THE ISBA BULLETIN

Vol. 22 No. 4 — December 2015

The official bulletin of the International Society for Bayesian Analysis

A MESSAGE FROM THE PRESIDENT

A MESSAGE FROM THE PRESIDENT

- Alexandra M. Schmidt -
ISBA President, 2015
alex@im.ufrj.br

"In the New Year, you carry all the experiences of the past years and that is the greatest power of every New Year! This year again, you are less student and more master!" - Mehmet Murat ildan

I came across this sentence and found it quite appropriate to start my last presidential address for the Bulletin. We Bayesians use our experiences to infer and make decisions. And this can be applied to our everyday lives. The more attention we give to our past experiences, the greater the chances of making choices that lead to happiness. For the New Year, I wish you health and strength to look carefully at your experiences, to be able to dream new dreams and, most importantly, to work to make them become true.

I am about to end my role as ISBA's President. Now it is time to look back, review where we were, and understand where we are heading. My aims as ISBA's President have been to ensure that ISBA kept involved in the activities set by former Presidents and to propose new goals and activities.

One of the challenges faced by ISBA during this year was the organization of ISBA 2016. The Program Council, chaired by Michele Guindani, together with the Local Organizing Committee, chaired by Stefano Cabras, have been doing a wonderful job to make sure that we have a wonderful meeting, following the Bayesian tradition of combining a superb scientific program with a astonishing location. For more details on ISBA 2016 see page 5.

Bayesian Analysis (BA), the **open-access** electronic journal sponsored by ISBA continues to perform really well. It ranked 7th in the JCR in 2014 (among 119 journals in Statistics & Probability). Marina Vanucci, the Editor-in-Chief of BA, and the Associate Editors have been doing a wonderful job keeping the high quality of the papers published there. The Executive Board of ISBA, understands the importance of BA and encourages all initiatives that strengthen the contributions of the journal to the scientific community. Starting next year there will be changes that we believe will improve BA's visibility. For details see page 12.

In this issue

'live'. The new production system allows readers to read articles in a timely manner and to cite them appropriately. All this of course has come with a price, as ISBA is now incurring increased costs for the new production system, in addition to those for hosting the Journal on Project Euclid. BA authors can help ISBA defray such costs by paying the voluntary article charges. I strongly encourage everybody to do so.

In closing, I wish to thank the Editors I have worked with for making the review process running smoothly: Ming-Hui Chen, David Dahl, David Heckerman, Lurdes Inoue, Antonietta Mira, Igor Pruenster, Bruno Sanso, Dan Spitzner, Mark Steel, and Kert Viele, as well as Valen Jonson and Sonia Petrone, who were also on the board in the first part of my 3-year term. The editors are supported by a larger number of associated editors and a vast number of referees. Kassie Fronczyk has been serving as the Managing Editor, and has done a great job chasing people up. Thank you all for your support and dedication to our Journal. You have made these 3 years very enjoyable! ▲

SOFTWARE HIGHLIGHT

A Menu-Driven Software Package for Bayesian Regression Analysis[1]

George Karabatsos
georgek@uic.edu

1 Introduction

Regression modeling is ubiquitous in empirical areas of scientific research. This is because most research questions can be asked in terms of how a dependent variable changes as a function of one or more covariates (predictors). Applications of regression modelling involve either prediction analysis, categorical data analysis, causal analysis, meta-analysis, survival analysis of censored data, spatial data analysis, time-series analysis, item response theory (IRT) analysis, and/or other types of regression analyses.

Bayesian Regression: Nonparametric and Parametric Models, is a free stand-alone, user-friendly, and fully menu-driven software package that can be used to perform data analysis using any one of over 80 Bayesian regression models, without having to write code. Currently, the software includes Bayesian infinite-mixture regression models, with mixture distribution assigned a prior distribution defined by either a Dirichlet process (Ferguson, 1973) (defining an ANOVA/linear DDP mixture model; De Iorio, et al. 2004; Müller, et al. 2005), the Pitman-Yor (1997) process, the normalized stable process (Kingman, 1975), the beta 2-parameter process (Ishwaran & Zarepour, 2000), the normalized inverse-Gaussian process (Lijoi et al., 2005), and processes defined either by geometric mixture weights (Fuentes-Garcia, et al. 2009) or covariate-dependent, ordered-probits regression mixture weights (Karabatsos & Walker, 2012b). The software also provides various parametric (finite-dimensional) Bayesian normal regression models, including normal linear models, and normal mixture models that define either 2-level or 3-level random-effects models (or HLMs: Hierarchical linear models). All of the nonparametric and parametric mixture regression models can handle multi-level data, and perform mixing on either the intercept parameter, or on the intercept and slope coefficient parameters.

All the nonparametric and parametric regression models of the software can handle either continuous, binary, ordinal dependent variables (including probit and logit models), as well as continuous dependent variables subject to either left, right, and/or interval censoring. The software allows the user to perform, as a function of covariates, various posterior predictive inferences of the dependent variable, including the mean, median, quantiles, variance, probability density function (p.d.f.), cumulative distribution function (c.d.f.), survival function, hazard function, and cumulative hazard function. Therefore, the software not

[1] Supported by NSF-MMS research grant SES-1156372.

only provides a traditional mean-based regression analysis, but also provides median regression, quantile regression, density regression, and survival analysis. Many of these models are assigned spike-and-slab priors to provide automatic variable (covariate) selection inference from the posterior distribution (e.g., George & McCulloch, 1997). Finally, the software also includes versions of the Bayesian infinite-mixture models for density estimation.

The software implements Markov chain Monte Carlo (MCMC) sampling methods to perform inference of the posterior distribution and posterior functionals. MCMC procedures for the infinite-mixture models are based on the slice-sampler of Kalli et al. (2011). Inference for all regression models are based on standard Gibbs and Metropolis sampling MCMC methods for the normal linear model (e.g., Denison et al., 2002; see Karabatsos & Walker, 2012a, 2012b).

2 Regression Models Provided by the Software

As is well-known, the Bayesian linear model, with conjugate multivariate normal (N) inverse-gamma (IG) prior, is defined by $y_i \,|\, \mathrm{x}_i \sim \mathrm{N}(\mathrm{x}_i^t\beta, \sigma^2)$, with $\beta \sim \mathrm{N}(0, \sigma^2\mathrm{diag}(\mathrm{v}_\beta))$, and $\sigma^2 \sim \mathrm{IG}(a, b)$. This model can be extended by the multilevel, normal mixture model. For example, for N_h groups of observations respectively indexed by $h = 1, \ldots, N_h$, the 2-level normal random effects linear model may be specified by $Y_{i(h)} \,|\, \mathrm{x}_{i(h)} \sim \mathrm{N}(y_{i(h)} \,|\, \mathrm{x}_{i(h)}^t\beta + \mathrm{x}_{i(h)}^t u_h, \sigma^2)$ for $i(h) = 1, \ldots, n_h$, with prior distributions $\beta \sim \mathrm{N}(0, \sigma^2\mathrm{diag}(\mathrm{v}_\beta))$, $u_h \,|\, T \sim \mathrm{N}(0, T)$, $\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2)$, and $T \sim \mathrm{IW}(p+3, s_0\mathrm{I}_{p+1})$ (IW: inverted-Wishart distribution). The assumptions of normal linear models, namely, the linearity of covariate effects, and the normality (or unimodality) of regression error and random effects distributions, can be violated by data, affecting model fit and posterior inferences.

Therefore, the Bayesian nonparametrics (BNP) field has developed many infinite-mixture regression models that relax these assumptions (e.g., Hjort, et al. 2010; Karabatsos & Walker, 2012a, 2012b; Mitra & Müller, 2015). A highly-flexible, BNP infinite-mixture regression model has the general form:

$$\begin{aligned} f_{G_\mathrm{x}}(y \,|\, \mathrm{x}; \zeta) &= \int f(y \,|\, \mathrm{x}; \psi, \theta(\mathrm{x}))\mathrm{d}G_\mathrm{x}(\theta) \\ &= \sum_{j=1}^{\infty} f(y \,|\, \mathrm{x}; \psi, \theta_j(\mathrm{x}))\omega_j(\mathrm{x}), \quad (1) \\ G_\mathrm{x}(B) &= \sum_{j=1}^{\infty} \omega_j(\mathrm{x})\delta_{\theta_j(\mathrm{x})}(B), \quad (2) \\ & \quad \forall B \in \mathcal{B}(\Theta), \end{aligned}$$

given a covariate (x) dependent, discrete mixing distribution G_x; kernel (component) densities $f(y \,|\, \mathrm{x}; \psi, \theta_j(\mathrm{x}))$ with component indices $j = 1, 2, \ldots$, respectively; with fixed parameters ψ; and with component parameters $\theta_j(\mathrm{x})$ having sample space Θ; and given mixing weights $(\omega_j(\mathrm{x}))_{j=1}^{\infty}$ that sum to 1 at every $\mathrm{x} \in \mathcal{X}$, with $\mathcal{X}$ the covariate space. In the model (1), the covariate-dependent mixing distribution is a random probability measure that has the general form given by (2), and is therefore an example of a species sampling model (Pitman, 1995), where $\delta_z(\cdot)$ denotes the degenerate probability measure $\delta_z(z) = 1$, with $\delta_z(B) = 1$ if $z \in B$; and $\mathcal{B}(\Theta)$ denotes the Borel σ-field of the space Θ of θ. The infinite-mixture mixture model (1) is completed by the specification of a prior distribution $\Pi(\zeta)$ on the space $\Omega_\zeta = \{\zeta\}$ of the model parameter, given by $\zeta = (\psi, (\theta_j(\mathrm{x}), \omega_j(\mathrm{x}))_{j=1}^{\infty}, \mathrm{x} \in \mathcal{X})$.

For example, if G_x is assigned a Dirichlet process prior distribution with $G \sim \mathcal{DP}(\alpha, G_0)$ (assuming $G_\mathrm{x} := G$), then the Bayesian model (1), with prior distribution $\Pi(\zeta)$ on all model parameters, is called a Dirichlet process mixture (DPM) model (Lo, 1984) where the mixture distribution (2) has stick-breaking mixture weights of the form $\omega_j = \upsilon_j \prod_{l=1}^{j-1}(1 - \upsilon_l)$, with $\upsilon_j \,|\, \alpha \sim \mathrm{Beta}(1, \alpha)$ and $\theta_j \,|\, G_0 \sim G_0$ for $j = 1, 2, \ldots$ (assuming $\theta_j(\mathrm{x}) := \theta_j$) (Sethuraman, 1994). Furthermore, if the kernel density functions are specified by normal density functions, with $f(y \,|\, \mathrm{x}; \psi, \theta_j(\mathrm{x})) := \mathrm{N}(y \,|\, \mathrm{x}^\mathrm{t}\beta_j, \sigma^2)$ for $j = 1, 2, \ldots$, then the model defines an ANOVA/linear DDP model, a DPM mixture of random intercept linear regression models. As alternative stick-breaking priors for G, a Pitman-Yor process assumes $\upsilon_j \sim \mathrm{Beta}(1 - a, b + ja)$ with $0 \leq a < 1$ and $b > -a$; the normalized σ-stable process assumes $\upsilon_j \sim \mathrm{Beta}(1 - a, b + ja)$; the beta 2-parameter process assumes $\upsilon_j \sim \mathrm{Beta}(a, b)$ for $a, b > 0$; the normalized inverse-Gaussian process assumes $\upsilon_j = \upsilon_{1j}/(\upsilon_{1j} + \upsilon_{0j})$, $\upsilon_{1j} \sim \mathrm{GIG}(c^2/\{\prod_{l=1}^{j-1}(1 - \upsilon_l)\}^{1(j>1)}, 1, -j/2)$, and $\upsilon_{0j} \sim \mathrm{IG}(1/2, 2)$, with

GIG the generalized inverse-Gaussian distribution (Favaro, et al., 2012); and the geometric weights prior assumes $\upsilon_j := \upsilon$ and $\upsilon \sim \text{Be}(a, b)$ (Fuentes-Garcia, et al. 2009). A covariate-dependent process prior for G_{x} is defined by ordinal regression mixture weights

$$\omega_j(\text{x}) = \Phi\left(\frac{j - \text{x}^{\text{t}}\beta_\omega}{\sqrt{\exp(\text{x}^{\text{t}}\lambda_\omega)}}\right) - \Phi\left(\frac{j - \text{x}^{\text{t}}\beta_\omega - 1}{\sqrt{\exp(\text{x}^{\text{t}}\lambda_\omega)}}\right), \tag{3}$$

where $j = 0, \pm1, \pm2, \ldots$ with $(\beta_\omega, \lambda_\omega) \sim \text{N}(\mu, \Sigma)$, and $\Phi(\cdot)$ the standard normal c.d.f. (Karabatsos & Walker, 2012b).

3 Using the Software

After starting the *Bayesian Regression* software, the user may click the File menu to import a comma-delimited data file. Before running a Bayesian regression analysis of the data set, the software user can mouse-click menu options: (1) to inspect, describe, and explore the data variables via basic descriptive statistics (e.g., means, standard deviations, quantiles/percentiles) and graphs (e.g., scatter plots, box plots, normal Q-Q plots, etc.); and (2) to pre-process the data of the dependent variable and/or the covariate(s) before including the variable(s) into the BNP regression model for data analysis. Examples of data pre-processing includes constructing new dummy indicator (0 or 1) variables, two-way interaction variables, and (e.g., thin-plate) spline covariates from variables in the data set (e.g., to set up a spatial data analysis); constructing lagged dependent variables (of chosen lag order) as covariates to set up a Bayesian auto-regression time series analysis; constructing propensity score covariates to set up a causal analysis with a regression model; and to perform other variable transformations (e.g. z-score transformations). Finally, the user can perform a nearest neighbor hot-deck imputation of missing data.

Next, the user can then click menu options to select a Bayesian regression model for data analysis, and the model's prior distribution parameters, dependent variable, and covariates. If necessary, the user can, for her/his chosen model, select the (level-2 and possibly level-3) grouping variables (if s/he chose a multilevel model); select the observation weights variable (e.g., to set up a meta-analysis); and/or select the variables that indicate whether the dependent variable is left-censored, right-censored, interval-censored, or uncensored (e.g., to set up a survival analysis). Finally, after the user makes all of these menu-selections for her/his chosen Bayesian model for data analysis, the software immediately presents a graphic of the model, along with all the variables that were selected for this model (e.g., lists of dependent variables, covariate(s)).

Then, the software user can click a button to run the MCMC sampling algorithm for the menu-selected Bayesian model, for a user-chosen number of MCMC sampling iterations. After all the MCMC sampling iterations have completed, the software automatically opens a text output file that summarizes the basic results of the data analysis (calculated from the generated MCMC samples). Results include point-estimates of the (marginal) posterior distributions of the model's parameters (e.g., posterior mean, median, variance, quantiles, etc.), and summaries of the model's predictive fit to the data. The user can also click other menu options to produce graphical and text output of the results. They include density plots, box plots, scatter plots, trace plots, and various plots of (marginal) posterior distributions of model parameters and fit statistics. Certain menu options allow the user to perform MCMC convergence analyses, via univariate trace plots of MCMC samples of model parameters, and via batch means (or sub-sampling) analyses to calculate the 95% Monte Carlo confidence intervals for the posterior point estimates. Other menu options allow the user, as a function of one or more covariates of interest, to graph and tabulate various posterior predictive estimates of the dependent variable, including the dependent variable mean, median, quantiles, variance, p.d.f., c.d.f., survival function, hazard function, and cumulative hazard function.

4 Software Access

The Bayesian Regression software can be downloaded and installed for use, from: http://georgek.people.uic.edu/BayesSoftware.html. This web-page also provides links to a user's manual (paper) for the software, which includes a textbook-style review of Bayesian statistical inference, and the MCMC methods, as well as software-based illustrations and exercises of

Bayesian regression modeling for the analysis of real data. The Help menu provides user instructions, and describes all the 83 models and example data sets that are currently provided by the software.

References

[1] DeIorio, M., Müller, P., Rosner, G., & MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99, pp. 205–215.

[2] Denison, D., Holmes, C., Mallick, B., & Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* New York: John Wiley and Sons.

[3] Favaro, S., Lijoi, A., & Prünster, I. (2012). On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, 99, pp. 663–674.

[4] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, pp. 209–230.

[5] Fuentes-Garcia, R., Mena, R., & Walker, S. (2009). A nonparametric dependent process for Bayesian regression. *Statistics and Probability Letters*, 79, pp. 1112–1119.

[6] George, E., & McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, pp. 339–373.

[7] Hjort, N., Holmes, C., Müller, P., & Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

[8] Ishwaran, H., & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, pp. 371–390.

[9] Kalli, M., Griffin, J., & Walker, S. (2011). Slice sampling mixture models. *Statistics and Computing*, 21, pp. 93–105.

[10] Karabatsos, G., & Walker, S. (2012a). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics*, 6, pp. 2038–2068.

[11] Karabatsos, G., & Walker, S. (2012b). Bayesian nonparametric mixed random utility models. *Computational Statistics and Data Analysis*, 56, pp. 1714–1722.

[12] Kingman, J. (1975). Random discrete distributions. *Journal of the Royal Statistical Society, Series B*, 37, pp. 1–22.

[13] Lijoi, A., Mena, R., & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100, pp. 1278–1291.

[14] Lo, A. (1984). On a class of Bayesian nonparametric estimates. *Annals of Statistics*, 12, pp. 351–357.

[15] Mitra, R., & Müller, P. (2015). *Nonparametric Bayesian Methods in Biostatistics*. New York: Springer-Verlag.

[16] Müller, P., Rosner, G., De Iorio, M., & MacEachern, S. (2005). A nonparametric Bayesian model for inference in related studies. *Applied Statistics*, 54, pp. 611–626.

[17] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102, pp. 145–158.

[18] Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25, pp. 855–900.

[19] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, pp. 639–650. ▲