Behav Res (2017) 49:335–362
DOI 10.3758/s13428-016-0711-7

A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation

George Karabatsos[1]

Published online: 8 March 2016

Abstract Most of applied statistics involves regression analysis of data. In practice, it is important to specify a regression model that has minimal assumptions which are not violated by data, to ensure that statistical inferences from the model are informative and not misleading. This paper presents a stand-alone and menu-driven software package, Bayesian Regression: Nonparametric and Parametric Models, constructed from MATLAB Compiler. Currently, this package gives the user a choice from 83 Bayesian models for data analysis. They include 47 Bayesian nonparametric (BNP) infinite-mixture regression models; 5 BNP infinite-mixture models for density estimation; and 31 normal random effects models (HLMs), including normal linear models. Each of the 78 regression models handles either a continuous, binary, or ordinal dependent variable, and can handle multi-level (grouped) data. All 83 Bayesian models can handle the analysis of weighted observations (e.g., for meta-analysis), and the analysis of left-censored, right-censored, and/or interval-censored data. Each BNP infinite-mixture model has a mixture distribution assigned one of various BNP prior distributions, including priors defined by either the Dirichlet process, Pitman-Yor process (including the normalized stable process), beta (two-parameter) process, normalized inverse-Gaussian process, geometric weights prior, dependent Dirichlet process, or the dependent infinite-probits prior. The software user can mouse-click to select a Bayesian model and perform data analysis via Markov chain Monte Carlo (MCMC) sampling. After the sampling completes, the software automatically opens text output that reports MCMC-based estimates of the model's posterior distribution and model predictive fit to the data. Additional text and/or graphical output can be generated by mouse-clicking other menu options. This includes output of MCMC convergence analyses, and estimates of the model's posterior predictive distribution, for selected functionals and values of covariates. The software is illustrated through the BNP regression analysis of real data.

Keywords Bayesian · Regression · Density estimation

Introduction

Regression modeling is ubiquitous in empirical areas of scientific research, 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 modeling involve either prediction analysis (e.g., Dension et al., 2002; Hastie et al. 2009), categorical data analysis (e.g., Agresti, 2002), causal analysis (e.g., Imbens, 2004; Imbens & Lemieux, 2008; Stuart, 2010), meta-analysis (e.g., Cooper, et al. 2009), survival analysis of censored data (e.g., Klein & Moeschberger, 2010), spatial data analysis (e.g., Gelfand, et al., 2010), time-series analysis (e.g., Prado & West, 2010), item response theory (IRT) analysis (e.g., van der Linden, 2015), and/or other types of regression analyses.

These applications often involve either the normal random-effects (multi-level) linear regression model (e.g., hierarchical linear model; HLM). This general model assumes that the mean of the dependent variable changes

✉ George Karabatsos
gkarabatsos1@gmail.com

[1] University of Illinois, Chicago, IL, USA

linearly as a function of each covariate; the distribution of the regression errors follows a zero-mean symmetric continuous (e.g., normal) distribution; and the random regression coefficients are normally distributed over pre-defined groups, according to a normal (random-effects) mixture distribution. Under the ordinary linear model, this mixture distribution has variance zero. For a discrete dependent variable, all of the previous assumptions apply for the underlying (continuous-valued) latent dependent variable. For example, a logit model (probit model, resp.) for a binary-valued (0 or 1) dependent variable implies a linear model for the underlying latent dependent variable, with error distribution assumed to follow a logistic distribution with mean 0 and scale 1 (normal distribution with mean 0 and variance 1, resp.) (e.g., Dension et al., 2002).

If data violate any of these linear model assumptions, then the estimates of regression coefficient parameters can be misleading. As a result, much research has devoted to the development of more flexible, Bayesian nonparametric (BNP) regression models. Each of these models can provide a more robust, reliable, and rich approach to statistical inference, especially in common settings where the normal linear model assumptions are violated. Excellent reviews of BNP models are given elsewhere (e.g., Walker, et al., 1999; Ghosh & Ramamoorthi, 2003; Müller & Quintana, 2004; Hjort, et al., 2010; Mitra & Müller, 2015).

A BNP model is a highly-flexible model for data, defined by an infinite (or a very large finite) number of parameters, with parameter space assigned a prior distribution with large supports (Müller & Quintana, 2004). Typical BNP models have an infinite-dimensional, functional parameter, such as a distribution function. According to Bayes' theorem, a set of data updates the prior to a posterior distribution, which conveys the plausible values of the model parameters given the data and the chosen prior. Typically in practice, Markov chain Monte Carlo (MCMC) sampling methods (e.g., Brooks et al., 2011) are used to estimate the posterior distribution (and chosen functionals) of the model parameters.

Among the many BNP models that are available, the most popular models in practice are infinite-mixture models, each having mixture distribution assigned a (BNP) prior distribution on the entire space of probability measures (distribution functions). BNP infinite-mixture models are popular in practice, because they can provide a flexible and robust regression analysis of data, and provide posterior-based clustering of subjects into distinct homogeneous groups, where each subject cluster group is defined by a common value of the (mixed) random model parameter(s). A standard BNP model is defined by the Dirichlet process (infinite-) mixture (DPM) model (Lo, 1984), with mixture distribution assigned a Dirichlet process (DP) (Ferguson 1973) prior distribution on the space of probability measures. Also, often in practice, a BNP model is specified as an infinite-mixture of normal distributions. This is motivated by the well-known fact that any smooth probability density (distribution) of any shape and location can be approximated arbitrarily-well by a mixture of normal distributions, provided that the mixture has a suitable number of mixture components, mixture weights, and component parameters (mean and variance).

A flexible BNP infinite-mixture model need not be a DPM model, but may instead have a mixture distribution that is assigned another BNP prior, defined either by a more general stick-breaking process (Ishwaran & James, 2001; Pitman, 1996), such as the Pitman-Yor (or Poisson-Dirichlet) process (Pitman, 1996; Pitman & Yor, 1997), the normalized stable process (Kingman, 1975), the beta two-parameter process (Ishwaran & Zarepour, 2000); or a process with more restrictive, geometric mixture weights (Fuentes-García, et al., 2009, 2010); or defined by the normalized inverse-Gaussian process (Lijoi et al., 2005), a general type of normalized random measure (Regazzini et al., 2003).

A more general BNP infinite-mixture model can be constructed by assigning its mixture distribution a covariate-dependent BNP prior. Such a BNP mixture model allows the entire dependent variable distribution to change flexibly as a function of covariate(s). The Dependent Dirichlet process (DDP; MacEachern, 1999, 2000, 2001) is a seminal covariate-dependent BNP prior. On the other hand, the infinite-probits prior is defined by a dependent normalized random measure, constructed by an infinite number of covariate-dependent mixture weights, with weights specified by an ordinal probits regression with prior distribution assigned to the regression coefficient and error variance parameters (Karabatsos & Walker, 2012a).

The applicability of BNP models, for data analysis, depends on the availability of user-friendly software. This is because BNP models typically admit complex representations, which may not be immediately accessible to non-experts or beginners in BNP. Currently there are a few nice command-driven R software packages for BNP mixture modeling. The **DPpackage** (Jara et al., 2011) of R (the R Development Core Team, 2015) includes many BNP models, mostly DPM models, that provide either flexible regression or density estimation for data analysis. The package also provides BNP models having parameters assigned a flexible mixture of finite Pólya Trees BNP prior (Hanson, 2006). The **bspmma** R package (Burr, 2012) provides DPM normal-mixture models for meta-analysis. Newer packages have recently arrived to the scene. They include the **BNPdensity** R package (Barrios et al., 2015), which provides flexible mixture models for nonparametric density estimation via more general normalized random measures, with mixture distribution assigned a BNP prior defined by

either a normalized stable, inverse-Gaussian, and generalized gamma process. They also include the **PReMiuM** R package (Liverani et al., 2015) for flexible regression modeling via DPM mixtures and clustering.

The existing packages for BNP modeling, while impressive, still suggest room for improvements, as summarized by the following points.

1. While the existing BNP packages provide many DPM models, they do not provide a BNP infinite-mixture model with mixture distribution assigned any one of the other important BNP priors mentioned earlier. Priors include those defined by the Pitman-Yor, normalized stable, beta, normalized inverse-Gaussian process; or defined by a geometric weights or infinite-probits prior. As exceptions, the **DPpackage** provides a Pitman-Yor process mixture of regressions model for interval-censored data (Jara et al., 2010); whereas the **BNPdensity** package provides models defined by more general normalized random measures, but only for density estimation and not for regression.
2. The **bspmma** R package (Burr, 2012), for meta-analysis, is limited to DPM models that do not incorporate covariate dependence (Burr & Doss, 2005).
3. The **DPpackage** handles interval-censored data, but does not handle left- or right-censored data.
4. While both BNP packages use MCMC sampling algorithms to estimate the posterior distribution of the user-chosen model, each package does not provide options for MCMC convergence analysis (e.g., Flegal & Jones, 2011). A BNP package that provides its own menu options for MCMC convergence analysis would be, for the user, faster and more convenient, and would not require learning a new package (e.g., **CODA** R package; Plummer et al., 2006) to conduct MCMC convergence analyses.
5. Both BNP packages do not provide many options to investigate how the posterior predictive distribution (and chosen functionals) of the dependent variable, varies as a function of one or more covariates.
6. Generally speaking, command-driven software can be unfriendly, confusing, and time-consuming to beginners and to experts. This includes well-known packages for parametric Bayesian analysis including **BUGS** and **OpenBUGS** (Thomas, 1994; Lunn et al., 2009), **JAGS** (2015), **STAN** (2015), **NIMBLE** (2015), and **BIPS** (2015).

In this paper, we introduce a stand-alone and user-friendly software package for BNP modeling, which the author constructed using MATLAB Compiler (Natick, MA). This package, named: **Bayesian Regression: Nonparametric and Parametric Models** (Karabatsos, 2016), provides BNP data analysis in a fully menu-driven software environment that resembles SPSS (I.B.M., 2015).

The software allows the user to mouse-click menu options:

1. To inspect, describe, and explore the variables of the data set, via basic descriptive statistics (e.g., means, standard deviations, quantiles/percentiles) and graphs (e.g., scatter plots, box plots, normal Q-Q plots, kernel density plots, etc.);
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 include constructing new dummy indicator (0 or 1) variables and/or two-way interaction variables from the covariates (variables), along with other options to transform variables; and performing a nearest-neighbor hot-deck imputation (Andridge & Little, 2010) of missing data values in the variables (e.g., covariate(s)).
3. To use list and input dialogs to select, in the following order: the Bayesian model for data analysis; the dependent variable; covariate(s) (if a regression model was selected); parameters of the prior distribution of the model; the (level-2 and possibly level-3) grouping variables (for a multilevel model, if selected); the observation weights variable (if necessary; e.g., to set up a meta-analysis); and the variables describing the nature of the censored dependent variable observations (if necessary; e.g., to set up a survival analysis). The observations can either be left-censored, right-censored, interval-censored, or uncensored. Also, if so desired, the user can easily use point-and-click to quickly highlight and select a large list of covariates for the model, whereas command-driven software requires the user to carefully type (or copy and paste) and correctly-verify the long list of the covariates.

After the user makes these selections, the **Bayesian Regression** software immediately presents a graphic of the user-chosen Bayesian model in the middle of the computer screen, along with all of the variables that were selected for this model (e.g., dependent variables, covariate(s); see #3 above). The explicit presentation of the model is important because BNP models typically admit complex representations. In contrast, the command-driven packages do not provide immediate on-screen presentations of the BNP model selected by the user.

Then the software user can click a button to run the MCMC sampling algorithm for the menu-selected Bayesian model. The user clicks this button after entering a number of MCMC sampling iterations. Immediately 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 (derived from the generated MCMC samples). Results include point-estimates of the (marginal) posterior distributions of the model's parameters, and summaries of the model's predictive fit to the data. Then, the user can click other menu options to produce graphical 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. For each available BNP infinite-mixture model, the software implements standard slice sampling MCMC methods (Kalli et al., 2011) that are suitable for making inferences of the posterior distribution (and chosen functionals) of model parameters.

Next, after a few mouse-clicks of appropriate menu options, the user can perform a detailed MCMC convergence analysis. This analysis evaluates whether a sufficiently-large number of MCMC samples (sampling iterations of the MCMC algorithm) has been generated, in order to warrant the conclusion that these samples have converged to samples from the posterior distribution (and chosen functionals) of the model parameters. More details about how to use the software to perform MCMC convergence analysis is provided in "ANOVA-linear DDP model" and 5.2.

The software also provides menu options to investigate how the posterior predictive distribution (and functionals) of the dependent variable changes as a function of covariates. Functionals of the posterior predictive distribution include: the mean, median, and quantiles to provide a quantile regression analysis; the variance functional to prove a variance regression analysis; the probability density function (p.d.f.) and the cumulative distribution function (c.d.f.) to provide a density regression analysis; and the survival function, hazard function, and the cumulative hazard function, for survival analysis. The software also provides posterior predictive inferences for BNP infinite-mixture models that do not incorporate covariates and only focus on density estimation.

Currently, the **Bayesian Regression** software provides the user a choice from 83 Bayesian models for data analysis. Models include 47 BNP infinite-mixture regression models, 31 normal linear models for comparative purposes, and 5 BNP infinite normal mixture models for density estimation. Most of the infinite-mixture models are defined by normal mixtures.

The 47 BNP infinite-mixture regression models can each handle a dependent variable that is either continuous-valued, binary-valued (0 or 1), or ordinal valued ($c = 0, 1, \ldots, m$), using either a probit or logit version of this model for a discrete dependent variable; with mixture distribution assigned a prior distribution defined either by the Dirichlet process, Pitman-Yor process (including the normalized stable process), beta (2-parameter) process, geometric weights prior, normalized inverse-Gaussian process, or an infinite-probits regression prior; and with mixing done on either the intercept parameter, or on the intercept and slope coefficient parameters, and possibly on the error variance parameter. Specifically, the regression models with mixture distribution assigned a Dirichlet process prior are equivalent to ANOVA/linear DDP models, defined by an infinite-mixture of normal distributions, with a covariate-dependent mixture distribution defined by independent weights (DeIorio et al., 2004; Müller et al., 2005). Similarly, the models with mixture distribution, instead, assigned a different BNP prior distribution (process) mentioned above, implies a covariate-dependent version of that process. See "ANOVA-linear DDP model" for more details. Also, some of the infinite-mixture regression models, with covariate-dependent mixture distribution assigned a infinite-probits prior, have spike-and-slab priors assigned to the coefficients of this BNP prior, based on stochastic search variable selection (SSVS) (George & McCulloch, 1993, 1997). In addition, the 5 BNP infinite normal mixture models, for density estimation, include those with mixture distribution assigned a BNP prior distribution that is defined by either one of the 5 BNP process mentioned above (excluding infinite-probits).

Finally, the 31 Bayesian linear models of the **Bayesian Regression** software include ordinary linear models, 2-level, and 3-level normal random-effects (or HLM) models, for a continuous dependent variable; probit and logit versions of these linear models for either a binary (0 or 1) or ordinal ($c = 0, 1, \ldots, m$) dependent variable; and with mixture distribution specified for the intercept parameter, or for the intercept and slope coefficient parameters.

The outline for the rest of the paper is as follows. "Overview of Bayesian inference" reviews the Bayesian inference framework. Appendix A reviews the basic probability theory notation and concepts that we use. In "Key BNP regression models", we define two key BNP infinite-mixture regression models, each with mixture distribution assigned a BNP prior distribution on the space of probability measures. The other 50 BNP infinite-mixture models of the **Bayesian Regression** software are extensions of these two key models, and in that section we give an overview of the various BNP priors mentioned earlier. In that section we also describe the Bayesian normal linear model, and a Bayesian normal random-effects linear model (HLM). "Using the Bayesian regression software" gives step-by-step software instructions on how to perform data analysis using a menu-chosen, Bayesian model. "Real data example" illustrates the **Bayesian Regression** software through the analysis of a real data set, using each of the two key BNP models, and a Bayesian linear model.

Appendix B provides a list of exercises that the software user can work through in order to practice BNP modeling on several example data sets, available from the software. These data-analysis exercises address applied problems in prediction analysis, categorical data analysis, causal analysis, meta-analysis, survival analysis of censored data, spatial data analysis, time-series analysis, and item response theory analysis. The last section ends with conclusions.

Overview of Bayesian inference

In a given research setting where it is of interest to apply a regression data analysis, a sample data set is of the form $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$. Here, n is the sample size of the observations, respectively indexed by $i = 1, \ldots, n$, where y_i is the ith observation of the dependent variable Y_i, corresponding to an observed vector of p observed covariates[1] $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\intercal$. A constant (1) Term is included in $\mathbf{x}$ for future notational convenience.

A regression model assumes a specific form for the probability density (or p.m.f.) function $f(y \mid \mathbf{x}; \boldsymbol{\zeta})$, conditionally on covariates $\mathbf{x}$ and model parameters denoted by a vector, $\boldsymbol{\zeta} \in \Omega_{\boldsymbol{\zeta}}$, where $\Omega_{\boldsymbol{\zeta}} = \{\boldsymbol{\zeta}\}$ is the parameter space. For any given model parameter value $\boldsymbol{\zeta} \in \Omega_{\boldsymbol{\zeta}}$, the density $f(y_i \mid \mathbf{x}_i; \boldsymbol{\zeta})$ is the likelihood of y_i given $\mathbf{x}_i$, and $L(\mathcal{D}_n ; \boldsymbol{\zeta}) = \prod_{i=1}^n f(y_i \mid \mathbf{x}_i; \boldsymbol{\zeta})$ is the likelihood of the full data set $\mathcal{D}_n$ under the model. A Bayesian regression model is completed by the specification of a prior distribution (c.d.f.) $\Pi(\boldsymbol{\zeta})$ over the parameter space $\Omega_{\boldsymbol{\zeta}}$, and $\pi(\boldsymbol{\zeta})$ gives the corresponding probability density of a given parameter $\boldsymbol{\zeta} \in \Omega_{\boldsymbol{\zeta}}$.

According to Bayes' theorem, after observing the data $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$, the plausible values of the model parameter $\boldsymbol{\zeta}$ is given by the posterior distribution. This distribution defines the posterior probability density of a given parameter $\boldsymbol{\zeta} \in \Omega_{\boldsymbol{\zeta}}$ by:

$$\pi(\boldsymbol{\zeta} \mid \mathcal{D}_n) = \frac{\prod_{i=1}^n f(y_i \mid \mathbf{x}_i; \boldsymbol{\zeta}) \mathrm{d}\Pi(\boldsymbol{\zeta})}{\int_{\Omega_{\boldsymbol{\zeta}}} \prod_{i=1}^n f(y_i \mid \mathbf{x}_i; \boldsymbol{\zeta}) \mathrm{d}\Pi(\boldsymbol{\zeta})}. \tag{1}$$

Conditionally on a chosen value of the covariates $\mathbf{x} = (1, x_1, \ldots, x_p)^\intercal$, the posterior predictive density of a future observation y_{n+1}, and the corresponding posterior predictive c.d.f. ($F(y \mid \mathbf{x})$), mean (expectation, $\mathbb{E}$), variance ($\mathbb{V}$), median, uth quantile ($Q(u \mid \mathbf{x})$, for some chosen $u \in [0, 1]$, with $Q(.5 \mid \mathbf{x})$ the conditional median), survival function (S), hazard function (H), and cumulative hazard function (Λ), is given respectively by:

$$f_n(y \mid \mathbf{x}) = \int f(y \mid \mathbf{x}; \boldsymbol{\zeta}) \mathrm{d}\Pi(\boldsymbol{\zeta} \mid \mathcal{D}_n), \tag{2a}$$

$$F_n(y \mid \mathbf{x}) = \int_{Y \le y} f(y \mid \mathbf{x}; \boldsymbol{\zeta}) \mathrm{d}\Pi(\boldsymbol{\zeta} \mid \mathcal{D}_n), \tag{2b}$$

$$\mathbb{E}_n(Y \mid \mathbf{x}) = \int y \mathrm{d}F_n(y \mid \mathbf{x}), \tag{2c}$$

$$\mathbb{V}_n(Y \mid \mathbf{x}) = \int \{y - \mathbb{E}_n(Y \mid \mathbf{x})\}^2 \mathrm{d}F_n(y \mid \mathbf{x}), \tag{2d}$$

$$Q_n(u \mid \mathbf{x}) = F_n^{-1}(u \mid \mathbf{x}), \tag{2e}$$

$$S_n(y \mid \mathbf{x}) = 1 - F_n(y \mid \mathbf{x}), \tag{2f}$$

$$H_n(y \mid \mathbf{x}) = f_n(y \mid \mathbf{x}) / \{1 - F_n(y \mid \mathbf{x})\}, \tag{2g}$$

$$\Lambda_n(y \mid \mathbf{x}) = -\log\{1 - F_n(y \mid \mathbf{x})\}. \tag{2h}$$

Depending on the choice of posterior predictive functional from (2a–2h), a Bayesian regression analysis can provide inferences in terms of how the mean (2c), variance (2d), quantile (2e) (for a given choice $u \in [0, 1]$), p.d.f. (2a), c.d.f. (2b), survival function (2f), hazard function (2g), or cumulative hazard function (2h), of the dependent variable Y, varies as a function of the covariates $\mathbf{x}$. While the mean functional $\mathbb{E}_n(Y \mid \mathbf{x})$ is conventional for applied regression, the choice of functional $\mathbb{V}_n(Y \mid \mathbf{x})$ pertains to variance regression; the choice of function $Q_n(u \mid \mathbf{x})$ pertains to quantile regression; the choice of p.d.f. $f_n(y \mid \mathbf{x})$ or c.d.f. $F_n(y \mid \mathbf{x})$ pertains to Bayesian density (distribution) regression; and the choice of survival $S_n(y \mid \mathbf{x})$ or a hazard function ($H_n(y \mid \mathbf{x})$ or $\Lambda_n(y \mid \mathbf{x})$) pertains to survival analysis.

In practice, the predictions of the dependent variable Y (for a chosen functional from (2a–2h)), can be easily viewed (in a graph or table) as a function of a subset of only one or two covariates. Therefore, for practice we need to consider predictive methods that involve such a small subset of covariates. To this end, let $\mathbf{x}_{\mathcal{S}}$ be a focal subset of the covariates $(x_1, \ldots, x_p)$, with $\mathbf{x}_{\mathcal{S}}$ also including the constant (1) term. Throughout, the term "focal subset of the covariates" is a short phrase that refers to covariates that are of interest in a given posterior predictive analysis. Also, let $\mathbf{x}_{\mathcal{C}}$ be the non-focal, complement set of q (unselected) covariates. Then $\mathbf{x}_{\mathcal{S}} \cap \mathbf{x}_{\mathcal{C}} \neq \emptyset$ and $\mathbf{x} = \mathbf{x}_{\mathcal{S}} \cup \mathbf{x}_{\mathcal{C}}$.

It is possible to study how the predictions of a dependent variable Y vary as a function of the focal covariates $\mathbf{x}_{\mathcal{S}}$, using one of four automatic methods. The first two methods are conventional. They include the *grand-mean centering method*, which assumes that the non-focal covariates $\mathbf{x}_{\mathcal{C}}$ is defined by the mean in the data $\mathcal{D}_n$, with $\mathbf{x}_{\mathcal{C}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{\mathcal{C}i}$; and the *zero-centering method*, which assumes that the non-focal covariates are given by $\mathbf{x}_{\mathcal{C}} := \mathbf{0}_q$ where $\mathbf{0}_q$ is a

[1] For each Bayesian model for density estimation, from the software, we assume $\mathbf{x} = 1$.

vector of q zeros. Both methods coincide if the observed covariates $\{\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\intercal\}_{i=1}^n$ in the data $\mathcal{D}_n$ have average $(1, 0, \ldots, 0)^\intercal$. This is the case if the covariate data $\{x_{ik}\}_{i=1}^n$ have already been centered to have mean zero, for $k = 1, \ldots, p$.

The partial dependence method (Friedman, 2001, Section 8.2) is the third method for studying how the predictions of a dependent variable Y varies as a function of the focal covariates $\mathbf{x}_\mathcal{S}$. In this method, the predictions of Y, conditionally on each value of the focal covariates $\mathbf{x}_\mathcal{S}$, are averaged over data ($\mathcal{D}_n$) observations $\{\mathbf{x}_{\mathcal{C}i}\}_{i=1}^n$ (and effects) of the non-focal covariates $\mathbf{x}_\mathcal{C}$. Specifically, in terms of the posterior predictive functionals (2a–2h), the averaged prediction of Y, conditionally on a value of the covariates $\mathbf{x}_\mathcal{S}$, is given respectively by:

$$f_n(y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n f_n(y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i}), \tag{3a}$$

$$F_n(y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n F_n(y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i}), \tag{3b}$$

$$\mathbb{E}_n(Y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_n(Y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i}), \tag{3c}$$

$$\mathbb{V}_n(Y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_n(Y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i}), \tag{3d}$$

$$Q_n(u \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n F_n^{-1}(u \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i}), \tag{3e}$$

$$S_n(y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n \{1 - F_n(y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i})\}, \tag{3f}$$

$$H_n(y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n H_n(y \mid \mathbf{x}_\mathcal{S}, \mathbf{x}_{\mathcal{C}i}), \tag{3g}$$

$$\Lambda_n(y \mid \mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n -\log\{1 - F_n(y \mid \mathbf{x}_S, \mathbf{x}_{\mathcal{C}i})\}. \tag{3h}$$

The equations above give, respectively, the (partial dependence) posterior predictive density, c.d.f., mean, variance, quantile (at $u \in [0, 1]$), survival function, hazard function, and cumulative hazard function, of Y, conditionally on a value $\mathbf{x}_S$ of the focal covariates. As a side note pertaining to causal analysis, suppose that the focal covariates include a covariate, denoted T, along with a constant (1) term, so that $\mathbf{x}_S = (1, t)$. Also suppose that the covariate T is a binary-valued (0,1) indicator of treatment receipt, versus non-treatment receipt. Then the estimate of a chosen (partial-dependence) posterior predictive functional of Y under treatment ($T = 1$) from (3a–3h), minus that posterior predictive functional under control ($T = 0$), provides an estimate of the causal average treatment effect (CATE). This is true provided that the assumptions of unconfoundedness and overlap hold (Imbens, 2004).

The partial-dependence method can be computationally-demanding, as a function of sample size (n), the dimensionality of $\mathbf{x}_S$, the number of $\mathbf{x}_S$ values considered when investigating how Y varies as a function of $\mathbf{x}_S$, and the number of MCMC sampling iterations performed for the estimation of the posterior distribution (density (1)) of the model parameters. In contrast, the *clustered partial dependence method*, the fourth method, is less computationally-demanding. This method is based on forming K-means cluster centroids, $\{\mathbf{x}_{\mathcal{C}t}\}_{t=1}^K$, of the data observations $\{\mathbf{x}_{\mathcal{C}i}\}_{i=1}^n$ of the non-focal covariates $\mathbf{x}_\mathcal{C}$, with $K = \text{floor}(\sqrt{n/2})$ clusters as a rule-of-thumb. Then the posterior predictions of Y, conditionally on chosen value of the covariate subset $\mathbf{x}_\mathcal{S}$, is given by any one of the chosen posterior functionals (3a–3h) of interest, after replacing $\frac{1}{n}\sum_{i=1}^n$ with $\frac{1}{K}\sum_{t=1}^K$, and $\mathbf{x}_{\mathcal{C}i}$ with $\mathbf{x}_{\mathcal{C}t}$.

The predictive fit of a Bayesian regression model, to a set of data, $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$, can be assessed on the basis of the posterior predictive expectation (2c) and variance (2d). First, the standardized residual fit statistics of the model are defined by:

$$r_i = \frac{y_i - \mathbb{E}_n(Y \mid \mathbf{x}_i)}{\sqrt{\mathbb{V}_n(Y \mid \mathbf{x}_i)}}, \quad i = 1, \ldots, n. \tag{4}$$

An observation y_i can be judged as an outlier under the model, when its absolute standardized residual $|\, r_i \,|$ exceeds 2 or 3. The proportion of variance explained in the dependent variable Y, by a Bayesian model, is measured by the R-squared statistic:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n (y_i - \mathbb{E}_n[Y \mid \mathbf{x}_i])^2}{\sum_{i=1}^n \left\{ y_i - \left(\frac{1}{n}\sum_{i=1}^n y_i \right) \right\}^2} \right). \tag{5}$$

Also, suppose that it is of interest to compare M regression models, in terms of predictive fit to the given data set $\mathcal{D}_n$. Models are indexed by $\underline{m} = 1, \ldots, M$, respectively. For each model $\underline{m}$, a global measure of predictive fit is given by the mean-squared predictive error criterion:

$$D(\underline{m}) = \sum_{i=1}^n \{y_i - \mathbb{E}_n(Y \mid \mathbf{x}_i, \underline{m})\}^2 + \sum_{i=1}^n \mathbb{V}_n(Y \mid \mathbf{x}_i, \underline{m}) \tag{6}$$

(Laud & Ibrahim, 1995; Gelfand & Ghosh, 1998). The first term in Eq. 6 measures model goodness-of-fit to the data $\mathcal{D}_n$, and the second term is a model complexity penalty. Among a set of M regression models compared, the model with the best predictive fit for the data $\mathcal{D}_n$ is identified as the one that has the smallest value of $D(\underline{m})$.

MCMC methods

In practice, a typical Bayesian model does not admit a closed-form solution for its posterior distribution (density function of the form Eq. 1). However, the posterior distribution, along with any function of the posterior distribution of interest, can be estimated through the use of Monte Carlo methods. In practice, they usually involve Markov chain Monte Carlo (MCMC) methods (e.g., Brooks et al., 2011). Such a method aims to construct a discrete-time

Harris ergodic Markov chain $\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^{S}$ with stationary (posterior) distribution $\Pi(\boldsymbol{\zeta} \mid \mathcal{D}_n)$, and ergodicity is ensured by a proper (integrable) prior density function $\pi(\boldsymbol{\zeta})$ (Robert & Casella, 2004, Section 10.4.3). A realization $\boldsymbol{\zeta}^{(s)}$ from the Markov chain can be generated by first specifying partitions (blocks) $\boldsymbol{\zeta}_b$ $(b = 1, \ldots, B)$ of the model's parameter $\boldsymbol{\zeta}$, and then simulating a sample from each of the full conditional posterior distributions $\Pi(\boldsymbol{\zeta}_b \mid \mathcal{D}_n, \boldsymbol{\zeta}_c, c \neq b)$, in turn for $b = 1, \ldots, B$. Then, as $S \to \infty$, the Markov (MCMC) chain $\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^{S}$ converges to samples from the posterior distribution $\Pi(\boldsymbol{\zeta} \mid \mathcal{D}_n)$. Therefore, in practice, the goal is to construct an MCMC chain (samples) $\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^{S}$ for a sufficiently-large finite S.

MCMC convergence analyses can be performed in order to check whether a sufficiently-large number (S) of sampling iterations has been run, to warrant the conclusion that the resulting samples ($\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^{S}$) have converged (practically) to samples from the model's posterior distribution. Such an analysis may focus only on the model parameters of interest for data analysis, if so desired. MCMC convergence can be investigated in two steps (Geyer, 2011). One step is to inspect, for each of these model parameters, the univariate trace plot of parameter samples over the MCMC sampling iterations. This is done to evaluate MCMC mixing, i.e., the degree to which MCMC parameter samples explores the parameter's support in the model's posterior distribution. Good mixing is suggested by a univariate trace plot that appears stable and "hairy" over MCMC iterations.[2] The other step is to conduct, for each model parameter of interest, a batch means (or subsampling) analysis of the MCMC samples, in order to calculate 95 % Monte Carlo Confidence Intervals (95 % MCCIs) of posterior point-estimates of interest (such as marginal posterior means, variances, quantiles, etc., of the parameter) (Flegal & Jones, 2011). For a given (marginal) posterior point-estimate of a parameter, the 95 % MCCI half-width size reflects the imprecision of the estimate due to Monte Carlo sampling error. The half-width becomes smaller as number of MCMC sampling iterations grows. In all, MCMC convergence is confirmed by adequate MCMC mixing and practically-small 95 % MCCIs half-widths (e.g., .10 or .01) for the (marginal) posterior point-estimates of parameters (and chosen functionals) of interest. If adequate convergence cannot be confirmed after an MCMC sampling run, then additional MCMC sampling iterations can be run until convergence is obtained for the (updated) total set of MCMC samples.

[2]The CUSUM statistic, which ranges between 0 and 1, is a measure of the "hairiness" of a univariate trace plot of a model parameter (see Brooks, 1998). A CUSUM value of .5 indicates optimal MCMC mixing.

For each BNP infinite-mixture model, the **Bayesian Regression** software estimates the posterior distribution (and functionals) of the model on the basis of a general slice-sampling MCMC method, which can handle the infinite-dimensional model parameters (Kalli et al., 2011). This slice-sampling method does so by introducing latent variables into the likelihood function of the infinite-mixture model, such that, conditionally on these variables, the model is finite-dimensional and hence tractable by a computer. Marginalizing over the distribution of these latent variables recovers the original likelihood function of the infinite-mixture model.

We now describe the MCMC sampling methods that the software uses to sample from the full conditional posterior distributions of the parameters, for each model that the software provides. For each DPM model, the full conditional posterior distribution of the unknown precision parameter (α) is sampled from a beta mixture of two gamma distributions (Escobar & West, 1995). For each BNP infinite-mixture model based on a DP, Pitman-Yor process (including the the normalized stable process), or beta process prior, the full conditional posterior distribution of the mixture weight parameters are sampled from appropriate beta distributions (Kalli et al., 2011). Also, for the parameters of each of the 31 linear models, and for the linear parameters of each of the BNP infinite-mixture models, the software implements (direct) MCMC Gibbs sampling of standard full conditional posterior distributions, derived from the standard theories of the Bayesian normal linear, probit, and logit models, as appropriate (Evans, 1965; Lindley & Smith, 1972; Gilks et al., 1993; Albert & Chib, 1993; Bernardo & Smith, 1994; Denison et al., 2002; Cepeda & Gamerman, 2001; O'Hagan & Forster, 2004; Holmes & Held, 2006; George & McCulloch, 1997; e.g., see Karabatsos & Walker, 2012a, b). When the full conditional posterior distribution of the model parameter(s) is non-standard, the software implements a rejection sampling algorithm. Specifically, it implements an adaptive random-walk Metropolis-Hastings (ARWMH) algorithm (Atchadé & Rosenthal, 2005) with normal proposal distribution, to sample from the full conditional posterior distribution(s) of the mixture weight parameter of a BNP geometric weights infinite-mixture model; the mixture weight parameter of a BNP normalized inverse-Gaussian process mixture model, using the equivalent stick-breaking representation of this process (Favaro et al., 2012). Also, for BNP infinite-mixture models and normal random-effects models that assign a uniform prior distribution to the variance parameter for random intercepts (or means), the software implements the slice sampling (rejection) algorithm with stepping-out procedure (Neal, 2003), in order to sample from the full conditional posterior distribution of this parameter. Finally, for computational speed considerations, we use the ARWMH algorithm instead of Gibbs

sampling, in order to sample from the full conditional posterior distributions for the random coefficient parameters (the intercepts u_{0h}; and possibly the $u_{kh}, k = 0, 1, \ldots, p$, as appropriate, for groups $h = 1, \ldots, H$) in a normal random-effects (or random intercepts) HLM; and for the random coefficients ($\boldsymbol{\beta}_j$) or random intercept parameters (β_{0j}) in a BNP infinite-mixture regression model, as appropriate (Karabatsos & Walker, 2012a, b).

The given data set ($\mathcal{D}_n$) may consist of censored dependent variable observations (either left-, right-, and/or interval-censored). If the software user indicates the censored dependent variable observations (see "Running the (12 steps for data analysis)", Step 6), then the software adds a Gibbs sampling step to the MCMC algorithm, that draws from the full-conditional posterior predictive distributions (density function (2a)) to provide multiple MCMC-based imputations of these missing censored observations (Gelfand et al., 1992; Karabatsos & Walker, 2012a).

Finally, the software implements Rao-Blackwellization (RB) methods (Gelfand & Mukhopadhyay, 1995) to compute estimates of the linear posterior predictive functionals from (2a–2h) and (3a–3h). In contrast, the quantile functional $Q_n(u \mid \mathbf{x})$ is estimated from order statistics of MCMC samples from the posterior predictive distribution of Y given $\mathbf{x}$. The 95 % posterior credible interval of the quantile functional $Q(u \mid \mathbf{x})$ can be viewed in a PP-plot (Wilk & Gnanadesikan, 1968) of the 95 % posterior interval of the c.d.f. $F(u \mid \mathbf{x})$, using available software menu options. The hazard functional $H_n(y \mid \mathbf{x})$ and the cumulative hazard functional $\Lambda_n(y \mid \mathbf{x})$ are derived from RB estimates of the linear functionals $f_n(y \mid \mathbf{x})$ and $F_n(y \mid \mathbf{x})$. The same is true for the partial-dependence functionals $Q_n(u \mid \mathbf{x}_{\mathcal{S}})$, $H_n(y \mid \mathbf{x}_{\mathcal{S}})$, and $\Lambda_n(y \mid \mathbf{x}_{\mathcal{S}})$.

Key BNP regression models

A BNP infinite-mixture regression model has the general form:

$$f_{G_{\mathbf{x}}}(y \mid \mathbf{x}; \boldsymbol{\zeta}) = \int f(y \mid \mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta}(\mathbf{x})) \mathrm{d}G_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} f(y \mid \mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta}_j(\mathbf{x})) \omega_j(\mathbf{x}), \quad (7)$$

given a covariate ($\mathbf{x}$) dependent, discrete mixing distribution $G_{\mathbf{x}}$; kernel (component) densities $f(y \mid \mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta}_j(\mathbf{x}))$ with component indices $j = 1, 2, \ldots$, respectively; with fixed parameters $\boldsymbol{\psi}$; and with component parameters $\boldsymbol{\theta}_j(\mathbf{x})$ having sample space Θ; and given mixing weights $(\omega_j(\mathbf{x}))_{j=1}^{\infty}$ that sum to 1 at every $\mathbf{x} \in \mathcal{X}$, with $\mathcal{X}$ the covariate space.

In the infinite-mixture model (7), the covariate-dependent mixing distribution is a random probability measure that has the general form,[3]

$$G_{\mathbf{x}}(B) = \sum_{j=1}^{\infty} \omega_j(\mathbf{x}) \delta_{\boldsymbol{\theta}_j(\mathbf{x})}(B), \quad \forall B \in \mathcal{B}(\Theta), \quad (8)$$

and is therefore an example of a species sampling model (Pitman, 1995).

The mixture model (7) is completed by the specification of a prior distribution $\Pi(\boldsymbol{\zeta})$ on the space $\Omega_{\boldsymbol{\zeta}} = \{\boldsymbol{\zeta}\}$ of the infinite-dimensional model parameter, given by:

$$\boldsymbol{\zeta} = (\boldsymbol{\psi}, (\boldsymbol{\theta}_j(\mathbf{x}), \omega_j(\mathbf{x}))_{j=1}^{\infty}, \mathbf{x} \in \mathcal{X}). \quad (9)$$

The BNP infinite-mixture regression model (7)–(8), completed by the specification of a prior distribution $\Pi(\boldsymbol{\zeta})$, is very general and encompasses, as special cases: fixed- and random-effects linear and generalized linear models (McCullagh & Nelder, 1989; Verbeke & Molenberghs, 2000; Molenberghs & Verbeke, 2005), finite-mixture latent-class and hierarchical mixtures-of-experts regression models (McLachlan & Peel, 2000; Jordan & Jacobs, 1994), and infinite-mixtures of Gaussian process regressions (Rasmussen et al., 2002).

In the general BNP model (7)–(8), assigned prior $\Pi(\boldsymbol{\zeta})$, the kernel densities $f(y \mid \mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta}_j(\mathbf{x}))$ may be specified as covariate independent, with: $f(y \mid \mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\theta}_j(\mathbf{x})) := f(y \mid \boldsymbol{\psi}, \boldsymbol{\theta}_j)$; and may not contain fixed parameters $\boldsymbol{\psi}$, in which case $\boldsymbol{\psi}$ is null. Also for the model, covariate dependence is not necessarily specified for the mixing distribution, so that $G_{\mathbf{x}} := G$. No covariate dependence is specified for the mixing distribution if and only if both the component parameters and the mixture weights are covariate independent, with $\boldsymbol{\theta}_j(\mathbf{x}) := \boldsymbol{\theta}_j$ and $\omega_j(\mathbf{x}) := \omega_j$. The mixing distribution $G_{\mathbf{x}}$ is covariate dependent if the component parameters $\boldsymbol{\theta}_j(\mathbf{x})$ or the mixture weights $\omega_j(\mathbf{x})$ are specified as covariate dependent.

Under the assumption of no covariate dependence in the mixing distribution, with $G_{\mathbf{x}} := G$, the Dirichlet process (Ferguson, 1973) provides a standard and classical choice of BNP prior distribution on the space of probability measures $\mathcal{G}_{\Theta} = \{G\}_{\Theta}$ on the sample space Θ. The Dirichlet process is denoted $\mathcal{DP}(\alpha, G_0)$ with precision parameter α and baseline distribution (measure) G_0. We denote $G \sim \mathcal{DP}(\alpha, G_0)$ when the random probability measure G is assigned a $\mathcal{DP}(\alpha, G_0)$ prior distribution on $\mathcal{G}_{\Theta}$. Under the

[3] Throughout, $\delta_{\boldsymbol{\theta}}(\cdot)$ denotes a degenerate probability measure (distribution) with point mass at $\boldsymbol{\theta}$, such that $\boldsymbol{\theta}^* \sim \delta_{\boldsymbol{\theta}}$ and $\Pr(\boldsymbol{\theta}^* = \boldsymbol{\theta}) = 1$. Also, $\delta_{\boldsymbol{\theta}}(B) = 1$ if $\boldsymbol{\theta} \in B$ and $\delta_{\boldsymbol{\theta}}(B) = 0$ if $\boldsymbol{\theta} \notin B$, for $\forall B \in \mathcal{B}(\Theta)$.

$\mathcal{DP}(\alpha, F_0)$ prior, the (prior) mean and variance of G are given respectively by Ferguson (1973):

$$\mathbb{E}[G(B)\,|\,\alpha, G_0] = \frac{\alpha G_0(B)}{\alpha} = G_0(B), \tag{10a}$$

$$\mathbb{V}[G(B)\,|\,\alpha, G_0] = \frac{G_0(B)[1 - G_0(B)]}{\alpha + 1}, \quad \forall B \in \mathcal{B}(\Theta). \tag{10b}$$

For the $\mathcal{DP}(\alpha, G_0)$ prior, Eq. 10a shows that the baseline distribution G_0 represents the prior mean (expectation) of G, and the prior variance of G is inversely proportional to the precision parameter α, as shown in Eq. 10b. The variance of G is increased (decreased, resp.) as α becomes smaller (larger, resp.). In practice, a standard choice of baseline distribution $G_0(\cdot)$ is provided by the normal $\mathrm{N}(\mu, \sigma^2)$ distribution. The $\mathcal{DP}(\alpha, G_0)$ can also be characterized in terms of a Dirichlet (Di) distribution. That is, if $G \sim \mathcal{DP}(\alpha, G_0)$, then:

$$(G(B_1), \ldots, G(B_k))\,|\,\alpha, G_0 \sim \mathrm{Di}(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)), \tag{11}$$

for every choice of $k \geq 1$ (exhaustive) partitions $B_1, \ldots, B_k$ of the sample space, Θ.

The $\mathcal{DP}(\alpha, G_0)$ can also be characterized as a particular "stick-breaking" stochastic process (Sethuraman, 1994; Sethuraman and Tiwari, 1982). A random probability measure (G) that is drawn from the $\mathcal{DP}(\alpha, G_0)$ prior, with $G \sim \mathcal{DP}(\alpha, G_0)$, is constructed by first taking independently and identically distributed (i.i.d.) samples of $(\upsilon, \boldsymbol{\theta})$ from the following beta (Be) and baseline (G_0) distributions:

$$\upsilon_j\,|\,\alpha \sim \mathrm{Be}(1, \alpha), \quad j = 1, 2, \ldots, \tag{12a}$$

$$\boldsymbol{\theta}_j\,|\,G_0 \sim G_0, \quad j = 1, 2, \ldots, \tag{12b}$$

and then using the samples $(\upsilon_j, \boldsymbol{\theta}_j)_{j=1}^{\infty}$ to construct the random probability measure by:

$$G(B) = \sum_{j=1}^{\infty} \omega_j \delta_{\boldsymbol{\theta}_j}(B), \quad \forall B \in \mathcal{B}(\Theta). \tag{12c}$$

Above, the ω_js are mixture weights, particularly, stick-breaking weights constructed by:

$$\omega_j = \upsilon_j \prod_{l=1}^{j-1}(1 - \upsilon_l), \text{ for } j = 1, 2, \ldots, \tag{12d}$$

and they sum to 1 (i.e., $\sum_{j=1}^{\infty}\omega_j = 1$).

More in words, a random probability measure, G, drawn from a $\mathcal{DP}(\alpha, G_0)$ prior distribution on $\mathcal{G}_\Theta = \{G\}_\Theta$, can be represented as infinite-mixtures of degenerate probability measures (distributions). Such a random distribution is discrete with probability 1, which is obvious because the degenerate probability measure ($\delta_{\boldsymbol{\theta}_j}(\cdot)$) is discrete. The locations $\boldsymbol{\theta}_j$ of the point masses are a sample from G_0. The random weights ω_j are obtained from a stick-breaking procedure, described as follows. First, imagine a stick of length 1. As shown in Eq. 12d, at stage $j = 1$ a piece is broken from this stick, and then the value of the first weight ω_1 is set equal to the length of that piece, with $\omega_1 = \upsilon_1$. Then at stage $j = 2$, a piece is broken from a stick of length $1 - \omega_1$, and then the value of the second weight $\omega_2 = \upsilon_2(1 - \upsilon_1)$ is set equal to the length of that piece. This procedure is repeated for $j = 1, 2, 3, 4, \ldots$, where at any given stage j, a piece is broken from a stick of length $1 - \sum_{l=1}^{j-1}\omega_j$, and then the value of the weight ω_j is set equal to the length of that piece, with $\omega_j = \upsilon_j \prod_{l=1}^{j-1}(1 - \upsilon_l)$. The entire procedure results in weights $(\omega_j)_{j=1}^{\infty}$ that sum to 1 (almost surely).

The stick-breaking construction (12a–12d) immediately suggests generalizations of the $\mathcal{DP}(\alpha, G_0)$, especially by means of increasing the flexibility of the prior (12a) for the random parameters $(\upsilon_j)_{j=1}^{\infty}$ that construct the stick-breaking mixture weights (12d). One broad generalization is given by a general stick-breaking process (Ishwaran & James, 2001), denoted $\mathcal{SB}(\mathbf{a}, \mathbf{b}, G_0)$ with positive parameters $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{b} = (b_1, b_2, \ldots)$, which gives a prior on $\mathcal{G}_\Theta = \{G\}_\Theta$. This process replaces the i.i.d. beta distribution assumption in Eq. 12a, with the more general assumption of independent beta (Be) distributions, with

$$\upsilon_j\,|\,a_j, b_j \sim \mathrm{Be}(a_j, b_j), \text{ for } j = 1, 2, \ldots. \tag{13}$$

In turn, there are many interesting special cases of the $\mathcal{SB}(\mathbf{a}, \mathbf{b}, G_0)$ process prior, including:

1. The Pitman-Yor (Poisson-Dirichlet) process, denoted $\mathcal{PY}(a, b, G_0)$, which assumes $a_j = 1 - a$ and $b_j = b + ja$, for $j = 1, 2, \ldots$, in Eq. 13, with $0 \leq a < 1$ and $b > -a$ (Perman et al., 1992; Pitman & Yor, 1997).
2. The beta two-parameter process, which assumes $a_j = a$ and $b_j = b$ in Eq. 13 (Ishwaran & Zarepour, 2000).
3. The normalized stable process (Kingman, 1975), which is equivalent to the $\mathcal{PY}(a, 0, G_0)$ process, with $0 \leq a < 1$ and $b = 0$.
4. The Dirichlet process $\mathcal{DP}(\alpha, G_0)$, which assumes $a_j = 1$ and $b_j = \alpha$ in Eq. 13, and with is equivalent to the $\mathcal{PY}(0, \alpha, G_0)$ process.
5. The geometric weights prior, denoted $\mathcal{GW}(a, b, G_0)$, which assumes in Eq. 13 the equality restriction $\upsilon = \upsilon_j$ for $j = 1, 2, \ldots$, leading to mixture weights (12d) that can be re-written as $\omega_j = \upsilon\,(1 - \upsilon)^{j-1}$, for $j = 1, 2, \ldots$ (Fuentes-García et al. 2009, 2010). These mixture weights may be assigned a beta prior distribution, with $\upsilon \sim \mathrm{Be}(a, b)$.

Another generalization of the $\mathcal{DP}(\alpha, G_0)$ is given by the mixture of Dirichlet process (MDP), defined by the stick-breaking construction (12a–12d), after sampling from prior distributions $\alpha \sim \Pi(\alpha)$ and $\boldsymbol{\vartheta} \sim \Pi(\boldsymbol{\vartheta})$ for the precision and baseline parameters (Antoniak, 1974).

A BNP prior distribution on $\mathcal{G}_\Theta = \{G\}_\Theta$, defined by a Normalized Random Measure (NRM) process, assumes that a discrete random probability measure G, given by Eq. 12c, is constructed by mixture weights that have the form

$$\omega_j = \frac{I_j}{\sum_{l=1}^{\infty} I_l}, \quad j = 1, 2, \ldots; \quad \omega_j \geq 0, \quad \sum_{l=1}^{\infty} \omega_l = 1. \quad (14)$$

The $I_1, I_2, I_3, \ldots$ are the jump sizes of a non-Gaussian Lévy process whose sum is almost surely finite (see e.g. James et al., 2009), and are therefore stationary independent increments (Bertoin, 1998). The $\mathcal{DP}(\alpha, G_0)$ is a special NRM process which makes the gamma (Ga) distribution assumption $\sum_{j=1}^{\infty} I_j \sim \text{Ga}(\alpha, 1)$ (Ferguson, 1973, pp. 218–219).

An important NRM is given by the normalized inverse-Gaussian $\mathcal{NIG}(c, G_0)$ process (Lijoi et al., 2005), which can be characterized as a stick-breaking process (Favaro et al., 2012), defined by the stick-breaking construction (12a–12d), after relaxing the i.i.d. assumption (12a), by allowing for dependence among the v_j distributions, with:

$$v_j = \frac{v_{1j}}{v_{1j} + v_{0j}}, \quad j = 1, 2, \ldots \quad (15a)$$

$$v_{1j} \sim \text{GIG}(c^2/\{\prod_{l=1}^{j-1}(1 - V_l)\}^{\mathbf{1}(j>1)}, 1, -j/2), \quad (15b)$$

$$v_{0j} \sim \text{IG}(1/2, 2). \quad (15c)$$

The random variables (15a) follow normalized generalized inverse-Gaussian (GIG) distributions, with p.d.f. given by equation (4) in Favaro et al. (2012), and Eqs. 15b–15c refer to GIG and inverse-gamma (IG) distributions.

Stick-breaking process priors can be characterized in terms of the clustering behavior that it induces in the posterior predictive distribution of $\boldsymbol{\theta}$. Let $\{\boldsymbol{\theta}_c^* : c = 1, \ldots, k_n \leq n\}$ be the $k_n \leq n$ unique values (clusters) among the n observations of a data set. Let $\Upsilon_n = \{\mathcal{C}_1, \ldots, \mathcal{C}_c, \ldots, \mathcal{C}_{k_n}\}$ be the random partition of the integers $\{1, \ldots, n\}$. Each cluster is defined by $\mathcal{C}_c = \{i : \boldsymbol{\theta}_i = \boldsymbol{\theta}_c^*\} \subset \{1, \ldots, n\}$, and has size $n_c = |\mathcal{C}_c|$, with cluster frequency counts $\mathbf{n}_n = (n_1, \ldots, n_c, \ldots, n_{k_n})$ and $\sum_{c=1}^{k_n} n_c = n$.

When G is assigned a Pitman-Yor $\mathcal{PY}(a, b, G_0)$ process prior, the posterior predictive probability of a new observation $\boldsymbol{\theta}_{n+1}$ is defined by:

$$P(\boldsymbol{\theta}_{n+1} \in B \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \frac{b + ak_n}{b + n} G_0(B) + \sum_{c=1}^{k_n} \frac{n_c - a}{b + n} \delta_{\boldsymbol{\theta}_c^*}(B), \quad \forall B \in \mathcal{B}(\Theta). \quad (16)$$

That is, $\boldsymbol{\theta}_{n+1}$ forms a new cluster with probability $(b + ak_n)/(b + n)$, and otherwise with probability $(n_c - a)/(b + n)$, $\boldsymbol{\theta}_{n+1}$ is allocated to old cluster $\mathcal{C}_c$, for $c = 1, \ldots, k_n$. Recall that the normalized stable process (Kingman, 1975) is equivalent to the $\mathcal{PY}(a, 0, G_0)$ process with $0 \leq a < 1$ and $b = 0$; and the $\mathcal{DP}(\alpha, G_0)$ is the $\mathcal{PY}(0, b, G_0)$ process with $a = 0$ and $b = \alpha$.

Under the $\mathcal{NIG}(c, G_0)$ process prior, the posterior predictive distribution is defined by the probability function,

$$P(y_{n+1} \in B \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = w_0^{(n)} G_0(B) + w_1^{(n)} \sum_{c=1}^{k_n} (n_c - .5)\delta_{y_c^*}(B), \quad \forall B \in \mathcal{B}(\Theta), \quad (17a)$$

with:

$$w_0^{(n)} = \frac{\sum_{l=0}^{n} \binom{n}{l} (-c^2)^{-l+1} \Gamma(k_n + 1 + 2l - 2n\,;\, c)}{2n \sum_{l=0}^{n-1} \binom{n-1}{l} (-c^2)^{-l} \Gamma(k_n + 2 + 2l - 2n\,;\, c)}, \quad (17b)$$

$$w_1^{(n)} = \frac{\sum_{l=0}^{n} \binom{n}{l} (-c^2)^{-l+1} \Gamma(k_n + 2l - 2n\,;\, c)}{n \sum_{l=0}^{n-1} \binom{n-1}{l} (-c^2)^{-l} \Gamma(k_n + 2 + 2l - 2n\,;\, c)}, \quad (17c)$$

where $\Gamma(\cdot\,;\,\cdot)$ is the incomplete gamma function (Lijoi et al., 2005, p. 1283). Finally, exchangeable partition models (e.g., Hartigan, 1990; Barry & Hartigan, 1993; Quintana & Iglesias, 2003) also give rise to random clustering structures of a form (17a–17c), and therefore coincide with the family of Gibbs-type priors, which include the $\mathcal{PY}(a, b, G_0)$ and $\mathcal{NIG}(c, G_0)$ processes and their special cases. More detailed discussions on the clustering behavior induced by various BNP priors are given by DeBlasi et al. (2015).

So far, we have described only BNP priors for the mixture distribution (8) of the general BNP regression model (7), while assuming no covariate dependence in the mixing distribution, with $G_{\mathbf{x}} := G$. We now consider dependent BNP processes. A seminal example is given by the Dependent Dirichlet process ($\mathcal{DDP}(\alpha_{\mathbf{x}}, G_{0\mathbf{x}})$) (MacEachern, 1999, 2000, 2001), which models a covariate ($\mathbf{x}$) dependent process $G_{\mathbf{x}}$, by allowing either the baseline distribution $G_{0\mathbf{x}}$, the stick-breaking mixture weights $\omega_j(\mathbf{x})$, and/or the precision parameter $\alpha_{\mathbf{x}}$ to depend on covariates $\mathbf{x}$. In general terms, a random dependent probability

measure $G_{\mathbf{x}} \,|\, \alpha_{\mathbf{x}}, G_{0\mathbf{x}} \sim \mathcal{DDP}(\alpha_{\mathbf{x}}, G_{0\mathbf{x}})$ can be represented by Sethuraman's (1994) stick-breaking construction, as:

$$G_{\mathbf{x}}(B) = \sum_{j=1}^{\infty} \omega_j(\mathbf{x}) \delta_{\theta_j(\mathbf{x})}(B), \quad \forall B \in \mathcal{B}(\Theta), \tag{18a}$$

$$\omega_j(\mathbf{x}) = \upsilon_j(\mathbf{x}) \prod_{k=1}^{j-1} (1 - \upsilon_k(\mathbf{x})), \quad (\upsilon_j(\mathbf{x}) : \mathcal{X} \to [0,1]), \tag{18b}$$

$$\upsilon_j \sim Q_{\mathbf{x}j}, \quad \theta_j(\mathbf{x}) \sim G_{0\mathbf{x}}. \tag{18c}$$

Next, we describe an important BNP regression model, with a dependent mixture distribution $G_{\mathbf{x}}$ assigned a specific $\mathcal{DDP}(\alpha_{\mathbf{x}}, G_{0\mathbf{x}})$ prior.

ANOVA-linear DDP model

Assume that the data $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ can be stratified into N_h groups, indexed by $h = 1, \ldots, N_h$, respectively. For each of group h, let $y_{i(h)}$ be the ith dependent observation of group h, and let $\mathbf{y}_h = (y_{i(h)})_{i(h)=1}^{n_h}$ be the column vector of n_h dependent observations, corresponding to an observed design matrix $\mathbf{X}_h = (\mathbf{x}_{1(h)}^{\intercal}, \ldots, \mathbf{x}_{i(h)}^{\intercal}, \ldots, \mathbf{x}_{n_h}^{\intercal})$ of n_h rows of covariate vectors $\mathbf{x}_{i(h)}^{\intercal}$ respectively. Possibly, each of the N_h groups of observations has only one observation (i.e., $n_h = 1$), in which case $N_h = n$.

The ANOVA-linear DDP model (DeIorio et al., 2004; Müller et al., 2005) is defined as:

$$(y_{i(h)})_{i(h)=1}^{n_h} \,|\, \mathbf{X}_h \sim f(\mathbf{y}_h \,|\, \mathbf{X}_h; \boldsymbol{\zeta}), \quad h = 1, \ldots, N_h \tag{19a}$$

$$f(\mathbf{y}_h \,|\, \mathbf{X}_h; \boldsymbol{\zeta}) = \sum_{j=1}^{\infty} \left\{ \prod_{i(h)=1}^{n_h} \mathrm{n}(y_{i(h)} \,|\, \mathbf{x}_{i(h)}^{\intercal} \boldsymbol{\beta}_j, \sigma^2) \right\} \omega_j \tag{19b}$$

$$\omega_j = \upsilon_j \prod_{l=1}^{j-1} (1 - \upsilon_l) \tag{19c}$$

$$\upsilon_j \,|\, \alpha \sim \mathrm{Be}(1, \alpha) \tag{19d}$$

$$\boldsymbol{\beta}_j \,|\, \boldsymbol{\mu}, \boldsymbol{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{T}) \tag{19e}$$

$$\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2) \tag{19f}$$

$$\boldsymbol{\mu}, \boldsymbol{T} \sim \mathrm{N}(\boldsymbol{\mu} \,|\, \mathbf{0}, r_0 \mathbf{I}_{p+1}) \mathrm{IW}(\boldsymbol{T} \,|\, p+3, s_0 \mathbf{I}_{p+1}) \tag{19g}$$

$$\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha), \tag{19h}$$

where $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{T})$ and $\mathrm{N}(\boldsymbol{\mu} \,|\, \mathbf{0}, r_0 \mathbf{I}_{p+1})$ each refers to a multivariate normal distribution, and IW refers to the inverted-Wishart distribution. Therefore, all the model parameters are assigned prior distributions, which together, define the joint prior p.d.f. for $\boldsymbol{\zeta} \in \Omega_{\boldsymbol{\zeta}}$ by:

$$\pi(\boldsymbol{\zeta}) = \prod_{j=1}^{\infty} \mathrm{be}(\upsilon_j \,|\, 1, \alpha) \mathrm{n}(\boldsymbol{\beta}_j \,|\, \boldsymbol{\mu}, \boldsymbol{T}) \mathrm{ig}(\sigma^2 \,|\, a_0/2, a_0/2) \tag{19a}$$

$$\times \mathrm{n}(\boldsymbol{\mu} \,|\, \mathbf{0}, r_0 \mathbf{I}_{p+1}) \mathrm{iw}(\boldsymbol{T} \,|\, p+3, s_0 \mathbf{I}_{p+1}) \mathrm{ga}(\alpha \,|\, a_\alpha, b_\alpha), \tag{19b}$$

with beta (be), multivariate normal (n), inverse-gamma (ig), inverted-Wishart (iw), and gamma (ga) p.d.f.s together defining the prior distributions in the ANOVA-linear DDP model (19a–19h). As shown, this model is based on a mixing distribution $G(\boldsymbol{\beta})$ assigned a $\mathcal{DP}(\alpha, G_0)$ prior, with precision parameter α and multivariate normal baseline distribution, $G_0(\cdot) := \mathrm{N}(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{T})$. Prior distributions are assigned to $(\alpha, \boldsymbol{\mu}, \boldsymbol{T})$ in order to allow for posterior inferences to be robust to different choices of the $\mathcal{DP}(\alpha, G_0)$ prior parameters.

The ANOVA-linear DDP model (19a–19h) is equivalent to the BNP regression model (7), with normal kernel densities $\mathrm{n}(y_i \,|\, \mu_j, \sigma^2)$ and mixing distribution $G_{\mathbf{x}}(\mu)$ (8) assigned a $\mathcal{DDP}(\alpha, G_{0\mathbf{x}})$ prior, where:

$$G_{\mathbf{x}}(B) = \sum_{j=1}^{\infty} \omega_j \delta_{\mathbf{x}^{\intercal}\boldsymbol{\beta}}(B), \ \forall B \in \mathcal{B}(\Theta), \tag{20}$$

with $\boldsymbol{\beta}_j \,|\, \boldsymbol{\mu}, \boldsymbol{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{T})$ and $\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2)$ (i.e., $G_0(\cdot) = \mathrm{N}(\boldsymbol{\beta} \,|\boldsymbol{\mu}, \boldsymbol{T}) \mathrm{IG}(\sigma^2 \,|a_0/2, a_0/2)$), and with the ω_j stick-breaking weights (19c) (DeIorio et al., 2004).

A menu option in the **Bayesian Regression** software labels the ANOVA-linear DDP model (19a–19h) as the "Dirichlet process mixture of homoscedastic linear regressions model" (for Step 8 of a data analysis; see next section). The software allows the user to analyze data using any one of many variations of the model (19a–19h). Variations of this DDP model include: "mixture of linear regressions" models, as labeled by a menu option of the software, with mixing distribution $G(\boldsymbol{\beta}, \sigma^2)$ for the coefficients and the error variance parameters; "mixture of random intercepts" models, with mixture distribution $G(\beta_0)$ for only the intercept parameter β_0, and with independent normal priors for the slope coefficient parameters $(\beta_k)_{k=1}^p$; mixture models having mixture distribution G assigned either a Pitman-Yor $\mathcal{PY}(a, b, G_0)$ (including the normalized stable process prior), beta process, geometric weights, or normalized inverse-Gaussian process $\mathcal{NIG}(c, G_0)$ prior, each implying, respectively, a dependent BNP prior for a covariate-dependent mixing distribution (20) (using similar arguments made for the DDP model by De Iorio et al., 2004); and mixed-logit or mixed-probit regression models for a binary (0 or 1) or ordinal ($c = 0, 1, \ldots, m$) dependent variable. Also, suppose that the ANOVA-linear DDP model (19a–19h) is applied to time-lagged dependent variable data (which can be set up using a menu option in the software; see "Installing the software", Step 3, and "Modify data set menu options"). Then this model is defined by an infinite-mixture of autoregressions, with mixture distribution assigned a time-dependent DDP (Lucca et al., 2012). The Help menu of the software provides a full list of models that are available from the software.

Infinite-probits mixture linear model

As mentioned, typical BNP infinite-mixture models assume that the mixture weights have the stick-breaking form (12d). However, a BNP model may have weights with a different form. The infinite-probits model is a Bayesian nonparametric regression model (7)–(8), with prior $\Pi(\boldsymbol{\zeta})$, and with mixture distribution (8) defined by a dependent normalized random measure (Karabatsos and Walker, 2012a).

For data, $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$, a Bayesian infinite-probits mixture model can be defined by:

$$y_i \,|\, \mathbf{x}_i \sim f(y \,|\, \mathbf{x}_i; \boldsymbol{\zeta}), \quad i = 1, \ldots, n \tag{21a}$$

$$f(y \,|\, \mathbf{x}; \boldsymbol{\zeta}) = \sum_{j=-\infty}^{\infty} \mathrm{n}(y \,|\, \mu_j + \mathbf{x}^\intercal \boldsymbol{\beta}, \sigma^2)\omega_j(\mathbf{x}) \tag{21b}$$

$$\omega_j(\mathbf{x}) = \Phi\left(\frac{j - \mathbf{x}^\intercal \boldsymbol{\beta}_\omega}{\sigma_\omega}\right) - \Phi\left(\frac{j - 1 - \mathbf{x}^\intercal \boldsymbol{\beta}_\omega}{\sigma_\omega}\right) \tag{21c}$$

$$\mu_j \,|\, \sigma_\mu^2 \sim \mathrm{N}(0, \sigma_\mu^2) \tag{21d}$$

$$\sigma_\mu \sim \mathrm{U}(0, b_{\sigma\mu}) \tag{21e}$$

$$\beta_0 \,|\, \sigma^2 \sim \mathrm{N}(0, \sigma^2 v_{\beta 0} \to \infty) \tag{21f}$$

$$\beta_k \,|\, \sigma^2 \sim \mathrm{N}(0, \sigma^2 v), \quad k = 1, \ldots, p \tag{21g}$$

$$\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2) \tag{21h}$$

$$\boldsymbol{\beta}_\omega \,|\, \sigma_\omega^2 \sim \mathrm{N}(\mathbf{0}, \sigma_\omega^2 v_\omega \mathbf{I}) \tag{21i}$$

$$\sigma_\omega^2 \sim \mathrm{IG}(a_\omega/2, a_\omega/2), \tag{21j}$$

with $\Phi(\cdot)$ the normal N(0, 1) c.d.f., and model parameters $\boldsymbol{\zeta} = ((\mu_j)_{j=1}^\infty, \sigma_\mu^2, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\beta}_\omega, \sigma_\omega)$ assigned a prior $\Pi(\boldsymbol{\zeta})$ with p.d.f.:

$$\pi(\boldsymbol{\zeta}) = \prod_{j=-\infty}^{\infty} \mathrm{n}(\mu_j \,|\, 0, \sigma_\mu^2)\mathrm{u}(\sigma_\mu \,|\, 0, b_{\sigma\mu})\mathrm{n}(\boldsymbol{\beta} \,|\, \mathbf{0}, \sigma^2 \mathrm{diag}(v_{\beta 0} \to \infty, v\mathbf{J}_p)) \tag{22a}$$

$$\times \mathrm{ig}(\sigma^2 \,|\, a_0/2, a_0/2)\mathrm{n}(\boldsymbol{\beta}_\omega \,|\, \mathbf{0}, \sigma_\omega^2 v_\omega \mathbf{I}_{p+1})\,\mathrm{ig}(\sigma_\omega^2 \,|\, a_\omega/2, a_\omega/2), \tag{22b}$$

where $\mathbf{J}_p$ denotes a $p \times 1$ vector of 1s, and $\mathrm{u}(\sigma_\mu \,|\, 0, b)$ refers to the p.d.f. of the uniform distribution with minimum 0 and maximum b.

The **Bayesian Regression** software labels the BNP model (21a–21j) as the "Infinite homoscedastic probits regression model," in a menu option (in Step 8 of a data analysis; see next section). This model is defined by a highly-flexible robust linear model, an infinite mixture of linear regressions (21b), with random intercept parameters μ_j modeled by infinite covariate-dependent mixture weights (21c). The model (21a–21j) has been extended and applied to prediction analysis (Karabatsos & Walker, 2012a), meta-analysis (Karabatsos et al., 2015), (test) item-response analysis (Karabatsos & Walker, 2015b), and causal analysis (Karabatsos & Walker, 2015a).

The covariate-dependent mixture weights $\omega_j(\mathbf{x})$ in Eq. 21c, defining the mixture distribution (8), are modeled by a probits regression for ordered categories $j = \ldots, -2, -1, 0, 1, 2, \ldots$, with latent location parameter $\mathbf{x}^\intercal \boldsymbol{\beta}_\omega$, and with latent standard deviation σ_ω that controls the level of modality of the conditional p.d.f. $f(y \,|\, \mathbf{x}; \boldsymbol{\zeta})$ of the dependent variable Y. Specifically, as $\sigma_\omega \to 0$, the conditional p.d.f. $f(y \,|\, \mathbf{x}; \boldsymbol{\zeta})$ becomes more unimodal. As σ_ω gets larger, $f(y \,|\, \mathbf{x}; \boldsymbol{\zeta})$ becomes more multimodal (see Karabatsos & Walker, 2012a).

The **Bayesian Regression** software allows the user to analyze data using any one of several versions of the infinite-probits regression model (21). Versions include models where the kernel densities are instead specified by covariate independent normal densities $\mathrm{n}(y \,|\, \mu_j, \sigma_j^2)$, and the mixture weights are modeled by:

$$\omega_j(\mathbf{x}) = \Phi\left(\frac{j - \mathbf{x}^\intercal \boldsymbol{\beta}_\omega}{\sqrt{\exp(\mathbf{x}^\intercal \boldsymbol{\lambda}_\omega)}}\right) - \Phi\left(\frac{j - \mathbf{x}^\intercal \boldsymbol{\beta}_\omega - 1}{\sqrt{\exp(\mathbf{x}^\intercal \boldsymbol{\lambda}_\omega)}}\right), \quad \text{for } j = 0, \pm 1, \pm 2, \ldots; \tag{23}$$

include models where either the individual regression coefficients $\boldsymbol{\beta}$ in the kernels, or the individual regression coefficients $(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega)$ in the mixture weights (23) are assigned spike-and-slab priors using the SSVS method (George & McCulloch, 1993, 1997), to enable automatic variable (covariate) selection inferences from the posterior distribution; and include mixed-probit regression models for binary (0 or 1) or ordinal ($c = 0, 1, \ldots, m$) dependent variables, each with inverse-link function c.d.f. modeled by a covariate-dependent, infinite-mixture of normal densities (given by Eq. 21b, but instead for the continuous underlying latent dependent variable; see Karabatsos & Walker, 2015a).

Some linear models

We briefly review two basic Bayesian normal linear models from standard textbooks (e.g., O'Hagan & Forster, 2004; Denison et al., 2002).

First, the Bayesian normal linear model, assigned a (conjugate) normal inverse-gamma prior distribution to the coefficients and error variance parameters, $(\boldsymbol{\beta}, \sigma^2)$, is defined by:

$$y_i \,|\, \mathbf{x}_i \sim f(y \,|\, \mathbf{x}_i), \quad i = 1, \ldots, n \tag{24a}$$

$$f(y \,|\, \mathbf{x}) = \mathrm{n}(y \,|\, \mathbf{x}^\intercal \boldsymbol{\beta}, \sigma^2) \tag{24b}$$

$$\beta_0 \,|\, \sigma^2 \sim \mathrm{N}(0, \sigma^2 v_{\beta 0} \to \infty) \tag{24c}$$

$$\beta_k \,|\, \sigma^2 \sim \mathrm{N}(0, \sigma^2 v_\beta), \quad k = 1, \ldots, p \tag{24d}$$

$$\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2). \tag{24e}$$

An extension of the model (24a–24e) is provided by the Bayesian 2-level normal random-effects model (HLM). Again, let the data $\mathcal{D}_n = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ be stratified into

N_h groups, indexed by $h = 1, \ldots, N_h$. Also, for each group h, let $y_{i(h)}$ be the ith dependent observation, and let $\mathbf{y}_h = (y_{i(h)})_{i(h)=1}^{n_h}$ be the column vector of n_h dependent observations, corresponding to an observed design matrix $\mathbf{X}_h = (\mathbf{x}_{1(h)}^\mathsf{T}, \ldots, \mathbf{x}_{i(h)}^\mathsf{T}, \ldots, \mathbf{x}_{n_h}^\mathsf{T})$ of n_h rows of covariate vectors $\mathbf{x}_{i(h)}^\mathsf{T}$ respectively. Then a Bayesian 2-level model (HLM) can be represented by:

$$y_{i(h)} \mid \mathbf{x}_{i(h)} \sim f(y \mid \mathbf{x}_{i(h)}), \quad i(h) = 1, \ldots, n_h \tag{25a}$$
$$f(y \mid \mathbf{x}_{i(h)}) = \mathrm{n}(y \mid \mathbf{x}_{i(h)}^\mathsf{T}\boldsymbol{\beta}_{Rh}, \sigma^2) \tag{25b}$$
$$\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_{Rh} = \mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \mathbf{x}^\mathsf{T}\boldsymbol{u}_h \tag{25c}$$
$$\beta_0 \mid \sigma^2 \sim \mathrm{N}(0, \sigma^2 v_{\beta 0} \to \infty) \tag{25d}$$
$$\beta_k \mid \sigma^2 \sim \mathrm{N}(0, \sigma^2 v_\beta), \quad k = 1, \ldots, p \tag{25e}$$
$$\boldsymbol{u}_h \mid \boldsymbol{T} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{T}), \quad h = 1, \ldots, N_h \tag{25f}$$
$$\sigma^2 \sim \mathrm{IG}(a_0/2, a_0/2) \tag{25g}$$
$$\boldsymbol{T} \sim \mathrm{IW}(p + 3, s_0\mathbf{I}_{p+1}). \tag{25h}$$

This model (25a–25h), as shown in (25f), assumes that the random coefficients $\boldsymbol{u}_h$ (for $h = 1, \ldots, N_h$) are normally distributed over the N_h groups.

Both linear models above, and the different versions of these models mentioned in "Introduction", are provided by the Bayesian Regression software. See the Help menu for more details.

Using the Bayesian regression software

Installing the software

The **Bayesian Regression** software is a stand-alone package for a 64-bit Windows computer.[4] To install the software on your computer, take the following steps:

1. Go the **Bayesian Regression** software web page: http://www.uic.edu/~georgek/HomePage/BayesSoftware.html. Then click the link on that page to download the **Bayesian Regression** software installation file, named BayesInstaller_web64bit.exe (or BayesInstaller_web32bit.exe).
2. Install the software by clicking the file BayesInstaller_webXXbit.exe. This will include a web-based installation of MATLAB Compiler Runtime, if necessary. As you install, select the option "Add a shortcut to the desktop," for convenience. (To install, be connected to the internet, and temporarily disable any firewall or proxy settings on your computer).

Then start the software by clicking the icon BayesRegXXbit.exe.

The next subsection provides step-by-step instructions on how to use the **Bayesian Regression** software to perform a Bayesian analysis of your data set. The software provides several example data files, described under the Help menu. You can create them by clicking the File menu option: "Create Bayes Data Examples file folder." Click the File menu option again to import and open an example data set from this folder. The next subsection illustrates the software through the analysis of the example data set PIRLS100.csv.

The **Bayesian Regression** software, using your menu-selected Bayesian model, outputs the data analysis results into space- and comma-delimited text files with time-stamped names, which can be viewed in free *NotePad++*. The comma-delimited output files include the posterior samples (.MC1), model fit residual (*.RES), and the model specification (*.MODEL) files. The software also outputs the results into graph (figure *.fig) files, which can then be saved into a EPS (*.eps), bitmap (*.bmp), enhanced metafile (*.emf), JPEG image (*.jpg), or portable document (*.pdf) file format. Optionally you may graph or analyze a delimited text output file after importing it into spreadsheet software (e.g., *OpenOffice*) or into the R software using the command line: `ImportedData = read.csv(file.choose()).`

Running the software (12 steps for data analysis)

You can run the software for data analysis using any one of many Bayesian models of your choice. A data analysis involves running the following 12 basic steps (required or optional).

In short, the 12 steps are as follows:

(1) Import or open the data file (Required);
(2) Compute basic descriptive statistics and plots of your data (Optional);
(3) Modify the data set (e.g., create variables) to set up your data analysis model (Optional);
(4) Specify a new Bayesian model for data analysis (Required);
(5) Specify observation weights (Optional);
(6) Specify the censored observations (Optional);
(7) Set up the MCMC sampling algorithm model posterior estimation (Required);
(8) Click the **Run Posterior Analysis** button (Required);
(9) Click the **Posterior Summaries** button to output data analysis results (Required);
(10) Check MCMC convergence (Required);
(11) Click the **Posterior Predictive button** to run model predictive analyses (Optional);
(12) Click the **Clear button** to finish your data analysis project.

Then you may run a different data analysis. Otherwise, you may then Exit the software and return to the same data analysis project later, after re-opening the software.

[4] An older version of the software can run on a 32-bit computer.

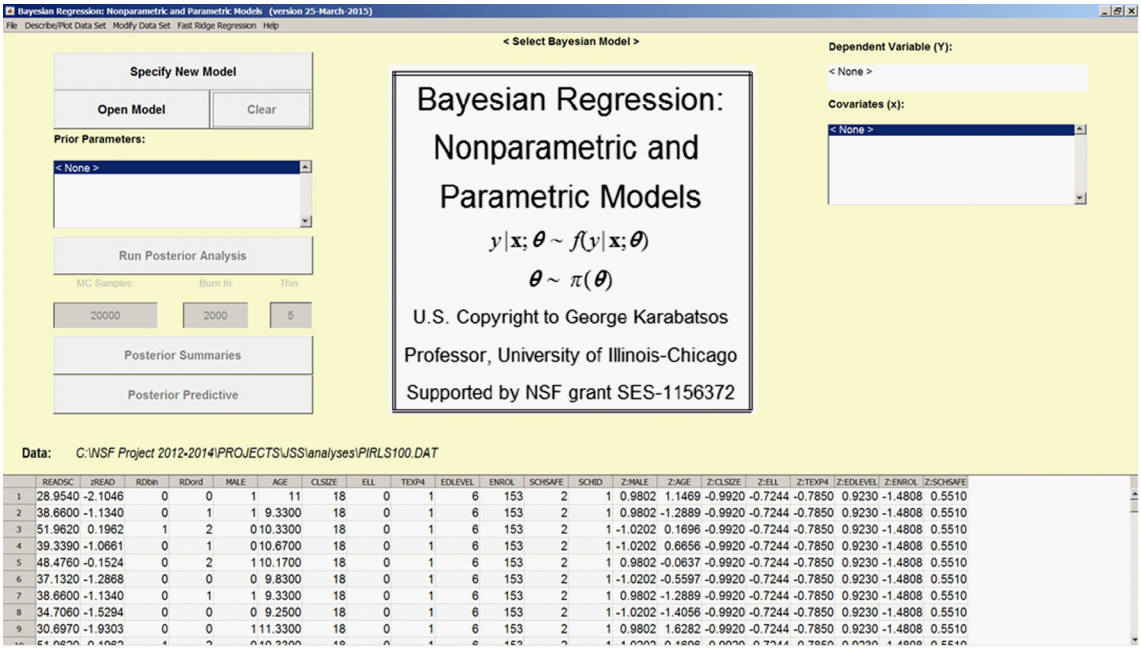


Fig. 1 A view of the Bayesian Regression software interface

Below, we give more details on the 12 steps of data analysis.

1. **(Required)** Use the **File menu** to **Import or open the data file** for analysis. Specifically, the data file that you import must be a comma-delimited file, with name having the .csv extension. (Or, you may click the File menu option to open an existing (comma-delimited) data (*.DAT) file). In the data file, the variable names are located in the first row, with numeric data (i.e., non-text data) in all the other rows. For each row, the number of variable names must equal the number of commas minus 1. The software allows for missing data values, each coded as NaN or as an empty blank. After you select the data file to import, the software converts it into a comma-delimited data (*.DAT) file. Figure 1 shows the interface of the **Bayesian Regression** software. It presents the PIRLS100.DAT data set at the bottom of the interface, after the PIRLS100.csv file has been imported.
2. **(Optional)** Use the **Describe/Plot Data Set menu option(s)** to compute basic descriptive statistics and plots of the data variables. Statistics and plots include the sample mean, standard deviation, quantiles, frequency tables, cross-tabulations, correlations, covariances, univariate or bivariate histograms,[5] stem-and-leaf plots, univariate or bivariate kernel density estimates,[6] quantile-quantile (Q-Q) plots, two- or three-dimensional scatter plots, scatter plot matrices, (meta-analysis) funnel plots (Egger et al., 1997), box plots, and plots of kernel regression estimates with automatic bandwidth selection.[7]
3. **(Optional)** Use the **Modify Data Set menu option(s)** to set up a data analysis. The menu options allow you to construct new variables, handle missing data, and/or to perform other modifications of the data set. Then the new and/or modified variables can be included in the Bayesian model that you select in Step 4. Figure 1 presents the PIRLS100.DAT data at the bottom of the software interface, and shows the data of the variables MALE, AGE, CLSIZE, ELL, TEXP4, EDLEVEL, ENROL, and SCHSAFE in the last 8 data columns, respectively, after taking z-score transformations and adding "Z:" to each variable name. Such transformations are done with the menu option: Modify Data Set

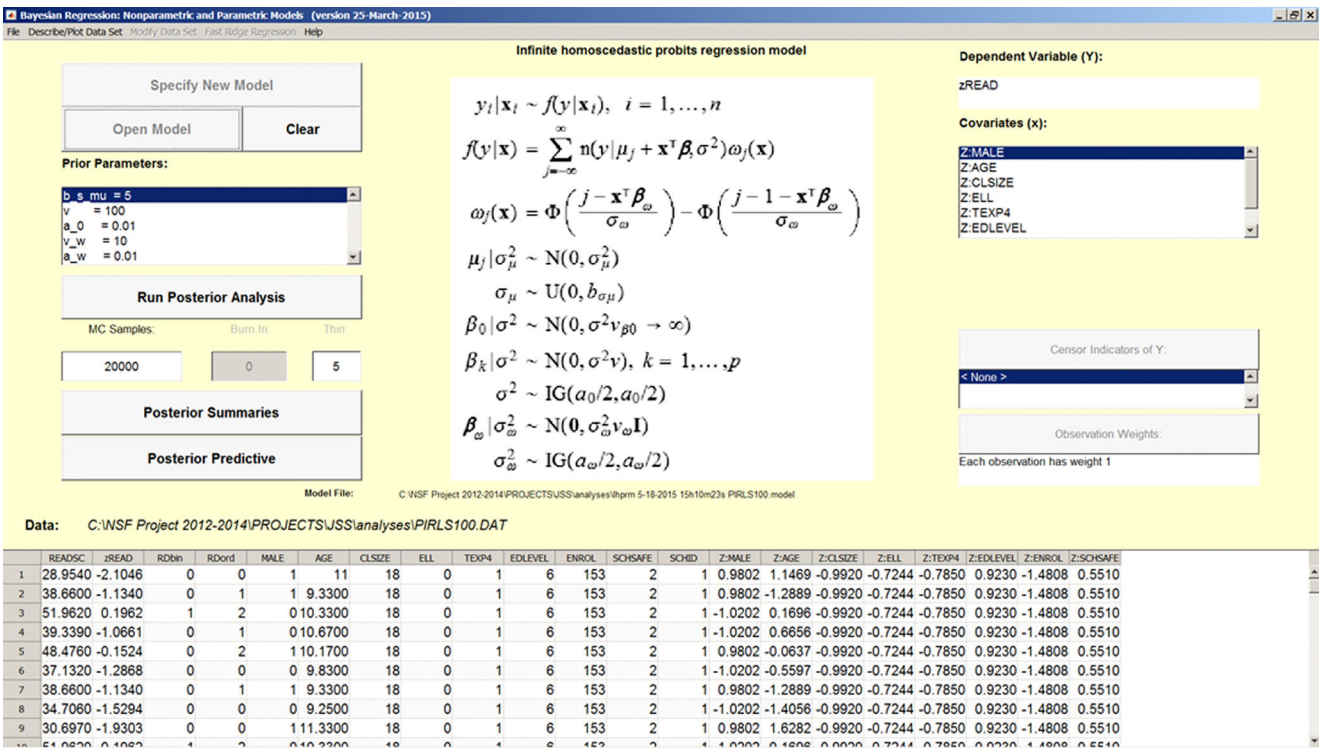
[5]For the univariate histogram, the bin size (h) is defined by the Freedman and Diaconis (1981) rule, with $h = 2(\text{IQR})n^{-1/3}$, where IQR is the interquartile range of the data, and n is the sample size. For the bivariate histogram, the automatic bin sizes are given by $h_k = 3.5\widehat{\sigma}_k n^{-1/4}$, where $\widehat{\sigma}_k$, $k = 1, 2$, is the sample standard deviation for the two variables (Scott, 1992).

[6]The univariate kernel density estimate assumes normal kernels, with automatic bandwidth (h) given by the normal reference rule, defined by $h = 1.06\widehat{\sigma}n^{-1/5}$, where $\widehat{\sigma}$ is the data standard deviation, and n is the sample size (Silverman, 1986, p. 45). The bivariate kernel density estimate assumes normal kernels, with automatic bandwidth determined by the equations in Botev et al. (2010).

[7]Kernel regression uses normal kernels, with an automatic choice of bandwidth (h) given by $\widehat{h} = \sqrt{\widehat{h}_x \widehat{h}_y}$, where $\widehat{h}_x = \text{med}(|\mathbf{x} - \text{med}(\mathbf{x})|)/c$, and $\widehat{h}_y = \text{med}(|\mathbf{y} - \text{med}(\mathbf{y})|)/c$, where $(\mathbf{x}, \mathbf{y})$ give the vectors of X data and Y data (resp.), med$(\cdot)$ is the median, $c = .6745 * (4/3/n)^{0.2}$, and n is the sample size (Bowman and Azzalini, 1997, p. 31).

> Simple variable transformations > Z score. "Modify data set menu options" provides more details about the available Modify Data Set menu options.

4. **(Required) Click the Specify New Model** button to **select a Bayesian model** for data analysis, and then for the model select: the dependent variable; the covariate(s) (predictor(s)) (if selected a regression model); the level-2 (and possibly level-3) grouping variables (if a multi-level model); and the model's prior distribution parameters. Figure 2 shows the software interface, after selecting the Infinite homoscedastic probits regression model, along with the dependent variable, covariates, and prior parameters.
5. **(Optional)** To weight each data observation (row) differently under your selected model, **click the Observation Weights button** to select a variable containing the weights (must be finite, positive, and non-missing). (This button is not available for a binary or ordinal regression model). By default, the observation weights are 1. For example, observation weights are used for meta-analysis of data where each dependent variable observation y_i represents a study-reported effect size (e.g., a standardized mean difference in scores between a treatment group and a control group, or a correlation coefficient estimate). Each reported effect size y_i has sampling variance $\widehat{\sigma}_i^2$, and observation weight $1/\widehat{\sigma}_i^2$ that is proportional to the sample size for y_i. Details about the various effect size measures, and their sampling variance formulas, are found in meta-analysis textbooks (e.g., Cooper et al., 2009). "Modify data set menu options" mentions a Modify Data Set menu option that computes various effect size measures and corresponding variances.
6. **(Optional) Click the Censor Indicators of Y button**, if the dependent variable consists of **censored observations** (not available for a binary or ordinal regression model). Censored observations often appear in survival data, where the dependent variable Y represents the (e.g., log) survival time of a patient. Formally, an observation, y_i, is *censored* when it is only known to take on a value from a known interval $[Y_{LBi}, Y_{UBi}]$; is *interval-censored* when $-\infty < Y_{LBi} < Y_{UBi} < \infty$; is *right censored* when $-\infty < Y_{LBi} < Y_{UBi} \equiv \infty$; and *left censored* when $-\infty \equiv Y_{LBi} < Y_{UBi} < \infty$ (e.g., Klein & Moeschberger, 2010). After clicking the **Censor Indicators of Y button**, select the two variables that describe the (fixed) censoring lower-bounds (LB) and upper-bounds (UB) of the dependent variable observations. Name these variables LB and UB. Then for each interval-censored observation y_i, its LB_i and UB_i values must be finite, with $LB_i < UB_i$, $y_i \leq UB_i$, and $y_i \geq LB_i$. For each



	READSC	zREAD	RDbin	RDord	MALE	AGE	CLSIZE	ELL	TEXP4	EDLEVEL	ENROL	SCHSAFE	SCHID	Z:MALE	Z:AGE	Z:CLSIZE	Z:ELL	Z:TEXP4	Z:EDLEVEL	Z:ENROL	Z:SCHSAFE
1	28.9540	-2.1046	0	0	1	11	18	0	1	6	153	2	1	0.9802	1.1469	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
2	38.6600	-1.1340	0	1	1	9.3300	18	0	1	6	153	2	1	0.9802	-1.2889	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
3	51.9620	0.1962	1	2	0	10.3300	18	0	1	6	153	2	1	-1.0202	0.1696	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
4	39.3390	-1.0661	0	1	0	10.6700	18	0	1	6	153	2	1	-1.0202	0.6656	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
5	48.4760	-0.1524	0	2	1	10.1700	18	0	1	6	153	2	1	0.9802	-0.0637	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
6	37.1320	-1.2868	0	0	0	9.8300	18	0	1	6	153	2	1	-1.0202	-0.5597	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
7	38.6600	-1.1340	0	1	1	9.3300	18	0	1	6	153	2	1	0.9802	-1.2889	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
8	34.7060	-1.5294	0	0	0	9.2500	18	0	1	6	153	2	1	-1.0202	-1.4056	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510
9	30.6970	-1.9303	0	0	1	11.3300	18	0	1	6	153	2	1	0.9802	1.6282	-0.9920	-0.7244	-0.7850	0.9230	-1.4808	0.5510

Fig. 2 A view of the Bayesian regression software interface, after the selection of the infinite probits mixture regression model, dependent variable, and covariates

right-censored observation y_i, its LB_i value must be finite, with $y_i \geq LB_i$, and set $UB_i = -9999$. For each left-censored observation y_i, its UB_i value must be finite, with $y_i \leq UB_i$, and set $LB_i = -9999$. For each uncensored observation y_i, set $LB_i = -9999$ and $UB_i = -9999$.

7. **(Required) Enter: the total number** (S) **of MC Samples**, i.e., MCMC sampling iterations (indexed by $s = 1, \ldots, S$, respectively)**;** the number $(s_0 \geq 1)$ of the initial **Burn-In period** samples; and the **Thin** number k, to retain every k^{th} sampling iterate of the S total MCMC samples. The MCMC samples are used to estimate the posterior distribution (and functionals) of the parameters of your selected Bayesian model. Entering a thin value $k > 1$ represents an effort to have the MCMC samples be (pseudo-) independent samples from the posterior distribution. The burn-in number (s_0) is your estimate of the number of initial MCMC samples that are biased by the software's starting model parameter values used to initiate the MCMC chain at iteration $s = 0$.
8. **(Required) Click the Run Posterior Analysis button** to run the MCMC sampling algorithm, using the selected MC Samples, Burn-In, and Thin numbers. A wait-bar will then appear and display the progress of the MCMC sampling iterations. After the MCMC sampling algorithm finishes running, the software will create an external: (a) model (*.MODEL) text file that describes the selected model and data set; (b) Monte Carlo (*.MC1) samples file which contains the generated MCMC samples; (c) residual (*.RES) file that contains the model's residual fit statistics; and (d) an opened, text output file of summaries of (marginal) posterior point-estimates of model parameters and other quantities, such as model predictive data-fit statistics. The results are calculated from the generated MCMC samples aside from any burn-in or thinned-out samples. Model fit statistics are calculated from all MCMC samples instead. The software puts all output files in the same subdirectory that has the data (*.DAT) file.
9. **(Required)** Click the **Posterior Summaries button** to select menu options for additional data analysis output, such as: text output of posterior quantile estimates of model parameters and 95 % Monte Carlo Confidence Intervals (see Step 10); trace plots of MCMC samples; 2-dimensional plots and 3-dimensional plots of (kernel) density estimates, univariate and bivariate histograms, distribution function, quantile function, survival function, and hazard functions, box plots, Love plots, Q-Q plots, and Wright maps, of the (marginal) posterior distribution(s) of the model parameters; posterior correlations and covariances of model parameters; and plots and tables of the model's standardized fit residuals. The software creates all text output files in the same subdirectory that contains the data (*.DAT) file. You may save any graphical output in the same directory.
10. **(Required)** Click the **Posterior Summaries** button for menu options to **check the MCMC convergence of parameter point-estimates**, for every model parameter of interest for data analysis. Verify: (1) that the univariate trace plots present good mixing of the generated MCMC samples of each parameter; and (2) that the generated MCMC samples of that parameter provide sufficiently-small half-widths of the 95 % Monte Carlo Confidence Intervals (95 % MCCIs) for parameter posterior point-estimates of interest (e.g., marginal posterior mean, standard deviation, quantiles, etc.). If for your model parameters of interest, either the trace plots do not support adequate mixing (i.e., plots are not stable and "hairy"), or the 95 % MCCI half-widths are not sufficiently small for practical purposes, then the MCMC samples of these parameters have not converged to samples from the model's posterior distribution. In this case you need to generate additional MCMC samples, by clicking the **Run Posterior Analysis button** again. Then re-check for MCMC convergence by evaluating the updated trace plots and the 95 % MCCI half-widths. This process may be repeated until MCMC convergence is reached.
11. **(Optional) Click the Posterior Predictive button**[8] to generate model's predictions of the dependent variable Y, conditionally on selected values of one or more (focal) covariates (predictors). See "Overview of Bayesian inference" for more details. Then select the posterior predictive functionals of Y of interest. Choices of functionals include the mean, variance, quantiles (to provide a quantile regression analysis), probability density function (p.d.f.), cumulative distribution function (c.d.f.), survival function, hazard function, the cumulative hazard function, and the probability that $Y \geq 0$. Then select one or more focal covariate(s) (to define $\mathbf{x}_S$), and then enter their values, in order to study how predictions of Y varies as a function of these covariate values. For example, if you chose the variable Z:CLSIZE as a focal covariate, then you may enter values like -1.1, .02, 3.1, so

[8] This button is not available for Bayesian models for density estimation. However, for these models, the software provides menu options to output estimates of the posterior predictive distribution, after clicking the **Posterior Summaries** button. They include density and c.d.f. estimates. See Step 9.

that you can make predictions of Y conditionally on these covariate values. Or you may base predictions on an equally-spaced grid of covariate (Z:CLSIZE) values, like $-3, -2.5, -2, \ldots 2, 2.5, 3$, by entering $-3 : .5 : 3$. If your data set observations are weighted (optional Step #5), then specify a weight value for the Y predictions. Next, if your selected focal covariates do not constitute all model covariates, then select among options to handle the remaining (non-focal) covariates. Options include the *grand-mean centering method*, the *zero-centering method*, the *partial dependence method*, and the *clustered partial dependence method.* After you made all the selections, the software will provide estimates of your selected posterior predictive functionals of Y, conditionally on your selected covariate values, in graphical and text output files, including comma-delimited files. (The software generates graphs only if you selected 1 or 2 focal covariates; and generates no output if you specify more than 300 distinct values of the focal covariate(s)). All analysis output is generated in the same subdirectory that contains the data (*.DAT) file. You may save any graphical output in the same directory.

12. **(Required)** After completing the Bayesian data analysis, you may click the **Clear button**. Then you may start a different data analysis with another Bayesian model, or exit the software. Later, you may return to and continue from a previous Bayesian data analysis (involving the same model and data set) by using menu options to generate new MCMC samples and/or new data analysis output. To return to the previous analysis, go to the File menu to open the (relevant) data file (if necessary), and then click the **Open Model button** to open the old model (*.MODEL) file. Then, the software will load this model file along with the associated MCMC samples (*.MC1) file and residual (*.RES) files. (Returning to a previous Bayesian regression analysis is convenient if you have already stored the data, model, MC samples, and residual files all in the same file subdirectory). Then after clicking the **Run Posterior Analysis button**, the newly generated MCMC samples will append the existing MCMC samples (*.MC1) file and update the residual (*.RES) file.

Finally, the software provides a **Fast Ridge Regression** menu option that performs a Bayesian data analysis using the ridge (linear) regression model (Hoerl & Kennard, 1970), with parameters estimated by a fast marginal maximum likelihood algorithm (Karabatsos, 2014). This menu option can provide a fast analysis of an ultra-large data set, involving either a very large sample size and/or number of covariates (e.g., several thousands). At this point we will not elaborate on this method because it is currently the subject of ongoing research.

Modify data set menu options

Some **Modify Data Set menu** options allow you to construct new variables from your data. These new variables may be included as either covariates and/or the dependent variable for your Bayesian data analysis model. Methods for constructing new variables include: simple transformations of variables (e.g., z-score, log, sum of variables); transforming a variable into an effect size dependent variable for meta-analysis (Borenstein, 2009; Fleiss & Berlin, 2009; Rosenthal, 1994); the creation of lagged dependent variables as covariates for a Bayesian autoregression time-series analysis (e.g., Prado & West, 2010); dummy/binary coding of variables; the construction of new covariates from other variables (covariates), via transformations including: polynomials, two-way interactions between variables, univariate or multivariate thin-plate splines (Green & Silverman, 1993) or cubic splines (e.g., Denison et al., 2002); spatial-weight covariates (Stroud et al., 2001; or thin-plate splines; Nychka, 2000) from spatial data (e.g., from latitude and longitude variables) for spatial data analysis.

Now we briefly discuss Modify Data Set menu options that can help set up a causal analysis of data from a non-randomized (or randomized) study. First, a propensity score variable, included as a covariate in a regression model; or as a dummy-coded covariate that stratifies each subject into one of 10 (or more) ordered groups of propensity scores; or as observations weights (entered as the inverse of the propensity scores); can help reduce selection bias in the estimation of the causal effect (slope coefficient) of a treatment-receipt (versus non-treatment/control) indicator covariate on a dependent variable of interest (Rosenbaum & Rubin, 1983a, 1984; Imbens, 2004; Lunceford & Davidian, 2004; Schafer & Kang, 2008; Hansen, 2008). As an alternative to using propensity scores, we may consider a regression discontinuity design analysis (Thistlewaite & Campbell, 1960; Cook, 2008). This would involve specifying a regression model, with dependent (outcome) variable of interest regressed on covariates that include an assignment variable, and a treatment assignment variable that indicates (0 or 1) whether or not the assignment variable exceeds a meaningful threshold. Then under mild conditions (Hahn et al., 2001; Lee & Lemieux, 2010), the slope coefficient estimate for the treatment variable is a causal effect estimate of the treatment (versus non-treatment) on the dependent variable. For either the propensity score or regression discontinuity design approach, which can be set up using appropriate Modify Data Set menu options,

Univariate Descriptive Statistics (for the 100 rows of data).

Results:

Variable	VarID	#obs	#miss	#distinct	Mean	SD
zREAD	2	100	0	58	-0.467	0.931
MALE	5	100	0	2	0.510	0.502
AGE	6	100	0	32	10.214	0.689
CLSIZE	7	100	0	13	22.130	4.184
ELL	8	100	0	13	6.015	8.346
TEXP4	9	100	0	8	2.840	2.356
EDLEVEL	10	100	0	2	5.540	0.501
ENROL	11	100	0	21	555.100	272.909
SCHSAFE	12	100	0	3	1.620	0.693

Fig. 3 Univariate descriptive statistics for the PIRLS100 data

causal effects can be expressed in terms of treatment versus non-treatment comparisons of general posterior predictive functionals of the dependent variable (e.g., Karabatsos & Walker, 2015a).

In some settings, it is of interest to include a multivariate dependent variable (multiple dependent variables) in the Bayesian regression model (Step 4). You can convert a multivariate regression problem into a univariate regression problem (Gelman et al., 2004, Ch. 19) by clicking a menu option that "vectorizes" or collapses the multiple dependent variables into a single dependent variable. This also generates new covariates in the data set, having a block design that is suited for the (vectorized) multiple dependent variables. To give an example involving item response theory (IRT) modeling (e.g., van der Linden, 2015), each examinee (data set row) provides data on responses to J items (data columns) on a test (e.g., examination or questionnaire). Here, a menu option can be used to vectorize (collapse) the J item response columns into a single new dependent variable (data column), and to create corresponding dummy item indicator (-1 or 0) covariates. In the newly-revised data, each examinee occupies J rows of data. Then a Bayesian regression model of the software, which includes the new dependent variable, item covariates, and the examinee identifier as a grouping variable (or as dummy (0, 1) indicator covariates), provides an Item Response Theory (IRT) model for the data (van der Linden, 2015).

Similarly, it may be of interest to include a categorical (polychotomous) dependent variable in your Bayesian regression model. Assume that the variable takes on $m + 1$ (possibly unordered) categorical values, indexed by $c = 0, 1, \ldots, m$, with $c = 0$ the reference category. In this

Posterior Predictive Model Fit Statistics

		Stat.	95% MCCIhw
Model posterior predictive SSE	D(m) =	31.959	0.412
Model SSE fit to data	Gof(m) =	3.646	0.192
Penalty (predictive variance)	P(m) =	28.313	0.223
	R squared =	0.958	0.002

Standardized Residuals (z_i) of Dependent Variable Responses:

Min	5%	10%	25%	50%	75%	90%	95%	Max
-0.675	-0.494	-0.391	-0.259	-0.057	0.262	0.388	0.499	0.665

0 (0%) of all n = 100 observations are outliers (with |z_i| > 2).
Range of 95% MCCI sizes of residuals: [0.004 , 0.021].
A report of the residuals, by individual, is provided upon request.

95%MCCIhw: Half-Width of 95% Monte Carlo Confidence Interval.

Fig. 4 Posterior predictive fit statistics for the infinite probits mixture model

Posterior Summary Estimates

Parameter	Mean	Med	SD	25%	75%	2.5%	97.5%
beta0	-0.456	-0.424	0.249	-0.552	-0.311	-1.122	-0.065
beta:Z:MALE	-0.061	-0.068	0.129	-0.142	-0.014	-0.232	0.264
beta:Z:AGE	-0.148	-0.151	0.121	-0.219	-0.092	-0.358	0.134
beta:Z:CLSIZE	0.206	0.187	0.174	0.061	0.325	-0.038	0.587
beta:Z:ELL	-0.056	0.015	0.174	-0.078	0.043	-0.599	0.092
beta:Z:TEXP4	-0.040	-0.024	0.086	-0.082	0.002	-0.221	0.136
beta:Z:EDLEVEL	0.075	0.062	0.094	-0.001	0.137	-0.067	0.285
beta:Z:ENROL	0.304	0.292	0.090	0.242	0.361	0.153	0.500
beta:Z:SCHSAFE	-0.198	-0.201	0.082	-0.237	-0.160	-0.382	-0.029
sigma^2	0.134	0.034	0.210	0.006	0.111	0.003	0.690
sigma^2_mu	0.820	0.695	1.004	0.539	0.879	0.000	2.269
beta_w0	0.122	0.017	0.567	-0.219	0.482	-0.867	1.341
beta_w:Z:MALE	-1.115	-0.990	0.734	-1.782	-0.457	-2.415	-0.083
beta_w:Z:AGE	-3.908	-4.099	2.256	-5.823	-2.509	-7.515	0.067
beta_w:Z:CLSIZE	-2.986	-3.402	1.462	-4.058	-2.410	-4.902	0.083
beta_w:Z:ELL	0.436	0.381	0.379	0.184	0.651	-0.236	1.282
beta_w:Z:TEXP4	-0.373	-0.304	0.538	-0.775	0.008	-1.426	0.664
beta_w:Z:EDLEVEL	1.917	1.915	1.288	0.902	3.184	-0.058	3.828
beta_w:Z:ENROL	2.402	2.795	1.442	1.563	3.502	-0.137	4.456
beta_w:Z:SCHSAFE	1.037	1.124	0.861	0.306	1.720	-0.608	2.470
sigma^2_w	1.366	1.053	1.248	0.316	2.210	0.003	4.203

Fig. 5 Marginal posterior parameter estimates of the infinite probits mixture model

case, you may run a menu option that recodes the categorical dependent variables into $m + 1$ binary (e.g., 0 or 1) variables, then vectorizes (collapses) these multiple binary variables (data columns) into a single binary dependent variable (column), and reformats the covariate data into a block design suitable for the multiple binary variables. Then for data analysis, you may specify a binary regression model that includes the (collapsed) binary dependent variable and the reformatted covariates (Begg & Gray, 1996; see Wu et al., 2004).

The Modify Data Set menu also provides menu options to reduce the number of variables for dimension reduction, including: principal components, multidimensional scaling, or scaling by the true score test theory and Cronbach's alpha reliability analysis of test items, and propensity scoring. There are also menu options that handle missing data, including: nearest-neighbor hot-deck imputation of missing data (Andridge & Little, 2010); the processing of multiple (missing data) imputations or plausible values obtained externally; and the assignment of missing (NaN) values based on specific missing data codes (e.g., 8, 9, or −99, etc.) of variables. Finally, there are Modify Data Set menu options for more basic data editing and reformatting, including: adding data on a row-identification (ID) variable; aggregating (e.g., averaging) variables by a group identification variable; the selection of cases (data rows) for inclusion in a reduced data set, including random selection or selection by values of a variable; the sorting of cases; the deletion of variables; changing the variable name; and moving the variables into other columns of the data file.

Real data example

We illustrate the software through the analysis of data set file PIRLS100.csv, using the software steps outlined in the previous section. The data are based on a random sample of 100 students from 21 low-income U.S. elementary schools, who took part of the 2006 Progress in International Reading Literacy Study (PIRLS).

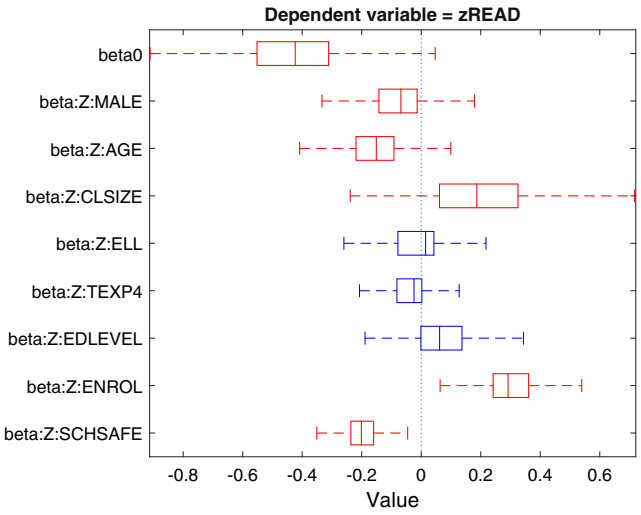


Fig. 6 *Box plots* of the marginal posterior distributions of the intercept (beta0) and slope coefficients of the infinite probits mixture model. *Center vertical line*: posterior median. *Thick box*: inter-quartile range (50 % interval). *Horizontal lines* (whiskers): 95 % credible interval (.025 and .975 marginal posterior quantiles)

The dependent variable is student literacy score (zREAD). The eight covariates are: student male status (Z:MALE) indicator (0 or 1) and age (AGE); student's class size (SIZE) and class percent English language learners (ELL); student's teacher years of experience (TEXP4) and education level (TEXP4 = 5 if bachelor's degree; TEXP4 = 6 if at least master's degree); student's school enrollment (ENROL) and safety rating (SAFE = 1 is high; SAFE = 3 is low). Figure 3 presents the univariate descriptive statistics of these variables, from a text output file obtained by the menu option: Describe/Plot Data Set > Summaries and Frequencies > Univariate descriptives. Data is also available on a school identifier variable (SCHID).

The data for each of the 8 covariates were transformed into z-scores having mean 0 and variance 1, using the menu option: Modify Data Set > Simple variable transformations > Z-score. These transformations will make the slope coefficients of the 8 covariates (resp.) interpretable on a common scale, in a regression analysis.

The BNP infinite-probits mixture model (21a–21j) was fit to the PIRLS100 data, using the dependent variable and the 8 z-scored covariates. For the model's prior density function (22a–22b), the following software defaults were specified for the prior parameters: $b_{\sigma\mu} = 5$, $v = 100$, $a_0 = 0.01$, $v_\omega = 10$, and $a_\omega = 0.01$. This represents an attempt to specify a rather non-informative prior distribution for the model parameters. In order to estimate the model's posterior distribution, 50,000 MCMC sampling iterations were run, using an initial burn-in of 2,000 and thinning interval of 5. In the software, these sampling iterations MCMC completed in less than one minute on a modern computer, including the automatic opening of the text output (e.g., see Figs. 4 and 5).

Figure 4 presents the model's predictive data-fit statistics (including their 95 % MCCI half-widths), from text output that was opened by the software immediately after the 50K MCMC sampling iterations were run. The model obtained a $D(m)$ statistic of 32, with an R-squared of .96, and had no outliers according to standardized residuals that ranged within −2 and 2.

Figure 5 presents the (marginal) posterior point-estimates of all the model parameters, calculated from the 50K MCMC samples (aside from the burn-in and thinned-out MCMC samples). Estimates include the marginal posterior mean, median, standard deviation, 50 % credible interval (given by the 25 % (.25 quantile) and 75 % (.75 quantile) output columns), and the 95 % credible interval (2.5 %, 97.5 % columns). Figure 5 is part of the same text output that reported the model fit statistics (Fig. 4).

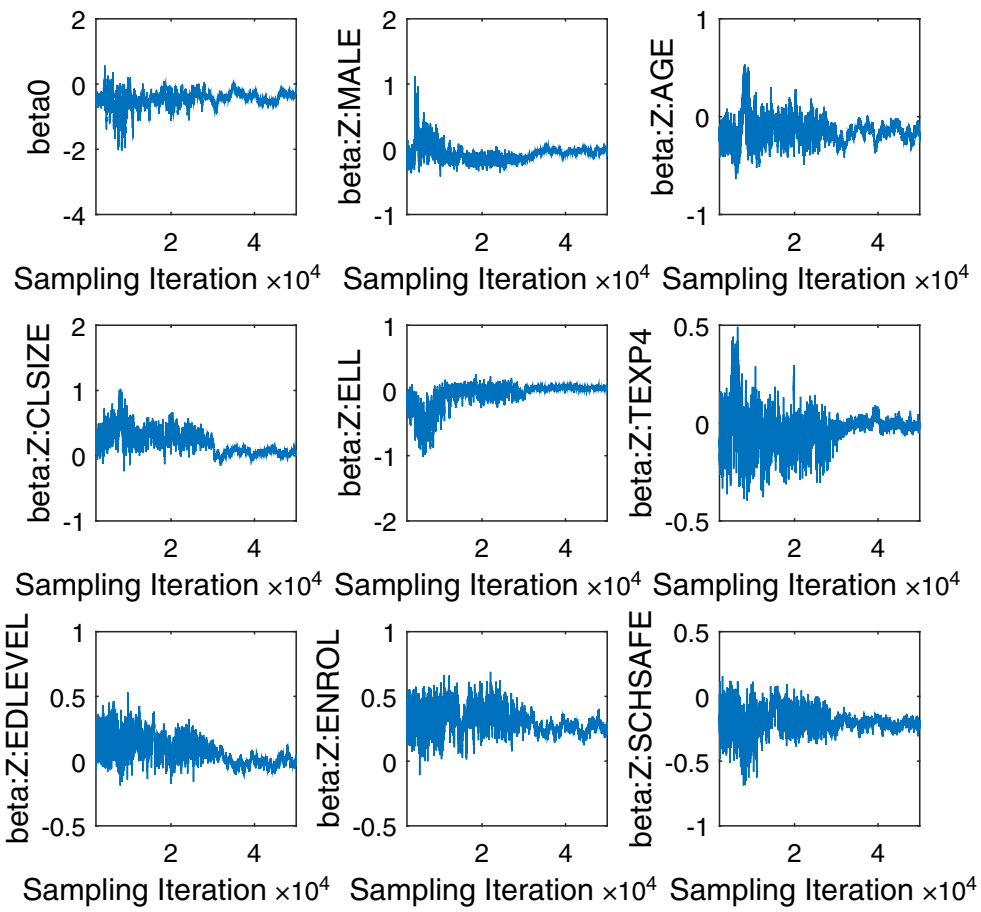


Fig. 7 Trace plots of MCMC samples of the intercept and slope coefficients for the infinite probots mixture model

Posterior Summary Estimates: 95% Monte Carlo Confidence Interval (MCCI) Half-Widths

Parameter	Mean	Med	SD	25%	75%	2.5%	97.5%
beta0	0.056	0.049	0.050	0.084	0.040	0.162	0.064
beta:Z:MALE	0.040	0.037	0.026	0.032	0.055	0.032	0.087
beta:Z:AGE	0.027	0.025	0.020	0.023	0.038	0.032	0.060
beta:Z:CLSIZE	0.062	0.061	0.020	0.051	0.074	0.033	0.100
beta:Z:ELL	0.065	0.064	0.022	0.080	0.051	0.112	0.031
beta:Z:TEXP4	0.020	0.020	0.017	0.026	0.020	0.042	0.035
beta:Z:EDLEVEL	0.030	0.030	0.011	0.024	0.036	0.015	0.050
beta:Z:ENROL	0.021	0.021	0.013	0.018	0.027	0.020	0.040
beta:Z:SCHSAFE	0.018	0.018	0.014	0.021	0.020	0.034	0.030
sigma^2	0.080	0.079	0.019	0.070	0.091	0.057	0.114

Fig. 8 Ninety-five percent MCCI half-widths of the marginal posterior point-estimates of the intercept and slope parameters of the infinite-probits mixture model

Figure 6 is a box plot of the (marginal) posterior quantile point-estimates of the intercept and slope coefficient parameters for the 8 covariates (i.e., parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\intercal$). A red box (blue box, resp.) flags a coefficient parameter that is (not, resp.) significantly different than zero, according to whether or not the 50 % (marginal) posterior interval (box) includes zero. The box plot menu option is available after clicking the Posterior Summaries button.

MCMC convergence, for the parameters of interest for data analysis, can be evaluated by clicking certain menu options after clicking the Posterior Summaries button. Figure 7 presents the trace plots for the model's intercept and slope parameters sampled over the 50K MCMC sampling iterations. The trace plots appear to support good mixing for these parameters, as each trace plot appears stable and "hairy." (According to the trace plot results, a larger burn-in may have been selected. This would lead to discarding additional MCMC samples, but at the cost of estimation efficiency in the posterior analysis; see MacEachern & Berliner, 1994). Figure 8 presents the 95 % MCCI half-widths of the (marginal) posterior coefficient point-estimates of Fig. 5. The half-widths are nearly all less than .10, and thus these posterior point-estimates are reasonably accurate in terms of Monte Carlo standard error. If necessary, the software can re-click the Posterior Analysis button, to run additional MCMC sampling iterations, to obtain more precise posterior point-estimates of model parameters (as would be indicated by smaller 95 % MCCI half-widths).

The user can click the Posterior Predictive button to investigate how chosen features (functionals) of the posterior predictive distribution varies as a function of one or more selected (focal) covariates, through graphical and text output.

Figure 9 provides a quantile and mean regression analysis, by showing the estimates of the mean, and the .1, .25, .5 (median), .75, and .9 quantiles of the model's posterior predictive distribution of zREAD, conditionally on selected values $-2, -1.5, \ldots, 1, 1.5, 2$ of the Z:CLSIZE covariate, and on zero for all the 7 other covariates (using the zero-centering method). It presents nonlinear relationships between literacy performance (zREAD) and class size

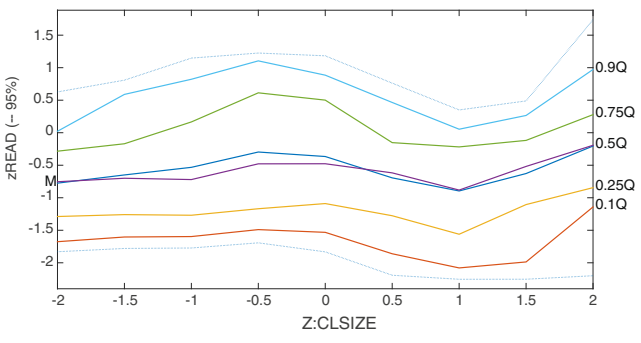


Fig. 9 For the infinite probits mixture model, the posterior predictive mean and quantiles of zREAD, over chosen values of the covariate Z:CLSIZE. Specifically, the .1, .25, .5 (median), .75, and .90 quantiles of zREAD

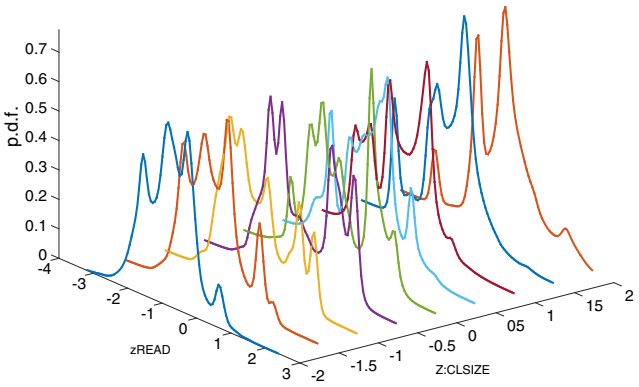


Fig. 10 For the infinite probits mixture model, the Rao-Blackwellized estimate of the posterior predictive p.d.f. of zREAD, as a function of Z:CLSIZE

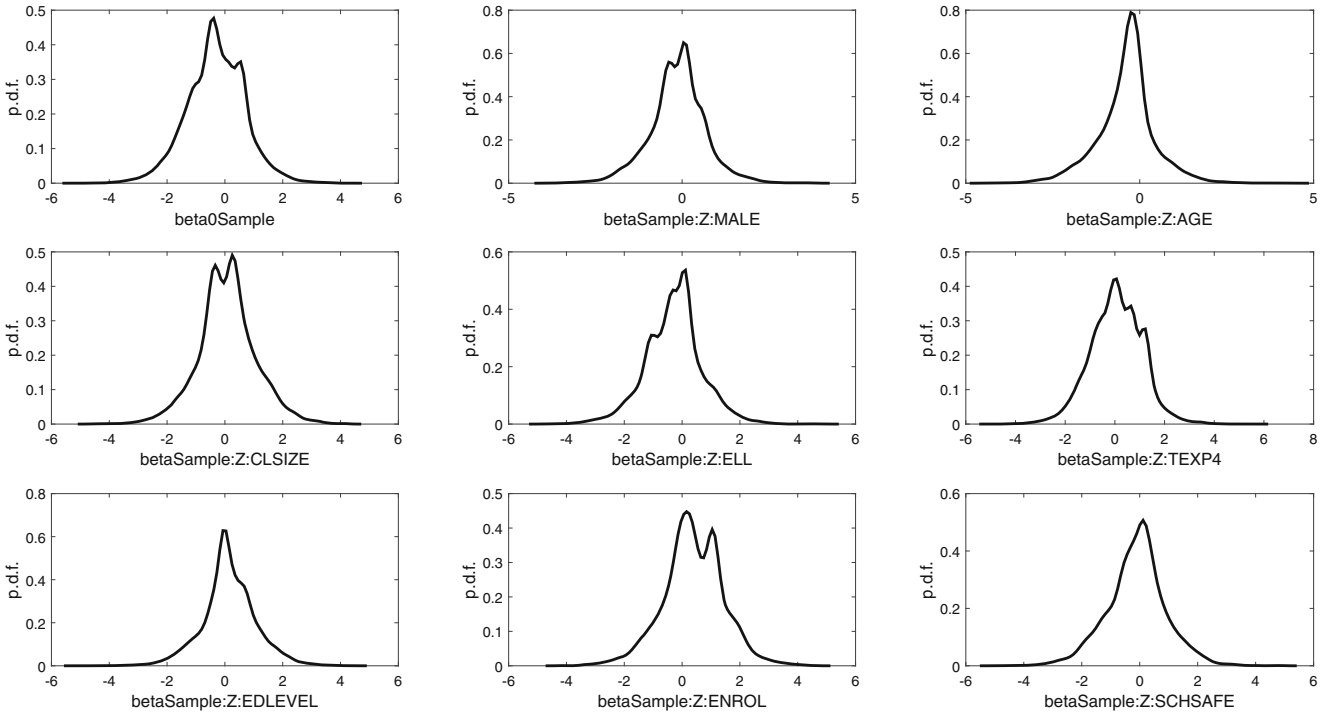


Fig. 11 Marginal posterior density estimates of the intercept and slope coefficients, under the ANOVA-linear DDP model

(Z:CLSIZE). In particular, for higher literacy-performing students (i.e., .75 and .90 quantiles of zREAD), there is a highly nonlinear relationship between zREAD and class size. For lower performing students (i.e., .1 and .25 quantiles of zREAD), there is a flatter relationship between zREAD and class size. For "middle" performing students (mean and .5 quantiles of zREAD), there is a rather flat relationship between zREAD and class size.

Figure 10 shows a 3-dimensional plot of (Rao-Blackwellized) estimates of the model's posterior predictive p.d.f. of zREAD, conditionally on the same values of the 8 covariates mentioned earlier. As shown, both the location and shape of the zREAD distribution changes as a function of class size. For each posterior predictive p.d.f. estimate (panel), a homogeneous cluster of students (in terms of a shared zREAD score) is indicated by a bump (or mode) in the p.d.f.

The ANOVA-linear DDP model (19a–19h) was fit to the PIRLS100 data. For this model's prior (19a–19b), the following parameters were chosen: $r_0 = 10$, $s_0 = 10$, $a_0 = 2$, $a_\alpha = 1$, and $b_\alpha = 1$. This was an attempt to specify a rather non-informative prior distribution for the model parameters. This model was estimated using 50,000 MCMC sampling iterations, with burn-in of 2,000 and thinning interval of 5. MCMC convergence was confirmed according to the results of the univariate trace plots together with the results of small 95 % MCCI half-widths for (marginal) posterior point-estimates (nearly all less than .1). Figure 11 presents the marginal posterior densities (distributions) of the intercept and slope parameters, from the mixture distribution of the model. Each of these distributions (densities) are skewed and/or multimodal, unlike normal distributions. Each mode (bump) in a distribution (density) refers to a homogeneous cluster of schools, in terms of a common value of the given (random) coefficient parameter.

Finally, recall that for the PIRLS100 data, the infinite-probits mixture model (21a–21j) obtained a $D(m)$ statistic of 32, an R-squared of .96, and no outliers according to all the absolute standardized residuals being less than 2 (Fig. 4). The ANOVA-linear DDP model (19a–19h) obtained a $D(m)$ statistic of 113, an R-squared of .64, and no outliers. The ordinary linear model obtained a $D(m)$ statistic of 139, an R-squared of .24 and 5 outliers. This linear model assumed rather non-informative priors, with $\beta_0 \sim \mathrm{N}(0, v_0 \to \infty)$, $\beta_k \sim \mathrm{N}(0, 1000)$, $k = 1, \ldots, 8$, and $\sigma^2 \sim \mathrm{IG}(.001, .001)$.

Conclusions

We described and illustrated a stand-alone, user-friendly and menu-driven software package for Bayesian data analysis. Such analysis can be performed using any one of many Bayesian models that are available from the package,

including BNP infinite-mixture regression models and normal random-effects linear models. As mentioned in the data analysis exercises of Appendix B, the software can be used to address a wide range of statistical applications that arise from many scientific fields. In the future, new Bayesian mixture models will be added to the software. They will include mixture models defined by other kernel density functions, and models with parameters assigned a mixture of Pólya Trees prior.

Acknowledgments This paper is supported by NSF-MMS research grant SES-1156372. The author appreciates the helpful comments made by the Associate Editor and two anonymous referees.

Appendix A: Technical preliminaries

We use the following notation for random variables and probability measures (functions) (e.g., Berger & Casella, 2002; Schervish, 1995). A random variable is denoted by an italicized capital letter, such as Y, with Y a function from a sample space $\mathcal{Y}$ to a set of real numbers. A realization of that random variable is denoted by a lower case, with $Y = y$. Also, $F(\cdot)$ (or $P(\cdot)$) is the probability measure (function) that satisfies the Kolmogorov probability axioms, having sample space domain $\mathcal{Y}$ and range (0, 1). Then $F(B)$ (or $P(B)$) denotes the probability of any event $B \subset \mathcal{Y}$ of interest. Throughout, a probability measure is denoted by a capital letter such as F (or G or Π resp., for example), and $f(y)$ ($g(y)$ and $\pi(y)$, resp.) is the corresponding probability density of y, if Y is discrete or continuous.

If Y is a continuous random variable, with sample space $\mathcal{Y} \subset \mathbb{R}^d$ ($d \in \mathbb{Z}_+$) and Borel σ-field $\mathcal{B}(\mathcal{Y})$, and with a probability density function (p.d.f.) f defined on $\mathcal{Y}$ such that $\int f(y)\mathrm{d}y = 1$, then the probability measure of f is given by $F(B) = \int_B f(y)\mathrm{d}y$ for all $B \in \mathcal{B}(\mathcal{Y})$, with $f(y) = \mathrm{d}F(y)/\mathrm{d}y \geq 0$. If Y is a discrete random variable with a countable sample space $\mathcal{Y} = \{y_1, y_2, \ldots\}$ and Borel σ-field $\mathcal{B}(\mathcal{Y})$, and with probability mass function (p.m.f.) f defined on $\mathcal{Y}$ such that $\sum_{k=1}^{\infty} f(y_k) = 1$, then the probability measure of f is given by $F(B) = \sum_{\{k:y_k \in B\}} f(y_k)$ for all $B \in \mathcal{B}(\mathcal{Y})$, with $0 \leq f(y) = P(Y = y) \leq 1$. Also, a cumulative distribution function (c.d.f.) is the probability measure $F(y) := F(B) = P(Y \leq y)$, where $B = \{y' : -\infty < y' \leq y\}$. The c.d.f. $F(y)$ corresponds to a p.d.f. $f(y) = \mathrm{d}F(y)/\mathrm{d}y$ if Y is continuous; and corresponds to a p.m.f. $f(y) = P(Y = y)$ if Y is discrete. A multidimensional random variable, $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$, has c.d.f. $F(y_1, \ldots, y_d) = P(Y_1 \leq y_1, \ldots, Y_k \leq y_d)$.

$Y \sim F$ means that the random variable Y has a distribution defined by the probability measure F. Thus F is also called a distribution. Notation such a $Y \sim F(y \,|\, x)$ (or $Y \sim F(y \,|\, \mathbf{x})$, resp.) means that the random variable $Y \,|\, x$ (the random variable, $Y \,|\, \mathbf{x}$, resp.) has a distribution defined by a probability measure $F(\cdot \,|\, x)$ conditionally on the value x of a variable X (or has a distribution $F(\cdot \,|\, \mathbf{x})$ conditionally on the value $\mathbf{x} = (x_1, \ldots, x_p)$ of p variables). Sometimes a probability distribution (measure) is notated without the $\cdot\,|$ symbol.

We denote $\mathrm{N}(\cdot \,|\, \mu, \sigma^2)$, $\Phi(\cdot) = \mathrm{N}(\cdot \,|\, 0, 1)$, $\mathrm{Ga}(\cdot \,|\, a, b)$, $\mathrm{IG}(\cdot \,|\, a, b)$, $\mathrm{U}(\cdot \,|\, a, b)$, $\mathrm{Be}(\cdot \,|\, a, b)$, $\mathrm{GIG}(\cdot \,|\, a, b, q)$, $\mathrm{N}(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathrm{IW}(\cdot \,|\, d, \boldsymbol{S})$, and $\mathrm{Di}(\cdot \,|\, \boldsymbol{\alpha})$, as the probability measures of a normal distribution with mean and variance parameters (μ, σ^2); the standard normal distribution; the gamma distribution with shape and rate parameters (a, b); the inverse-gamma distribution with shape and rate parameters (a, b); the uniform distribution with minimum and maximum parameters (a, b); the beta distribution with shape parameters (a, b); the inverse-Gaussian distribution with mean $\mu > 0$ and shape $\lambda > 0$; the generalized inverse-Gaussian distribution with parameters $(a, b, q) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$; the multivariate normal distribution for a random vector $(Y_1, \ldots, Y_d)$, with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\intercal$ and $d \times d$ variance-covariance matrix $\boldsymbol{\Sigma}$; the inverted-Wishart distribution with degrees of freedom d and scale matrix $\boldsymbol{S}$; and of the Dirichlet distribution with precision parameters $\boldsymbol{\alpha}$, respectively. These distributions are continuous and define p.d.f.s denoted by lower-case letters, with $\mathrm{n}(\cdot \,|\, \mu, \sigma^2)$, $\mathrm{n}(\cdot \,|\, 0, 1)$, $\mathrm{ga}(\cdot \,|\, a, b)$, $\mathrm{ig}(\cdot \,|\, a, b)$, $\mathrm{u}(\cdot \,|\, a, b)$, $\mathrm{be}(\cdot \,|\, a, b)$, $\mathrm{gig}(\cdot \,|\, a, b, q)$, $\mathrm{n}(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathrm{iw}(\cdot \,|\, d, \boldsymbol{S})$, and $\mathrm{di}(\cdot \,|\, \boldsymbol{\alpha})$, respectively. The p.d.f. equations are found in statistical distribution textbooks (e.g., Johnson et al., 1994).

Appendix B: BNP data analysis exercises

You may use the Bayesian regression software to answer the following data analysis exercises. Relevant data sets are available from the software, and are described under the Help menu.

1. **(Regression)** Perform regression analyses of the PIRLS100.csv data from the 2006 Progress in International Reading Literacy Study (PIRLS). Data are from a random sample of 100 students who each attended one of 21 U.S. low income schools. The dependent variable, zREAD, is a continuous-valued student literacy score. Also, consider eight covariates (predictors): a binary (0,1) indicator of male status (MALE), student age (AGE), size of student's class (CLSIZE), percent of English language learners in the student's classroom (ELL), years of experience of the student's teacher (TEXP), years of education of the student's teacher (EDLEVEL), the number of students enrolled in the student's school (ENROL), and a 3-point rating

of the student's school safety (SCHSAFE = 1 is highest; SCHSAFE = 3 is lowest). There is also data on a school ID (SCHID) grouping variable.

Analyze the data set using the regression models listed below, after performing a z-score transformation of the 8 covariates using the menu option: Modify Data Set > Simple variable transformations > Z-score. The SCHID grouping variable may be included in a multi-level regression model.

- ANOVA/linear DDP model (Dirichlet process (DP) mixture of linear regressions).
- An infinite-probits model.
- Another Bayesian nonparametric (infinite) mixture of regression models.

(a) Now, describe (represent) each of the models, mathematically and in words.
(b) Summarize the results of your data analyses, while focusing on the relevant results. Describe the relationship between the dependent variable and each covariate. For the infinite-probits model, describe the clustering behavior in the posterior predictive distribution (density) of the dependent variable, conditionally on values of one or two covariates of your choice. For the DDP model, graphically describe the clustering behavior of the random intercept and slope parameters.
(c) Evaluate how well the data support the assumptions of other parametric models, such as the linearity of regression, the normality of the regression errors, and the normality of the random intercept and slope coefficients. Fit a normal linear regression and normal random-effects (multi-level HLM) models to the data, and compare the predictive fit between the linear model(s) and the Bayesian nonparametric models above.

2. **(Binary Regression)** Analyze the data using the following binary regression models, with binary (0,1) dependent variable READPASS, and the 8 z-scored covariates.

- ANOVA/linear DDP logit (or probit) model (a DP mixture of regressions).
- An infinite-probits model for binary regression.
- Another Bayesian nonparametric (infinite) mixture of regression models.

(a) Describe (represent) each of the models, mathematically and in words.
(b) Summarize the results of your data analyses, while focusing on the relevant results. Describe the relationship between the dependent variable and each covariate. For the infinite-probits model, describe the clustering behavior in the posterior predictive distribution (density) of the (latent) dependent variable, conditionally on values of 1 or 2 covariates of your choice. For the DDP model, graphically describe the clustering behavior of the random intercept and slope parameters.
(c) Evaluate how well the data support assumptions of other parametric models, namely, the unimodality of the (latent) regression errors, and the normality of the random intercept and slope coefficients. This will require fitting a probit (or logit) linear and/or random-effects regression model to the data. Compare the fit between the probit (or logit) model(s) and the Bayesian nonparametric models listed above.

3. (**Causal Analysis**) Analyze the PIRLS100.csv data using a Bayesian nonparametric regression model. Investigate the causal effect of large (versus small) class size (**CLSIZE**) on reading performance (**zREAD**), in the context of a regression discontinuity design (e.g., Cook, 2008). For the data set, use a Modify Data Set menu option to construct a new (treatment) variable defined by a (0, 1) indicator of large class size, named **LARGE**, where **LARGE** = 1 if **CLSIZE** $\geq$ 21, and zero otherwise. Perform a regression of **zREAD** on the predictors (**LARGE**, **CLSIZE**), in order to infer the causal effect of large class size (**LARGE** = 1) versus small class size (**LARGE** = 0), on the variable **zREAD**, conditionally on **CLSIZE** = 21. Do so by inferring the coefficient estimates of the covariate **LARGE**, and by performing posterior predictive inferences of **zREAD**, conditionally on **LARGE** = 1 versus **LARGE** = 0. Under relatively mild conditions, such comparisons provide inferences of causal effects (of large versus small class size), as if the treatments were randomly assigned to subjects associated with class sizes in a small neighborhood around **CLSIZE** = 21. A key condition (assumption) is that students have imperfect control over the **CLSIZE** variable, around the value of **CLSIZE** = 21 (Lee, 2008; Lee & Lemieux, 2010).
4. (**3-level data**) Analyze the Classroom data set (classroom300.csv) using a Bayesian nonparametric regression model. In the data, students (Level 1) are nested within classrooms (Level 2) nested within schools (Level 3). A classroom identifier (classid) provides a level-2 grouping variable. A school identifier (schoolid) provides a level-3 grouping variable. Model mathgain as the dependent variable, defined by student gain in mathematical knowledge. Also, for the model, include at least three covariates: student socioeconomic status (ses), the level of experience of the student's teacher (yearsteaching), and student house poverty level (housepov).

5. **(Longitudinal Data Analysis)** Fit a Bayesian nonparametric regression model to analyze the GPA.csv data. Investigate the relationship between gpa and the covariates of time, job type, gender status, and possibly student ID. Consider performing an auto-regressive time-series analysis of the data, using the model. For this, you can use a Modify Data Set menu option to construct lag-terms of the dependent variable (for selected lag order) and then include them as additional covariates in the regression model.
6. **(Meta Analysis)** Fit a Bayesian nonparametric regression model to perform a meta-analysis of the Calendar.csv data. The dependent variable is **Effect** size, here defined by the unbiased standardized difference (Hedges, 1981) between mean student achievement in schools that follow a year-round calendar, and (minus) the mean student achievement in schools that follow the traditional nine-month calendar. The covariates are standard error of the effect size (**SE_ES**), and study publication **Year**; and the observation weights are given by the variable **weight** (= 1/**Var**). The overall effect-size, over studies, is represented by the model's intercept parameter. In your report of the data analysis results, common on the impact of publication bias on estimates of the overall effect size. Publication bias may be assessed by inferring the slope coefficient estimate of the **SE_ES** covariate; and/or by posterior predictive inferences of the Effect dependent variable, conditionally over a range of values of **SE_ES**. Such inferences provide regression analyses for a funnel plot (Thompson & Sharp, 1999).
7. **(Survival Analysis of censored data)** Use a Bayesian nonparametric regression model to perform a survival analysis of the larynx.csv data or the bcdeter.csv data. The larynx data set has dependent variable **logyears**; and covariates of patient **age**, cancer **stage**, and diagnosis year (**diagyr**). The observations of **logyears** are either right-censored or uncensored, as indicated by the variables **LB** and **UB**. The bcdeter data set has dependent variable **logmonths**, and treatment type indicator covariates (**radioth**, **radChemo**). The **logmonths** observations are either right-censored, interval-censored, or uncensored, as indicated by the variables **logLBplus1** and **logUBplus1**.
8. (**Item Response Analysis**) Use a BNP model to perform an item response (IRT) analysis of either the NAEP75.csv data set, or the Teacher49.csv data set. The NAEP75 data has binary item-response scores (0 = Incorrect, 1 = Correct). The Teacher49 data has ordinal item response scores (0, 1, or 2). Each data set has information about an examinee, per data row, with examinee item responses in multiple data columns. For either data set, use a "vectorize" Modify Data Set menu option, to collapse the multiple item response variables (columns) into a single column dependent variable, and to construct item dummy (0 or −1) covariates. Report the (marginal) posterior estimates of examinee ability, item difficulty (for each test item), as well as the estimates of the other model parameters.

References

Agresti, A. (2002). *Categorical data analysis.* Wiley.

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Andridge, R.R., & Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*, 40–64.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, *2*, 1152–1174.

Atchadé, Y., & Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, *11*, 815–828.

Barrios, E., Lijoi, A., Nieto-Barajas, L., & Prünster, I. (2015). Package BNPdensity: Ferguson-Klass type algorithm for posterior normalized random measures.

Barry, D., & Hartigan, J. (1993). A Bayesian-analysis for change point problems. *Journal of the American Statistical Association*, *88*, 309–319.

Begg, C., & Gray, R. (1996). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71*, 11–18.

Bernardo, J., & Smith, A. (1994). *Bayesian theory*. Chichester: Wiley.

Bertoin, J. (1998). *Lévy Processes*. Cambridge: Cambridge University Press.

BIPS Development Team (2015). BIPS: Bayesian inference for the physical sciences.

Borenstein, M. (2009). Effect sizes for continuous data. In Cooper, H., Hedges, L., & Valentine, J. (Eds.) *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York: Russell Sage Foundation.

Botev, Z., Grotowski, J., & Kroese, D. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, *38*, 2916–2957.

Bowman, A., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations*. Oxford: Oxford University Press.

Brooks, S. (1998). Quantitative convergence assessment for Markov chain Monte Carlo via cusums. *Statistics and Computing*, *8*, 267–274.

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC.

Burr, D. (2012). bspmma: An R package for Bayesian semiparametric models for meta analysis. *Journal of Statistical Software*, *50*, 1–23.

Burr, D., & Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, *100*, 242–251.

Casella, G., & Berger, R. (2002). *Statistical inference*, 2nd edn. Duxbury, Pacific Grove: CA.

Cepeda, E., & Gamerman, D. (2001). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, *14*, 207–221.

Cook, T. (2008). Waiting for life to arrive: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*, 636–654.

Cooper, H., Hedges, L., & Valentine, J. (2009). *The handbook of research synthesis and meta-analysis (2nd Ed.).* New York: Russell Sage Foundation.

DeBlasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I., & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*, 212–229.

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

Denison, D., Holmes, C., Mallick, B., & Smith, A. (2002). *Bayesian methods for nonlinear classification and regression*. New York: Wiley.

Development Team (2015). Stan: A C++ library for probability and sampling.

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

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

Evans, I. (1965). Bayesian estimation of parameters of a multivariate normal distribution. *Journal of the Royal Statistical Society, Series B*, *27*, 279–283.

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

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

Flegal, J., & Jones, G. (2011). Implementing Markov chain Monte Carlo: Estimating with confidence. In Brooks, S., Gelman, A., Jones, G., & Meng, X. (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 175–197). Boca Raton, FL: CRC.

Fleiss, J., & Berlin, J. (2009). Effect size for dichotomous data. In Cooper, H., Hedges, L., & Valentine, J. (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 237–253). New York: Russell Sage Foundation.

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: l2 theory. *Probability Theory and Related Fields*, *57*, 453–476.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

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

Fuentes-García, R., Mena, R., & Walker, S. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics*, *39*, 669–682.

Gelfand, A., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. Boca Raton: Chapman and Hall/CRC.

Gelfand, A., & Ghosh, J. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, *85*, 1–11.

Gelfand, A., & Mukhopadhyay, S. (1995). On nonparametric Bayesian inference for the distribution of a random sample. *Canadian Journal of Statistics*, *23*, 411–420.

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

Gelman, A., Carlin, A., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*, 2nd edn. Boca Raton, Florida: Chapman and Hall.

George, E., & McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

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

Geyer, C. (2011). Introduction to MCMC. In Brooks, S., Gelman, A., Jones, G., & Meng, X. (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 3–48). Boca Raton, FL: CRC.

Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York: Springer.

Gilks, W., Wang, C., Yvonnet, B., & Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, *49*, 441–453.

Green, P., & Silverman, B. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach CRC press*. Los Altos: CA.

Hahn, J., Todd, P., & der Klaauw, W.V. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*, 201–209.

Hansen, B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*, 481–488.

Hanson, T. (2006). Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, *101*, 1548–1565.

Hartigan, J. (1990). Partition models. *Communications in Statistics: Theory and Methods*, *19*, 2745–2756.

Hastie, T., Tibshiriani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*, 2nd edn. New York: Springer-Verlag.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128.

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

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Holmes, C., & Held, K. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, *1*, 145–168.

Corp, I.B.M. (2015). *IBM SPSS Statistics for windows, version 22.0 IBM corp.* Armonk: NY.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics*, *86*, 4–29.

Imbens, G.W., & Lemieux, T. (2008). Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, *142*, 615–635.

Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

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

JAGS Development Team (2015). JAGS: Just Another Gibbs Sampler.

James, L., Lijoi, A., & Prünster, I. (2009). Posterior analysis for norMalized random measures with independent increments. *Scandinavian Journal of Statistics*, *36*, 76–97.

Jara, A., Hanson, T., Quintana, F., Müller, P., & Rosner, G. (2011). DPPackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, *40*, 1–20.

Jara, A., Lesaffre, E., DeIorio, M., & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics*, *4*, 2126–2149.

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions (vol. 1)*. New York: Wiley.

Jordan, M., & Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.

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

Karabatsos, G. (2014). Fast Marginal Likelihood Estimation of the Ridge Parameter in Ridge Regression. Technical report. *ArXiv preprint*, *1409*, 2437.

Karabatsos, G. (2016). A menu-driven software package for Bayesian regression analysis. *The ISBA Bulletin*, *22*(4), 13–16.

Karabatsos, G., Talbott, E., & Walker, S. (2015). A Bayesian nonparametric meta-analysis model. *Research Synthesis Methods*, *6*, 28–44.

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

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

Karabatsos, G., & Walker, S. (2015a). A Bayesian Nonparametric Causal Model for Regression Discontinuity Designs. In Müller, P., & Mitra, R. (Eds.) *Nonparametric Bayesian Methods in Biostatistics and Bioinformatics*. ArXiv preprint 1311.4482) (pp. 403–421). New York: Springer-Verlag.

Karabatsos, G., & Walker, S. (2015b). Bayesian nonparametric irt. In Van der Linden, W. (Ed.), *Handbook Of Item Response Theory, Volume 1: Models, Statistical Tools, and Applications*. New York: Taylor and Francis.

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

Klein, J., & Moeschberger, M. (2010). *Survival analysis*, 2nd edn. New York: Springer-Verlag.

Laud, P., & Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B*, *57*, 247–262.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, *142*, 675–697.

Lee, D., & Lemieux, T. (2010). Regression discontinuity designs in economics. *The Journal of Economic Literature*, *48*, 281–355.

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

Lindley, D., & Smith, A. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 1–41.

Liverani, S., Hastie, D., Azizi, L., Papathomas, M., & Richardson, S. (2015). PREmium: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, *64*(7), 1–30.

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

Lucca, M.D., Guglielmi, A., Müller, P., & Quintana, F. (2012). A simple class of Bayesian nonparametric autoregression models. *Bayesian Analysis*, *7*(3), 771–796.

Lunceford, J., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, *23*, 2937–2960.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067.

MacEachern, S. (1999). Dependent nonparametric processes. *Proceedings of the Bayesian Statistical Sciences Section of the American Statistical Association*, 50–55.

MacEachern, S. (2000). *Dependent Dirichlet processes Technical report*. The Ohio State University: Department of Statistics.

MacEachern, S. (2001). Decision theoretic aspects of dependent nonparametric processes. In George, E. (Ed.), *Bayesian Methods with Applications to Science, Policy and Official Statistics* (pp. 551–560): International Society for Bayesian Analysis.

MacEachern, S., & Berliner, L. (1994). Subsampling the Gibbs sampler. *The American Statistician*, *48*(3), 188–190.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*, 2nd edn. London: Chapman and Hall.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley and Sons.

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

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer-Verlag.

Müller, P., & Quintana, F. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*, 95–110.

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

Neal, R. (2003). Slice sampling (with discussion). *Annals of Statistics*, *31*, 705–767.

NIMBLE Development Team (2015). NIMBLE: An R package for programming with BUGS models.

Nychka, D. (2000). Spatial-process estimates as smoothers. In *Smoothing and Regression: Approaches, Computation, and Application* (pp. 393–424). New York: Wiley.

O'Hagan, A., & Forster, J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference volume 2B*. London: Arnold.

Perman, M., Pitman, J., & Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, *92*, 21–39.

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

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T., Shapeley, L., & MacQueen, J. (Eds.), *Statistics, probability and game theory. Papers in honor of David Blackwell* (pp. 245–268). Hayward: Institute of Mathematical Sciences.

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

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

Prado, R., & West, M. (2010). *Time series*. Modeling: Computation, and Inference. Chapman and Hall/CRC.

Quintana, F., & Iglesias, P. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B*, *65*, 557–574.

Development Core Team, R. (2015). *R: a language and environment for statistical computing*. Vienna Austria: R Foundation for statistical computing.

Rasmussen, C., Ghahramani, Z., Becker, S., & Ghahramani, Z. (2002). Infinite mixtures of Gaussian process experts. In Diettrich, T. (Ed.), *Advances in Neural Information Processing Systems* (Vol. 14, pp. 881–888). Cambridge, MA: The MIT Press.

Regazzini, E., Lijoi, A., & Prünster, I. (2003). Distributional results for means of norMalized random measures with independent increments. *Annals of Statistics*, *31*, 560–585.

Robert, C., & Casella, G. (2004). *Monte carlo statistical methods*, 2nd edn. New York: Springer.

Rosenbaum, P., & Rubin, D. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenthal, R. (1994). Parametric measures of effect size. In Cooper, H., & Hedges, L. (Eds.), *The Handbook of Research Synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Schervish, M. (1995). *Theory of statistics*. New York: Springer-Verlag.

Scott, D. (1992). *Multivariate density estimation: Theory, practice and visualization*. New York: John Wiley and Sons.

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

Sethuraman, J., & Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameters. In Gupta, S., & Berger, J. (Eds.), *Statistical Decision Theory and Related Topics III: Proceedings of the Third Purdue Symposium, Volume* (pp. 305–315). New York: Academic Press.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Boca Raton Florida: Chapman and hall.

Stroud, J., Müller, P., & Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society: Series B*, *63*, 673–689.

Stuart, E. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, *25*, 1–21.

Thistlewaite, D., & Campbell, D. (1960). Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*, 309–317.

Thomas, A. (1994). BUGS: A statistical modelling package. *RTA/BCS Modular Languages Newsletter*, *2*, 36–38.

Thompson, S., & Sharp, S. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, *18*, 2693–2708.

van der Linden, W. (2015). *Handbook of modern item response theory*, 2nd edn. Monterey: CTB McGraw-Hill.

Verbeke, G., & Molenbergs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Walker, S., Damien, P., Laud, P., & Smith, A. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B*, *61*, 485–527.

Wilk, M., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1–17.

Wu, T.-F., Lin, C.-J., & Weng, R. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*, 975–z1005.