Europe PMC Funders Group
Author Manuscript
Electron J Stat. Author manuscript; available in PMC 2018 April 03.

Published in final edited form as:
Electron J Stat. 2016 July 18; 10(2): 1807–1828. doi:10.1214/15-EJS1032.

Scalable Bayesian nonparametric regression via a Plackett-Luce model for conditional ranks

Tristan Gray-Davies,
Department of Statistics, University of Oxford

Chris C. Holmes, and
Department of Statistics, University of Oxford

François Caron
Department of Statistics, University of Oxford

Abstract

We present a novel Bayesian nonparametric regression model for covariates X and continuous response variable $Y \in \mathbb{R}$. The model is parametrized in terms of marginal distributions for Y and X and a regression function which tunes the stochastic ordering of the conditional distributions $F(y/x)$. By adopting an approximate composite likelihood approach, we show that the resulting posterior inference can be decoupled for the separate components of the model. This procedure can scale to very large datasets and allows for the use of standard, existing, software from Bayesian nonparametric density estimation and Plackett-Luce ranking estimation to be applied. As an illustration, we show an application of our approach to a US Census dataset, with over 1,300,000 data points and more than 100 covariates.

Keywords and phrases

Bayesian nonparametrics; Composite likelihood; Plackett-Luce; Pólya Tree; Dirichlet process mixtures

1 Introduction

Bayesian nonparametric regression offers a flexible and robust way of modeling the dependence between covariates $x \in \mathcal{X}$ and a response variable $Y \in \mathbb{R}$ by using models with larger support than their parametric counterparts. Nonparametric statistical models are motivated by robustness and their ability to capture effects such as outliers, strong nonlinearities or multimodalities, while providing probabilistic measures of predictive uncertainty. Bayesian nonparametric regression methods are largely underpinned by one of two random probability measures namely, Dirichlet process mixtures (Ferguson, 1973; Lo, 1984) and Pólya trees (Lavine, 1992, 1994). These approaches, widely applied to density estimation problems (see e.g. Hjort *et al.*, 2010), have been used as building blocks of various nonparametric regression models through a number of different approaches.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

One approach, called the conditional approach, considers the covariates as fixed, and models directly the conditional distribution $f(y/x)$ of the response given the covariate. This conditional distribution may be constructed in a semiparametric or fully nonparametric way. The semiparametric conditional approach typically assumes that

$$Y = \eta(x) + \epsilon \quad (1)$$

where η is some unknown flexible mean function and ϵ is the residual. Regression models (priors) have been proposed for the mean function η such as Gaussian processes (see e.g. Rasmussen, 2006), basis function representations such as splines or kernels (Denison *et al.*, 2002; Müller & Quintana, 2004) or Bayesian regression trees (Chipman *et al.*, 2010). More generally, Kottas & Gelfand (2001) and Lavine & Mockus (1995) proposed to use Dirichlet process mixtures for the distribution of the residuals, while Pati & Dunson (2014) jointly model the mean function and residual distribution using Gaussian processes and probit stick-breaking processes (Chung & Dunson, 2009). The fully nonparametrikc conditional approach considers that $f(y|x) = \int_{\Theta} f(y|x,\theta)P_x(d\theta)$ takes the form of a mixture model with unknown mixing distribution P_x for θ. A prior is set on the family of probability distributions $(P_x)_{x \in \mathcal{X}}$. In particular, following the seminal work of MacEachern (1999), various dependent Dirichlet process models have been proposed in the literature (Gelfand & Kottas, 2003; Griffin & Steel, 2006; Dunson *et al.*, 2007; Caron *et al.*, 2007, 2008; Dunson & Park, 2008). Similarly, Trippa *et al.* (2011) define a class of dependent random probability distributions using Pólya trees.

An alternative to the conditional approach is to treat the covariates as random variables and to build a joint statistical model for (X, Y). In this way, one can cast the regression problem as a density estimation one. For example, Müller *et al.* (1996) proposed to use Dirichlet process mixtures for the joint distribution of (X, Y). This approach was later extended by Shahbaba & Neal (2009), Hannah *et al.* (2011) and Wade *et al.* (2014).

A major drawback of current Bayesian methods for semi or nonparametric regression is that many methods do not scale well with the number of samples and/or with the dimensionality of the covariates. In this paper, we propose a novel joint Bayesian nonparametric regression model $F_{X,Y}$ that affords an approximation which can scale easily to large data applications. The model is parameterized in terms of the marginal distributions of the response F_Y and covariates F_X, and then a conditional regression model that utilises the two marginal distributions,

$$F_X \sim \mathbb{P}_X \quad (2)$$

$$F_Y \sim \mathbb{P}_Y \quad (3)$$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$$\beta \sim \pi_\beta \quad (4)$$

$$F_{X,Y}(x,y) = C_{\lambda_\beta}\left(F_X(x), F_Y(y)\right) \quad (5)$$

where $\mathbb{P}_X$ and $\mathbb{P}_Y$ are some nonparametric prior over probability distributions, $\lambda_\beta : \mathcal{X} \to \mathbb{R}_+$ is some parametric regression function of the covariates, and C_{λ_β} plays a role similar to a copula in that it takes marginal distributions as inputs and characterises the dependence between them using the function λ_β. In particular we consider a Plackett-Luce model for ranks for the regression structure. This construction, detailed in Section 3, builds on the original Plackett-Luce model (Luce, 1959; Plackett, 1975) for ranking. The positive function λ_β tunes the stochastic ordering of the responses given the covariates, the ratio $\lambda_\beta(X_i)/(\lambda_\beta(X_i) + \lambda_\beta(X_j))$ representing the conditional probability, $Pr(Y_i < Y_j | X_i, X_j)$, that response Y_i is less than response Y_j given knowledge of $\{X_i, X_j\}$. There is thus a natural interpretation of the parameters: λ_β tunes the relative ordering of the responses at different covariate values, and F_Y sets the marginal distribution of the responses. This strong interpretability is an important feature as it provides a good vehicle for specifying prior beliefs.

For inference we propose to use a marginal composite likelihood approach, which we show allows the model to scale tractably to large data applications and allows for the use of standard, existing, software from Bayesian nonparametric density estimation and Plackett-Luce ranking estimation to be applied. As an illustration, we show an application of our approach to a US Census dataset, with over 1,300,000 data points and more than 100 covariates.

The paper is organized as follows. Section 2 provides background on Dirichlet process mixtures and Pólya trees for density estimation. Section 3 describes the Plackett-Luce copula model. The marginal composite likelihood approach for scalable inference is presented in Section 4. Section 5 presents some results of our approach on simulated data and on the US Census dataset.

2 Bayesian nonparametric density estimation

The appeal of Bayesian nonparametric models is the large support and probabilistic inference provided by such priors. This both safeguards against model misspecification and enables highly flexible estimation of distributions. This has lead to particular popularity of Bayesian nonparametric priors in density estimation.

In the simple case of density estimation for a real valued random variable many nonparametric priors exist - see Hjort *et al.* (2010) for a recent review. A popular class of model is the Dirichlet Process Mixture (Lo (1984)), whereby a Dirichlet process prior is placed on the distribution of the parameters of a parametric family. The result is an “infinite mixture model”. Precisely:

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$$f_Y(y) = \int K(y|\theta)dP(\theta)$$
$$P \sim \mathrm{DP}\left(c, P_0\right)$$

where K is the density of the chosen parametric family, $c > 0$ is a scale parameter and P_0 is a base measure. Since draws from a Dirichlet Process are almost surely atomic measures, there is positive probability of observations sharing a parameter value given the random measure P. The result is an effect of clustering within a sample, with a random, limitless number of clusters. This has proved to be an extremely popular model as it models heterogeneity within a sample well, and provides a highly flexible support. Efficient MCMC schemes (Escobar & West (1995); MacEachern & Müller (1998); Neal (2000)) have lead to the widespread use of the Dirichlet Process Mixture (DPM) in density estimation.

Pólya trees provide another flexible nonparametric prior for density estimation (Ferguson (1974); Lavine (1992, 1994); Mauldin *et al.* (1992)). They are defined as follows: Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k) \in E^k = \{0, 1\}^k$, and define a sequence of embedded partitions of $\mathbb{R}$ to be $\Gamma_k = \{B_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E^k\}$, where the $B_{\boldsymbol{\epsilon}}$ are defined recursively, such that $B_{\boldsymbol{\epsilon}0} \cup B_{\boldsymbol{\epsilon}1} = B_{\boldsymbol{\epsilon}}$. Now let $E^* = \cup_{k\geq 1} E^k$, the set of all countable sequences of zeros and ones, and let $\mathscr{A} = \left\{\alpha_{\boldsymbol{\epsilon}} : \boldsymbol{\epsilon} \in E^*\right\}$ be a set of nonnegative real numbers. Then, a random probability measure P is a Pólya tree process with respect to $\Gamma = \{\Gamma_k : k \geq 1\}$ and $\mathscr{A}$ if $P(B_{\boldsymbol{\epsilon}0} \mid B_{\boldsymbol{\epsilon}}) \sim \mathrm{Beta}(\alpha_{\boldsymbol{\epsilon}0}, \alpha_{\boldsymbol{\epsilon}1})$, independently for all $\boldsymbol{\epsilon} \in E^*$. There are two properties of the Pólya tree process that are appealing for density estimation: Pólya trees are conjugate, meaning that both the prior and the posterior have the same functional form, and, for certain choices of $\mathscr{A}$, realizations are absolutely continuous probability distributions, almost surely. It is worth pointing out that empirically the model can depend heavily on the defined sequence of partitions Γ, although a mixture of Pólya trees proposed by Lavine (1992) can smooth out this dependence over multiple partitions. In what follows we make use of these nonparametric models to specify priors for the marginal distributions of covariates and response variables.

3 The statistical model

Let (X_i, Y_i), $i = 1, \ldots, n$ be the covariates and responses and regression function $\lambda_\beta : \mathscr{X} \to \mathbb{R}_+$. To build the dependence we introduce a latent random variable Z_i that is used to capture the underlying relative level of the response via,

$$Z_i | X_i = x_i \sim \mathrm{Exp}\left(\lambda_\beta\left(x_i\right)\right) \quad (6)$$

where Exp(a) denotes the standard exponential distribution of rate a. The latent variable Z_i may be interpreted as an "arrival time" of individual i. The arrival times then define a conditional ranking of the predicted response variables $Y_1, \ldots, Y_n$.

The model can be summarized as follows, for $i = 1, \ldots, n$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$$X_i \overset{\text{iid}}{\sim} F_X \quad (7)$$

$$Z_i | X_i, \beta \overset{\text{ind}}{\sim} \text{EXP}\left(\lambda_\beta \left(X_i \right) \right) \quad (8)$$

$$Y_i = F_Y^{-1}\left(F_Z\left(Z_i\right)\right) \quad (9)$$

where

$$\begin{aligned} F_Z(z) &= \int_{\mathcal{X}} F_{Z|X=x}(z) dF_X(x) \\ &= \int_{\mathcal{X}} \left(1 - e^{-\lambda_\beta(x) z} \right) dF_X(x) . \end{aligned}$$

Figure (1) shows the correspondence between the conditional exponential random variables, $Z|X$, shown in 1(a) for differing covariate values, and the resulting predictive distributions in 1(b), where the marginal F_Y is a Gaussian mixture model shown as the black line. We can see visually that the distributions in 1(b) are stochastically ordered under the model. The coloured points shown in (a) are mapped to the points shown in (b), where again ordering is preserved.

As F_Y and F_Z are cumulative density functions, $F_Y^{-1} \circ F_Z$ is a monotonically increasing function and

$$\mathbb{P}\left(Y_i \leq Y_j\right) = \mathbb{P}\left(Z \leq Z_j\right) = \frac{\lambda_\beta\left(x_i\right)}{\lambda_\beta\left(x_i\right) + \lambda_\beta\left(x_j\right)} .$$

This clarifies the role of the regression function. More generally, given an ordering $\nu = (\nu_1, \ldots, \nu_n)$ (a permutation of $\{1, 2, \ldots, n\}$), we have

$$\mathbb{P}\left(Y_{\nu_1} \leq Y_{\nu_2}, \ldots, \leq Y_{\nu_n}\right) = \mathbb{P}\left(Z_{\nu_1} \leq Z_{\nu_2}, \ldots, \leq Z_{\nu_n}\right) = \prod_{i=1}^{n} \frac{\lambda_\beta\left(x_{\nu_i}\right)}{\sum_{j \geq i} \lambda_\beta\left(x_{\nu_j}\right)} .$$

The above model is the Plackett-Luce model (Luce, 1959; Plackett, 1975), popular in the ranking literature, and also corresponds to the partial likelihood used for Cox proportional hazards models (Cox, 1972).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

By construction $F_Z(Z_i)$ is marginally uniformly distributed on [0, 1]. Thus, $Y_i = F_Y^{-1}(F_Z(Z_i))$ is marginally distributed from F_Y. The joint distribution $F_{X,Y}$ can thus be described in terms of marginals F_X and F_Y and a Plackett-Luce copula C_{λ_β} such that

$$F_{X,Y}(x,y) = C_{\lambda_\beta}\left(F_X(x), F_Y(y)\right).$$

The Plackett-Luce copula takes the following form

$$C_{\lambda_\beta}(u_1,u_2) = u_1 - \int_{\omega=0}^{u_1} \exp\left(-\lambda_\beta(\omega) F_Z^{-1}(u_2)\right) d\omega. \quad (10)$$

Figure 2 shows illustration of the copula for different functions λ_β.

The conditional distribution function can then be expressed as

$$F_{Y \mid X=x}(y) = 1 - \exp(-\lambda_\beta(x) F_Z^{-1}(F_Y(y))).$$

Given λ_β, the random variables $Y/X = x$ are stochastically ordered. For x_1, x_2 such that $\lambda_\beta(x_1) \leq \lambda_\beta(x_2)$

$$F_{Y \mid X=x_1}(y) \leq F_{Y \mid X=x_2}(y) \quad \forall_y \in \mathbb{R}.$$

If F_Y has a density with respect to Lebesgue measure, f_Y, then we can use a change of variables to calculate the conditional density as follows:

$$\begin{aligned} f_{Y \mid X=x}(y) &= f_Y(y) \frac{f_{Z \mid X=x}(z(y))}{f_Z(z(y))} \\ &= f_Y(y) \frac{f_{Z \mid X=x}(F_Z^{-1}(F_Y(y)))}{f_Z(F_Z^{-1}(F_Y(y)))} \\ &= f_Y(y) \frac{\lambda_\beta(x) \exp[-\lambda_\beta(x) F_Z^{-1}(F_Y(y))]}{\int_{\mathcal{X}} \lambda_\beta(x') \exp[-\lambda_\beta(x') F_Z^{-1}(F_Y(y))] dF_X(x')}. \end{aligned}$$

It can be seen from this representation that the conditional density of Y_i, given X_i is simply the marginal density of Y_i, re-weighted across its quantiles ($F_Y(y)$) by a function of X_i.

We end the construction of the model by assuming a prior over the finite-dimensional parameter β and Bayesian nonparametric prior over the marginal distributions F_X and F_Y

$$\beta \sim \pi_\beta \quad (11)$$

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

$$F_Y \sim \mathbb{P}_Y \quad (12)$$

$$F_X \sim \mathbb{P}_X \quad (13)$$

where π_β is some parametric prior and $\mathbb{P}_X$ and $\mathbb{P}_Y$ may be a Dirichlet process mixture or a Pólya tree prior, as described in Section 2.

4 Approximations for posterior inference and prediction

Assume that both F_X and F_Y admit a density with respect to Lebesgue measure, noted f_X and f_Y. The unknown quantities for our regression model are therefore (f_Y, β, f_X). Given data $(x_{1:n}, y_{1:n})$, where $x_{1:n} = (x_1, \ldots, x_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$, we have the following likelihood:

$$L(f_Y, \beta, f_X; (x_{1:n}, y_{1:n})) = \prod_{i=1}^{n} f_Y(y_i) \frac{\lambda_\beta(x_i)\exp[-\lambda_\beta(x_i)F_Z^{-1}(F_Y(y_i))]}{\int_{\mathcal{X}} \lambda_\beta(x')\exp[-\lambda_\beta(x')F_Z^{-1}(F_Y(y_i))]dF_X(x')} f_X(x_i).$$

(14)

Inference could proceed using numerical methods such as MCMC but for large datasets this is cumber-some. Hence we consider here a Bayesian composite marginal likelihood approach (Lindsay, 1988; Cox & Reid, 2004; Varin *et al.*, 2011; Pauli *et al.*, 2011; Ribatet *et al.*, 2012) that we show offers computational tractability and the use of standard Bayesian methods. Define $y^*_{1:n}$ to be $y_{1:n}$ ordered from lowest to highest, and let $\nu_{1:n} = (\nu_1, \ldots, \nu_n)$ be a vector representing the order of $y_{1:n}$, so that $y^*_i = y_{\nu_i}$. Then we can re-write our data $\{y_{1:n}, x_{1:n}\}$ equivalently as $\{y^*_{1:n}, \nu_{1:n}, x_{1:n}\}$.. Now let L_C denote the composite marginal likelihood based on $\{y^*_{1:n}\}$ and $\{\nu_{1:n}, x_{1:n}\}$. That is the product of the likelihood terms associated with each of these terms:

$$\begin{aligned} L_C(f_Y, \beta, f_X; \{y_{1:n}, x_{1:n}\}) &= L(f_Y, \beta, f_X; \{y^*_{1:n}\}) \times L(f_Y, \beta, f_X; \{\nu_{1:n}, x_{1:n}\}) \\ &= n! \left[\prod_{i=1}^{n} f_Y(y_i)\right] \times \left[\prod_{i=1}^{n} \frac{\lambda_\beta(x_{\nu_i})}{\sum_{j \geq i} \lambda_\beta(x_{\nu_j})}\right] \times \left[\prod_{i=1}^{n} f_X(x_i)\right]. \end{aligned} \quad (15)$$

We can see that this composite likelihood approach factors the likelihood into separate terms involving f_Y, β and f_X, leading to the following pseudo posterior distribution

$$\pi_C(f_Y, \beta, f_X \mid \{y_{1:n}, x_{1:n}\}) = \pi_C(f_Y \mid y^{\star}_{1:n})\pi_C(f_X \mid x_{1:n})\pi_C(\beta \mid \nu_{1:n}, x_{1:n}) \quad (16)$$

Inference over the parameters f_Y, β, f_X can thus be carried out independently under the composite likelihood approach. Standard software for Bayesian nonparametric univariate density estimation can be used for f_Y and f_X, and software for fitting Plackett-Luce/Cox proportional hazard can be used for fitting β. Overall the advantages of the approximate composite likelihood approach include computational tractability and scalable inference using standard software, hence good numerical reproducibility, and high interpretability as the components in the composite likelihood have explicit form and meaning. This latter point aids in prior elicitation as it allows the analyst to separate out and represent their beliefs on the marginal distributions, which are simpler to specify than the full conditionals, and then consider the dependence given the marginals.

The Bayesian composite likelihood approach has attracted some attention over recent years (Pauli *et al.*, 2011; Varin *et al.*, 2011; Ribatet *et al.*, 2012). In particular, Ribatet *et al.* (2012) considered two adjustements to the marginal likelihood approach in order to retain some of the desirable properties of the usual likelihood. However, their adjustments apply to a specific form of composite likelihood, where it factorizes as a product of composite likelihoods for each observation: $L_c^{total}(y|\theta) = \prod_{i=1}^{n} L_c(y_i|\theta)$ where $L_c(y_i|\theta)$ is the composite likelihood for observation i. Our composite likelihood approach does not fit in this framework, as we do not have this product form over the observations, and we cannot therefore apply the adjustments suggested by Ribatet *et al.* (2012). Extending the adjustment of Ribatet *et al.* (2012) to our framework is an interesting direction, but beyond the scope of this article.

4.1 Asymptotics for the marginal composite posteriors

Consider first the pseudo-posterior for f_Y:

$$\begin{aligned}\pi_C(f_Y \mid \{y_{1:n}, x_{1:n}\}) &\propto \pi(f_Y) L_C(f_Y; \{y_{1:n}, x_{1:n}\}) \\ &\propto \pi(f_Y) \prod_{i=1}^{n} f_Y(y_i).\end{aligned}$$

So our pseudo-posterior is exactly the posterior based on the i.i.d sample $\{y_{1:n}\}$, where $y_{1:n} \sim F_Y$. This is the standard setting for posterior inference, so we can apply consistency results from Bayesian nonparametric inference for F_Y, see for example Ghosal & Van der Vaart (2013). The same is true for f_X. Now consider the log-linear form for λ: $\lambda(x) = \exp(-\beta x)$. Then, we have the pseudo-posterior:

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

$$\pi_C(\beta \mid \{y_{1:n}, x_{1:n}\}) \propto \pi(\beta) L_C(\beta; \{y_{1:n}, x_{1:n}\}) \propto \pi(\beta) \prod_{i=1}^{n} \frac{e^{\beta x_{\nu_i}}}{\sum_{j \geq i} e^{\beta x_{\nu_j}}}.$$

This is exactly the posterior considered by Kim (2006) in a different setting where a Bernstein-Von Mises theorem is proven, which can be applied here.

4.2 Posterior predictive

We can use simulation methods such as MCMC to easily generate samples $\{F_Y^{(j)}, \beta^{(j)}\}_{j=1}^{m}$ from the pseudo-posterior (16); the predictive distribution can then be approximated by

$$p(y' \mid x', \{y_{1:n}, x_{1:n}\}) \simeq \frac{1}{m} \sum_{j=1}^{m} p(y' \mid x', \beta^{(j)}, F_Y^{(j)}).$$

To simulate from this distribution, we can use the forward generating process of our model, given $X = x'$:

$$Z' \mid X' = x' \sim \text{Exp}(\lambda_\beta(x')) \quad (17)$$

$$Y' = F_Y^{-1}(F_Z(Z')). \quad (18)$$

In many applications, modeling F_X might be cumbersome, and not the primary object of interest. In this case we propose to use an empirical Bayes approach by setting $F_X = \hat{F}_X$ at the empirical CDF. So, to generate a posterior predictive sample, given a posterior sample $\{F_Y^{(j)}, \beta^{(j)}\}_{j=1}^{m}$, Eq. (18) becomes:

$$Y'^{(j)} = F_Y^{-1(j)}\left(1 - \frac{1}{n} \sum_{i=1}^{n} e^{-Z'^{(j)} \lambda_{\beta^{(j)}}^{(x_i)}}\right)$$

where we note that $Z'^{(j)}$ is conditional on $X' = x'$, and the CDF inversion is tractable, depending on the form of F_Y. Alternately one can use Monte Carlo to draw samples from the predictive, which is trivial when F_Y can be sampled from. Some particular examples are discussed in Appendix A.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

5 Illustrations

In this Section we apply our method to two examples. The first is a simulation example where we generate from a multi-modal conditional and explore the ability of our method to fit the data. The second is a large real-world application in the regression analysis of US Census data.

5.1 Simulation example

In this section we apply the model to a dataset simulated from our model to consider how well we can recover known dependence. The marginal distribution of Y is set to a mixture of three Gaussian distributions, with means 3, 9 and 15, standard deviations of 2, 0.5 and 1 with mixture weights of 0.5, 0.2 and 0.3 respectively. β is set to 0.25, with $\lambda_\beta(x) = \exp(\beta x)$. $X \sim \text{Unif}(0, 20)$ and $n = 500$. The data is shown in Figure 3(a).

Clearly any type of linear or non-linear regression with a parametric noise distribution will be inappropriate here. The conditional distribution of Y given x is multi-modal, rendering many popular regression models inappropriate.

We compared our approach to a linear dependent Dirichlet process mixture of normals (LDDPM) (De Iorio *et al.*, 2004), using the R package DPpackage (Jara *et al.*, 2011; Jara, 2007). This model specifies that

$$Y_i \mid x_i \sim \int \mathcal{N}(y_i; x_i\beta, \sigma^2) G(d\beta, d\sigma^2)$$
$$G \mid \alpha, \mu_b, s_b \sim \text{DP}(\alpha G_0)$$

where $G_0 = \mathcal{N}(\mu_b, s_b)\,\text{Gamma}(\tau_1/2,\ \tau_2/2)$ and

$$s_b \mid \nu, \psi \sim IW(\nu, \psi)$$

with $\alpha = 1$, $\mu_b = (9, 0)^T$, $\nu = 4$, $\tau_1 = 1$, $\tau_2 = 2$, $\psi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $s_b = \begin{pmatrix} 36 & 0 \\ 0 & 36 \end{pmatrix}$.

We apply our model, modeling the marginal as a Dirichlet Process mixture of Gaussian distributions using $\alpha = 1$ and a normal-inverted-Wishart distribution for the base measure. That is, our base measure $G_0(\mu, \sigma^2) = \mathcal{N}(\mu \mid \mu_1, \frac{\sigma^2}{\kappa_1}) IW(\sigma^2 \mid \nu_1, \psi_1)$, where $\mu_1 = 9$, $\kappa_1 = 0.5$, $\nu_1 = 4$ and $\psi_1 = 1$. A Gaussian prior centered at 0 with unit variance is used for β.

In Figure 3(b) the simulated data is shown, with the 80% highest posterior density (HPD) intervals of the predictive distribution at each value of x. Qualitatively we see that the model can capture the nonlinearities in the data and demonstrates the flexibility to model the multi-modal conditional response. In Figure 4 we show the predictive marginal, $\hat{F}_Y$ and the posterior distribution for β. Clearly the marginal distribution for Y is very well recovered from the data. This parameterization of the model in terms of the marginal distribution for the response allows this to be estimated from the complete dataset, without reliance on other

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

aspects of the model. The strength of information available is apparent in the quality of the fit to the sampling distribution. The posterior for the parameter β shows reasonable support around the true value, being slightly pulled towards 0 by the prior.

We can further inspect how these come together in the posterior predictive conditional distribution for Y given x. Consider this distribution for $x = 5$ and $x = 12$, for both our model and the linear DDP mixture model, as shown in Figure 5. Again, our model provides a reasonable fit. The predictive distribution is not as accurate as the marginal distribution for Y, but this is to be expected, since the conditional distribution is a product of the whole model, compounding uncertainties from both β and the marginal distribution for y. Nonetheless, the fit is good and noticeably better than the flexible linear DDP mixture, as you would expect, given that the sampling distribution is within the support of our model. Concretely, the L1-distance between the estimated conditional and the true conditional distribution can be calculated in each case. When $x = 5$ the distance to our prediction is 0.00869, whereas the distance to the linear DDP is 0.0214, and when $x = 12$ the distance to our prediction is 0.0127 and the distance to the linear DDP is 0.0146.

A point of note is that these posterior predictive plots are smoothed kernel density estimates of MCMC samples. Therefore, Gaussian shapes are slightly exaggerated. Whilst not entirely clear from the plot, both our predictive and the sampling distribution comprise of slightly skewed Gaussian distributions, since the conditional distribution is the marginal distribution for Y weighted across the quantiles.

To illustrate that the model is capable of modeling a range of distributions, we consider data sampled from a Gaussian linear model. The covariates are simulated uniformly on [0, 10], with $Y \sim N(3 + 2x, 2)$ and $n = 300$. We use our model, modeling the marginal for Y with a Pólya tree prior whose partition is set on a Gaussian distribution with mean 12.5 and standard deviation 6, and $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$. A Gaussian prior centered at 0 with variance 8 is used for for β. The posterior predictive 80% HPD intervals display a reasonable fit of the linear data, shown in Figure 6. The variance seems slightly inflated, but this is a consequence of the large support of the model.

5.2 US Census application

We apply the methodology to a regression task using US census data[1] for personal annual income.

We use the American Community Survey data from 2013, which comprises of responses to questions on the survey given to a 1% sample of the US population. Since we are interested in income, the subset of 1, 371, 401 employed civilians over the age of 16 is used. We have used a relevant, linearly independent subset of the data as covariates, excluding highly informative questions such as occupation, which would almost completely explain the response. This leaves 15 explanatory variables, 10 of which are categorical variables, some of which have many levels. The result is a 1, 371, 401 × 114 design matrix.

[1] http://www.census.gov/acs/www/data_documentation/pums_data/

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

The covariates are: US state (Texas as a baseline), weight, age, class of worker (employee of private for-profit company as a baseline), travel time to work, means of transportation to work (works from home as a baseline), language other than English spoken at home (no as a baseline), marital status (married as a baseline), educational attainment (regular high school diploma as a baseline), gender (male as a baseline), hours worked a week, weeks worked last year, disability status (without a disability as a baseline), quarter of birth (first quarter as a baseline), and world area of birth (United States of America as a baseline).

The levels of annual income shown in Figure (7) can be seen to be heavy tailed, which requires a flexible model to capture. Another noticeable feature of the data is that the income levels are discontinuous, with large spikes in frequency at particular income levels. This could in part be due to standardized salary structures resulting in certain salary levels becoming common. This motivates the use of a nonparametric approach as it is difficult to imagine how a parametric density could conditionally capture the features shown in Figure (7). However, standard Bayesian nonparametric models simply cannot be applied to a problem of this scale. Attempting to apply existing methods in this literature, such as the linear dependent Dirichlet process mixture, failed to run due to the dimensionality and scale of the data.

For the analysis, we consider both the empirical distribution function and a Pólya tree prior for the marginal distribution of y. The partition of the Pólya tree is set on the quantiles of a Gaussian distribution with mean 35,000 and standard deviation 20,000, and $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$. We use a log-linear regression function $\lambda(x) = \exp(\beta x)$ and place independent Gaussian priors with mean 0 and unit variance on the coefficients in β.

5.2.1 Predictive performance—We compare the out-of-sample predictive performance of our model with three competing non-Bayesian approaches namely, a standard linear regression model, a median regression model and a LASSO[2]. For our model we investigated three distinct priors for the marginal distribution of the response: a Pólya tree centred on a Gaussian, a Pólya tree centred on a Laplace, and an empirical Bayes approach using the empirical CDF. To compare methods we use repeated random subsets of 1000 test samples and train each model on the remaining data, with 10 repeats. Predictive accuracy is judged by mean squared-error (MSE), mean absolute error (MAE) and qualitatively via a qq-plot. To create the qq-plots we compute the predictive distribution function $F(y|x)$ evaluated at the observed value for each of these test samples. Under the assumption that we have a perfect predictive distribution, these values should be independent uniform random variables. A deviation from this distribution implies a mis-match of the posterior predictive and the actual distribution. We are unable to apply this approach to the median regression model, as it does not provide a predictive distribution and would require fitting the model for a large number of quantiles. In the case of the linear model we used maximum likelihood estimates for prediction, rather than a fully Bayesian approach. With such a large dataset the strength of any reasonable default prior would be significantly diminished, so this should mimic a Bayesian approach well.

[2]These models were fitted in R using the functions lm, rq (from the quantreg package) and lars (from the lars package). For LASSO the regularization parameter was chosen using cv.lars. Default settings were used for each.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Summary statistics of predictive fit are shown in Table 1. Perhaps unsurprisingly on such a large data set the linear model targeting the conditional mean does best on MSE but this is at the expense of the median under MAE. In addition, studying the predictive qq-plot in Figure(8b) shows the inadequacy of the linear model to provide calibrated predictions. The LASSO performs relatively poorly suggesting most covariates are influential for prediction, whereas the median regression whilst, as expected, providing relative accuracy on the MAS does so at the expense of MSE and as mentioned above suffers from the lack of a fully predictive model. The Bayesian nonparametric methods perform relatively well on against both summary measures, with perhaps that based on the Laplace marginal showing greatest accuracy. In Figure(8a) we show the predictive qq-plot from this model, demonstrating that the full predictive distribution is captured well. There is some evidence of misfit in the upper tail of the Pólya tree model due perhaps due to the lack of heavy tails in the Gaussian base measure. These diagnostics suggest that even non-linear regression models with parametric noise would not provide a satisfactory fit for the data, since the unusual conditional distribution of the response cannot be captured by such models. This highlights the benefit of our nonparametric approach.

We next consider inference for covariate effects. In order to gain a measure of the relevance of each covariate we quantified the concentration of the posterior probability measure away from the prior "null" centring of $\beta_j = 0$. To do this we estimated the Bayesian sign-probability from the posterior marginal for each covariate as,

$$PrSign_j = \max\left[\int_{\beta<0} \pi(\beta_j \mid \cdot)d\beta, \int_{\beta>0} \pi(\beta_j \mid \cdot)d\beta\right] \quad (19)$$

where $\pi(\beta_j|\cdot)$ is the posterior marginal for β_j. This measures the relative tail area in the posterior marginal laying to the left or right of 0. A large value of *PrSign* suggests there is strong evidence against $\beta_j = 0$. In certain respects this is akin to a Bayesian marginal version of a p-value, and is trivially calculated from MCMC output, or from normal approximations to the posterior distribution. Table 2 shows the most relevant covariates as ranked by this measure.

Unsurprisingly, hours worked a week and weeks worked last year show high certainly of a positive effect on income. After these, educational achievement measured via degrees unsurprisingly imply higher earnings compared to the regular high school diploma. Since these are part of the same variable it is simple to compare the effects due to these degrees. Despite Bachelor's degree providing the most certainty of a positive effect, a further Professional degree beyond bachelor's has the highest posterior mean effect. The ranking in Table 2 reflects the greater evidence in the data for a non-zero Bachelor effect, due to a much higher number of observations of those with Bachelor's degrees, and hence lower variance in the effect size compared with those with a higher degree. There is also strong evidence for Female workers earning less than their male counterparts, as well as increasing income with age and even travel time to work.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Finally, we show it is simple to provide the full posterior predictive distribution of annual income of somebody in the test sample, using the Pólya tree model. We choose as a hypothetical person a 57 year old female from North Carolina, who is self employed, married, 140 lbs bodyweight, who works from home, speaks English at home, went to college but for less than a year, who usually works 30 hours a week, for 43.5 weeks last year, was born in the first quarter of the year in the USA. The structure and shape of the posterior predictive, represented in Figure 9, match that of the marginal distribution for Y in the data, just on a narrower range.

6 Discussion

We introduced a new Bayesian semiparametric regression model that is designed to scale to large data applications. In doing so we make use of an interpretable model for ranks, via a Plackett-Luce copula method, and nonparametric density models for the marginals. We used a composite marginal likelihood approximation that leads to a number of advantages. It affords computationally tractability, aids in the interpretation of the model, and makes prior specification explicit on known objects.

The key to the scalability of the method is the use of the composite likelihood approximation, which splits the inference into two simpler tasks. The use of the Laplace approximation for the covariate effect and the Pólya tree for the marginal response allow for fast posterior inference, without requiring any MCMC sampling methods. In fact, sampling methods are only used for prediction, which is by far the slowest part of the inference procedure.

Going forward, it would be interesting to see if theoretical bounds on the approximation error as a function of sample size could be derived. It may also be possible to apply results such as those found in Kim (2006) to provide further guarantees of asymptotic behavior such as properties of the predictive distribution. In addition, it would be interesting to explore non-linear models for the regression function $\lambda(x)$, such as those based on a random forests methodology. In fact, random forests applied to the US Census dataset (with restricted node size to enable application to this scale) gives a highly competitive MSE to our tested models. This might be because random forests is able to capture interaction terms between covariates, which seem highly plausible *a priori* in this particular dataset. It will be interesting to incorporate such flexibility into a Bayesian nonparametric approach using Plackett-Luce regression functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Caron, F., Davy, M., Doucet, A. Generalized Polya urn for time-varying Dirichlet process mixtures. 23rd Conference on Uncertainty in Artificial Intelligence (UAI'2007); 2007.

Caron F, Davy M, Doucet A, Duflos E, Vanheeghe P. Bayesian inference for linear dynamic models with Dirichlet process mixtures. IEEE Transactions on Signal Processing. 2008; 56(1):71–84.

Europe PMC Funders Author Manuscripts

Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010; 4(1):266–298.

Chung Y, Dunson DB. Nonparametric Bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association. 2009; 104(488)

Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972; 34(2):187–220.

Cox DR, Reid N. A note on pseudolikelihood constructed from marginal densities. Biometrika. 2004; 91(3):729–737.

De Iorio, Maria, Müller, Peter, Rosner, Gary L., MacEachern, Steven N. An ANOVA model for dependent random measures. Journal of the American Statistical Association. 2004; 99(465):205–215.

Denison, DGT., Holmes, CC., Mallick, BK., Smith, AFM. Bayesian methods for nonlinear classification and regression. John Wiley & Sons; 2002.

Dunson DB, Park J-H. Kernel stick-breaking processes. Biometrika. 2008; 95(2):307–323. [PubMed: 18800173]

Dunson DB, Pillai N, Park J-H. Bayesian density regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(2):163–183.

Escobar MD, West M. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association. 1995; 90(430):577–588.

Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics. 1973; 1(2):209–230.

Ferguson TS. Prior Distributions on Spaces of Probability Measures. The Annals of Statistics. 1974; 2(4):615–629.

Gelfand A, Kottas A. Bayesian semiparametric regression for median residual life. Scandinavian Journal of Statistics. 2003; 30(4):651–665.

Ghosal, S., Van der Vaart, AW. Fundamentals of nonparametric Bayesian inference. Cambridge University Press; New York: 2013.

Griffin JE, Steel MFJ. Order-Based Dependent Dirichlet Processes. Journal of the American Statistical Association. 2006; 101(473):179–194.

Hannah LA, Blei D, Powell WB. Dirichlet process mixtures of generalized linear models. The Journal of Machine Learning Research. 2011; 12:1923–1953.

Hjort, NL., Holmes, CC., Müller, P., Walker, SG. Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press; 2010.

Jara A. Applied Bayesian Non- and Semi-parametric Inference Using DPpackage. R News. 2007; 7(3): 17–26.

Jara A, Hanson T, Quintana F, Müller P, Rosner G. DPpackage: Bayesian Semi- and Nonparametric Modeling in R. Journal of Statistical Software. 2011; 40(5):1–30.

Kim Y. The Bernstein–von Mises theorem for the proportional hazard model. The Annals of Statistics. 2006; 34(4):1678–1700.

Kottas A, Gelfand AE. Bayesian semiparametric median regression modeling. Journal of the American Statistical Association. 2001; 96(456):1458–1468.

Lavine M. Some Aspects of Polya Tree Distributions for Statistical Modelling. The Annals of Statistics. 1992; 20(3):1222–1235.

Lavine M. More Aspects of Polya Tree Distributions for Statistical Modelling. The Annals of Statistics. 1994; 22(3):1161–1176.

Lavine M, Mockus A. A nonparametric Bayes method for isotonic regression. Journal of Statistical Planning and Inference. 1995; 46(2):235–248.

Lindsay BG. Composite likelihood methods. Contemporary Mathematics. 1988; 80(1):221–39.

Lo AY. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. The Annals of Statistics. 1984; 12(1):351–357.

Luce, RD. Individual Choice Behavior: A Theoretical Analysis. John Wiley and sons; 1959.

MacEachern, SN. Dependent Nonparametric Processes. Proceedings of the Bayesian Statistical Science Section. American Statistical Association; 1999.

MacEachern SN, Müller P. Estimating Mixture of Dirichlet Process Models. Journal of Computational and Graphical Statistics. 1998; 7(2):223–238.

Mauldin, R. Daniel, Sudderth, William D., Williams, SC. Polya Trees and Random Distributions. The Annals of Statistics. 1992; 20(3):1203–1221.

Müller P, Quintana FA. Nonparametric Bayesian data analysis. Statistical Science. 2004; 19(1):95–110.

Müller P, Erkanli A, West M. Bayesian curve fitting using multivariate normal mixtures. Biometrika. 1996; 83(1):67–79.

Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics. 2000; 9(2):249–265.

Pati D, Dunson DB. Bayesian nonparametric regression with varying residual density. Annals of the Institute of Statistical Mathematics. 2014; 66(1):1–31. [PubMed: 24465053]

Pauli F, Racugno W, Ventura L. Bayesian composite marginal likelihoods. Statistica Sinica. 2011; 21(1):149.

Plackett RL. The Analysis of Permutations. Journal of the Royal Statistical Society Series C (Applied Statistics). 1975; 24(2):193–202.

Rasmussen, CE. Gaussian processes for machine learning. MIT Press; 2006.

Ribatet M, Cooley D, Davison AC. Bayesian inference from composite likelihoods, with an application to spatial extremes. Statistica Sinica. 2012; 22:813–845.

Shahbaba B, Neal R. Nonlinear models using Dirichlet process mixtures. Journal of Machine Learning Research. 2009; 10:1829–1850.

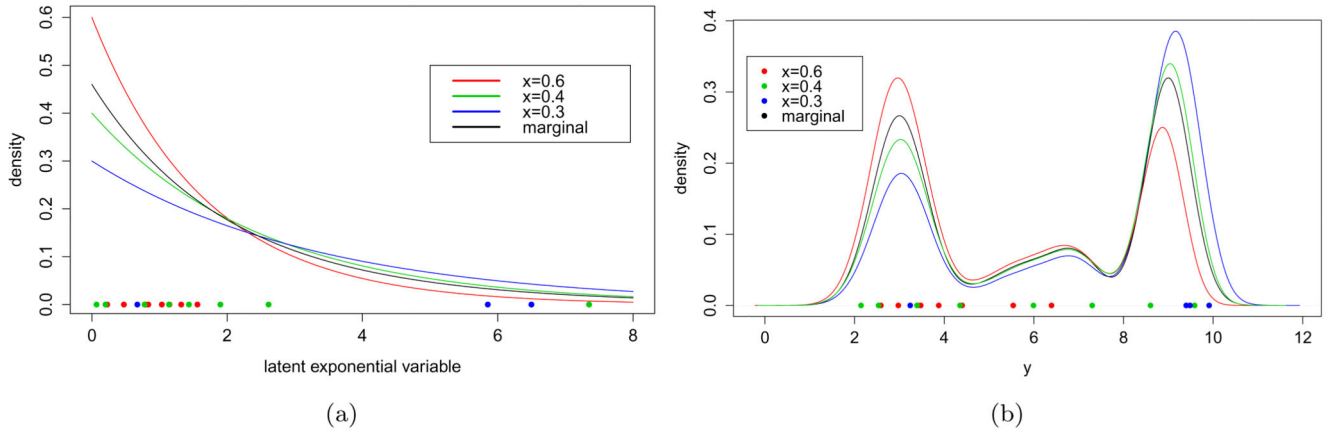
Trippa L, Müller P, Johnson W. The multivariate beta process and an extension of the Polya tree model. Biometrika. 2011; 98(1):17–34. [PubMed: 23956460]

Varin C, Reid N, Firth D. An overview of composite likelihood methods. Statistica Sinica. 2011; 21(1): 5–42.

Wade S, Dunson DB, Petrone S, Trippa L. Improving Prediction from Dirichlet Process Mixtures via Enrichment. Journal of Machine Learning Research. 2014; 15:1041–1071.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



(a) (b)

Fig 1.

Illustration of the latent variable used to capture the regression dependence. In (a) we show the distribution of the conditional latent variable, Z, at various points in X assuming a log-linear dependence. In (b) we see the corresponding predictive distributions using a Gaussian mixture model for the marginal, F_Y, shown as the black line. The points in Z shown in (a) are mapped to the points in Y shown in (b) where the ordering is preserved.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

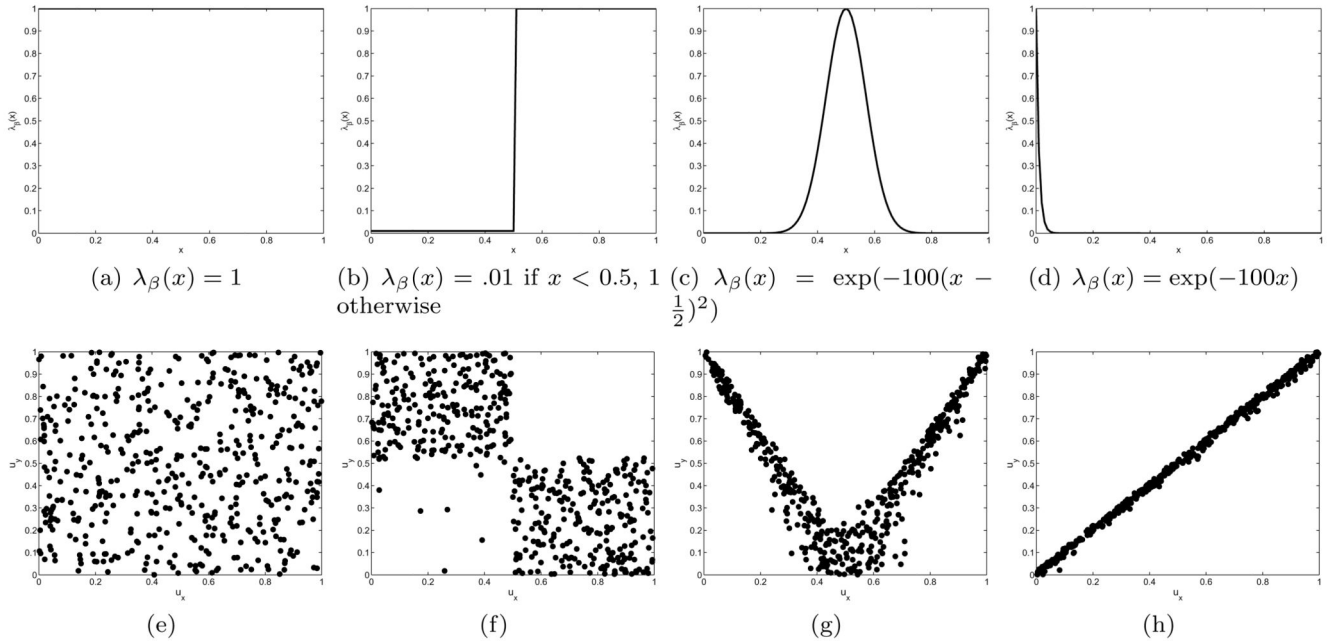


Fig 2.

Examples of the Plackett-Luce copula for different functions λ_β. The top figures (a-d) plot the different functions λ_β. The bottom figures (e-h) represent samples from the copula $C_{\lambda_\beta}(u_x, x_y)$, where $X \in [0, 1]$ and F_X is uniform.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

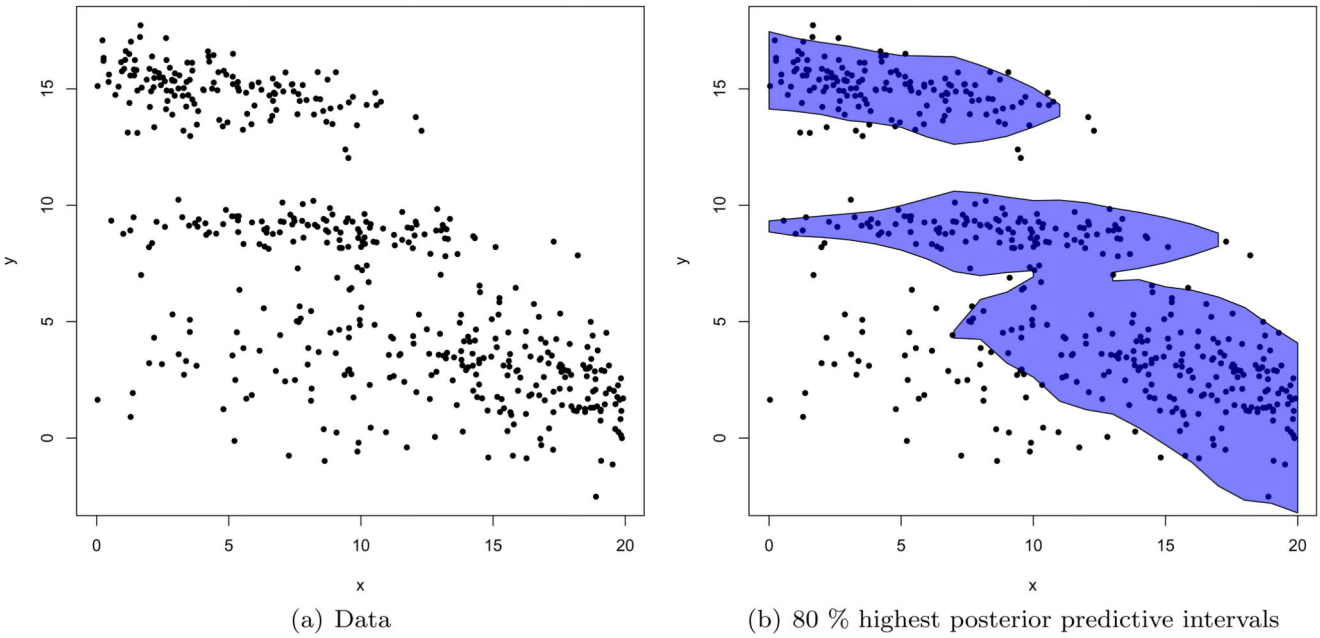


Fig 3.

(a) Data simulated from the model with mixture of three Gaussians marginal distribution for Y. (b) 80% highest posterior density intervals of the predictive distribution at each value of x.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

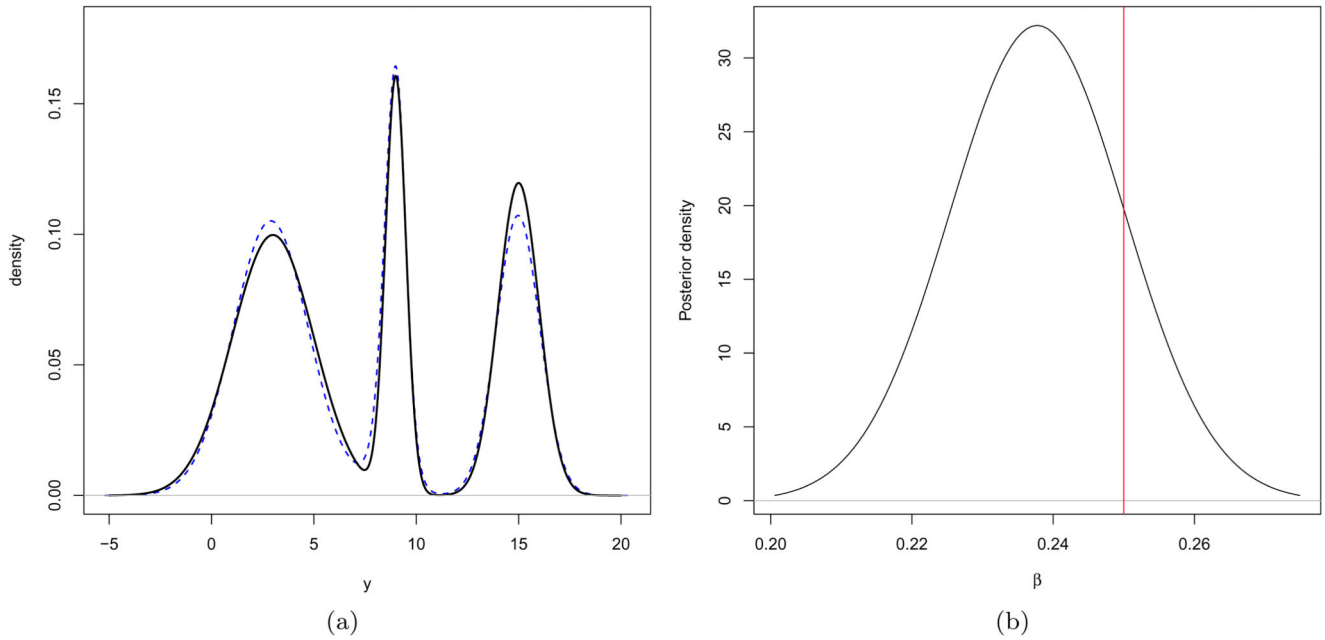
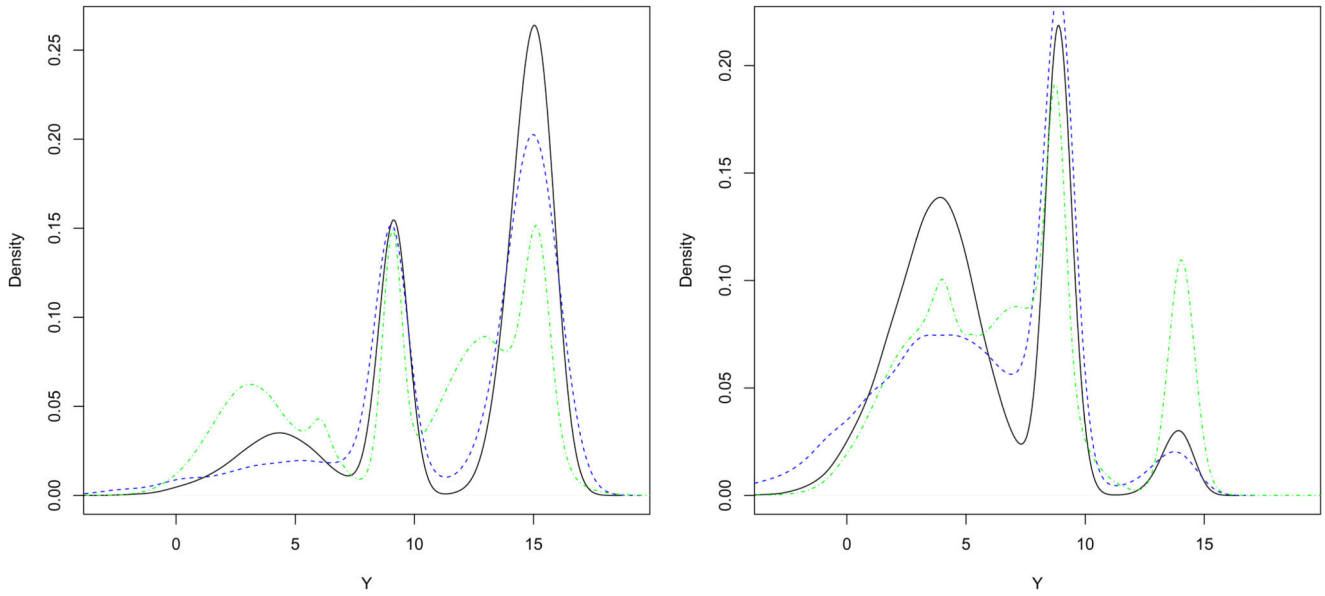


Fig 4.

(a) The posterior predictive marginal for y under our model in blue, compared to the actual sampling distribution in black. (b) The posterior distribution for β, compared to the true value of 0.25 marked in red.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



(a) $x = 5$ (b) $x = 12$

Fig 5.

Predictive densities for (a) $x = 5$ and (b) $x = 12$. The true predictive is shown in black, the predictive distributions under our model in blue, and the predictive under the linear DDP mixture in green.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

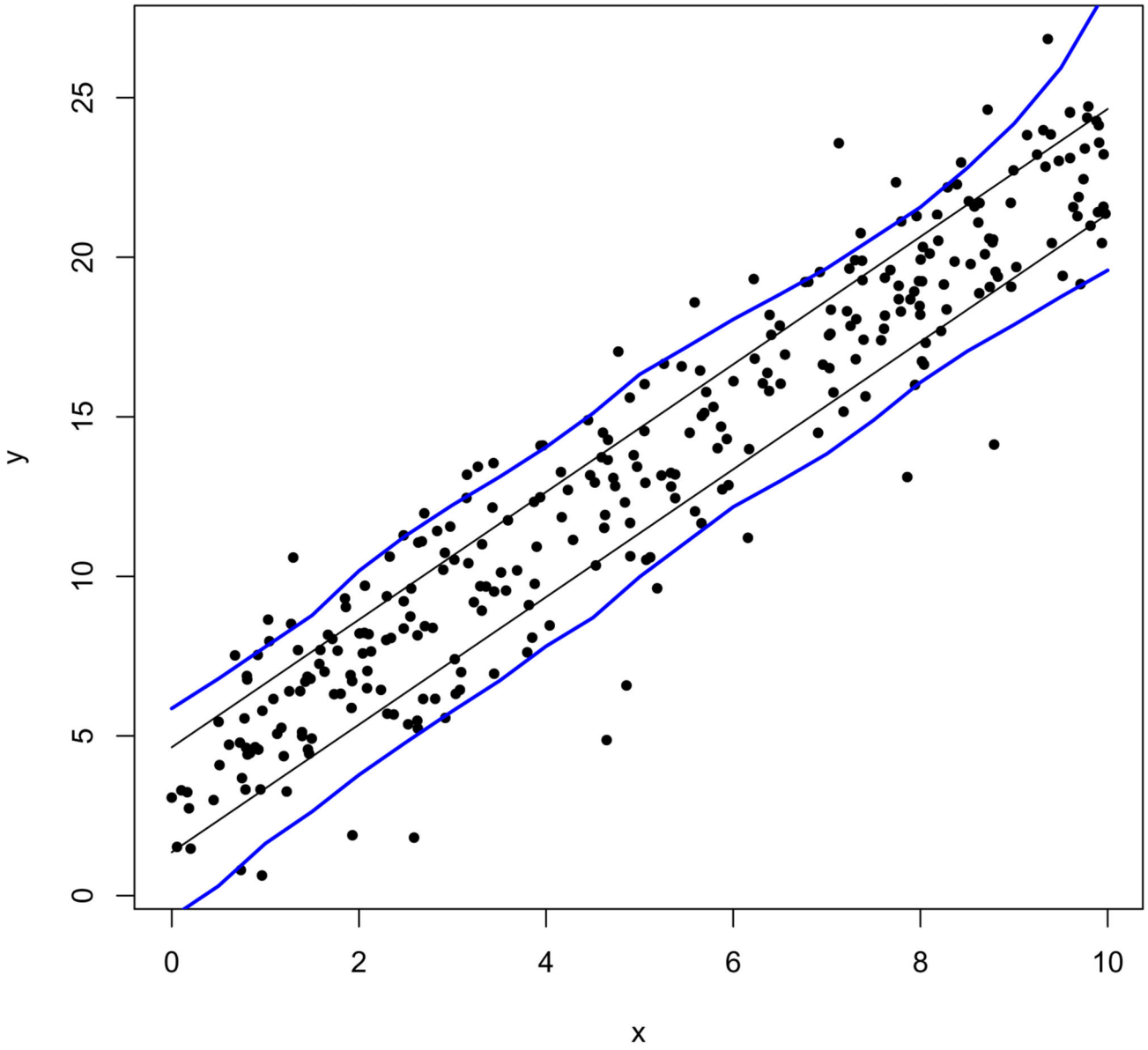


Fig 6.

Simulated data from a linear model with Gaussian residuals (black dots). 80% HDP intervals of the predictive distribution of our model at each value of x are represented in blue, and true HDP interval in black

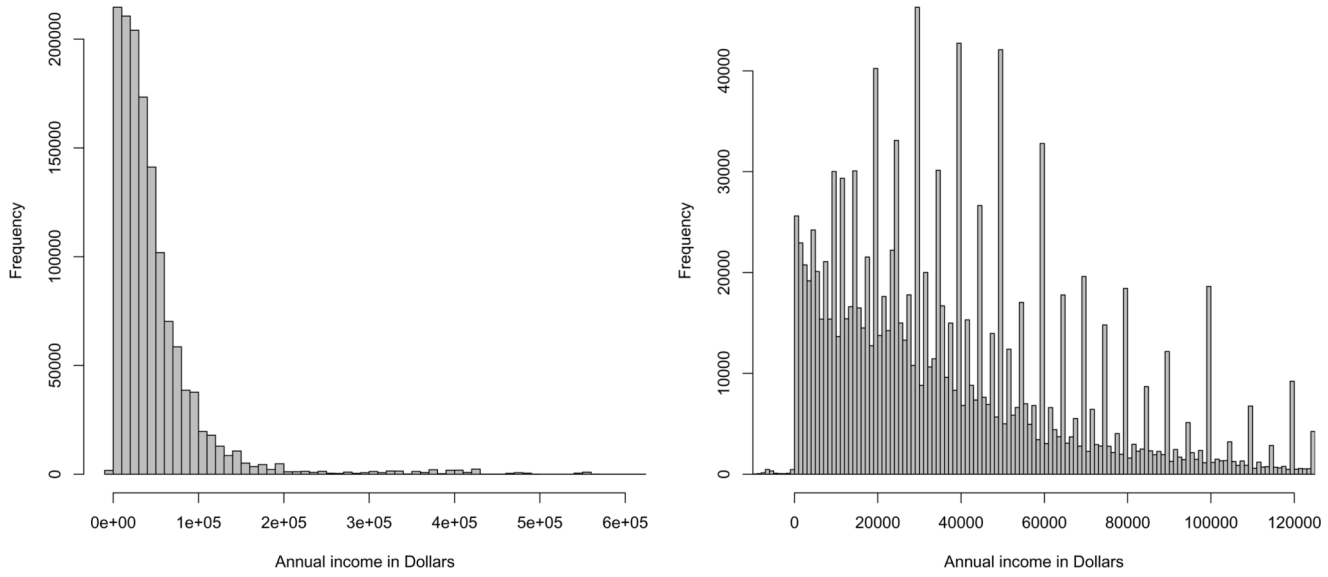
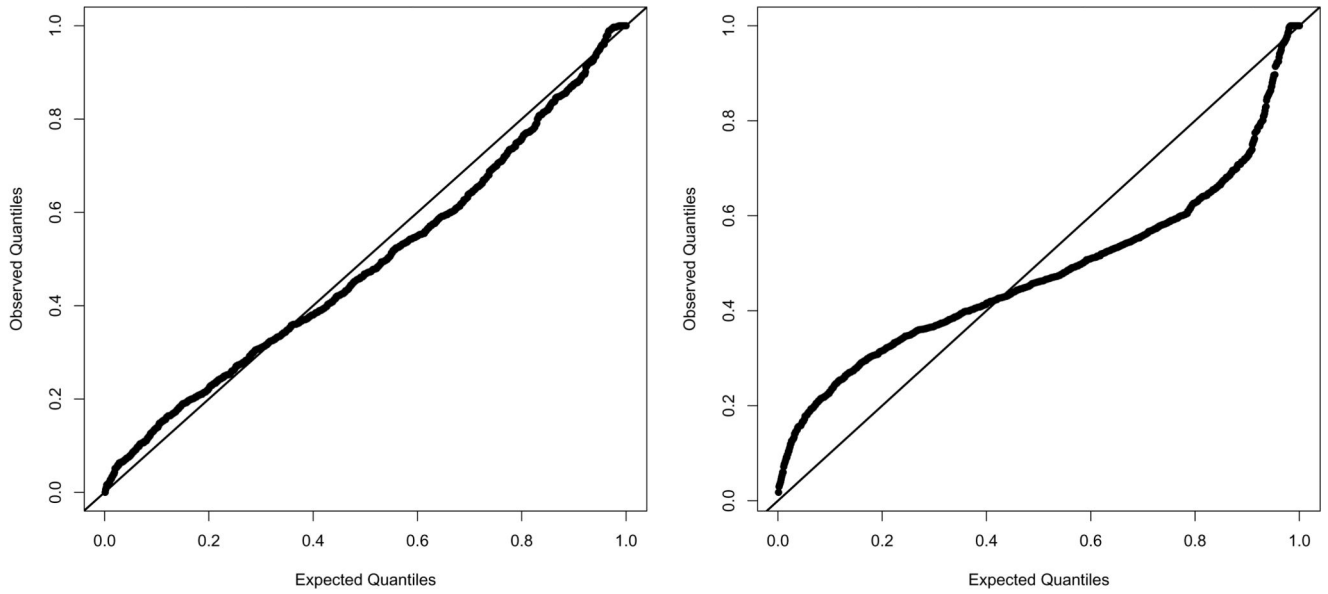


Fig 7.

Histograms of annual income on different scales. Right hand plot is zoomed in on incomes up to 120,000.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



(a) Proposed model with Pólya tree prior (b) Linear model

Fig 8.

qq-plots (a) under our model using a Pólya tree prior centred on Laplace for the marginal distribution of the response and (b) using a standard linear model.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

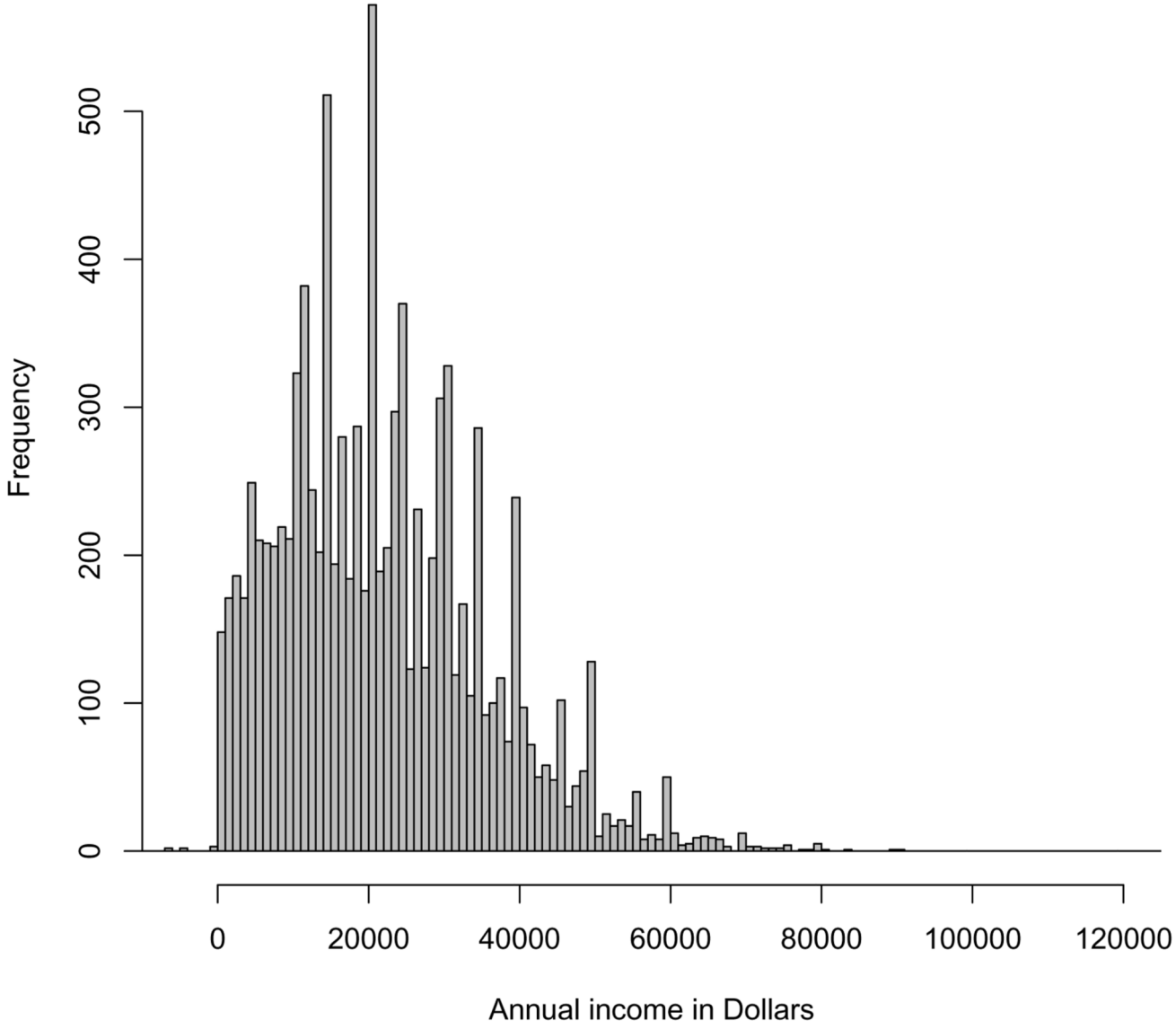


Fig 9.

Posterior predictive distribution using a Pólya tree for the marginal distribution, zoomed in on incomes up to 120,000

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Mean out-of-sample prediction errors with standard deviation of this error after 10 repetitions

	Mean square error (10^9)	Mean absolute error(10^4)
Empirical model*	2.79 ± 0.51	2.44 ± 0.15
Pólya tree (Gaussian)*	2.81 ± 0.64	2.41 ± 0.16
Pólya tree (Laplace)*	2.71 ± 0.52	2.44 ± 0.14
Linear model	2.66 ± 0.59	2.67 ± 0.16
LASSO	2.99 ± 0.65	2.81 ± 0.16
Median regression	2.99 ± 0.68	2.48 ± 0.17

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 2

Top covariate parameters ranked by (19): the log posterior probability of the parameter being a different sign to the posterior mean. A negative parameter value has a positive effect on income.

	Log probability of different sign	Posterior mean
Hours worked a week	-1.1×10^5	-0.044
Weeks worked last year	-1.0×10^5	-0.045
Bachelor's degree	-4.2×10^4	-0.80
Master's degree	-4.16×10^4	-1.0
Professional degree beyond a bachelor's degree	-2.7×10^4	-1.4
Age	-2.6×10^4	-0.018
Female	-1.8×10^4	0.35
Doctorate degree	-1.5×10^4	-1.2
Never Married	-1.3×10^4	0.37
Associate's degree	-6.5×10^3	-0.39
Travel time to work	-3.2×10^3	-0.0035
1 or more years of college credit, no degree	-2.0×10^3	-0.18
Self employed (incorporated)	-2.0×10^3	-0.30
Grade 11 in school	-1.8×10^3	0.38
Walks to work	-1.5×10^3	0.36
Disabled	-1.3×10^3	0.19