a copula and we have the following representation:

$$C(u_1, u_2) = P(F_1(Y_1) \leqslant u_1, F_2(Y_2) \leqslant u_2) = P(Y_1 \leqslant F_1^{-1}(u_1), Y_2 \leqslant F_2^{-1}(u_2)) =$$

$$= H_Y(F_1^{-1}(u_1), F_2^{-1}(u_2)),$$

where $F_i^{-1}(t) = \inf\{x \in \mathbb{R} : F_i(y) \geqslant t\}$ denotes the generalized inverse of F_i and a copula C follows from the expression:

$$H(y_1, y_2) = P(Y_1 \leqslant y_1, Y_2 \leqslant y_2) = P(F_1(Y_1) \leqslant F_1(y_1), F_2(Y_2) \leqslant F_2(y_2)) =$$

$$= C(F_1(y_1), F_2(y_2)).$$

Definition 2.2.2. *A copula* $C : [0, 1]^2 \to [0, 1]$ *has the following properties:*

1. C *is grounded, i.e. for every* $\boldsymbol{u} = (u_1, u_2) \in [0, 1]^2$, $C(\boldsymbol{u}) = 0$ *if at least one coordinate* $u_i = 0$, $i = 1, 2$;

2. C *is* $2-$*increasing, i.e. for every* $\boldsymbol{u} \in [0, 1]^2$ *and* $\boldsymbol{v} \in [0, 1]^2$ *such that* $\boldsymbol{u} \leqslant \boldsymbol{v}$, *the C-volume* $V_C([\boldsymbol{u}, \boldsymbol{v}])$ *of the box* $[\boldsymbol{u}, \boldsymbol{v}]$ *is non-negative;*

3. $C(1, u_2) = u_2$ *and* $C(u_1, 1) = u_1$ *for all* $u_i \in [0, 1]^2$.

After the definition of the properties of a copula, we can explain the main result regarding the theory of copula (Sklar (1959)).

Theorem 2.2.1 (Sklar's Theorem)**.** *Let* H *be the joint distribution function with marginals* F_1, F_2. *Then there exists a copula* C *such that for all* $(y_1, y_2) \in [-\infty, \infty]^2$,

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2)).$$

If the marginals are all continuous, then C is unique.

Proof. Let $Y = (Y_1, Y_2)$ be a random vector on a probability space $(\Omega, \mathcal{A}, P)$ with distribution H and let $V \sim \mathcal{U}(0,1)$ be independent of Y. Considering the distirbutional transforms $U_i = F_i(Y_i, V)$, we have that $U_i \stackrel{d}{=} \mathcal{U}(0,1)$ and $Y_i = F_i^{-1}(U_i)$ almost surely, $i = 1, 2$. Thus defining C to be the distribution of $U = (U_1, U_2)$ we obtain:

$$\begin{aligned} H(y_1, y_2) &= P(Y \leqslant \mathbf{y}) = P(F_i^{-1}(U_i) \leqslant y_i, i = 1, 2) \\ &= P(U_i \leqslant F_i(y_i), i = 1, 2) = C(F_1(y_1), F_2(y_2)), \end{aligned}$$

and we can conclude that C is a copula. □

When $H(\cdot)$ and $C(\cdot)$ are differentiable, the equation

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2)),$$

for the joint cumulative distribution function implies that the joint probability density function satisfies:

$$\frac{h(y_1, y_2)}{f_1(y_1) f_2(y_2)} = c[F_1(y_1), F_2(y_2)],$$

where $c(\cdot)$ is the probability density function of the copula distribution:

$$c(u_1, u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2).$$

In the above part of the section we have discussed the properties of a copula, while in the following we will analyse different classes of copula that are of interest for our analysis: the Archimedean and the Gaussian copulas. The latter one is a class of copula, which allows for a variety of different dependence structures and it is fully described in Nelsen (2006) and in Genest and MacKay (1986).

Definition 2.2.3. *Let $\varphi : [0,1] \to [0,\infty)$ be a continuous, strictly decreasing function such that $\varphi(1) = 0$. The pseudo-inverse of φ is the function $\varphi^{[-1]} : [0,\infty) \to [0,1]$ given by:*

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leqslant t \leqslant \varphi(0), \\ 0, & \varphi(0) \leqslant t \leqslant \infty. \end{cases}$$

Note that $\varphi^{[-1]}$ is continuous and decreasing on $[0,\infty)$ and strictly decreasing on $[0,\varphi(0)]$. Hence, $\varphi^{[-1]}(\varphi(u)) = u$ on $[0,1]$ and

$$\varphi(\varphi^{[-1]}(t)) = \begin{cases} t, & 0 \leqslant t \leqslant \varphi(0), \\ \varphi(0), & \varphi(0) \leqslant t \leqslant \infty. \end{cases}$$

If we assume $\varphi(0) = \infty$, then we have that $\varphi^{[-1]} = \varphi^{-1}$. The following theorem will introduce the archimedean copulas and the proof can be found in Nelsen (2006).

Theorem 2.2.2. *Let $\varphi : [0,1] \to [0,\infty)$ be a continuous, strictly decreasing function such that $\varphi(1) = 0$ and let $\varphi^{[-1]}$ be the pseudo-inverse of φ. Let $C : [0,1]^2 \to [0,1]$ be the function given by:*

$$C(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)). \tag{2.1}$$

Then C is a copula if and only if φ is convex.

The copula defined in (2.1) is called Archimedean copula, where the function φ is called the generator of the copula and it has the following properties:

Theorem 2.2.3. *Let C be an Archimedean copula with generator φ. Then*

1. *C is symmetric, i.e. $C(u,v) = C(v,u)$ for all $u, v \in [0,1]$.*

2. *C is associative, i.e. $C(C(u,v),w) = C(u,C(v,w))$ for all $u, v, w \in [0,1]$.*

In the last part of this section we describe briefly the two families of copula of our interest for the rest of the paper, the Clayton and the Frank copula and the Gaussian copula.

Example 1 (Clayton copula). *Let* $\varphi(t) = (t^{-\theta} - 1)/\theta$, *where* $\theta \in [-1, \infty)\backslash 0$, *then the Clayton copula is:*

$$C_\theta(u, v) = \max\left(\left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}, 0\right).$$

For values of $\theta > 0$ *the copula is strictly positive and the previous definition can be rewritten as:*

$$C_\theta(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}.$$

Example 2 (Frank Copula). *Let* $\phi(t) = -\ln \frac{e^{-\theta t}-1}{e^{-\theta}-1}$, *where* $\theta \in \mathbb{R}\backslash 0$, *then the Frank copula is given by:*

$$C_\theta(u, v) = -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right).$$

On the other hand, the last example of this section describes the Gaussian copula.

Example 3 (Gaussian Copula). *Let* $\Phi_\rho(y_1, y_2)$ *denote the standard bivariate normal distribution function of the form:*

$$\Phi_\rho(y_1, y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{2(1-\rho^2)}\right],$$

where $\rho \in (-1, 1)$ *is the correlation coefficient. Then the Gaussian copula is given by:*

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)).$$

Furthermore, in the next section, we will describe the conditional copula and their estimation.

2.2.2 The conditional copula

Let Y_1 and Y_2 be continuous variables of interest and X be a covariate that may affect the dependence between Y_1 and Y_2. Following Gijbels et al. (2012), Abegaz et al. (2012) and

Acar et al. (2010), we suppose that the conditional distribution of (Y_1, Y_2) given $X = x$ exists and we denote the corresponding conditional joint distribution function by

$$H_x(y_1, y_2) = P(Y_1 \leqslant y_1, Y_2 \leqslant y_2 | X = x).$$

If the marginals of H_x, denoted as

$$F_{1x}(y_1) = P(Y_1 \leqslant y_1 | X = x), \qquad F_{2x}(y_2) = P(Y_2 \leqslant y_2 | X = x),$$

are continuous, then according to Sklar's theorem there exists a unique copula C_x which equals

$$C_x(u, v) = H_x(F_{1x}^{-1}(u), F_{2x}^{-1}(v)), \tag{2.2}$$

where $F_{1x}^{-1}(u) = \inf\{y_1 : F_{1x} \geqslant u\}$ and $F_{2x}^{-1}(v) = \inf\{y_2 : F_{2x} \geqslant v\}$, are the conditional quantile functions and $u = F_{1x}(y_1)$ and $v = F_{2x}(y_2)$ are called pseudo-observations. The conditional copula C_x fully describes the conditional dependence structure of (Y_1, Y_2) given $X = x$. An alternative expression for (2.2) is

$$H_x(y_1, y_2) = C_x(F_{1x}(y_1), F_{2x}(y_2)). \tag{2.3}$$

2.2.3 Bayesian nonparametric copula density estimation

Let $\Phi_\rho(y_1, y_2)$ denote the standard bivariate normal distribution function with correlation coefficient ρ. Then, C_ρ is the copula corresponding to Φ_ρ, taking the form:

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \tag{2.4}$$

where Φ is the univariate standard normal distribution function. The Gaussian copula density is:

$$c_\rho(u, v) = |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\Phi^{-1}(u), \Phi^{-1}(v))(\Sigma^{-1} - \mathbf{I}) \begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{pmatrix}\right\} \tag{2.5}$$

where the correlation matrix is:

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Wu et al. (2015) proposed to use an infinite mixture of Gaussian copulas for the estimation of a copula density, as follows

$$c(u, v) = \sum_{j=1}^{\infty} w_j c_{\rho_j}(u, v) \tag{2.6}$$

where the weights w_j's sum up to 1 and the ρ_j's vary in $(-1, 1)$. Given a set of n observations $(u_1, v_1), \ldots, (u_n, v_n)$, their model can be described through a hierarchical specification, i.e.

$$\begin{aligned} (u_i, v_i) \mid \rho_i &\overset{\text{ind}}{\sim} c_{\rho_i}(u_i, v_i), \qquad i = 1, \ldots, n, \\ \rho_i \mid G &\overset{\text{iid}}{\sim} G, \\ G &\sim DP(\lambda, G_0), \end{aligned} \tag{2.7}$$

where G is a Dirichlet Process prior with total mass λ and base measure G_0. This proposal is motivated by the fact that bivariate density functions on the real plain can be arbitrarily well approximated by a mixture of a countably infinite number of bivariate normal distributions of the form

$$f(y_1, y_2) = \sum_{j=1}^{\infty} w_j N((y_1, y_2)|(\mu_{1j}, \mu_{2j}), \Sigma_j)$$

where $N((y_1, y_2)|(\mu_{1j}, \mu_{2j}), \Sigma_j)$ is the joint bivariate normal density with mean vector (μ_{1j}, μ_{2j}) and correlation matrix Σ_j (see Lo (1984) and Ferguson (1983)). Roughly speaking, the authors are mimicking the Dirichlet process mixture model in the copula setting (see Escobar (1994) and Escobar and West (1995)). The sampling strategy follows the

slice sampler of Walker (2007) and Kalli et al. (2011). The authors show that the Gaussian mixture is flexible enough to accurately approximate any bivariate copula density.

2.3 Conditional copula estimation with Dirichlet process priors

The data object of study requires a model which can take into account the effect of the covariate. We build on the model introduced by Wu et al. (2015) and illustrated in the previous section. The idea is to replace the Gaussian copula with a conditional version where the correlation is a function of the covariate, i.e.

$$c_\rho(u, v|x) = c_{\rho(x)}(u, v).$$

The function $\rho(x)$ can be modelled as preferred, for instance, with a generalized linear model or with a non-linear function. In any case, we have that $\rho(x)$ will depend on a vector of parameters $\boldsymbol{\beta}$, so that

$$c_{\rho(x)}(u, v) = c_{\rho(x|\boldsymbol{\beta})}(u, v).$$

We assume a Dirichlet process prior on the vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$. Following the model description provided in equation (2.7), we can summarize our model as follows,

$$\begin{aligned} (u_i, v_i) \mid \rho(x_i|\boldsymbol{\beta}_i) &\overset{\text{ind}}{\sim} c_{\rho(x_i|\boldsymbol{\beta}_i)}(u_i, v_i), \qquad i = 1, \ldots, n, \\ \boldsymbol{\beta}_i \mid G &\overset{\text{iid}}{\sim} G, \\ G &\sim DP(\lambda, G_0), \end{aligned} \tag{2.8}$$

where G is a Dirichlet process prior with total mass λ and base measure G_0. As in Wu et al. (2015), our model can be described as an infinite mixture of Normal distributions,

$$c_\rho(u, v|x) = \sum_{j=1}^{\infty} w_j c_{\rho(x|\boldsymbol{\beta}_j)}(u, v),$$

and hence suitable for implementing a slice sampling algorithm, as explained in the next section.

In order to model the function $\rho(x|\boldsymbol{\beta})$, we would like to follow some standard approaches in the literature. Abegaz et al. (2012) model the dependence of the parameter of interest, with respect to the covariate, through a *calibration function* $\theta(x|\boldsymbol{\beta})$. It is important to highlight that in many copula families the parameter space is restricted. In contrast, the calibration function $\theta(x|\boldsymbol{\beta})$ can assume any value on the real line. In our case, our parameter is restricted to the interval $(-1, 1)$ and we need a transformation which can link the calibration function $\theta(x|\boldsymbol{\beta})$ to $\rho(x|\boldsymbol{\beta})$. In this paper, we adopt the following transformation,

$$\rho(x|\boldsymbol{\beta}) = \frac{2}{|\theta(x|\boldsymbol{\beta})| + 1} - 1.$$

In our simulated and real data examples we focus on two particular calibration functions studied in the literature, which are

$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x^2$$

$$\theta(x|\boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 \exp\left(-\beta_4 x^2\right)$$

respectively, such that $\theta(x|\boldsymbol{\beta}) \in (-\infty, +\infty)$ and, consequently, $\rho(x|\boldsymbol{\beta}) \in (-1, 1)$.

2.4 Posterior sampling algorithm

Suppose that, given the observations (y_{1i}, y_{2i}), for $i = 1, \ldots, n$, the corresponding pseudo-observations (u_i, v_i) are calculated using a nonparametric rank-based estimation approach, where $u_i = r_{1i}/(n+1)$ and $v_i = r_{2i}/(n+1)$, for $i = 1, \ldots, n$, where r_{ki}, for $k = 1, 2$, denotes the rank of y_{ki} among all y_{kh}, with $h \in 1, ..., n$.

Following equation (2.6), given (u_i, v_i) for $i = 1, \ldots, n$, and the conditional variable x_i, the conditional copula density function for each pair (u_i, v_i) can be written as an infinite

mixture of conditional Gaussian copulas, such that:

$$c(u_i, v_i|x_i) = \sum_{j=1}^{\infty} w_j c_{\rho(x_i|\boldsymbol{\beta}_j)}(u_i, v_i) \tag{2.9}$$

where w_j's are the stick-breaking weights, i.e.

$$w_j = \pi_j \prod_{l=1}^{j-1}(1 - \pi_l)$$

where the π_j are distributed as a $\mathcal{B}e(1, \lambda)$, $\lambda > 0$. In order to sample from the infinite mixture displayed in equation (2.9), we use the slice sampling algorithm for mixture models proposed by Walker (2007) and Kalli et al. (2011). To reduce the dimensionality of the problem, the authors introduce a latent variable z_i for each i which allows us to write the infinite mixture model as follows:

$$c(u_i, v_i, z_i|x_i) = \sum_{j=1}^{\infty} \mathbb{I}(z_i < w_j) c_{\rho(x_i|\boldsymbol{\beta}_j)}(u_i, v_i). \tag{2.10}$$

The introduction of the slice variable z_i reduces the sampling complexity to the analogous of a finite mixture model. In particular, letting

$$A_w = \{j : z_i < w_j\}, \tag{2.11}$$

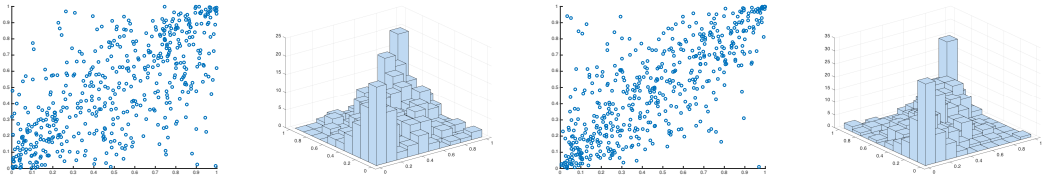
then it can be proved that the cardinality of the set A_w is almost surely finite. Consequently, there is a finite number of parameters to be estimated. By iterating the data augmentation principle further, we introduce another latent variable d_i, which is called allocation variable, allowing us to allocate each observation to one component of the mixture model. Then, the conditional copula density $c(u_i, v_i, z_i, d_i|x_i)$ takes the form:

$$c(u_i, v_i, z_i, d_i|x_i) = \mathbb{I}(z_i < w_{d_i}) c_{\rho(x_i|\boldsymbol{\beta}_{d_i})}(u_i, v_i) \tag{2.12}$$

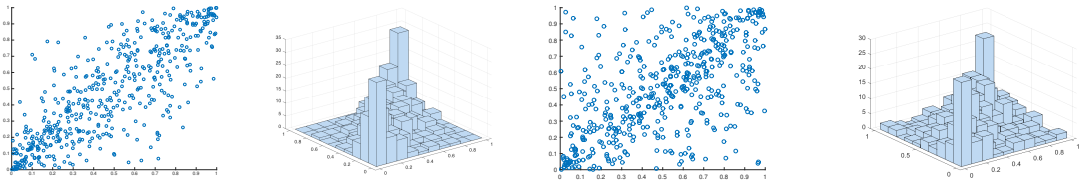
where $d_i \in \{1, 2, \dots\}$. Hence, the full likelihood function of the conditional copula model is:

$$\prod_{i=1}^{n} c(u_i, v_i, z_i, d_i | x_i) = \prod_{i=1}^{n} \mathbb{I}(z_i < w_{d_i}) c_{\rho(x_i|\boldsymbol{\beta}_{d_i})}(u_i, v_i). \tag{2.13}$$

We use the notation $(U, V) = \{i = 1, \dots, n : (u_i, v_i)\}$, $X = \{x_1, \dots, x_n\}$ to describe the pseudo-observations and the covariate values, respectively. We denote with $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots\}$ the vector of parameters and $D = \{d_1, \dots, d_n\}$, $Z = \{z_1, \dots, z_n\}$ and $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots\}$ the new variables that we have introduced in this Section.



(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.



(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGURE2.3: Gaussian copula with sample size $n = 500$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

Therefore, we used a Gibbs sampler to simulate iteratively from the posterior distribution function, according to the following steps:

1. The stick-breaking components $\boldsymbol{\pi}$ are updated given $[Z, D, \boldsymbol{\beta}, (U, V), X]$;

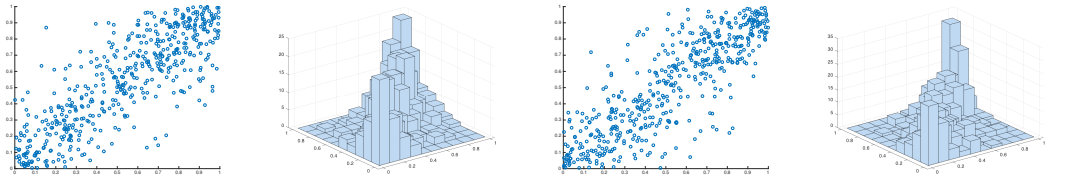
2. The latent slice variables Z are updated given $[\boldsymbol{\pi}, D, \boldsymbol{\beta}, (U, V), X]$;

3. The allocation variables D are updated given $[\boldsymbol{\pi}, Z, \boldsymbol{\beta}, (U, V), X]$;

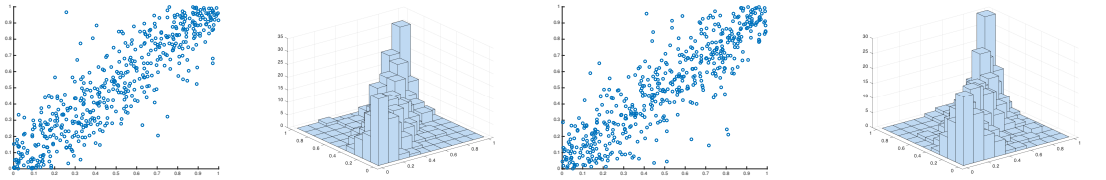
4. The vector of parameters $\boldsymbol{\beta}$ are updated given $[\boldsymbol{\pi}, Z, D, (U, V), X]$.

The Gibbs sampling details are explained in Appendix B.1.

Note that, once the marginals and the conditional copula are estimated according to the approach described above, the conditional joint distribution function $H_x(y_1, y_2)$ can be easily obtained from expression 2.3.



(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.



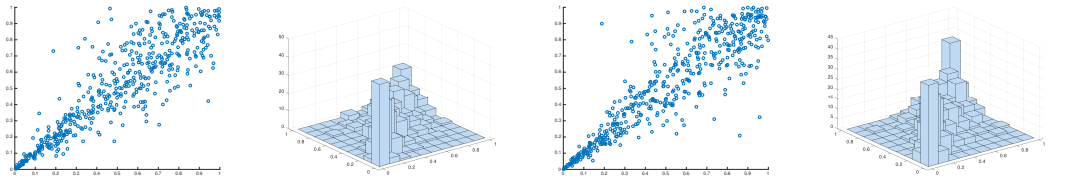
(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGURE2.4: Frank copula with sample size $n = 500$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

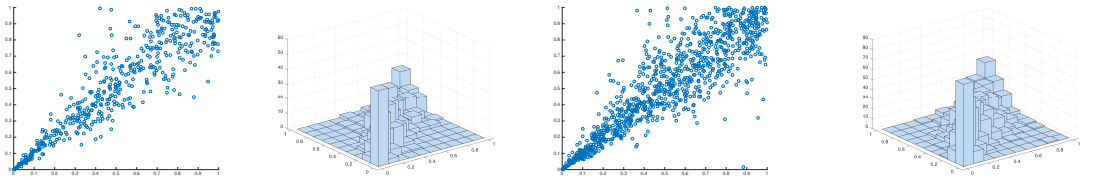
2.5 Simulation experiments

This section illustrates the performance of our Bayesian nonparametric conditional copula model with simulated data. We generate datasets (U, V) of sizes $n = 250, 500$ and 1000 from elliptical and archimedean copula families, such as the Gaussian, Frank and Double Clayton copula, which combines the regular Clayton copula with its 90◦ rotation, allowing positive

and negative dependence modelling. The copula dependence parameter is considered as a function of the exogenous variable X, which is simulated from a Uniform distribution in the interval $[-2, 2]$ and the base measure G_0 is a multivariate Normal distribution with a vector of zeros as mean and a variance-covariance matrix $\sigma^2 \cdot \mathbb{I}$ and σ^2 big enough.



(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.



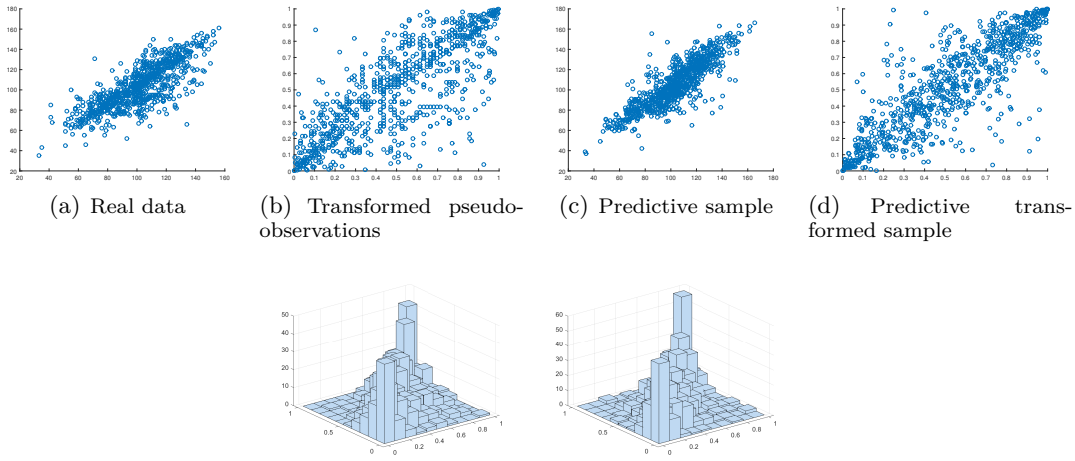
(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGURE2.5: Double Clayton copula with sample size $n = 500$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

We run the Gibbs sampler algorithm described in Section 2.4 for 4000 iterations with (i) 500 burn-in iterations and (ii) 3500 burn-in iterations. Aiming at a parsimonious representation of the results, we focussed on 3500 burn-in iterations, since 500 burn-in iterations gave very similar results.

Figures B.1, 2.3 and B.2 illustrate the results of the application of the Bayesian non-parametric conditional copula model to data simulated from a gaussian copula, with sample sizes $n = 250, 500$ and 1000, respectively. Figures B.3, 2.4 and B.4 illustrate similar results for the Frank copula; while Figures B.5, 2.5 and B.6 illustrate analogous results for the Double Clayton copula. In Figures 2-14, panels (a), (b), (c) and (d) show the scatter plots

and histograms of the simulated data and the predictive samples, respectively, obtained using the first calibration function; while panels (e), (f), (g) and (h) show the scatter plots and histograms of the simulated data and the predictive sample, respectively, obtained using the second calibration function. The comparison between the simulated and predictive outputs highlights the good fit of the Bayesian nonparametric conditional copula model using either calibration function and with different sample sizes. The model performance appears to be consistent across all three copula families, demonstrating that the approach is suitable to model different dependence patterns and tail structures.



(a) Real data (b) Transformed pseudo-observations (c) Predictive sample (d) Predictive transformed sample

(e) Histogram of real data (f) Predictive histogram

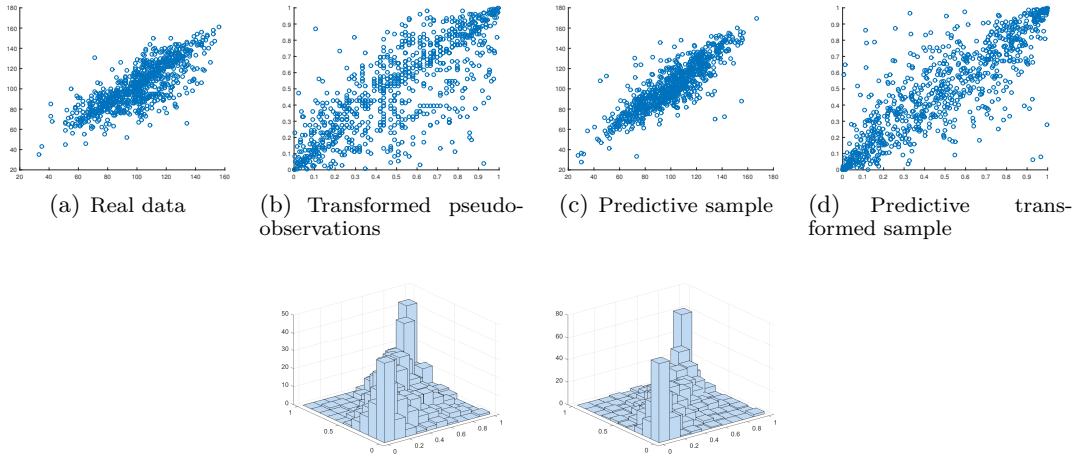
FIGURE2.6: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect to the mother's level of education; panels (c) and (d): scatterplots of the predictive and transformed predictive sample; panels (e) and (f): histograms of the real data and the predictive sample.

2.6 Real Data applications

We now apply the proposed Bayesian nonparametric conditional copula method to a sample of 839 adolescent twin pairs, which is a subset of the National Merit Twin Study (Loehlin and Nichols, 2009, 2014). The dataset contains questionnaire data from 17 years old twins

and their parents, where the twins were identified among 600.000 US high school juniors who took part to the National Merit Scholarship Qualifying Test (NMSQT). The NMSQT was designed to measure cognitive aptitude, that is students' readiness for future intellectual or educational pursuits. The participants to the test include identical twins and same-sex fraternal twins who were asked to fill in a complete questionnaire in order to understand their school performance and attitude. Our purpose is to examine whether the relationship between twins' cognitive ability, measured by the NMSQT, is influenced by their socioeconomic status, measured by parent education and parental income. The variables we considered from this study are the overall measures of each twin's performance at school (obtained as the sum of individual scores in English Usage, Mathematics Usage, Social Science Reading, Natural Science Reading and Word Usage/Vocabulary), the mother's and father's level of education and the family income. The overall scores range from 30 to 160, the education covariates range from 0 to 6, while the family income covariate ranges from 0 to 7. The levels of the education covariates correspond to: less than 8-th grade, 8-th grade, part high school, high school graduate, part college or junior college, college graduate, and graduate or professional degree beyond the bachelor's degree. The levels of the income covariate correspond to values going from less than $5000 per year to over than $25000 per year.

As discussed in Section 3.1, the scatterplots in Figure 2.1 clearly show that there is a positive correlation between the twins' school performance and the strength of dependence varies according to the values of a covariate, which is the mother's (panel (a)) or father's level of education (panel (b)) or the family income (panel (c)). In Figure 2.1 the effect of the covariates is illustrated by dots of different colours, where we notice that most of the light rose dots are grouped in the upper right corner, while the dark brown dots lie in the bottom left corner. Therefore, the higher the parents' education or family income,



(a) Real data (b) Transformed pseudo-observations (c) Predictive sample (d) Predictive transformed sample

(e) Histogram of real data (f) Predictive histogram

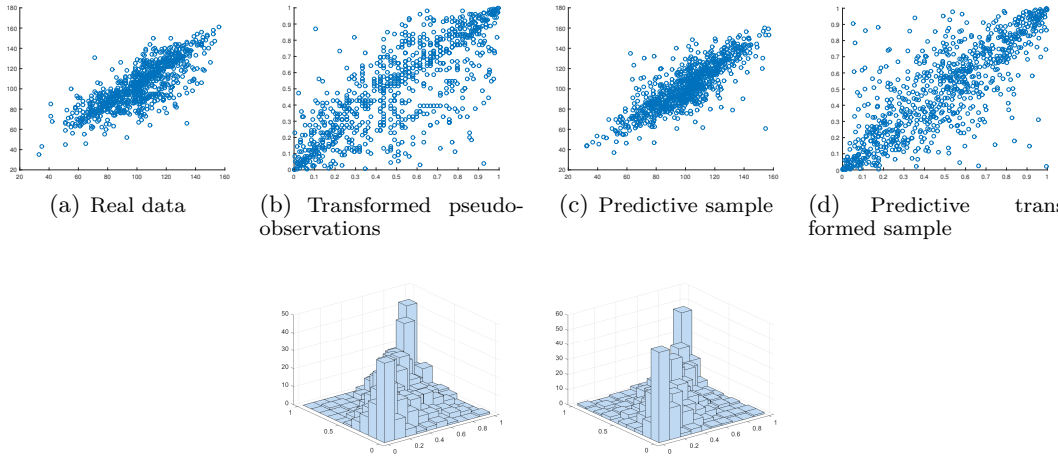
FIGURE2.7: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect to the father's level of education; panels (c) and (d): scatterplots of the predictive and transformed predictive sample; panels (e) and (f): histograms of the real data and the predictive sample.

the higher the twins' school performance. In order to model the effect of a covariate, such as the mother's and father's education and family income, on the dependence between the overall scores of the twins, we implement the Bayesian nonparametric conditional copula model.

Note that, with a different dataset, the methodology may be extended to include more than one covariate. However, model specification issues and increased computational costs must be carefully considered.

Note that the pseudo-observations are obtained using the nonparametric rank-based approach described in Section 2.4.

Adopting the same priors of the simulation studies, we run the Gibbs sampling algorithm described in Section 2.4 for 4000 iterations. Figures 2.6, 2.7 and 2.8 show, with respect to the mother's and father's education and family income, respectively, the scatterplots of the twins' overall scores using the real and transformed pseudo-observations



(a) Real data (b) Transformed pseudo-observations (c) Predictive sample (d) Predictive transformed sample

(e) Histogram of real data (f) Predictive histogram

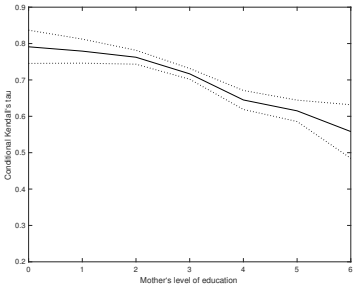
FIGURE2.8: Panels (a) and (b): scatterplots of the twins' overall scores for the real and pseudo-observations with respect to the family income; panels (c) and (d): scatterplots of the predictive and transformed predictive sample; panels (e) and (f): histograms of the real data and the predictive sample.

(panels (a) and (b)), the scatterplots of the predictive and transformed predictive samples (panels (c) and (d)) and the histograms of the real and the predictive samples (panels (e) and (f)). From the comparison between the scatterplots and histograms of the real and predictive samples obtained with the three different covariates, it emerges that the Bayesian nonparametrics conditional copula model accurately captures the tail structures and the dependence patterns between the twins' overall scores. We note that the good performance of this approach in tail modelling makes it suitable to various applications focussing on extremes. Figure 2.9 shows the conditional Kendall's tau[1] estimated from the model against the mother's (top panel) and father's level of education (middle panel) and

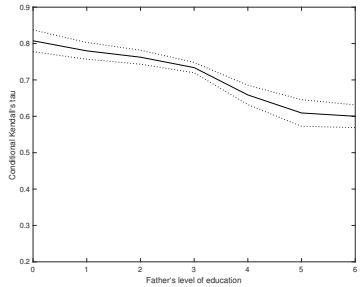
[1] The conditional Kendall's tau of (Y_1, Y_2) given $X = x$ is a nonparametric measures of correlation between two ranked variables (Y_1, Y_2) with respect to a covariate $X = x$ and has the following form:

$$\tau(x) = 2P((Y_1 - Y_1')(Y_2 - Y_2') > 0 | X = X' = x) - 1 = 4 \int\int C_x(u_1, u_2) dC_x(u_1, u_2) - 1,$$

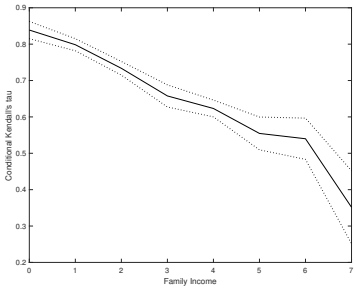
where C_x is the appropriate conditional copula and (Y_1', Y_2', X') is an independent copy of the random vector (Y_1, Y_2, X).



(a) Mother's level of education



(b) Father's level of education



(c) Family income

FIGURE2.9: Estimated Kendall's tau against the mother's (top panel) and father's level of education (middle panel) and the family income (bottom panel) and an approximate 95% confidence interval (dotted lines).

the family income (bottom panel). The plots clearly illustrate the negative effect of all three covariates on the dependence between the twins' overall scores. The effect is greater for the family income, where the Kendall's tau decreases from approximately 0.83 to 0.33, while for the parents' education levels the Kendall's tau decreases from approximately 0.8 to 0.6. Therefore, the higher the parents' education and family income, the better the socioeconomic status and the higher the differences between the twins' school performances. The cognitive aptitudes of twins from less advantaged families are more similar to each other than those from high income, highly educated families. Families of high socioeconomic status represent supportive and challenging environments, able to offer a wide range of opportunities and choices to their children, and allowing them to express themselves

freely. Hence, twins raised in wealthy families are encouraged to develop differences in their traits, and may show rather dissimilar cognitive abilities, albeit high on average.

On the contrary, families of low socioeconomic status offer scarce opportunities to their children and may represent limiting and restrictive environments. In less advantaged families, twins cannot develop their full potential and differences and both tend to show low cognitive abilities.

This might suggest, as in Loehlin et al. (2009), an interaction between genetic and environmental factors. Genes multiply environmental inputs that support intellectual growth such that an increased socioeconomic status raises the average cognitive ability but also magnifies individual differences in cognitive ability (see Bates et al. (2013)).

2.7 Conclusion

In this paper we proposed a Bayesian nonparametric conditional copula approach to model the strength and type of dependence between two variables of interest and we applied the methodology to the National Merit Twin Study. In order to capture the dependence structure between two variables, we introduced two different calibration functions expressing the functional form of a covariate variable. The statistical inference was obtained implementing a slice sampling algorithm, assuming an infinite mixture model for the copula. The methodology combines the advantages of the conditional copula approach with the modeling flexibility of Bayesian nonparametrics.

The simulation studies illustrated the good performance of our model with three distinct copula families and different sample sizes. The application to the twins data revealed the importance of the environment in the development of twins belonging to low socioeconomic classes and suggest that socioeconomic factors are more influential in families with lower social levels.

Although this paper focusses on bivariate copula models, the methodology can be extended to multivariate copulas including more than one covariate. However, the inclusion of multiple covariates needs special attention regarding the choice of variables prior to estimate the calibration functions. Moreover, the increasing computational cost due to the additional covariates should be taken carefully into consideration.

We are currently working on extending the framework of the conditional copula approach presented here by considering the hierarchical/nested copulas (see Segers and Uyttendaele (2014)) in order to allow a vast class of Archimedean copulas and study the tree copulas structure.

Acknowledgements

We would like to thank all the conference partecipants for helpful discussion at: "PGR Seminar" at the University of Kent; "10th Annual RCEA Bayesian Econometric Workshop" at Rimini Center for Economic Analysis; "3rd Bayesian Young Statistician Meeting" at University of Florence; "9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016)" at University of Seville.

Bibliography

Abegaz, F., Gijbels, I., and Veraverbeke, N. (2012), "Semiparametric estimation of conditional copulas," *Journal of Multivariate Analysis*, 110, 43–73.

Acar, E. F., Craiu, R. V., and Yao, F. (2010), "Dependence calibratin in conditional copulas: a nonparametric approach," *Biometrics*, 67, 445–453.

Andersen, E. (2005), "Two-stage estimation in copula models used in family studies," *Lifetime Data Analysis*, 11, 333–350.

Baker, S. (2016), "The latent class twin method," *Biometrics*, doi: 10.1111/biom.12460.

Bartram, S., Taylor, S., and Wang, Y. (2007), "The Euro and European financial market dependence," *Journal of Banking and Finance*, 31, 1461–1481.

Bates, T., Lewis, G., and Weiss, A. (2013), "Childhood Socioeconomic Status Amplifies Genetic Effects on Adult Intelligence," *Psychological Science*, 24, 2111–2116.

Braekers, R. and Veraverbeke, N. (2005), "A copula-graphic estimator for the conditional survival function under dependent censoring," *Canadian Journal of Statistics*, 33, 429–447.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004), "Copula methods in finance," *John Wiley and Sons*.

Craiu, R. V. and Sabeti, A. (2012), "In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes," *Journal of Multivariate Analysis*, 110, 106–120.

Emura, T. and Wang, W. (2012), "Nonparametric maximum likelihood estimation for dependent truncation data based on copulas," *Journal of Multivariate Analysis*, 110, 171–188.

Escobar, M. D. (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, vol. 66, Chapman & Hall.

Ferguson, T. (1983), "Bayesian density estimation by mixtures of normal distributions," in *Recent Advances in Statistics*, eds. H. Rizvi, J. Rustagi, and D. Siegmund, pp. 287–302, New York: Academic Press.

Genest, C. and MacKay, R. (1986), "Copules archimediennes et familles de lois bidimensionell dont les marges sont donnees," *The Canadian Journal of Statistics*, 14, 145–159.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012), "Multivariate and functional covariates and conditional copulas," *Electronic Journal of Statistics*, 6, 1273–1306.

Huang, X. and Zhang, N. (2008), "Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach," *Biometrics*, 64, 1090–1099.

Joe, H. (2014), *Dependence Modeling with Copulas*, Chapman & Hall.

Jondeau, E. and Rockinger, M. (2006), "The copula-GARCH model of conditional dependencies: An international stock market application," *Journal of International Money and Finance*, 25, 827–853.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Kolev, N., dos Anjos, U., and Vaz de Mendes, B. (2006), "Copulas: a review and recent developments," *Stochastic Models*, 22, 617–660.

Lo, A. (1984), "On a class of Bayesian nonparametric estimates I: density estimates," *Annals of Statistics*, 12, 351–357.

Loehlin, J. and Nichols, R. (2009), "The National Merit twin study," *Harvard Dataverse, V3, http://hdl.handle.net/1902.1/13913*.

Loehlin, J. and Nichols, R. (2014), "Heredity, Environment and Personality: A study of 850 sets of twins," *University of Texas Press*.

Loehlin, J., Harden, K., and Turkheimer, E. (2009), "The Effect of Assumptions about Parental Assortative Mating and Genotype–Income Correlation on Estimates of Genotype–Environment Interaction in the National Merit Twin study," *Behavior Genetics*, 39, 165–169.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer Series in Statistics.

Owzar, K., Jung, S.-H., and Sen, P. K. (2007), "A copula approach for detecting prognostic genes associated with survival outcome in microarray studies," *Biometrics*, 63, 1089–1098.

Papaspiliopoulous, O. and Roberts, G. O. (2008), "Retrospective Markov Chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Patton, A. J. (2006), "Modelling asymmetric exchange rate dependence," *International Economic Review*, 47, 527–556.

Segers, J. and Uyttendaele, N. (2014), "Nonparametric estimation of the tree structure of a nested Archimedean copula," *Computational Statistics and Data Analysis*, 72, 190–204.

Sklar, A. (1959), "Fonctions de réparation à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

Wang, X., Guo, X., He, M., and Zhang, H. (2011), "Statistical Inference in Mixed Models and Analysis of Twin and Family Data," *Biometrics*, 67, 987–995.

Wu, J., Wang, X., and Walker, S. (2015), "Bayesian nonparametric estimation of a copula," *Journal of Statistical Computation and Simulation*, 85, 103–116.

Appendix B

Technical Details of Chapter 2

B.1 Gibbs sampling details

Let $\mathcal{D}_j = \{i = 1, \ldots, n : d_i = j\}$ be the set of indexes of the observations allocated to the j-th component of the mixture, while $\mathcal{D} = \{j : \mathcal{D}_j \neq \varnothing\}$ is the set of indexes of non-empty mixtures components. Let $D^* = \sup\{\mathcal{D}\}$ be the number of stick-breaking components used in the mixture. As in Kalli et al. (2011), the sampling of infinite elements of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ is not necessary, since only the elements of the full conditional probability density functions of D are need.

The maximum number of stick-breaking components to be sampled is:

$$N^* = \max\{i = 1, \ldots, n | N_i^*\},$$

where N_i^* is the smallest integer such that $\sum_{j=1}^{N_i^*} w_j > 1 - z_i$.

B.1.1 Update of π

We update the stick-breaking components and consequently the weights w_j based on the equation $w_j = \pi_j \prod_{k<j}(1 - \pi_k)$. Assuming that π_j is distributed as a Beta ($\mathcal{B}e(1, \lambda)$), the

full conditional distribution of π_j is:

$$\pi_j|\cdots \sim \mathcal{B}e(1+\#\{d_i=j\}, \lambda+\#\{d_i>j\}), \tag{B.1}$$

where $\#\{d_i=j\}$ are the number of d_i equal to j and $\#\{d_i>j\}$ is the number of d_i greater than j for $j<D^*$.

On the other hand, if $j=D^*+1,\ldots,N^*$ we have that

$$\pi_j|\cdots \sim \mathcal{B}e(1,\lambda).$$

B.1.2 Update of Z

From the full likelihood function (2.13), z_i follows a uniform distribution

$$z_i|\cdots \sim \mathcal{U}(0, w_{d_i}) \tag{B.2}$$

and it is sampled accordingly.

B.1.3 Update of D

The allocation variable d_i values lie between 0 and N_i and the density of d_i satisfies

$$P(d_i=j|\ldots) \propto \mathbb{I}(z_i<w_{d_i}) c_{\rho(x_i|\boldsymbol{\beta}_{d_i})}(u_i,v_i). \tag{B.3}$$

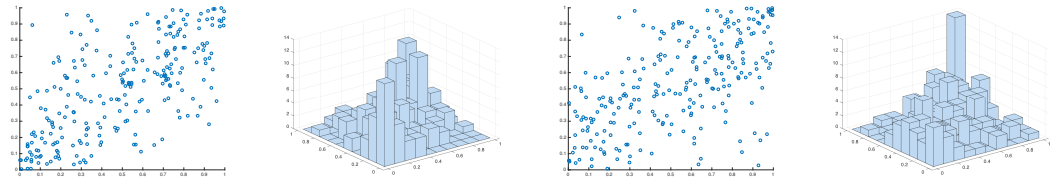
B.1.4 Update of β

The full conditional of the vector of parameters $\boldsymbol{\beta}_k$, for $k \geqslant 1$ is:

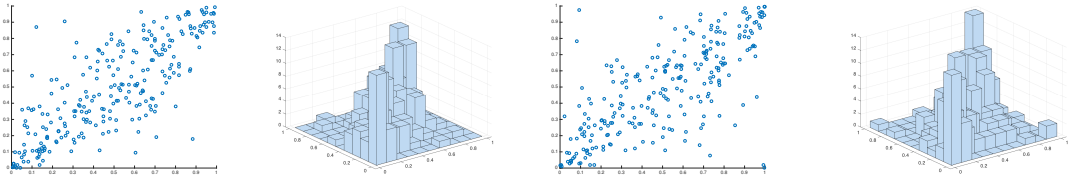
$$f(\boldsymbol{\beta}_k|\ldots) \propto \pi(\boldsymbol{\beta}_k) \prod_{d_i=k} c_{\rho(x_i|\boldsymbol{\beta}_k)}(u_i,v_i), \tag{B.4}$$

where $\pi(\boldsymbol{\beta}_k)$ is the prior on $\boldsymbol{\beta}$. Since the (B.4) is not a standard distribution, we used a Random Walk Metropolis Hastings.

B.2 Graphical part of the simulated examples

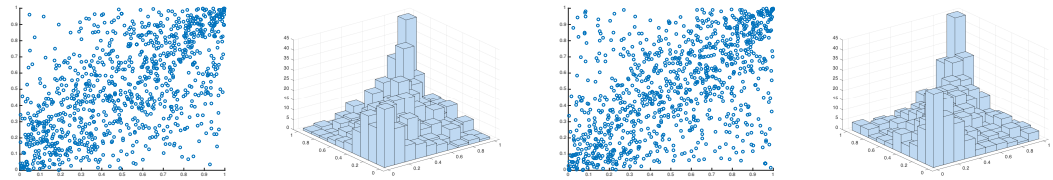


(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.

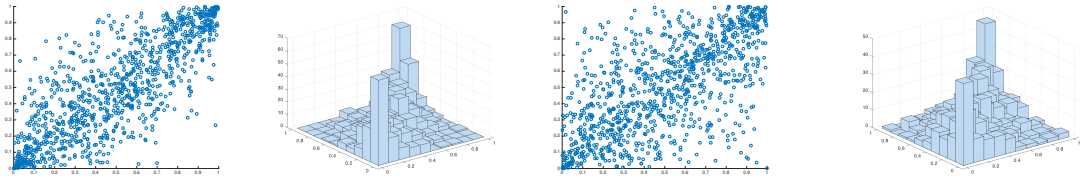


(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal. fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.1: Gaussian copula with sample size $n = 250$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

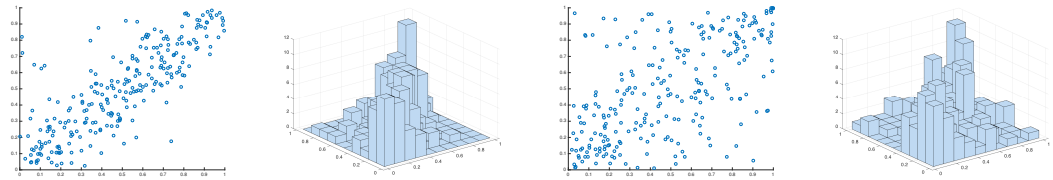


(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.

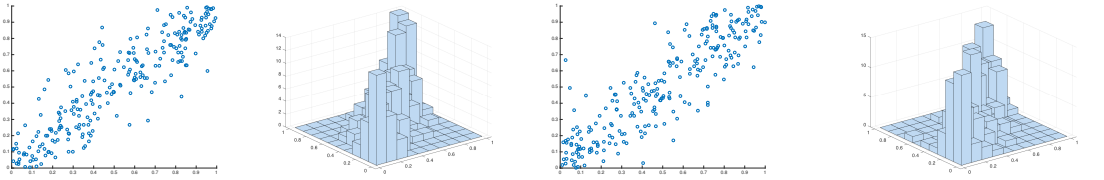


(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.2: Gaussian copula with sample size $n = 1000$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

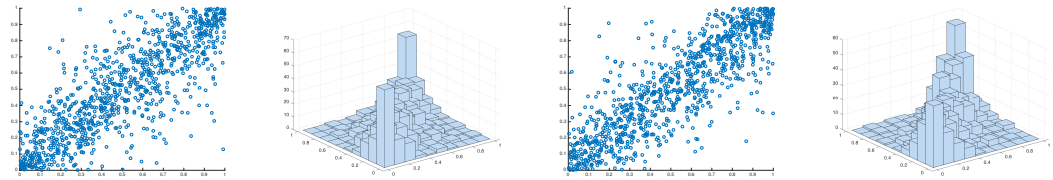


(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.

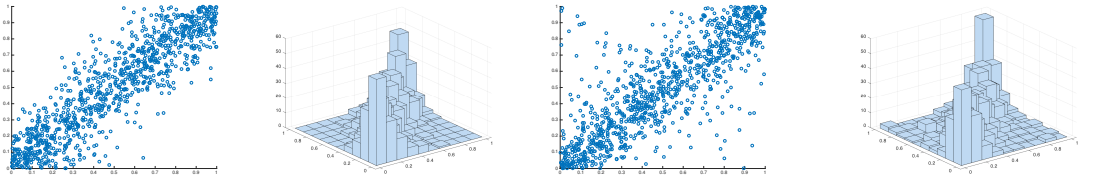


(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.3: Frank copula with sample size $n = 250$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

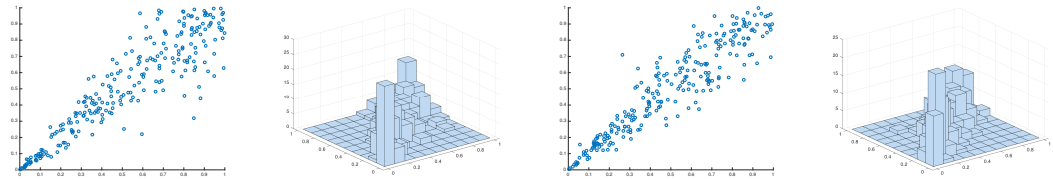


(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.

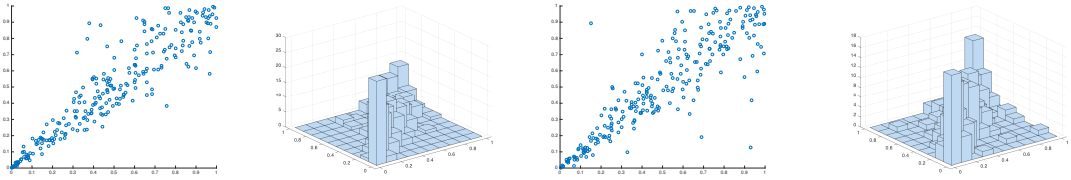


(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.4: Frank copula with sample size $n = 1000$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

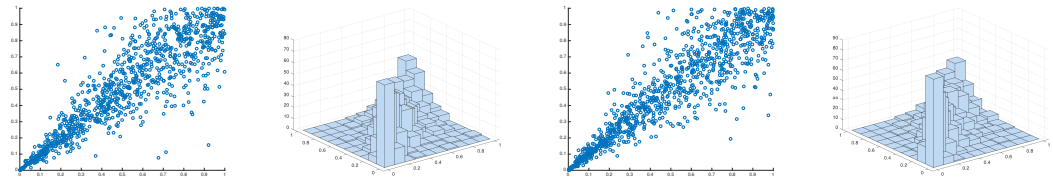


(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.

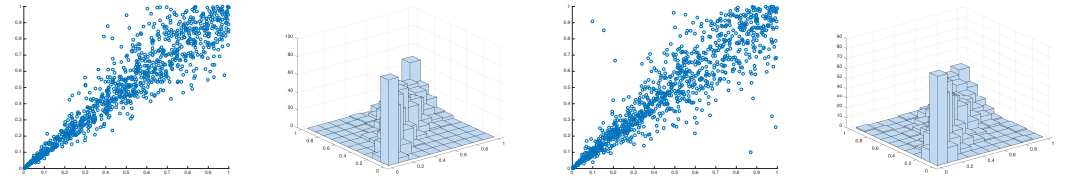


(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.5: Double Clayton copula with sample size $n = 250$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.



(a) Simulated, 1^{st} cal.fun. (b) Simulated, 1^{st} cal.fun. (c) Predictive, 1^{st} cal.fun. (d) Predictive, 1^{st} cal.fun.



(e) Simulated, 2^{nd} cal.fun. (f) Simulated, 2^{nd} cal.fun. (g) Predictive, 2^{nd} cal. fun. (h) Predictive, 2^{nd} cal. fun.

FIGUREB.6: Double Clayton copula with sample size $n = 1000$. Panels (a), (b), (c) and (d) depict the scatter plots and histograms, obtained with the first calibration function, of the simulated and predictive samples, respectively; panels (e), (f), (g) and (h) depict the scatter plots and histograms, obtained with the second calibration function, of the simulated and predictive sample, respectively.

Appendix C

Technical Details of Chapter 1 and of Chapter 2

C.1 Slice Sampling Representation

Walker (2007) and Kalli et al. (2011) proposed a new algorithm, called slice sampling, for sampling the mixture of Dirichlet process model. First of all, let us define the mixture of Dirichlet process model with Gaussian kernel as follow:

$$f_P(y) = \int \mathcal{N}(y|\mu, \sigma^2) dP(\phi)$$

where $P \sim \mathrm{DP}(M, P_0)$ means P follows a Dirichlet Process with concentration parameter M and based measure P_0 and $\phi = (\mu, \sigma^2)$ is the parameter of interest, where μ is the mean and σ^2 is the variance of the Normal distribution. The Dirichlet process, P has a stick-breaking representation as:

$$P = \sum_{j=1}^{\infty} w_j \delta_{\phi_j}$$

where δ_{ϕ_j} is the delta of Dirac with a point mass of 1 at ϕ_j and $\phi_1, \phi_2, \ldots$ are independent and identically distributed from P_0 and the weights follow from the stick-breaking

representation:

$$w_1 = v_1, \quad w_j = v_j \prod_{l<j}(1 - v_l)$$

where the variables $(v_1, v_2, \dots)$ are independent and identically distributed as a $\mathcal{B}e(1, M)$. Given the form of P, we can write

$$f_{v,\mu,\sigma^2}(y) = \sum_{j=1}^{\infty} w_j \mathcal{N}(y|\mu_j, \sigma_j^2)$$

and we need to find the finite number of variables need to be sampled to produce a stationary Markov chain. Therefore we introduce a latent variable u such that:

$$f_{v,\mu,\sigma^2}(y, u) = \sum_{j=1}^{\infty} \mathbb{I}(u < w_j)\mathcal{N}(y|\mu_j, \sigma_j^2) = \sum_{j=1}^{\infty} w_j \mathcal{U}(u|u, w_j)\mathcal{N}(y|\mu_j, \sigma_j^2) \tag{C.1}$$

and with probability w_j, we have that u and y are independent and are uniform and normal distributed. We introduce the set $A_w(u) = \{j : w_j > u\}$, which is a finite set for all $u > 0$, then we can reformulate (C.1) as:

$$f_{v,\mu,\sigma^2}(y, u) = \sum_{j \in A_w(u)} \mathcal{N}(y|\mu_j, \sigma_j^2)$$

The introduction of the latent variable u allows us to have a finite mixture model and we can introduce a new latent variable d, which will identify the component of the mixture from which y is to be taken. Hence the joint density has the following form:

$$f_{v,\mu,\sigma^2}(y, u, d) = \mathbb{I}(u < w_d)\mathcal{N}(y|\mu_d, \sigma_d^2)$$

Finally the joint posterior distribution is proportional to

$$\prod_{i=1}^{n} \mathbb{I}(u_i < w_{d_i})\mathcal{N}(y_i|\mu_{d_i}, \sigma_{d_i}^2). \tag{C.2}$$

The previous slice sampling mix often slowly due to the correlation between u and w and during the update of the latent variable u and of the finite set $A_w(u)$ can lead to simulation of more w's. For solving these problems, the positive sequence $\xi_1, \xi_2, \ldots$ is introduced and

$$f_{v,\mu,\sigma^2}(y, u, d) = \xi_d^{-1} \mathbb{I}(u < \xi_d) w_d \mathcal{N}(y|\mu_d, \sigma_d^2)$$

The choice of ξ depends on the rate at which the ratio $r_i = \mathbb{E}[w_i]/\xi_i$ increases with i. In fact, faster rates of increase are associated with better mixing but longer running times. In the Gibbs sampler, we need to update the following variables:

$$\{(\mu_j, \sigma_j^2, v_j), j = 1, 2, \ldots; (d_i, u_i), i = 1, \ldots, n\} \tag{C.3}$$

If the variables ξ and v are conditionally independent then the step of the Gibbs sampler are the following:

- $\pi(\mu_j, \sigma_j^2| \cdots) \propto p_0(\mu_j, \sigma_j^2) \prod_{d_i=j} \mathcal{N}(y_i|\mu_j, \sigma_j^2)$
- $\pi(v_j) \propto \mathcal{B}e(v_j|a_j, b_j)$ where $a_j = 1 + \sum_{i=1}^n \mathbb{I}(d_i = j)$ and $b_j = M + \sum_{i=1}^n \mathbb{I}(d_i > j)$;
- $\pi(u_i| \cdots) \propto \mathbb{I}(0 < u_i < \xi_{d_i})$;
- $P(d_i = k| \cdots) \propto \mathbb{I}(k|\xi_k > u_i) w_k/\xi_k \mathcal{N}(y_i|\mu_k, \sigma_k^2)$

The variables (μ_j, σ_j^2, v_j) need to be sample up to the integer $N = \max_i \{N_i\}$, where N_i is the largest integer l for which $\xi_l > u_i$. On the other hand, the retrospective sampler (Papaspiliopoulous and Roberts (2008)) is an alternative conditional methods which defines a Markov chain with the correct posterior of the infinite dimensional model. The principal difference is in the update of the allocation variable d_i, in fact in the slice sampling d_i is finite at each iteration of the Gibbs sampler, while in the retrospective sampler the value of d_i is computed through a Metropolis-Hastings update. This Metropolis-Hasting update

involves the potential simulation of extra variables n times per iteration, while the slice sampling only generates extra variables once per iteration.

Chapter 3

The Yule–Simon Distribution: an Objective Bayesian Analysis and a Posterior Inference

Abstract. The Yule–Simon distribution is usually employed in the analysis of frequency data. As the Bayesian literature, so far, ignored this distribution, here we show the derivation of two objective priors for the parameter of the Yule–Simon distribution and an explicit Gibbs sampling scheme when a Gamma prior is chosen for the shape parameter. In particular, we discuss the Jeffreys prior and a loss-based prior, which has recently appeared in the literature. We illustrate the performance of the derived priors and of the proposed algorithm through simulation studies and the analysis of real datasets.

Keywords: Kullback-Leibler divergence, Loss-based prior, Objective Bayes, Social Network daily returns, Text Analysis, Data Augmentation.

This chapter is based on:

- Leisen, F., Rossini, L. and Villa, C. (2016). *"Objective Bayesian Analysis of the Yule–Simon Distribution with Applications"*. Working paper available at http://arxiv.org/abs/1604.05661;
- Leisen, F., Rossini, L. and Villa, C. (2016). *"A Note on the Posterior Inference for the Yule–Simon Distribution"*. Forthcoming in the Journal of Statistical Computation and Simulation.

3.1 Introduction

In this work we aim to fill a gap in the Bayesian literature by proposing two objective priors for the parameter of the Yule–Simon distribution. The distribution was firstly discussed in Yule (1925) and then re-proposed in Simon (1955), and can be used in scenarios where the center of interest is some sort of frequency in the data. For example, Yule (1925) used it to model abundance of biological genera, while Simon (1955) exploited the distribution properties to model the addition of new words to a text. It goes without saying that other areas of applications can be considered where, for instance, frequencies represent the elementary unit of observation. For example, in this chapter we show the employment of the Yule–Simon distribution in modelling daily increments of social network stock options, surnames and 'superstar' success in the music industry.

Despite the wide range of applications, the literature on the Yule–Simon distribution appears to be limited. For example, Gallardo et al. (2016) highlight that the heavy-tailed property of the Yule–Simon distribution allows for extreme values even for small sample sizes. In particular, they claim that the above property is suitable to model short survival times which, due to the nature of the problem, happen with relatively high frequency. And, more surprisingly, to the best of our knowledge it seems that no attention has been given to the problem by the Bayesian community. Given the challenges that classical inference faces in estimating the parameter of the distribution (Garcia Garcia, 2011), the possibility of tackling the problem from a Bayesian perspective is, undoubtedly, appealing.

In addressing the estimation of the shape parameter of the Yule–Simon distribution by means of the Bayesian framework, we opted for an objective approach. We propose two priors: the first is the Jeffreys rule prior (Jeffreys, 1961), while the second is obtained by applying the loss-based approach discussed in Villa and Walker (2015). Although we

formally introduce the Yule–Simon distribution and its derivation in the next Section, it is important to give an anticipation of the general idea here, so to fully appreciate the gain in adopting an objective approach. As nicely illustrated in Chung and Cox (1994), the shape parameter of the distribution is linked via a one-to-one transformation to the probability that the next observation will not take a value previously observed. For example, if we have observed n words in a text, we wish to make inference on the probability that the $(n + 1)$ observation is a word not yet encountered in the text, assuming this probability to be constant. It is then clear that the Yule–Simon distribution models extremely large events. As such, the information in the data about these events is limited and a "wrongly" elicited prior could end up dominating the data. On the other hand, a prior with minimal information content would allow the data "to speak", resulting in a more robust inferential procedure. We do not advocate that in every circumstance an objective approach is the only suitable. In fact, if reliable prior information is available, an elicited prior would represent, in general, the natural choice. Alas, in the presence of phenomena with extremely rare events, the above information is often insufficient or incomplete, and an objective choice would then represent the most sensible one.

On the other hand, we propose an explicit Gibbs sampling scheme when a Gamma prior is chosen for the shape parameter. The algorithm we propose is based on a stochastic representation of the Yule–Simon distribution as a mixture of Geometric distributions. This naturally suggests a data augmentation scheme which can be employed to address Bayesian inference. In particular the choice of a Gamma prior leads to explicit full conditional distributions.

The chapters is organized as follows. In Section 3.2 we set the scene by introducing the Yule–Simon distribution and the notation that will be used throughout the chapter.

The proposed objective priors are derived and discussed in Section 3.3. Section 3.4 collects the analysis of the frequentist performances of the posterior distributions yielded by the proposed priors. Through a set of several simulation scenarios, we compare and analyse the inferential capacity of the objective priors here discussed. In Section 3.5 we illustrate the application of the priors to three real-data applications. In Section 3.6 we present the algorithm related to the data augmentation scheme and we illustrate it by means of simulations, where we consider both a single i.i.d. sample and a count data regression. Section 3.7 discusses applications to text analysis and comparison with the frequentist results in the literature. Finally, Section 3.8 is reserved to concluding remarks and points of discussion.

3.2 Preliminaries

The most known functional form of the Yule–Simon distribution, possibly, is the following:

$$f(k;\rho) = \rho\,\mathrm{B}(k,\rho+1), \qquad k = 1,2,\ldots \text{ and } \rho > 0, \tag{3.1}$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function and ρ is the shape parameter. The distribution in (3.1) was firstly proposed by Yule (1925) in the field of biology; in particular, to represent the distribution of species among genera in some higher taxon of biotic organisms. More recently, Simon (1955) noticed that the above distribution can be observed in other phenomena, which appear to have no connection among each others. These include, the distribution of word frequencies in texts, the distribution of authors by number of scientific articles published, the distribution of cities by population and the distribution of incomes by size. The derivation process followed by Simon (1955) was based on word frequencies, and it consisted of two assumptions:

(i) The probability that the $(n+1)$-th word is a word observed exactly k times in the

first n words, is proportional to $k \cdot h(k,n)$, which is the total number of occurences of all the words that have been observed exactly k times. In particular, $h(k,n)$ is the number of different words that have occured exactly k times in the first n words (e.g. if in a text there are 300 different words that have appeared once each, then $h(k,n) = 300$);

(ii) The probability that the $(n+1)$-th word is new (i.e. not being observed in the first n words) is constant and equal to $\alpha \in (0,1)$.

Simon (1955) shows that, under the condition of stationarity, the process defined by the above two assumptions yields (3.1) by setting $\rho = 1/(1-\alpha)$, obtaining:

$$f(k;\alpha) = \frac{1}{1-\alpha}\mathrm{B}\left(k, \frac{1}{1-\alpha}+1\right). \tag{3.2}$$

An important consequence of the above assumption (ii) is that the shape parameter ρ of the distribution takes values in $(1,+\infty)$. In other words, should we use the model as in Yule (1925), which includes the possibility that $0 < \rho \leqslant 1$, we would loose the interpretation of the generating process described by the two assumptions above. In fact, for $\rho < 1$, the probability of observing a new word would be negative; while for $\rho = 1$ the probability would be zero, rendering the process trivial (i.e. all the observed words will be equal to the first one observed). Furthermore, the expectation of the Yule–Simon distribution is defined only for values of the shape parameter larger than one, and this property is something one would expect in most applications. For all the above reasons, in the first part of the chapter we focus on the parametrization of the Yule-Simon given in (3.2), and we will discuss prior distributions for α.

In addition to the parametrization of the Yule–Simon distribution as in (3.2), we will also consider the possibility of having the parameter α discrete. This is a common finding in

literature, especially when implementations of the model are considered. See, for example, Simon (1955) and Garcia Garcia (2011). The discretization of α will be discussed in detail in Section 3.3.2.

On the other hand, the probability distribution defined in (3.1) can be seen as a mixture of Geometric distributions. Precisely, let W be an exponentially distributed random variable with parameter ρ, and let K be a Geometric distribution with probability of success equal to e^{-W}. Therefore, it is easy to see that the Yule-Simon distribution can be recovered as the marginal of the random vector (K, W), i.e.

$$f(k;\rho) = \int_0^\infty e^{-w}(1-e^{-w})^{k-1}\rho e^{-\rho w}dw. \tag{3.3}$$

The above description of the Yule-Simon distribution is crucial to define a data augmentation scheme in a Bayesian setting.

3.3 Objective Priors for the Yule-Simon distribution

This section is devoted to the derivation of two objective priors for the Yule-Simon distribution: the Jeffreys prior and loss-based prior. The former assumes that parameter space of α is continuous and it is based on the well-known invariance property proposed by Jeffreys (1961); the latter assumes the parameter space discrete and is based on Villa and Walker (2015).

Based on the previous description of the Yule–Simon distribution, we consider the following Bayesian model,

$$k_1, \ldots, k_n|\alpha \overset{i.i.d}{\sim} f(k;\alpha)$$

$$\alpha \ \sim \ \pi(\alpha)$$

where $f(k;\alpha)$ is the Yule–Simon distribution described in 3.2 and $\pi(\alpha)$ is the objective prior distribution of the parameter of interest α, either the Jeffreys or the loss-based prior.

The likelihood function of the above model, conditionally to the parameter α, is the following:

$$L(\mathbf{k};\alpha) = \prod_{i=1}^{n} f(k_i,\alpha) = \prod_{i=1}^{n} \frac{1}{1-\alpha}\mathrm{B}\left(k_i, \frac{1}{1-\alpha}+1\right),$$

where $\mathbf{k} = (k_1,\ldots,k_n)$ is the vector of observations. In order to compute the Bayesian analysis of the model, we consider the posterior distribution for α:

$$\pi(\alpha;\mathbf{k}) \propto L(\mathbf{k};\alpha)\pi(\alpha),$$

where $\pi(\alpha)$ is the objective prior as described in the following part of the section.

3.3.1 The Jeffreys Prior

The Jeffreys prior is defined in the following way (Jeffreys, 1961):

$$\pi(\alpha) \propto \sqrt{\mathcal{I}(\alpha)},$$

where $\mathcal{I}(\alpha) = \mathbb{E}_\alpha\left[-\frac{\partial^2 \log(f(k;\alpha))}{\partial \alpha^2}\right]$ is the Fisher Information. In the next Theorem (which proof is in the Appendix) an explicit expression of the Jeffreys prior for the Yule-Simon distribution is provided.

Theorem 3.3.1. *Let $f(k;\alpha)$ be the Yule-Simon distribution defined in equation (3.2), with $0 < \alpha < 1$. The Jeffreys prior for α is*

$$\pi(\alpha) \propto q(\alpha), \tag{3.4}$$

where

$$q(\alpha) = \frac{1}{1-\alpha}\sqrt{1 - \frac{1}{(2-\alpha)^2}\,{}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)},$$

with ${}_3F_2$ being the generalized hypergeometric function.

The Jeffreys prior stated in Theorem 3.3.1 is a proper prior. In fact, let

$$\pi(\alpha) = \frac{q(\alpha)}{K},$$

where

$$K = \int_0^1 \frac{1}{1-\alpha}\sqrt{1 - \frac{1}{(2-\alpha)^2}\,{}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)}\,d\alpha$$

is the normalizing constant of $\pi(\alpha)$. It is not difficult to prove that

$$K < \infty.$$

Indeed,

$$K \leqslant \int_0^1 \sqrt{\frac{3-\alpha}{1-\alpha}}\frac{1}{2-\alpha}d\alpha = \frac{1}{3}\pi - \ln(2-\sqrt{3}) < \infty.$$

The result above follows from the following inequality (see Figure 3.1)

$${}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right) \geqslant 1.$$

The properness of the prior in (3.4) ensures the properness of the yielded posterior distribution for α, as such suitable for inference.

3.3.2 The Loss-based Prior

Villa and Walker (2015) introduced a method for specifying an objective prior for discrete parameters. The idea is to assign a *worth* to each parameter value by objectively measuring what is lost if the value is removed, and it is the true one. The loss is evaluated by applying the well known result in Berk (1966) stating that, if a model is misspecified, the posterior distribution asymptotically accumulates on the model which is the nearest to the true one, in terms of the Kullback–Leibler divergence.

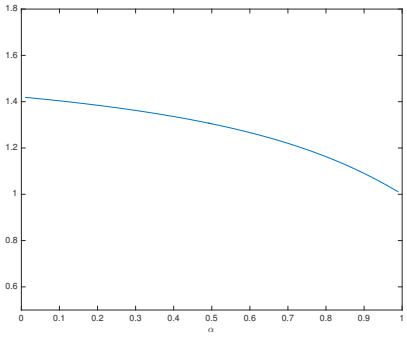


FIGURE3.1: Plot of the generalised hypergeometric function ${}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)$, where α takes values in $(0, 1)$.

Given that the parameter $\alpha \in (0, 1)$ of the Yule–Simon is in principle continuous, the above method can not be applied. However, the boundedness of the interval allows for an easy discretization, directly we can consider the set

$$\mathbb{D}_M = \left\{ \alpha = \frac{i}{M} : i = 1, \ldots, M-1 \right\}.$$

Therefore, the *worth* of the parameter value α is represented by the Kullback–Leibler divergence

$$D_{KL}(f(k|\alpha) \| f(k|\alpha')) = \int f(k|\alpha) \log \left\{ \frac{f(k|\alpha)}{f(k|\alpha')} \right\} dk,$$

where $\alpha' \neq \alpha$ is the parameter value that minimizes the divergence. To link the *worth* of a parameter value to the prior mass, Villa and Walker (2015) use the self-information loss function. This particular type of loss function measures the loss in information contained in a probability statement (Merhav and Feder, 1998). We can write the loss-based prior as utility functions, where in our special case $u_1(\alpha) = \log(\pi(\alpha))$ is the utility associated with the prior for our model $f(k|\alpha)$ and $u_2(\alpha)$ is the minimum divergence from $f(k|\alpha)$, i.e. the utility of keeping α in $\mathbb{D}_M$. The utility described above are 2 different ways of

measuring the same utility in α. Hence, $u_1(\alpha) \in (-\infty, 0]$ and $u_2(\alpha) \in [0, \infty)$, while we want $u_1(\alpha) = -\infty$ when $u_2(\alpha) = 0$. If we use an exponential transformations $\exp\{u_1(\alpha)\}$ and $\exp\{u_2\} - 1$, then the scales are matched. Therefore, we have

$$e^{u_1(\alpha)} = \pi(\alpha) \propto e^{g(u_2(\alpha)},$$

where $g(u) = \log(e^u - 1)$. As we now have, for each value of α, the loss in information measured in two different ways, we simply equate them obtaining the loss-based prior of Villa and Walker (2015):

$$\pi(\alpha) \propto \exp\left\{\min_{\alpha' \neq \alpha} D_{KL}(f(k|\alpha) \| f(k|\alpha'))\right\} - 1 \qquad \alpha, \alpha' \in \mathbb{D}_M, \tag{3.5}$$

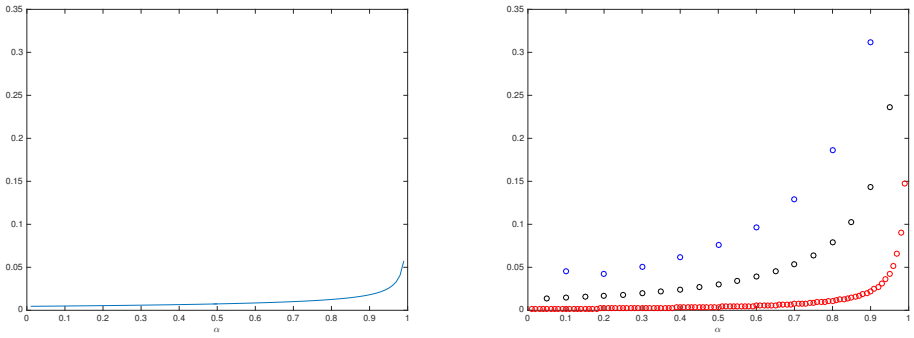
where

$$D_{KL}(f(k|\alpha) \| f(k|\alpha')) = \log\left(\frac{1-\alpha'}{1-\alpha}\right) + \mathbb{E}_\alpha\left\{\log\left[B\left(k; \frac{1}{1-\alpha} + 1\right)\right]\right\} - \mathbb{E}_\alpha\left\{\log\left[B\left(k; \frac{1}{1-\alpha'} + 1\right)\right]\right\}.$$

As the discretized parameter space is finite, no matter what value of M one chooses, the prior (3.5) is proper, hence, the yielded posterior will be proper as well.

An important aspect is that the value α' minimizing the Kullback–Leibler divergence can not be analytically determined, and the prior has to be computationally derived. However, even for large values of M, the computational cost is trifling compared to the whole Monte Carlo procedure necessary to simulate from the posterior distribution.

To have a feeling of the prior distributions derived above, we have plotted them in Figure 3.2. The behaviour of the priors is similar, in the sense that they tend to increase as α increases and, for increasing values of M, the two distributions seem to converge.



(a) (b)

FIGURE3.2: Prior distribution for α in panel obtained by applying, in panel (a), Jeffreys rule, while, in panel (b), the loss-based method with $M = 10$ (blue dots), with $M = 20$ (black dots) and with $M = 100$ (red dots).

However, we note that the Jeffreys prior is flatter than the loss-based priors for large values of the parameter, i.e. for α approximately greater than 0.8.

3.4 Simulation Study for objective priors

The objective priors defined in Section 3.3 are automatically derived by taking into consideration properties intrinsic to the Yule–Simon distribution. In other words, they do not depend on experts knowledge or previous observations. It is therefore necessary, in order to validate them, to assess the goodness of the priors by making inference on simulated data. This section is dedicated in performing a simulation study on the parameter α using observations obtained from fully known distributions.

We have considered different sample sizes, $n = 30$, $n = 100$ and $n = 500$, to analyse the behaviour of the prior distributions under different level of information coming from the data. Here we show the results for $n = 100$ only, as the sole differences in using $n = 30$ and $n = 500$ sample sizes are limited to the precision of the inferential results: relatively low for $n = 30$ and relatively high for $n = 500$, as one would expect. Besides that, the differences in

the performance of the two priors noted for $n = 100$ remain for the other sample sizes. As the loss-based prior depends on the discretization of the parameter space, for illustration purposes, we have considered $M = 10$ and $M = 20$, that is $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ and $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$, respectively.

Both the Jeffreys prior and the loss-based prior yield posterior distributions for α which are not analytically tractable, hence, it is necessary to use Monte Carlo methods. The reasons behind this choice are discussed in the following two remarks.

Remark 1. In the case of the Jeffreys' prior, one may use univariate numerical integration to compute quantities of interest, such as mean and credible intervals. However, in our experience, numerical integration may fail when the complexity on the integrand increases.

Remark 2. For the Loss based prior, one may be tempted to compute the posterior probabilities as

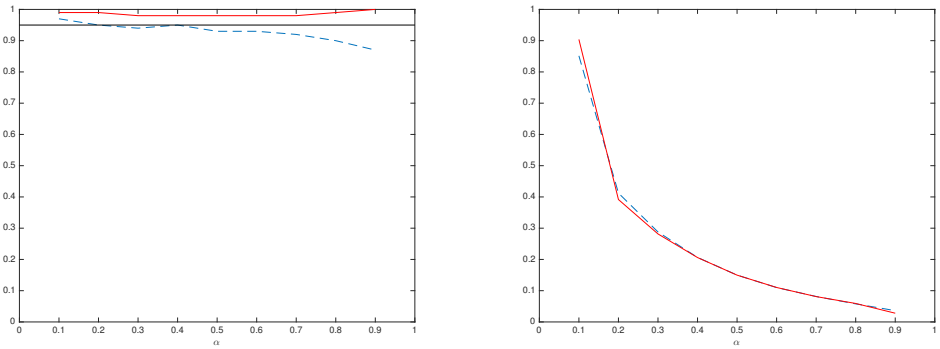
$$\pi(\alpha_j|\boldsymbol{k}) = \frac{L(\boldsymbol{k}|\alpha_j)\pi(\alpha_j)}{\sum_{\alpha_i \in \mathbb{D}_M} L(\boldsymbol{k}|\alpha_i)\pi(\alpha_i)}.$$

However, the Kullback-divergence in equation (3.5) may take values close to zero, leading to probabilities which can be interpreted as zero by the computational software employed. This is particularly true for moderately large values of M. Therefore, a Metropolis-Hastings in the logarithmic scale allows to overcome the above problem.

We have generated 100 samples from a Yule–Simon distribution with the parameter α set to every value in the discretization of the parameter space, 9 for $M = 10$ and 19 for $M = 20$. For each sample we have simulated from the posterior distribution of α, under both priors, by running $10,000$ iterations, with a burn-in period of $2,000$ iterations.

To evaluate the priors we have considered two frequentist measures. The first is the frequentist coverage of the 95% credible interval. That is, for each posterior, we compute

the interval between the 0.025 and 0.975 quantiles and see if the true value of α is included in it. Over repeated samples, one would expect a proportion of about 95% of the posterior intervals to contain the true parameter value. The second frequentist measure gives an idea of the precision of the inferential process, and it is represented by the square root of the mean squared error (MSE) from the mean, relative to the parameter value: $\sqrt{\text{MSE}(\alpha)}/\alpha$. We have considered the MSE from the median as well but, due to the approximate symmetry of the posterior, the results are very similar to the MSE from the mean. Figure 3.3

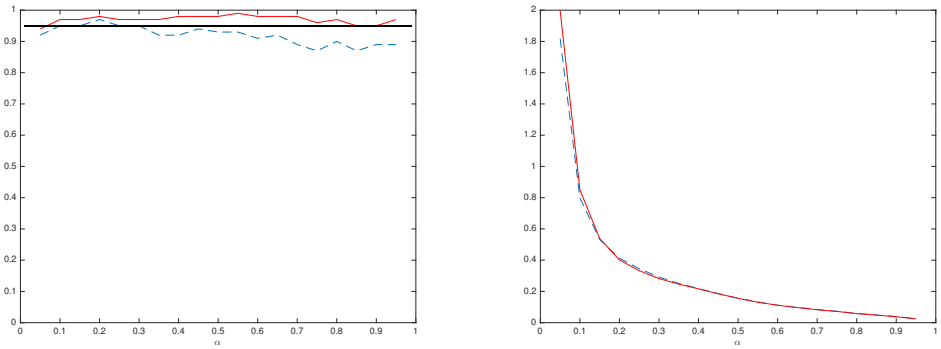


(a) (b)

FIGURE3.3: Frequentist properties of the Jeffreys prior (dashed line) and the loss-based prior (continuous line) for $n = 100$. The loss-prior is considered on the discretized parameter space with $M = 10$. The left plot shows the posterior frequentist coverage of the 95% credible interval, and the right plot represents the square root of the MSE from the mean of the posterior, relative to α.

details the results for the simulations with $n = 100$ and a parameter space for α discretized with increments of 0.1, that is $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. If we compare the coverage, we note that the loss-based prior tends to over-cover the credible interval, while the Jeffreys prior, although shows a better coverage for values of $\alpha < 0.5$, deteriorates in performance as the parameter tends to the upper bound of its space. Looking at the MSE, both priors appear to have very similar performance, and the (relative) error tends to decrease and

α increases. In Figure 3.4 we have compared the frequentist performance of the Jeffreys



(a) (b)

FIGURE3.4: Frequentist properties of the Jeffreys prior (dashed line) and the loss-based prior (continuous line) for $n = 100$. The loss-prior is considered on the discretized parameter space with $M = 20$. The left plot shows the posterior frequentist coverage of the 95% credible interval, and the right plot represents the square root of the MSE from the mean of the posterior, relative to α.

prior with the loss-based prior defined over a more densely discretized parameter space, i.e. $\alpha = \{0.05, 0.10, \ldots, 0.95\}$. We note a smoother behaviour of the priors compared to Figure 3.3, which is obviously due to the denser characterization considered. The coverage still reveals a tendency of the loss-based prior to over-cover, although less pronounced than the previous case. Jeffreys prior does not present any significant difference from the previous case, as one would expect. For what it concerns the MSE, the differences between the two priors are negligible, and the only aspect we note, as mentioned above, is a smoother decrease of the error as the parameter increases.

We look more into the details of the objective approach by analysing two i.i.d. samples. In particular, we consider a random sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.40$ and a sample, of the same size, from a Yule–Simon with $\alpha = 0.68$.

In both cases, since we do not have an explicit form of the posterior distribution,

we have sampled from the posterior distribution via Monte Carlo methods, in particular using a Metropolis–Hastings in logarithmic scale (with the proposal transition kernel for α distributed as a Beta for the Jeffreys prior and as a Discrete Uniform on the interval $(1/M, (M-1)/M)$ for the loss-based prior), with $10,000$ iterations and a burn-in period of $2,000$ iterations. Figure 3.5 shows the posterior samples and posterior histograms derived by applying the Jeffreys prior and the loss-based prior with two different discretizations, that is $M = 10$ and $M = 20$. The summary statistics of the three posteriors are reported in Table 3.1, where we have the mean, the median, and the 95% credible interval. By

Prior	Mean	Median	95% C.I.
Jeffreys	0.40	0.41	(0.23,0.53)
Loss-based ($M = 10$)	0.40	0.4	(0.2,0.5)
Loss-based ($M = 20$)	0.40	0.41	(0.22,0.56)

Table 3.1: Summary statistics of the posterior distributions for the parameter α of the simulated data from a Yule-Simon distribution with $\alpha = 0.40$.

comparing the mean of the posterior distributions, we see that they are all centered around the true parameter value. The credible interval yielded by the loss-based priors with the most dense discretization ($M = 20$) is larger than the other two intervals. However, the difference is very small and we can conclude that the three prior distributions result in posteriors which carry the same uncertainty. In other words, the three objective priors perform in the same way.

Similar considerations can be made for the case where we have sampled $n = 100$

Prior	Mean	Median	95% C. I.
Jeffreys	0.68	0.68	(0.57,0.77)
Loss-based ($M = 10$)	0.68	0.7	(0.6,0.8)
Loss-based ($M = 20$)	0.68	0.68	(0.55,0.79)

Table 3.2: Summary statistics of the posterior distributions for the parameter α of the simulated data from a Yule-Simon distribution with $\alpha = 0.68$.

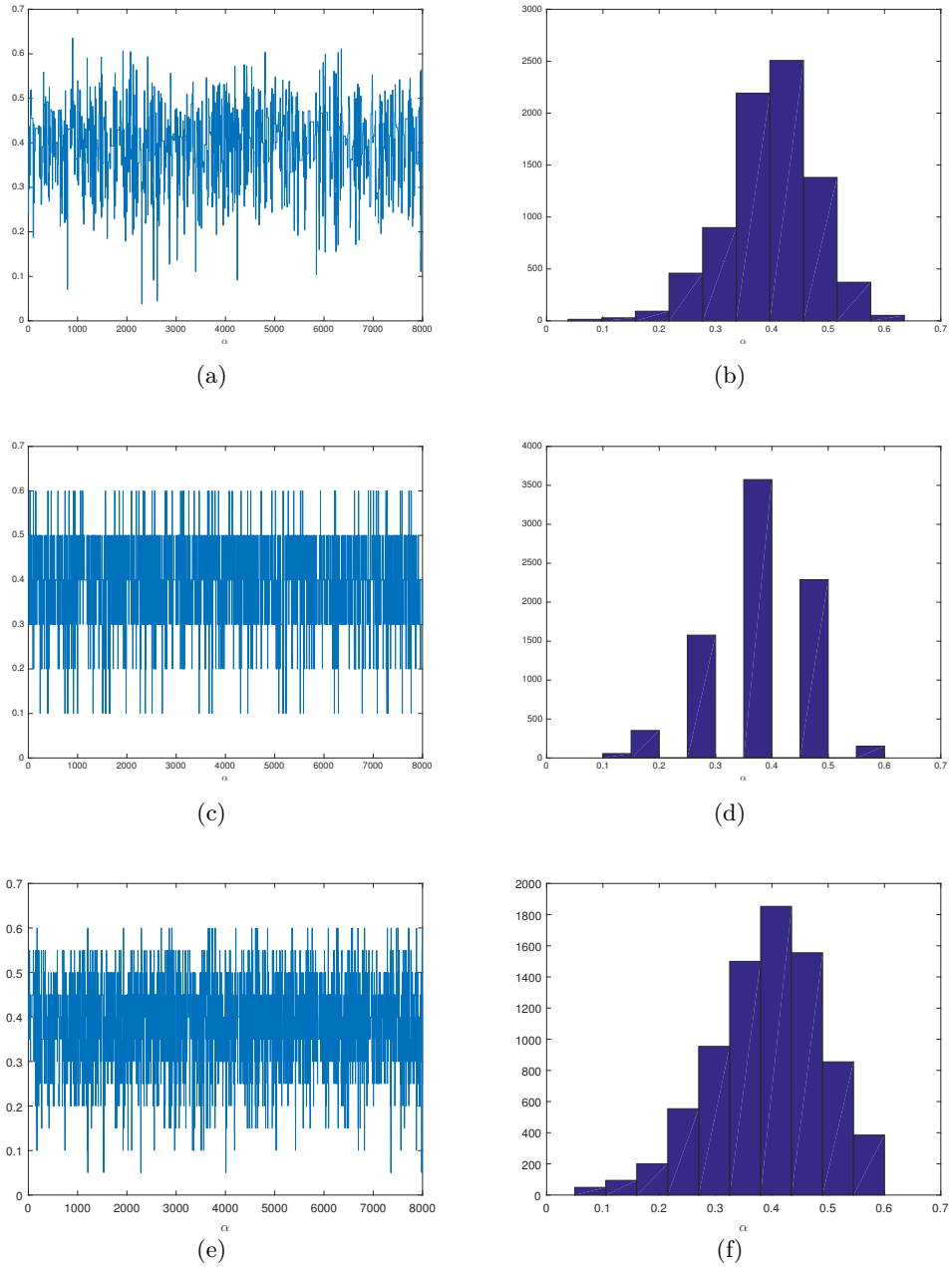


FIGURE3.5: Posterior samples (left) and histograms (right) of the analysis of an i.i.d. sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.40$. From top to bottom, we have Jeffreys prior, loss-based prior with $M = 10$ and loss-based prior with $M = 20$.

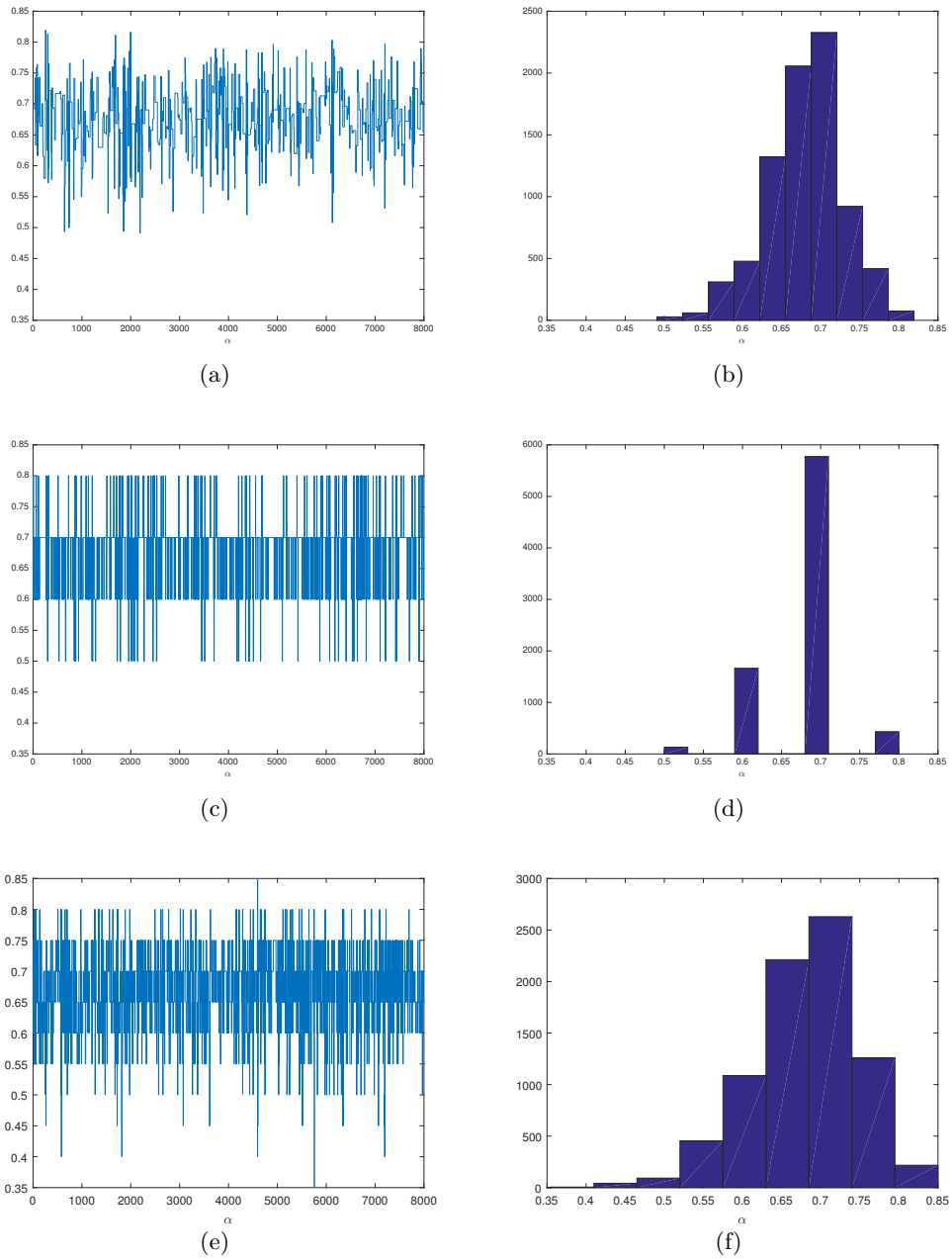


FIGURE3.6: Posterior samples (left) and histograms (right) of the analysis of an i.i.d. sample of size $n = 100$ from a Yule–Simon distribution with $\alpha = 0.68$. From top to bottom, we have Jeffreys prior, loss-based prior with $M = 10$ and loss-based prior with $M = 20$.

observations from a Yule–Simon distribution with $\alpha = 0.68$. By inspecting Figure 3.6 and Table 3.2, we note a very similar behaviour of the three priors, in the sense that the posterior distributions are still centered around the true value of α and that the credible intervals do not present important differences. Note that the choice of a true parameter value which would have not been included in any of the two discretized sample spaces, upon which the loss-prior is based, allows to show that the inferential process appears to be not affected by the discretization, hence motivating it.

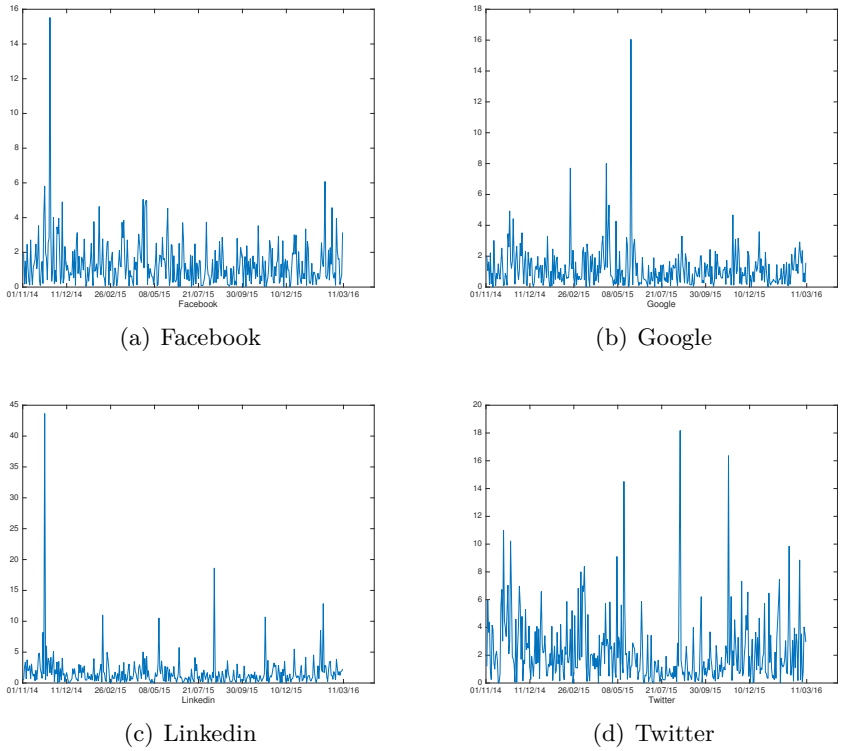
To conclude, the simulation study shows no tangible differences in the performance of the prior distributions, in the spirit of objective Bayesian analysis.

3.5 Real Data Application for objective priors

To illustrate the proposed priors, both the Jeffreys and the loss-based prior for the Yule-Simon distribution, we analyze three datasets. The first dataset concerns daily increments of four popular social networks stock indexes in the US market, the second contains the frequencies of surnames observed in the 1990 US Census, and the last dataset consists of 'number one' hits in the US music industry.

3.5.1 Social network stock indexes

We analyze different data in the social media marketing, in particular we focus on Facebook, Twitter, Linkedin and Google. These four major companies are the most powerful social networks in the world and are listed in the Wall Street exchange market (http://finance.yahoo.com). We analyze the daily increments for the stocks and, in particular, we consider the adjusted closing price from the 1^{st} of October 2014 to the 11^{th} of March 2016, for a total of $n = 365$ observations. The daily increments are obtained by applying $z_t = |r_t/r_{t-1} - 1| \cdot 100$, for $t = 2, \ldots, 365$, where r_t is the adjusted closing price for



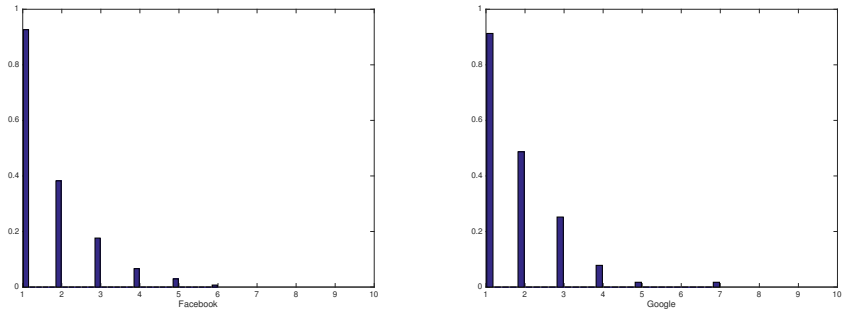
(a) Facebook (b) Google

(c) Linkedin (d) Twitter

FIGURE3.7: Daily increments for Facebook, Google, Linkedin and Twitter from the 1^{st} of October 2014 to the 11^{th} of March 2016.

the index at day t, and we built our frequency on it. These are shown in Figure 3.7, while Figure 3.8 shows the histogram of the frequencies of the discretized data. The discretization has been done by counting the number of times a daily return took a value truncated at the second decimal digit. For example, if two observed daily returns are 1.2494 and 1.2573, they were both considered as two occurrences of the same value. By inspecting the histograms in Figure 3.8 is seems that the (transformed) Yule–Simon distribution might be a suitable statistical model to represent the data. We apply the Bayesian framework and obtain the posterior distribution for the parameter of interest as

$$\pi(\alpha|\mathbf{k}) \propto L(\mathbf{k}|\alpha)\pi(\alpha),$$



(a) Facebook (b) Google

(c) Linkedin (d) Twitter

FIGURE3.8: Histograms of the discretized daily returns for Facebook, Google, Linkedin and Twitter.

where $\mathbf{k} = (k_1, \ldots, k_n)$ represents the set of observations, i.e. the frequencies of the discretized daily returns, $L(\mathbf{k}|\alpha)$ the likelihood function and $\pi(\alpha)$ the prior distribution which, in turn, has the form of the Jeffreys prior in (3.4) or the loss-based prior (3.5). We have obtained the posterior distributions for the parameter α of the transformed Yule-Simon distribution by Monte Carlo methods. We run 25,000 iterations with a burn-in period of $5,000$ iterations. We have reported the chain and the histogram of the posterior distributions in Figure 3.9 and in Figure 3.10, with the corresponding summary statistics in Table 3.3. Note that, with the purpose of limiting the amount of space used, we have included the plots of the Facebook and Google daily returns only.

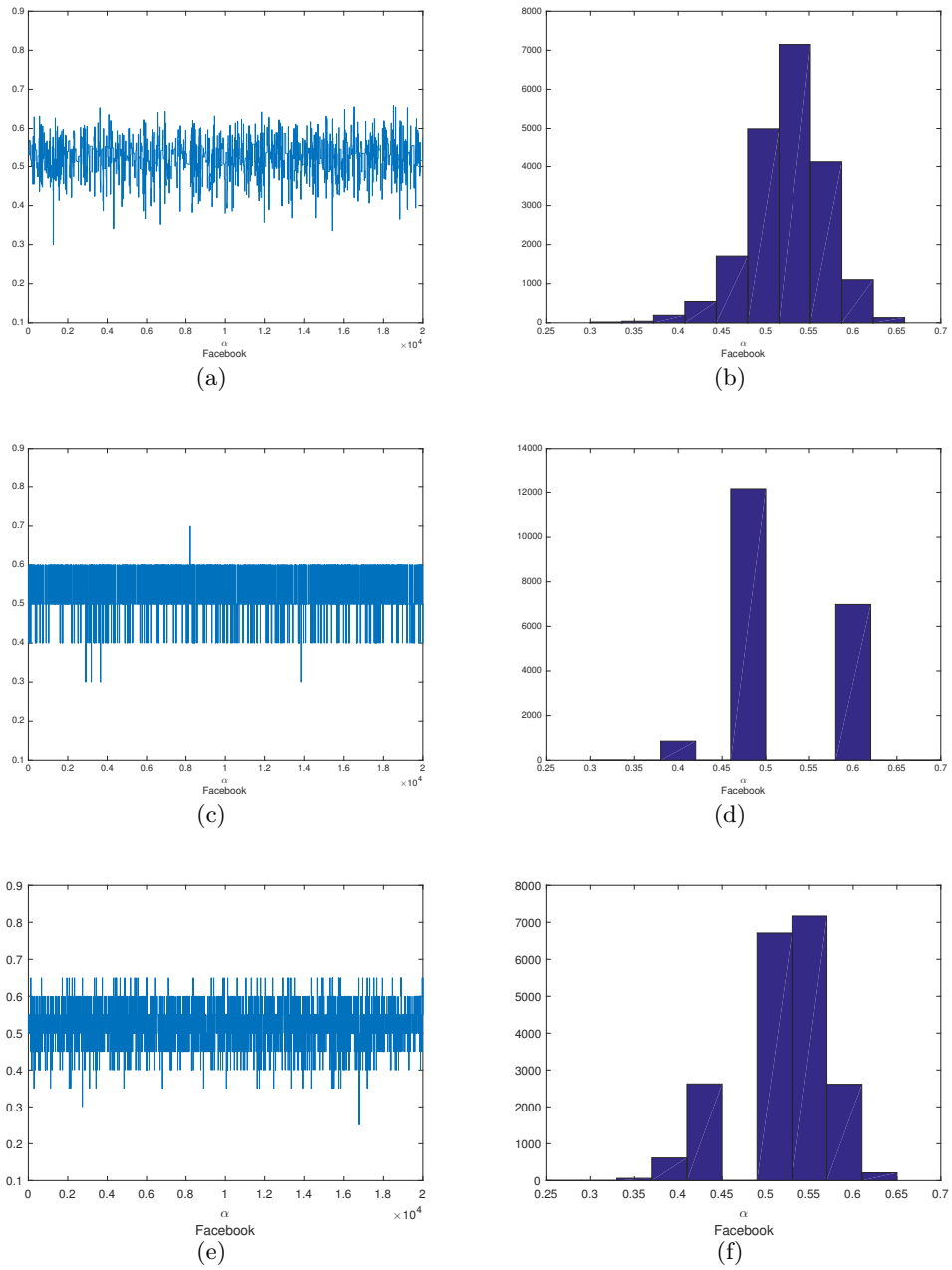


FIGURE3.9: Posterior samples (left) and posterior histograms (right) for the Facebook daily returns obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

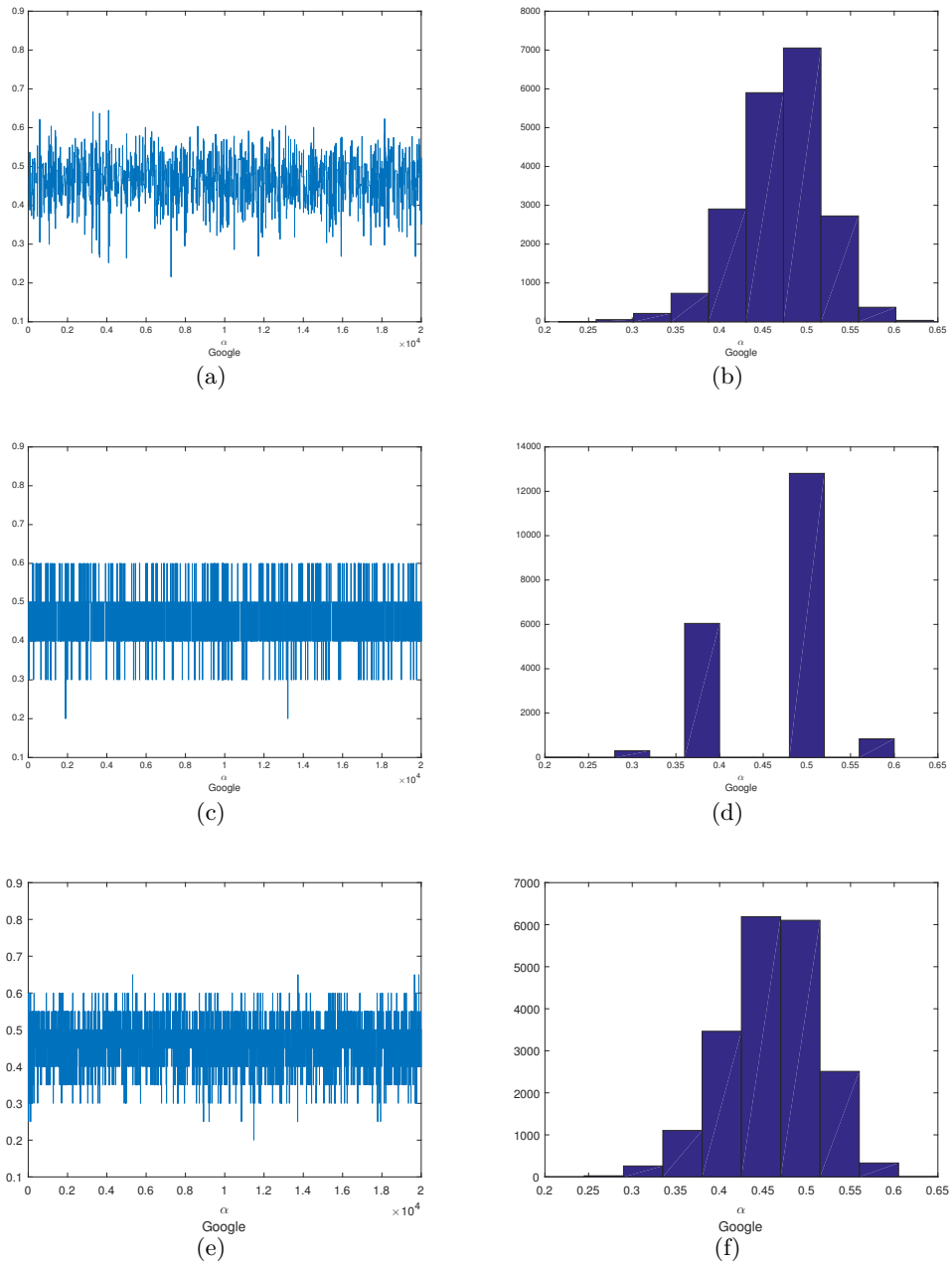


FIGURE3.10: Posterior samples (left) and posterior histograms (right) for the Google daily returns obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

Company	Prior	Mean	Median	95% C.I.
Facebook	Jeffreys	0.53	0.53	(0.43, 0.61)
Facebook	Loss-based ($M = 10$)	0.53	0.5	(0.4, 0.6)
Facebook	Loss-based ($M = 20$)	0.52	0.55	(0.40, 0.60)
Google	Jeffreys	0.47	0.47	(0.37, 0.55)
Google	Loss-based ($M = 10$)	0.47	0.5	(0.4, 0.6)
Google	Loss-based ($M = 20$)	0.47	0.46	(0.35, 0.55)
Linkedin	Jeffreys	0.56	0.57	(0.47, 0.64)
Linkedin	Loss-based ($M = 10$)	0.57	0.6	(0.5, 0.6)
Linkedin	Loss-based ($M = 20$)	0.56	0.55	(0.45, 0.65)
Twitter	Jeffreys	0.68	0.68	(0.62, 0.73)
Twitter	Loss-based ($M = 10$)	0.69	0.7	(0.6, 0.7)
Twitter	Loss-based ($M = 20$)	0.68	0.70	(0.60, 0.75)

Table 3.3: Summary statistics of the posterior distribution for the parameter α of the social network stock index data.

#	Surname	Frequency	#	Surname	Frequency
1	Smith	2502021	6	Davis	1193807
2	Johnson	2014550	7	Miller	1054530
3	Williams	1738482	8	Wilson	843126
4	Jones	1544488	9	Moore	775975
5	Brown	1544488	10	Taylor	773488

Table 3.4: Ten most common Surname in United States from the Census 1990 analysis.

For all the four assets we notice that the results for α are very similar, as can be inferred by the minimal (or absence of) difference between the means and the medians. The credible intervals, as well, are very similar, with a slight larger size for the case where the loss-based prior with ($M = 20$) is applied. One way of interpreting the results is as follows. The parameter α can be seen as the probability that the next observation is different from the ones observed so far, and therefore we note that Twitter has the highest chance to take a daily increment not yet observed, while Google has the smallest.

Prior	Mean	Median	95% C. I.
Jeffreys	0.53	0.54	(0.47, 0.58)
Loss-based ($M = 10$)	0.52	0.5	(0.5, 0.6)
Loss-based ($M = 20$)	0.53	0.55	(0.45, 0.60)

Table 3.5: Summary statistics of the posterior distributions for the parameter α of the Census surname analysis.

3.5.2 Census Data - Surname analysis

The second example we examine the frequency of surnames in the US (http://www.census.gov/en.html). From the population censuses (Maruka et al., 2010), we focus on the US Census completed in 1990 and consider the first 500 most common surnames. Refer to Table 3.4 for a list of the first 10 most frequent surnames. Briefly, the process followed by Maruka et al. (2010) to obtain the data converts the surname with Senior (SR), Junior (JR) or a number in the last name field (f.e. Moore Sr or Moore Jr or Moore III are converted to Moore) and, in addition, the authors examined each name entry for the possibility of an inversion (e.g. a first name appearing in the last name fields or vice-versa). However, as there is the possibility of having many surnames that also inverted can sound absolutely right, the authors considered also the surname of the spouse, obtaining additional information to invert the name field of the entire family.

The analysis has been performed by running both the Markov Chain Monte Carlo for 25,000 iterations, with a burn-in of 5,000 iterations.

The posterior samples and the posterior histograms are shown in Figure 3.11, with the corresponding summary statistics of the posterior distributions reported in Table 3.5. We again notice similarities to the simulation study and the analysis of daily increments, in the sense that means and medians are very similar for each prior, and the 95% credible interval obtained by applying the loss-based prior with $M = 20$ is slightly larger than the one obtained by using either the Jeffreys prior or the loss-based prior with $M = 10$.

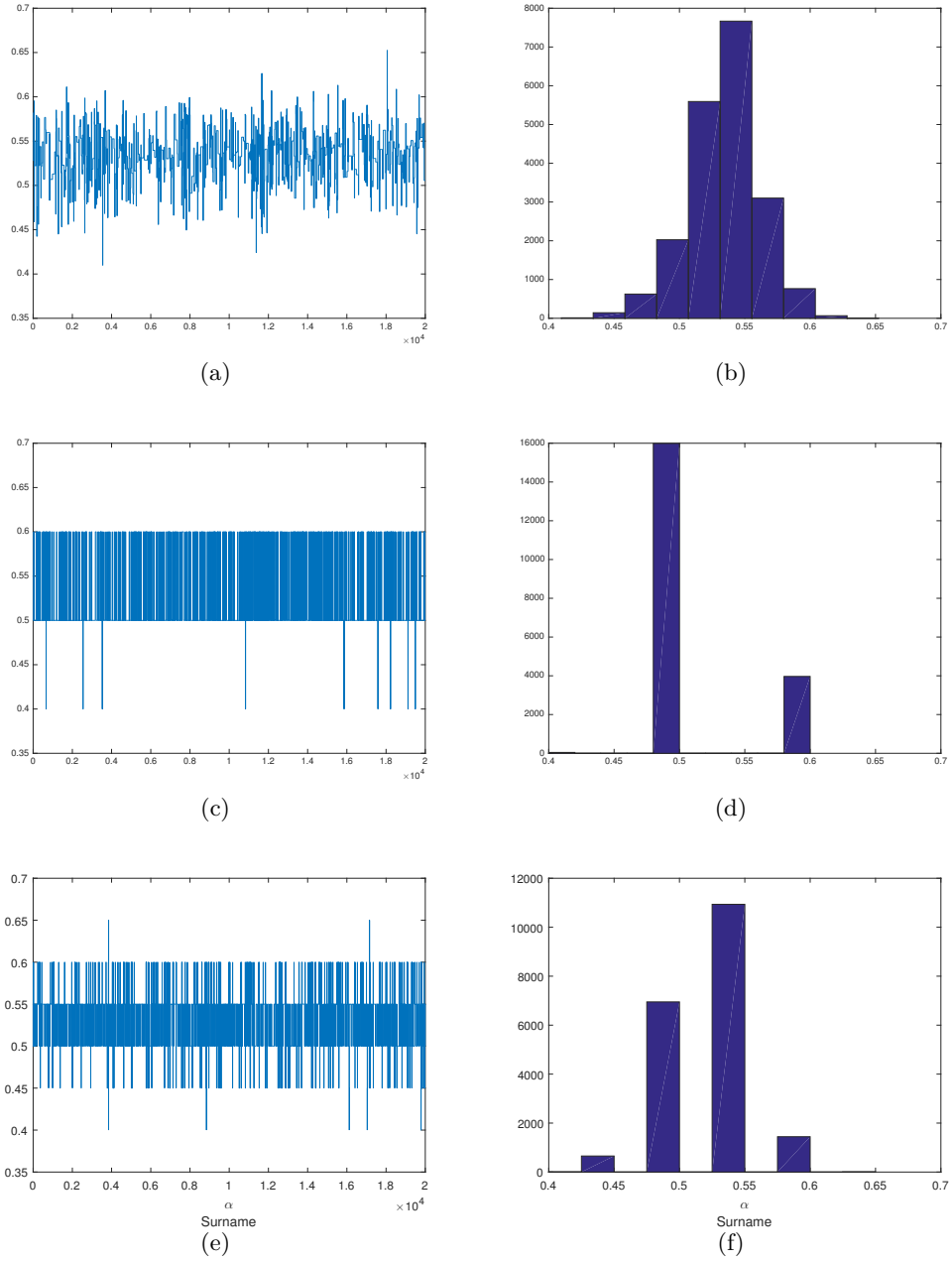


FIGURE3.11: Posterior sample (left) and posterior histogram (right) for the surname data set obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

The estimated value of α, on the basis of the 500 most common surnames in the US (and if we consider the posterior mean) is, roughly, 1/2. In other words, there are about 50% chances that the next observed surname is not in the list of the 500. Obviously, a larger sample size would yield a smaller posterior mean, as the number of surnames is finite and the more we observe, the harder is to find a "new" one.

3.5.3 'Superstardom' analysis

The last example consists in modelling the number of 'number one' hits a music artist had in the period 1955–2003 on the Billboard Hot 100 chart. The data, which is displayed in Table 3.6, has been used by Chung and Cox (1994) and Spierdijk and Voorneveld (2009) to show an apparent absence of correlation between talent and success in the music industry.

Hits	Observations	Hits	Observations
1	119	9	4
2	57	10	2
3	30	11	1
4	13	12	2
5	10	13	1
6	4	14	1
7	1	15	1
8	1	16	1

Table 3.6: Number of 'number one' hits per artist from 1955 to 2003.

We have run the Monte Carlo simulation for 25,000 iterations, with a burn in period of 5,000, for each of the considered priors. The posterior samples and histograms are shown in Figure 3.12, with the corresponding statistic summaries in Table 3.7.

This example of the music hits allows for some interesting points of dicussion. First, we note that the posterior distributions of for α are skewed; therefore, the posterior median represents a better centrality index than the posterior mean. Second, it is clear that the "true" value of α may be close to zero. As such, in order to explore better the parameter

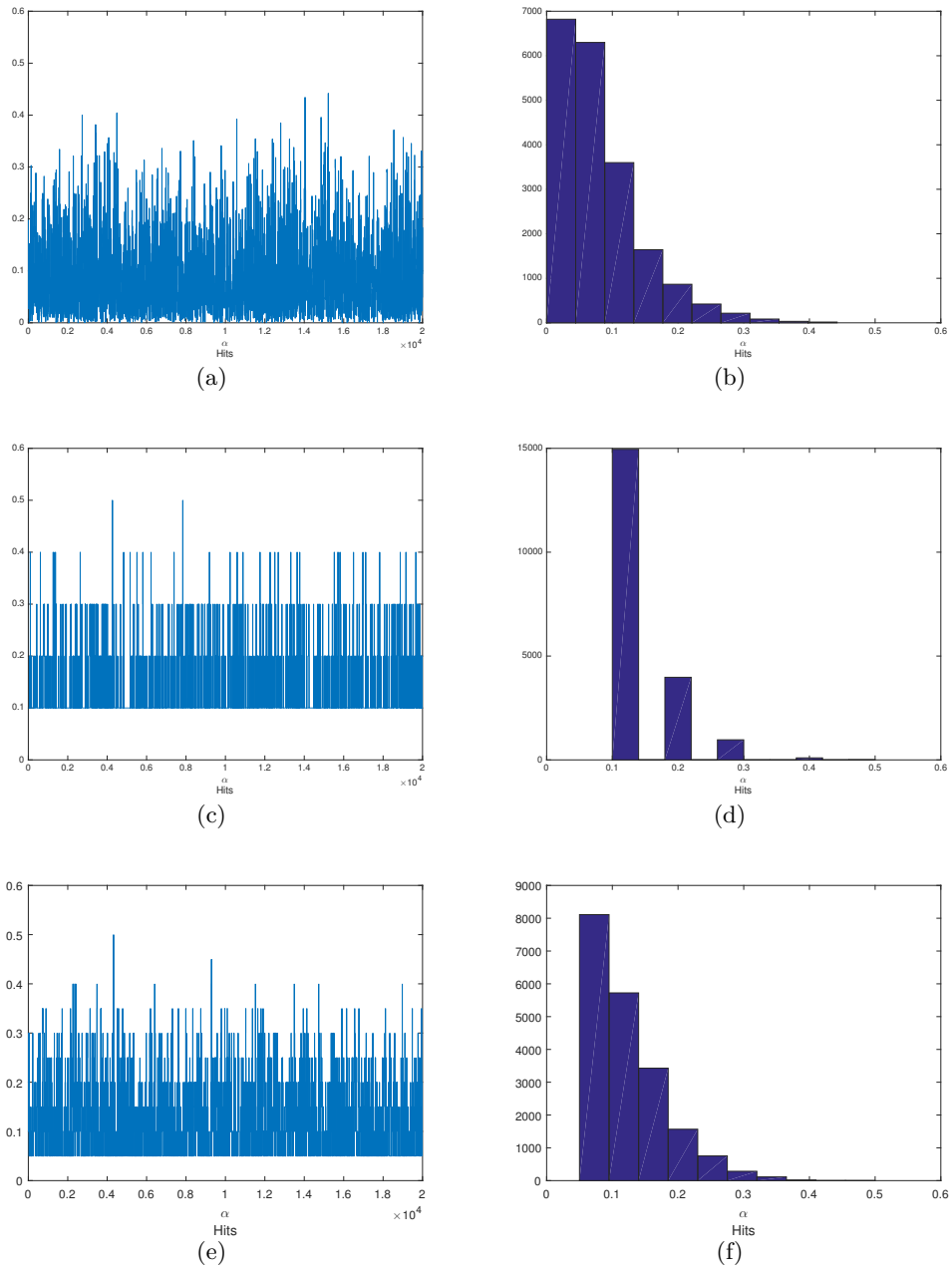


FIGURE3.12: Posterior sample (left) and posterior histogram (right) for the music 'number one' hits data set obtained by applying the Jeffreys prior (top), the loss-based prior with $M = 10$ (middle) and the loss-based prior with $M = 20$ (bottom).

Prior	Mean	Median	95% C.I.
Jeffreys	0.08	0.07	(0.004, 0.24)
Loss-based ($M = 10$)	0.13	0.1	(0.1, 0.3)
Loss-based ($M = 20$)	0.11	0.10	(0.05, 0.25)
Loss-based ($M = 100$)	0.10	0.08	(0.01, 0.29)

Table 3.7: Summary statistics of the posterior distribution for the parameter α of the analysis of the music 'number one' hits.

space when the loss-based prior is used, a denser discretization is more appropriate. We have then considered $M = 100$, resulting the posterior summary statistics in Table 3.7. We note now that the posterior median is similar to the one obtained using the Jeffreys prior. It is therefore recomendable that, when the inference on α indicates values near the parameter space boudaries, the level of discretization to be considered should be relatively dense.

3.6 Bayesian inference for Data Augmentation problem

In this section we will consider the data augmentation algorithm and the Gibbs sampling scheme when we choose a Gamma prior for the shape parameter. Based on (3.3), we can consider the following Bayesian model,

$$\begin{aligned} k_1, \ldots, k_n | \rho &\sim f(k; \rho) \\ \rho &\sim \text{Gamma}(a, b), \end{aligned} \tag{3.6}$$

where $f(k; \rho)$ is the Yule-Simon distribution defined in (3.1). The likelihood function of the above model, conditionally to the parameter ρ, is the following:

$$\begin{aligned} L(\mathbf{k}, \rho) &= \prod_{i=1}^{n} \int_0^{\infty} e^{-w_i}(1 - e^{-w_i})^{k_i - 1} \rho e^{-\rho w_i} dw_i \\ &= \int_{(0,\infty)^n} \prod_{i=1}^{n} e^{-w_i}(1 - e^{-w_i})^{k_i - 1} \rho e^{-\rho w_i} d\mathbf{w} \end{aligned}$$

$$= \int_{(0,\infty)^n} L(\mathbf{k}, \mathbf{w}, \rho) d\mathbf{w} \tag{3.7}$$

where $\mathbf{k} = (k_1, \ldots, k_n)$ is a vector of observations, $\mathbf{w} = (w_1, \ldots, w_n)$ is a vector of auxiliary variables, and

$$L(\mathbf{k}, \mathbf{w}, \rho) = \prod_{i=1}^{n} e^{-w_i}(1 - e^{-w_i})^{k_i - 1} \rho e^{-\rho w_i}. \tag{3.8}$$

In order to perform the Bayesian analysis of the model introduced in (3.6), we consider the following augmented version of the posterior distribution:

$$\pi(\rho, \mathbf{w}|\mathbf{k}) \propto L(\mathbf{k}, \mathbf{w}, \rho)\pi(\rho),$$

where $\pi(\rho) \propto \rho^{a-1}e^{-b\rho}$ is the Gamma prior. To sample from the posterior distribution we adopt a Gibbs sampling scheme and compute the full conditional distribution. It is straightforward to note that

$$p(w_i|w_{-i}, \mathbf{k}, \rho) \propto e^{-\rho w_i} e^{-w_i}(1 - e^{-w_i})^{k_i - 1}.$$

The change in variable $t_i = e^{-w_i}$, leads to a full-conditional distribution which is distributed as a Beta$(\rho + 1, k_i)$. On the other hand, the full-conditional distribution for ρ is

$$p(\rho|\mathbf{k}, \mathbf{w}) \propto \rho^{a+n-1} e^{-\rho(b + \sum_{i=1}^{n} w_i)} \sim \text{Gamma}\left(a + n, b + \sum_{i=1}^{n} w_i\right).$$

To sum up, the updating rule of the Gibbs sampler is as follows:

- Sample $t_i|\rho, k_i \sim \text{Beta}(\rho + 1, k_i)$, for $i = 1, \ldots, n$;
- Compute $w_i = -\log t_i$, for $i = 1, \ldots, n$;
- Sample $\rho|\mathbf{w}, \mathbf{k} \sim \text{Gamma}\,(a + n, b + \sum_{i=1}^{n} w_i)$.

In the following part we analyse the performance of the above algorithm by considering a i.i.d. sample generated from a Yule–Simon distribution (Section 3.6.1), and on a regression model for count data where the shape parameter of the Yule–Simon distribution is modelled in a similar fashion to the one in the classical Poisson regression (Section 3.6.2).

3.6.1 Single i.i.d. sample

This section is devoted to test the performance of the data augmentation algorithm on simulated data. To do this, we sample from a Yule–Simon distribution with two values of the parameter, $\rho = 0.8$ and $\rho = 5$. For each value of the parameter, we have simulated samples of different sizes, respectively $n = 30$, $n = 100$ and $n = 500$. Note that the choice of a relatively small sample size has the purpose to leverage on the Bayesian property of giving sensible results even when the information coming from the data is limited.

For the simulations, we have chosen a Gamma prior with shape parameter $a = 0.25$ and rate parameter $b = 0.05$. The choice was made with the intent of having a large variance in the prior, reflecting a fairly large prior uncertainty. The Gibbs sampler is run for $50,000$ iterations, with a burn-in period of $10,000$ iterations. This is repeated 20 times per sample to capture the variability in the procedure. Table 3.8 displays the summary statistics of the posteriors, that is, the mean, the median and mean square errors from these two indexes. Both in terms of central value and mean square error the simulation results are excellent, proving the soundness of the algorithm and, more in general, of the whole proposed approach.

As an example, in Figure 3.14 we show the posterior results for one simulation of the sample of size $n = 30$ from the Yule–Simon with $\rho = 5$, and one simulation from the same distribution with $n = 100$. We see that the chains exhibit a good mixing and that the means converge to the true value rather quickly. In detail, we have the posterior mean

ρ	n	Mean	Median	MSE Mean	MSE Median	Fixed-Point Alg
0.8	30	0.7955	0.7800	0.00002	0.00041	0.785
0.8	100	0.7606	0.7564	0.00160	0.00190	0.7582
0.8	500	0.8045	0.8035	0.00002	0.00001	0.8034
5	30	4.9800	4.5600	0.00046	0.19000	4.42
5	100	4.8200	4.7000	0.03600	0.10000	4.66
5	500	4.9000	4.8700	0.00990	0.01670	4.85

Table 3.8: Summary statistics of the posterior distributions for the parameter ρ of the simulated data from a Yule-Simon distribution with different values of $\rho = \{0.8, 5\}$ and sample sizes $n = \{30, 100, 500\}$ compared with the fixed-point algorithm of Garcia Garcia (2011).

equal to 4.98 for $n = 30$ and equal to 4.81 for $n = 100$, and the 95% credible intervals are, respectively, $(2.22, 10.17)$ and $(3.10, 7.30)$. As one would expect, the credible interval for the smaller sample size is larger than the one obtained with $n = 100$. This is reflected in the histogram in Figure 3.14 as well.

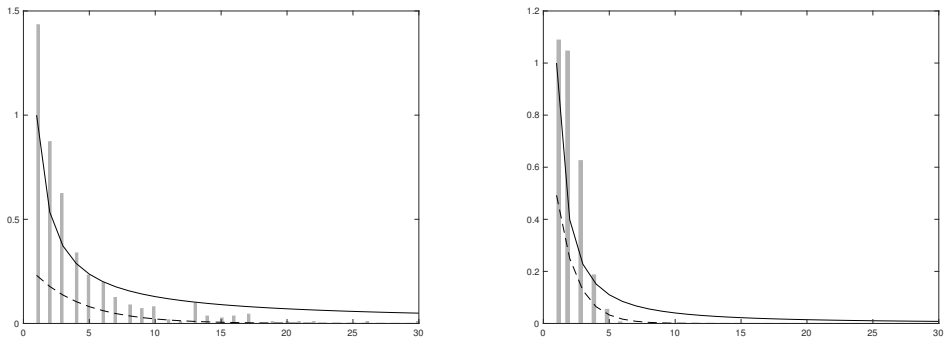


FIGURE3.13: Data (histogram), predictive distribution for Yule–Simon (solid line) and Geometric distribution (dashed line) for mixture of Geometric distributions (left) and for Poisson distribution (right).

Remark 3. As suggested, we compare the computational times to approximate the posterior for each of the three priors. We run a Gibbs sampler for $10,000$ iterations for a simulated example with $n = 100$. As expected, the posterior computation for the gamma prior

takes only 9.43 seconds for running $10,000$ iterations, while the computational time to approzimate the posterior for the objective priors takes 12.35 and 141 seconds for the loss-based prior and for the Jeffreys prior, respectively, due to the use of the M-H algorithm.

Remark 4. As regard to the predictive densities, we generate the data with $n = 500$ from a mixture of Geometric distribution $(0.75 \cdot \mathbb{G}e(0.4) + 0.25 \cdot \mathbb{G}e(0.1))$ and from a rescaled Poisson distribution with parameter $\lambda = 1$. We compare our Yule–Simon distribution with Gamma prior (solid lines in Figure A.7) with a standard Geometric distribution with conjugate beta prior ($\mathbb{B}e(a, b)$, with $a = b = 1$) for the probability of success (dashed lines in the same figure). Figure 3.13 shows that our approach approximate both heavy and light tails with respect to the Geometric distribution with beta prior.

3.6.2 Count data regression

In a count data regression model we are interested in the relations between the probability of a dependent variable k_i and the vector of independent variables x_i. The model is based on the following three assumptions:

1. the observation k_i follows the Yule–Simon distribution with parameter ρ_i, i.e.

$$f(k_i; \rho_i) = \rho_i B(k_i, \rho_i + 1), \quad k_i = 1, 2, \ldots, \quad \rho_i > 0;$$

2. the parameters of interest are modelled in the following way:

$$\rho_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}), \quad i = 1, \ldots, n,$$

   where $\boldsymbol{\beta}$ is a $(n_\beta \times 1)$ vector of parameters and $\mathbf{x}_i' = (1, x_{i2}, \ldots, x_{in_\beta})$ is a $(1 \times n_\beta)$ vector of regressors including a constant;

3. the observation pairs $(k_i, x_i), i = 1, \ldots, n$ are independently distributed.

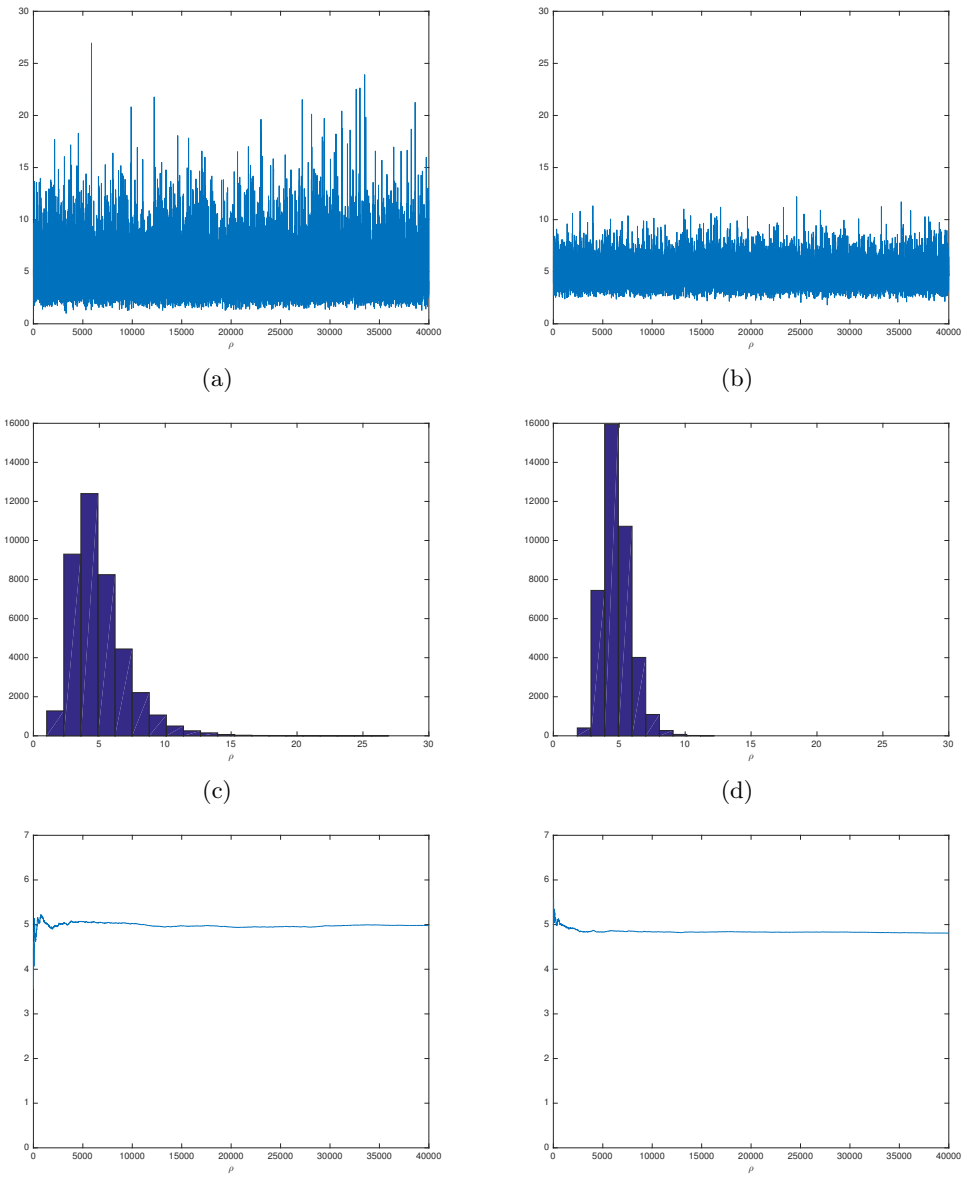


FIGURE3.14: Posterior sample (top), posterior histogram (middle) and progressive mean (bottom) for the simulation study of a Yule–Simon distribution with $\rho = 5$ and sample size $n = 30$ (left) and $n = 100$ (right).

For sake of illustration we focus on the case with one regressor only, although the arguments can easily be extended to include multiple regressors. Therefore, we have $\boldsymbol{\beta}' = (\beta_0, \beta_1)$, $\mathbf{x}_i' = (1, x_{i2})$ and $\rho_i = \exp\{\beta_0 + \beta_1 x_{i2}\}$. Assuming a standard bivariate normal prior for $\boldsymbol{\beta}$, we obtain the following augmented version of the posterior distribution:

$$\pi(\boldsymbol{\beta}, \mathbf{w}, \mathbf{x}|\mathbf{k}) \propto \left[\prod_{i=1}^{n} e^{-w_i}(1 - e^{-w_i})^{k_i - 1}\right] \exp\left\{\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\beta}\right\} \left[\prod_{i=1}^{n} e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}} w_i}\right] e^{-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}}.$$

Therefore, the full conditional distribution for the parameter of interest β is given by:

$$\pi(\boldsymbol{\beta}|\mathbf{w}, \mathbf{x}, \mathbf{y}) \propto \left[\prod_{i=1}^{n} \exp\left\{-e^{\mathbf{x}_i'\boldsymbol{\beta}} w_i\right\}\right] \exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\beta}\right\}. \tag{3.9}$$

As the expression in (3.9) is not an explicit known distribution, Monte Carlo methods have to be used. In particular, we adopt a Metropolis within Gibbs to obtain samples from the posterior distribution. We use a random walk proposal and the Gibbs sampler for the count data regression is as follows:

- Sample $t_i|\beta_0, \beta_1, x_i, k_i \sim \text{Beta}\left(\exp\{\beta_0 + \beta_1 x_{i2}\} + 1, k_i\right)$, for $i = 1, \ldots, n$;
- Compute $w_i = -\log t_i$, for $i = 1, \ldots, n$;
- Sample $\boldsymbol{\beta}|\mathbf{w}, \mathbf{k}, \mathbf{x}$ from the random walk Metropolis-Hastings algorithm.

We test the proposed data augmentation algorithm on two simulated data sets: for the first data set we have $(\beta_0, \beta_1) = (1.5, -1.0)$, and for the second one we have $(\beta_0, \beta_1) = (-0.5, 5.0)$. In both cases, the regressor values are sampled from a uniform $(0, 1)$. We ran $50,000$ iterations with a burn-in period of $10,000$ iterations, and this has been repeated 20 times per sample. For comparison purposes, we use the R function (*VGLM*) developed by Yee (2008, 2016) in the package VGAM. The function allows us to estimate the vector generalized linear model (see Yee (2014, 2015)), when we consider a Yule–Simon distribution.

Table 3.9 shows the posteriors mean, median, mean square errors and credible intervals for the two different scenarios. Overall, the results obtained by applying our algorithm are very close to the true parameter values. As noted in Section 3.6.1, the Bayesian approach outperforms the frequentist for small sample sized.

In both cases and for all the different sample sizes, the results are interesting for our approach and in particular, as seen in the previous simulated example, for small sample size the results are better from a Bayesian perspective with respect to the frequentist approach.

n	$\boldsymbol{\beta}$	Mean	Median	MSE Mean	95% C.I.	VGLM
30	$\beta_0 = -0.5$	-0.5	-0.5	0.0012	(-0.7,-0.2)	-0.2
	$\beta_1 = 5.0$	5.0	5.0	0.0014	(4.7,5.2)	7.7
30	$\beta_0 = 1.5$	1.6	1.6	0.0035	(1.3,1.8)	3.0
	$\beta_1 = -1.0$	-1.0	-1.0	0.0025	(-1.2,-0.7)	-0.9
100	$\beta_0 = -0.5$	-0.6	-0.6	0.0069	(-0.8,-0.4)	-0.7
	$\beta_1 = 5.0$	4.9	4.9	0.0071	(4.7,5.2)	4.8
100	$\beta_0 = 1.5$	1.4	1.4	0.0103	(1.2,1.6)	1.4
	$\beta_1 = -1.0$	-1.0	-1.0	0.0021	(-1.3,-0.8)	-1.2
500	$\beta_0 = -0.5$	-0.5	-0.5	0.0000	(-0.7,-0.3)	-0.5
	$\beta_1 = 5.0$	5.0	5.0	0.0029	(4.7,5.2)	5.1
500	$\beta_0 = 1.5$	1.5	1.5	0.0002	(1.3,1.7)	1.5
	$\beta_1 = -1.0$	-1.0	-1.0	0.0004	(-1.2,-0.8)	-0.9

Table 3.9: Summary statistics of the posterior distributions for the parameter (β_0, β_1) of the Yule–Simon regression with $(\beta_0, \beta_1) = \{(-0.5, 5.0); (1.5, -1.0)\}$ and sample sizes $n = \{30, 100, 500\}$ and VGLM estimators.

To better illustrate the performance we have simulated 300 observations for a case with $\beta_0 = 3.5$ and $\beta_1 = -2.2$. Figure 3.15 shows the posterior samples and the posterior histograms obtained with a Gibbs sampler run for 50,000 iterations with a burn-in period of 10,000. We see that for both parameters of the regression the chain has a good mixing, and the posterior means for β_0 and β_1 are, respectively, 3.40 and -2.195. The 95% credible intervals are, respectively, $(3.2, 3.6)$ and $(-2.4, -2.2)$ which comfortably contain the true values of the parameters.

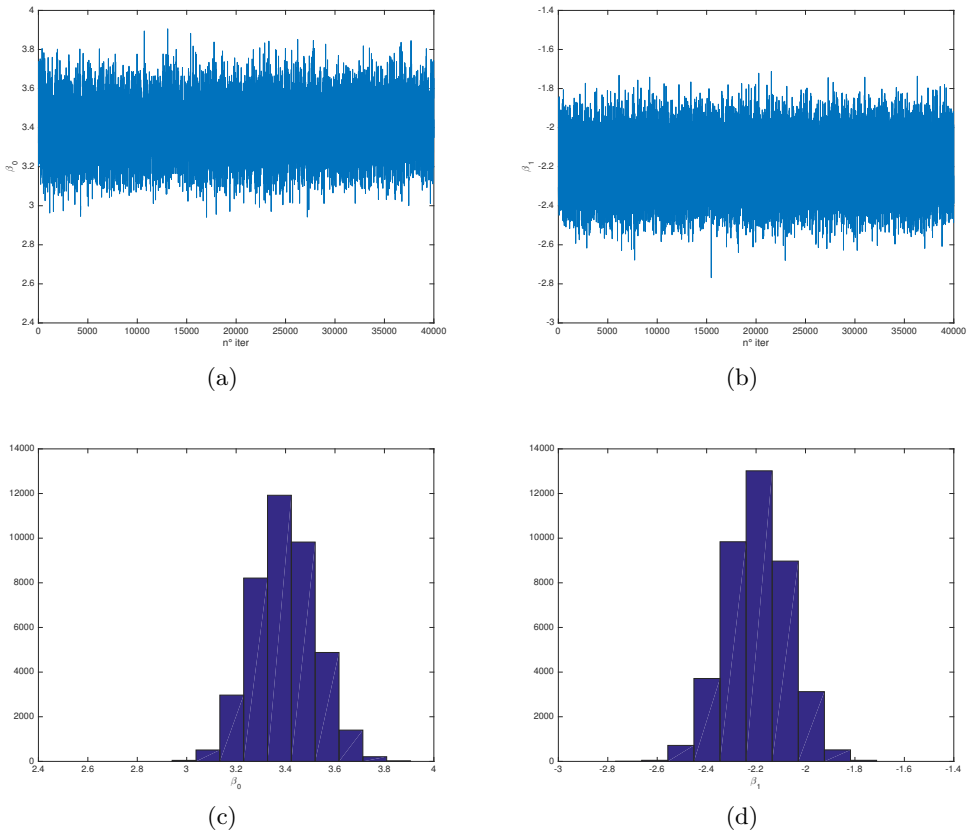


FIGURE3.15: Posterior sample (top) and posterior histogram (bottom) for the simulation study of a count data regression with $\beta_0 = 3.5$ (left) and $\beta_1 = -2.2$ (right) and sample size $n = 300$.

As above highlighted, the procedure can be applied to multiple regressors, Figure 3.16 shows the posterior samples and posterior histograms for a scenario with $\beta_0 = 1.5$, $\beta_1 = -1.0$ and $\beta_2 = 0.4$. For a sample of $n = 300$, and with the same setting of the Gibbs sampler used in the previous illustration, we see a good mixing of the chains as well as good inferential results. In particular, the three means for β_0, β_1 and β_2 are, respectively, 1.5, -0.9 and 0.4, with respective 95% credible intervals $(1.3, 1.7)$, $(-1.2, -0.7)$ and $(0.1, 0.6)$.

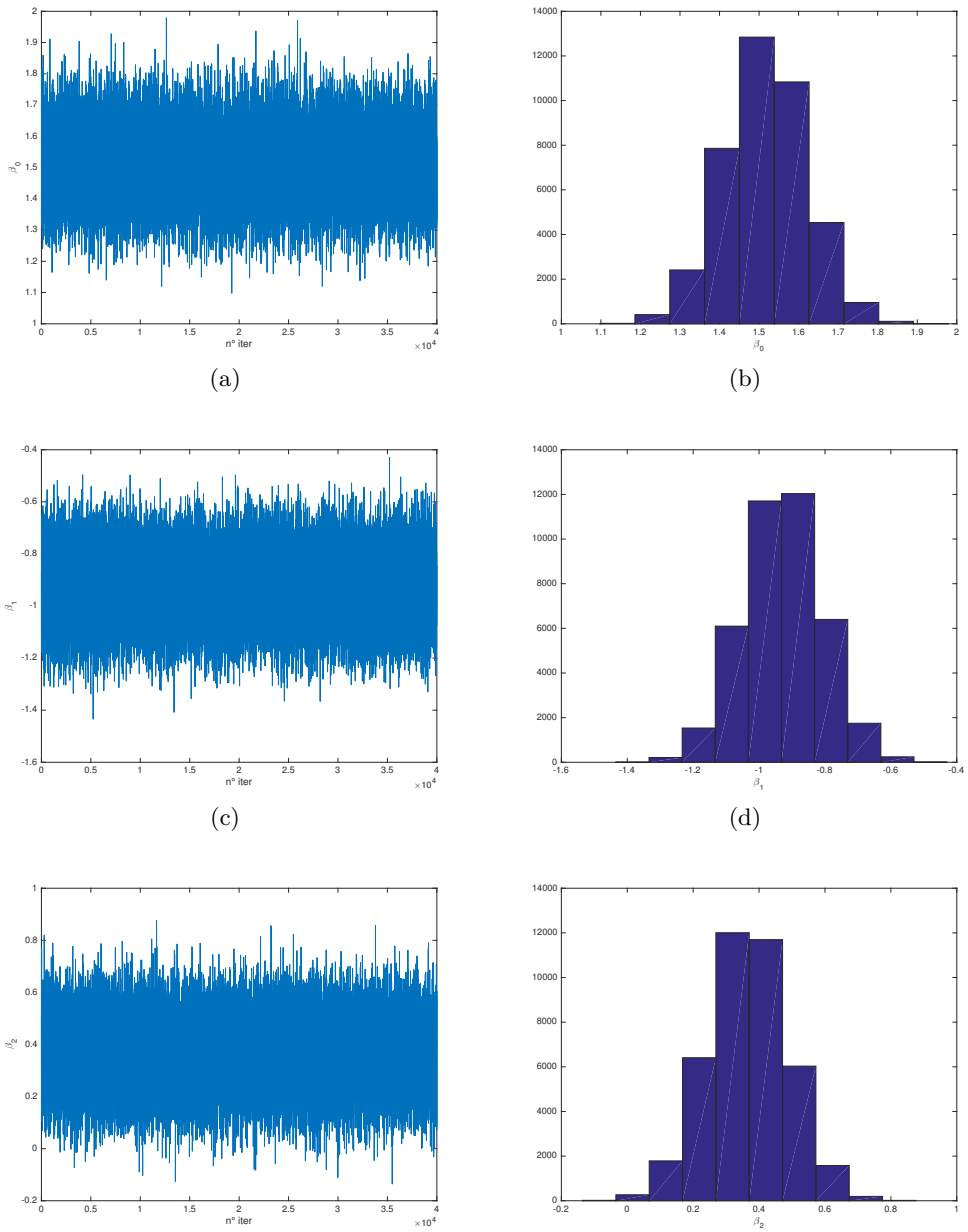


FIGURE3.16: Posterior sample (left) and posterior histogram (right) for the simulation study of a count data regression with $\beta_0 = 1.5$ (top), $\beta_1 = -1.0$ (middle) and $\beta_2 = 0.4$ (bottom) and sample size $n = 300$.

3.7 Applications to text analysis

To illustrate our data augmentation algorithm, we use the Yule-Simon distribution to model word frequency in five novels: *Ulysses* by James Joyce, *Don Quixote* by Miguel de Cervantes, *Moby Dick* by Herman Melville, *War and Peace* by Leo Tolstoi and *Les Miserables* by Victor Hugo. All texts are the English version present in the website of the Gutenberg Project (http://www.gutenberg.org). We have selected the above novels as they have been analysed in Garcia Garcia (2011), and we can compare our results with the author's.

The key information for each data set is n, the number in the text of distinct words in the text (see Table 3.10), and $\mathbf{k}$, the frequency at which each of the words appears in the text.

The inferential procedure consists in the Gibbs sampling algorithm introduced in Section 3.6. For each text, we run three chains, from different starting points, for 10,000 iterations and a burn-in period of 1,000 iterations. The convergence of the sampler has been assessed by graphical means (e.g. progressive means, Gelman and Rubin's plot) and numerical means, such as the Gelman and Rubin's convergence diagnostic and the Geweke's convergence diagnostic. The summary of a posterior for each text are shown in Table 3.10, where we have reported the posterior mean and median, and the 95% credible interval of the posterior. Figure 3.17 shows the posterior chain and the posterior histogram for two of the analysed texts: the Ulysses and the Don Quixote.

To support our conclusions, we compare our estimation results with the ones obtained by applying the fixed-point algorithm proposed by Garcia Garcia (2011). We have implemented the above algorithm on the data available to us, and the right column of Table 3.10 reports the maximum likelihood estimates for each text. First, we note that our fixed-

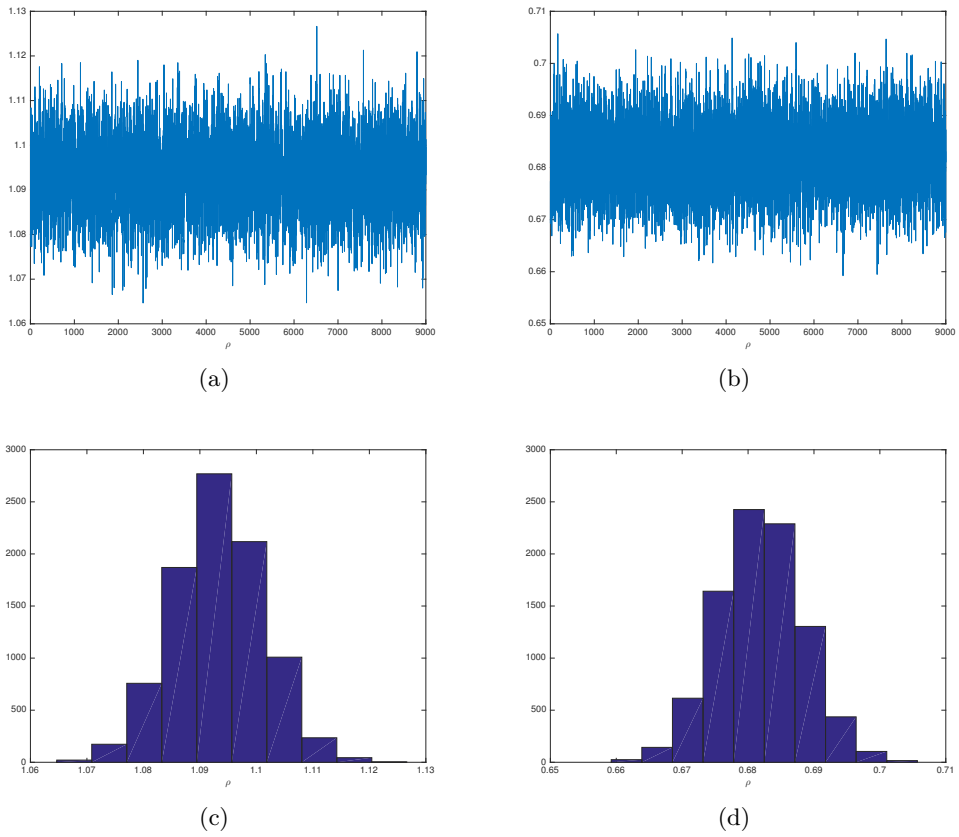


FIGURE3.17: Posterior sample and posterior histogram for the frequency of words analysis for the Ulysses (left) and the Don Quixote (right).

Novel	n	Mean	Median	95% C.I.	Fixed-Point Alg
Ulysses	29,841	1.09	1.09	(1.08,1.11)	1.09
Don Quixote	15,180	0.68	0.68	(0.67,0.70)	0.68
Moby Dick	17,221	0.88	0.88	(0.86,0.89)	0.88
War and Peace	18,239	0.63	0.63	(0.62,0.64)	0.63
Les Miserables	23,248	0.69	0.69	(0.68,0.70)	0.69

Table 3.10: Summary statistics of the posterior distributions for the parameter ρ for frequency of words compared with the fixed point algorithm.

point estimates are very similar to the results in Garcia Garcia (2011), with the exception of the Don Quixote where we have used a different version of the text. Second, and most important, the mean of our posterior is virtually identical to the estimate in Garcia Garcia (2011).

3.8 Discussions

It is surprising how, from time to time, the Bayesian literature presents gaps even for problems which appear to be straightforward. The Yule–Simon distribution has undoubtedly many possibilities of application, as the discussed examples and the refereed papers show, and therefore demanded for a satisfactory discussion within the Bayesian framework.

Given the importance that objective Bayesian analysis can have in applications, and not only (Berger, 2006), we have presented two priors which are suitable in scenarios with minimal prior information. The first prior is the Jeffreys prior which, as it is well known, has the appealing property of being invariant under monotone differentiable transformations of the parameter of interest. The second prior is derived considering the loss in information one would incur if the 'wrong' model was selected. Although the latter requires a discretization of the parameter space, we have shown through simulation studies that the performance of the yielded posterior are very similar, both between the Jeffreys and the loss-based prior, and between different structures of the discretized parameter space. This is not surprising as both priors, i.e. the Jeffreys and the loss-based, have a similar behaviour, in the sense that they increase as the parameter α increases.

We have limited our analysis to the case where the shape parameter of the Yule–Simon distribution, ρ, is strictly larger than one. Doing so, we allow for a more convenient parametrization of the distribution where the new parameter $\alpha = (\rho - 1)/\rho$ has the interpretation of being the probability that the next observation takes a value not observed

before.

Besides through a simulation study, we have compared the objective priors by applying them on three data sets: the first related to financial data, the second to surnames in the US and the third one on the number of hits in the music industry. All comparisons allowed to show that the two proposed objective priors lead to similar results, in terms of posterior distributions. For obvious reasons, we have not considered if the choice of the Yule–Simon is the *best* model to represent the data, but limited our analysis to make inference for the unknown parameter α.

On the other side, the data augmentation algorithm introduced in Section 3.6 performs an efficient and fast estimation of the shape parameter of the Yule–Simon distribution. The simulation study presented in Section 3.6, which discussed both a single i.i.d. sample and a count data regression sample, shows a clear out-performance of the Bayesian approach against the appropriate frequentist procedures. This is particularly true for relatively small sample sizes, rendering the Bayesian inference for the Yule–Simon distribution attractive to practitioners.

For the real data examples discussed in this note, where the sample sizes are large, we see equivalent results of the proposed data augmentation algorithm and of the fixed-point algorithm. This, somehow, validates both approaches.

We are interested, in the future works, to expand the literature related to the Yule–Simon distribution, in particular, to work with a multivariate expansion of it. On the other side, in the last years, the literature on integer time series has increased importance and the Yule–Simon distribution can be applied to Autoregressive models and integer GARCH models.

Bibliography

Berger, J. (2006), "The case for objective Bayesian analysis," *Bayesian Analysis*, 1, 385–402.

Berk, R. (1966), "Limiting behaviour of posterior distributions when the model is incorrect," *Ann. of Math. Statist.*, 37, 51–58.

Chung, K. and Cox, R. (1994), "A stochastic model of superstardom: an application of the Yule distribution," *Review of Economics and Statistics*, 76, 771–775.

Gallardo, D. I., Gomex, H. W., and Bolfarine, H. (2016), "A new cure rate model based on the Yule-Simon distribution with application to a melanoma data set," *Journal of Applied Statistics*, 10.1080/02664763.2016.1194385.

Garcia Garcia, J. M. (2011), "A fixed-point algorithm to estimate the Yule-Simon distribution parameter," *Applied Mathematics and Computation*, 217, 8560–8566.

Gradshteyn, I. and Ryzhik, I. (2007), *Table of Integrals, Series and Products*, Academic Press.

Jeffreys, H. (1961), *Theory of Probability*, London: Oxford University Press.

Maruka, Y. E., Shnerb, N. M., and Kessler, D. A. (2010), "Universal features of surname

distribution in a subsample of a growing population," *Journal of Theoretical Biology*, 262, 245–256.

Merhav, N. and Feder, M. (1998), "Universal prediction," *IEEE Trans. Inf. Theory*, 44, 2124–2147.

Simon, H. A. (1955), "On a class of skew distribution functions," *Biometrika*, 42, 425–440.

Spierdijk, L. and Voorneveld, M. (2009), "Superstars without Talent? The Yule Distribution Controversy," *The Review of Economics and Statistics*, 91, 648–652.

Villa, C. and Walker, S. (2015), "An Objective approach to prior mass function for discrete parameter spaces," *Journal of the American Statistical Association*, 110, 1072–1082.

Yee, T. W. (2008), "The VGAM Package," *R News*, 8, 28–39.

Yee, T. W. (2014), "Reduced-rank vector generalized linear models with two linear predictors," *Computational Statistics and Data Analysis*, 71, 889–902.

Yee, T. W. (2015), *Vector Generalized Linear and Additive Models: With an implementation in R*, New York, USA: Springer.

Yee, T. W. (2016), *Package VGAM for R*.

Yule, G. U. (1925), "A Mathematical theory of evolution, based on the conclusion of Dr. J.C. Willis," *Philosophical Transactions of the Royal Society of London, Series B*, 213, 21–87.

Appendix D

Technical Details of Chapter 3

D.1 Proof of Theorem

Proof of Theorem 3.3.1. First of all, we note that

$$\frac{\partial^2 \log(f(k;\alpha))}{\partial \alpha^2} = \frac{1}{(1-\alpha)^2} + \frac{2}{(1-\alpha)^3}\left[\psi^{(0)}\left(\frac{1}{1-\alpha}+1\right) - \psi^{(0)}\left(\frac{1}{1-\alpha}+k+1\right)\right] + \frac{1}{(1-\alpha)^4}\left[\psi^{(1)}\left(\frac{1}{1-\alpha}+1\right) - \psi^{(1)}\left(\frac{1}{1-\alpha}+k+1\right)\right],$$

where $\psi^{(i)}$ is the polygamma function:

$$\psi^{(i)}(x) = \frac{\partial^{i+1}}{\partial x^{i+1}} \log(\Gamma(x)) = (-1)^{i+1} i! \sum_{k=0}^{\infty} \frac{1}{(x+k)^{i+1}} \quad i = 1, 2, \ldots.$$

It's easy to see that

$$\psi^{(0)}\left(\frac{1}{1-\alpha}+1\right) - \psi^{(0)}\left(\frac{1}{1-\alpha}+k+1\right) = \sum_{j=1}^{k} \frac{1}{\frac{1}{1-\alpha}+j}$$

and

$$\psi^{(1)}\left(\frac{1}{1-\alpha}+1\right)-\psi^{(1)}\left(\frac{1}{1-\alpha}+k+1\right)=\sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}.$$

Therefore, we have that the Fisher information is:

$$\mathcal{I}(\alpha)=-\frac{1}{(1-\alpha)^2}+\frac{2}{(1-\alpha)^3}\mathbb{E}_\alpha\left[\sum_{j=1}^{k}\frac{1}{\frac{1}{1-\alpha}+j}\right]$$
$$-\frac{1}{(1-\alpha)^4}\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right]. \tag{D.1}$$

In order to compute the Jeffreys prior, we need compute the two expected value of equation (D.1) separately.

$$\mathbb{E}_\alpha\left[\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)}\right]=\sum_{k=1}^{\infty}\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right)$$
$$=\sum_{j=1}^{\infty}\frac{1}{(\frac{1}{1-\alpha}+j)}\sum_{k=j}^{\infty}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right). \tag{D.2}$$

The second summation with respect to k in equation (D.2) can be rewritten as:

$$\sum_{k=j}^{\infty}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right)=\sum_{k=j}^{\infty}\frac{1}{1-\alpha}\int_0^1 x^{k-1}(1-x)^{\frac{1}{1-\alpha}}dx$$
$$=\sum_{l=0}^{\infty}\frac{1}{1-\alpha}\int_0^1 x^{l}x^{j-1}(1-x)^{\frac{1}{1-\alpha}}dx$$
$$=\frac{1}{1-\alpha}\int_0^1 x^{j-1}(1-x)^{\frac{1}{1-\alpha}-1}dx$$
$$=\frac{\Gamma(j)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+j)}. \tag{D.3}$$

Finally, we have that :

$$\begin{aligned}
\mathbb{E}_\alpha\left[\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)}\right] &= \Gamma(\frac{1}{1-\alpha}+1)\sum_{j=1}^{\infty}\frac{\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j)(\frac{1}{1-\alpha}+j)} \\
&= \Gamma\left(\frac{1}{1-\alpha}+1\right)\sum_{j=1}^{\infty}\frac{\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)} \\
&= \sum_{j=1}^{\infty}\left(\frac{1-\alpha}{1-\alpha}\right)\frac{\Gamma(\frac{1}{1-\alpha}+1)\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)} \\
&= (1-\alpha)\sum_{j=1}^{\infty}\frac{1}{1-\alpha}\frac{\Gamma(\frac{1}{1-\alpha}+1)\Gamma(j)}{\Gamma(\frac{1}{1-\alpha}+j+1)} = \\
&= (1-\alpha), \qquad \text{(D.4)}
\end{aligned}$$

where the summation $\sum_{j=1}^{\infty}(\frac{1}{1-\alpha})B(j,\frac{1}{1-\alpha}+1)=1$, since we are summing over all the possible values of the probability function of the Yule-Simon distribution.

As we have done with the first expected value of (D.1), now we compute the second expected value of equation (D.1):

$$\begin{aligned}
\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right] &= \sum_{k=1}^{\infty}\sum_{j=0}^{k-1}\frac{1}{(1+\frac{1}{1-\alpha}+j)^2}\frac{1}{1-\alpha}\frac{\Gamma(k)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+1+k)} \\
&= \sum_{k=1}^{\infty}\sum_{j=1}^{k}\frac{1}{(\frac{1}{1-\alpha}+j)^2}\left(\frac{1}{1-\alpha}\right)\frac{\Gamma(k)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+1+k)} \\
&= \sum_{j=1}^{\infty}\frac{1}{(\frac{1}{1-\alpha}+j)^2}\sum_{k=j}^{\infty}\frac{1}{1-\alpha}B\left(k,\frac{1}{1-\alpha}+1\right) \\
&= \sum_{j=1}^{\infty}\frac{1}{(\frac{1}{1-\alpha}+j)^2}\frac{\Gamma(j)\Gamma(\frac{1}{1-\alpha}+1)}{\Gamma(\frac{1}{1-\alpha}+j)}, \qquad \text{(D.5)}
\end{aligned}$$

where the last equality follows from (D.3).Finally we obtain the following form:

$$\mathbb{E}_\alpha\left[\sum_{j=0}^{k-1}\frac{1}{(\frac{1}{1-\alpha}+1+j)^2}\right] = \sum_{j=1}^{\infty}\frac{B(j,\frac{1}{1-\alpha}+1)}{\frac{1}{1-\alpha}+j}. \qquad \text{(D.6)}$$

The equation (D.6) can be written in a more simple way as a function of an Hypergeometric function.

$$\sum_{j=1}^{\infty} \frac{B(j, \frac{1}{1-\alpha}+1)}{\frac{1}{1-\alpha}+j} = \sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} \int_0^1 x^{j-1}(1-x)^{\frac{1}{1-\alpha}} dx$$

$$= \int_0^1 (1-x)^{\frac{1}{1-\alpha}} \sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} x^{j-1} dx. \tag{D.7}$$

Looking at the summation we have:

$$\sum_{j=1}^{\infty} \frac{1}{\frac{1}{1-\alpha}+j} x^{j-1} dx = \sum_{l=0}^{\infty} \frac{x^l}{\frac{1}{1-\alpha}+l+1}. \tag{D.8}$$

But the denominator can be written as a ratio of Pochhammer representations:

$$\frac{1}{\frac{1}{1-\alpha}+l+1} = \frac{\left(\frac{1}{1-\alpha}+1\right)_l}{\left(\frac{1}{1-\alpha}+2\right)_l} \frac{1}{\frac{1}{1-\alpha}+1}.$$

Hence the equation (D.8) is written as:

$$\sum_{l=0}^{\infty} \frac{x^l}{\frac{1}{1-\alpha}+l+1} = \frac{1}{\frac{1}{1-\alpha}+1} \sum_{l=0}^{\infty} \frac{\left(\frac{1}{1-\alpha}+1\right)_l}{\left(\frac{1}{1-\alpha}+2\right)_l} x^l \frac{(1)_l}{l!} =$$

$$= \frac{1}{\frac{1}{1-\alpha}+1} \, {}_2F_1\left(1, \frac{1}{1-\alpha}+1, \frac{1}{1-\alpha}+2, x\right),$$

where ${}_2F_1(\alpha, \beta, \gamma, x)$ is the hypergeometric function. So we have that equation (D.7) can be rewritten as:

$$\int_0^1 \frac{(1-x)^{\frac{1}{1-\alpha}}}{\frac{1}{1-\alpha}+1} \, {}_2F_1\left(1, \frac{1}{1-\alpha}+1, \frac{1}{1-\alpha}+2, x\right) =$$

$$= \frac{(1-\alpha)^2}{(2-\alpha)^2} \, {}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right),$$

where the last equality follows from 7.512.5 of Gradshteyn and Ryzhik (2007). Summing up,

$$\mathcal{I}(\alpha) = -\frac{1}{(1-\alpha)^2} + \frac{2}{(1-\alpha)^3}(1-\alpha) +$$

$$-\frac{1}{(1-\alpha)^4}\frac{(1-\alpha)^2}{(2-\alpha)^2}\,{}_3F_2\left(1, \frac{1}{1-\alpha}+1, 1; \frac{1}{1-\alpha}+2, \frac{1}{1-\alpha}+2; 1\right)$$

and this concludes the proof. □