Density Regression with Bayesian Additive Regression Trees

Vittorio Orlandi, Jared Murray, Antonio Linero, Alexander Volfovsky

Density Regression

Density Regression

How the entire probability distribution of a response Y changes as a function of covariates $\mathbf{X}$.

Density Regression

How the entire probability distribution of a response Y changes as a function of covariates $\mathbf{X}$.

Generalizes density estimation, mean and quantile regression.

Density Regression

How the entire probability distribution of a response Y changes as a function of covariates X.

Generalizes density estimation, mean and quantile regression.

Response density can be...

Density Regression

How the entire probability distribution of a response Y changes as a function of covariates $\mathbf{X}$.

Generalizes density estimation, mean and quantile regression.

Response density can be...

Primary object of interest

- Studies of income inequality (Gerfin 1994; Daly et al. 2006; Angrist et al. 2006)

Density Regression

How the entire probability distribution of a response Y changes as a function of covariates $\mathbf{X}$.

Generalizes density estimation, mean and quantile regression.

Response density can be...

Primary object of interest

- Studies of income inequality (Gerfin 1994; Daly et al. 2006; Angrist et al. 2006)

Used in downstream analysis

- Using redshift distribution to estimate cosmological parameters (Wittman 2009)
- Pricing of financial derivatives (Fan et al. 2004)

Bayesian Density Regression

Jointly model $p(y, \mathrm{x} \mid \theta)$ to learn about conditional $p(y \mid \mathrm{x}, \theta)$

- West et al. 1993; Muller et al. 1996; Park et al. 2010; Shahbaba et al. 2009; Taddy et al. 2010; Molitor et al. 2010; Wade, Mongelluzzo, et al. 2011; Dunson and Bhattacharya 2011; Hannah et al. 2011; Wade, Dunson, et al. 2014
- Often influenced by joint distribution of covariates

Bayesian Density Regression

Focus explicitly on $p(y \mid \mathbf{x}, \theta)$

- *Dependent Dirichlet Processes*
  MacEachern 1999; MacEachern 2000; De Iorio et al. 2004; Griffin et al. 2006; Dunson and Peddada 2008; De Iorio et al. 2009; Wang et al. 2011
- *Covariate Dependent Mixtures*
  Jacobs et al. 1991; Jordan et al. 1994; Geweke et al. 2007; Villani et al. 2009
- *Miscellaneous*
  Tokdar et al. 2010; Trippa et al. 2011; Jara et al. 2011; Ma 2012; Shen and Ghosal 2014

Our Contribution

- Density regression model with good theoretical guarantees & finite sample performance
- Flexible model with easy to set priors
- Code & computational efficiency

The Model

A First Model

Begin without covariates

A First Model

Begin without covariates

Continuous latent variable model (Pati et al. 2011)

A First Model

Begin without covariates

Continuous latent variable model (Pati et al. 2011)

$$Y = f(U) + \epsilon, \quad U \sim U(0,1), \quad \epsilon \sim N(0, \sigma^2)$$

A First Model

Begin without covariates

Continuous latent variable model (Pati et al. 2011)

$$Y = f(U) + \epsilon, \quad U \sim U(0,1), \quad \epsilon \sim N(0, \sigma^2)$$

$$p(y) = \int_0^1 \phi_\sigma \left(y - f(u)\right) d\, u$$

A First Model

Begin without covariates

Continuous latent variable model (Pati et al. 2011)

$$Y = f(U) + \epsilon, \quad U \sim U(0,1), \quad \epsilon \sim N(0,\sigma^2)$$

$$p(y) = \int_0^1 \phi_\sigma\left(y - f(u)\right) d\, u$$

How general is this?

A First Model

What if f is a step function?
Consider $0 = \nu_0 < \nu_1 < \cdots < 1$ such that $\sum_{h=0}^{\infty}(\nu_{h+1} - \nu_h) = 1$.

Then if $f(u) = \mu_h$ for $u \in [\nu_h, \nu_{h+1})$:

$$
\begin{aligned}
p(y) &= \int_0^1 \phi_\sigma\left(y - \mu_h\right) \mathbb{1}(u \in [\nu_h, \nu_{h+1}))\, du \\
&= \sum_{h=1}^{\infty} (\nu_{h+1} - \nu_h) \phi_\sigma\left(y - \mu_h\right)
\end{aligned}
$$

which is a discrete location mixture model.

Can place priors on mixture-specific components...

A First Model

What if f is a step function?
Consider $0 = \nu_0 < \nu_1 < \cdots < 1$ such that $\sum_{h=0}^{\infty} (\nu_{h+1} - \nu_h) = 1$.

Then if $f(u) = \mu_h$ for $u \in [\nu_h, \nu_{h+1})$:

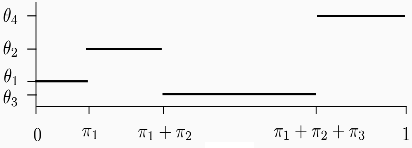
$$\begin{aligned} p(y) &= \int_0^1 \phi_\sigma \left(y - \mu_h\right) \mathbb{1}(u \in [\nu_h, \nu_{h+1}))\, du \\ &= \sum_{h=1}^{\infty} (\nu_{h+1} - \nu_h) \phi_\sigma \left(y - \mu_h\right) \end{aligned}$$

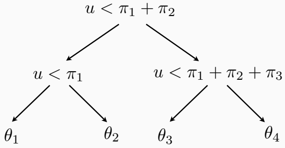
which is a discrete location mixture model.

Can place priors on mixture-specific components... or directly on f

Why Trees?

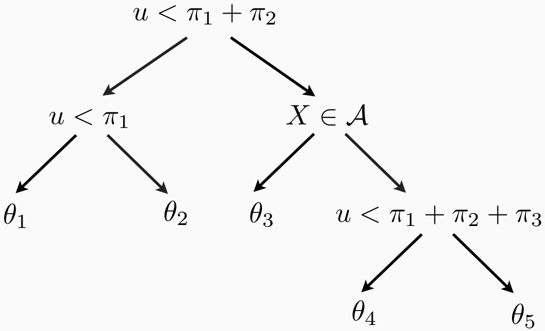
Any step function has a binary decision tree representation





Why Trees?

Easy to introduce **x**:



Why Trees?

Yields the implied conditional densities:

$$p(y \mid \mathbf{x} \in \mathcal{A}) = \pi_1\phi_\sigma(y-\theta_1) + \pi_2\phi_\sigma(y-\theta_2) + (\pi_3+\pi_4)\phi_\sigma(y-\theta_3)$$

$$p(y \mid \mathbf{x} \notin \mathcal{A}) = \pi_1\phi_\sigma(y-\theta_1) + \pi_2\phi_\sigma(y-\theta_2) + \pi_3\phi_\sigma(y-\theta_4) + \pi_4\phi_\sigma(y-\theta_5)$$

Borrowing of information across covariate space

Bayesian Additive Regression Trees (BART)

Chipman, George, McCulloch, 2010

Given response Y, covariates $\mathbf{x} = (x_1, \dots, x_p)$, model:

$$Y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma_0^2)$$

Model the mean function f as a sum of decision trees $\{g_j\}_{j=1}^m$:

$$f(x) = \sum_{j=1}^{m} g_j(x)$$

Constrain each tree to be a 'weak learner'

Density Regression BART (DR-BART)

Specify the following latent-variable density regression model:

$$Y = f(\mathbf{x}, U) + \epsilon$$
$$U \sim U(0,1), \quad \epsilon \sim N(0, \sigma^2)$$

where $f \sim \text{BART}$.

Density Regression BART (DR-BART)

Specify the following latent-variable density regression model:

$$
\begin{aligned}
Y &= f(\mathbf{x}, U) + \exp[v(\mathbf{x}, U)/2]\epsilon \\
U &\sim U(0,1), \quad \epsilon \sim N(0, \sigma^2)
\end{aligned}
$$

where $f, v \sim \text{BART}$.

Theory

Overview

Demonstrate posterior concentration with respect to the integrated Hellinger distance:

$$h(p, q) = \left(\int \left(\sqrt{p(y \mid \mathrm{x})} - \sqrt{q(y \mid \mathrm{x})} \right)^2 \, dy \, F_{\mathrm{X}}(d\mathrm{x}) \right)^{1/2}$$

Overview

Demonstrate posterior concentration with respect to the integrated Hellinger distance:

$$h(p, q) = \left(\int \left(\sqrt{p(y \mid \mathrm{x})} - \sqrt{q(y \mid \mathrm{x})} \right)^2 \, dy \, F_{\mathrm{X}}(d\mathrm{x}) \right)^{1/2}$$

Proof based off variants of Ghosal et al. 2007a sufficient conditions (Ghosal et al. 2007a; Ghosal et al. 2007b; Shen, Tokdar, et al. 2013; Li et al. 2022+):

1. Prior thickness
2. Entropy bound
3. Support condition

Overview

Demonstrate posterior concentration with respect to the integrated Hellinger distance:

$$h(p,q) = \left(\int \left(\sqrt{p(y \mid \mathrm{x})} - \sqrt{q(y \mid \mathrm{x})} \right)^2 dy \, F_{\mathrm{x}}(d\mathrm{x}) \right)^{1/2}$$

Extend:

1. Jeong et al. 2020 *regression* concentration results for BART
2. Pati et al. 2011 density *estimation* concentration results for a continuous latent variable model

Posterior Concentration

Assumptions:

- Extensions of Jeong et al. 2020 BART prior assumptions
- $\log p_0(y \mid \mathbf{x})$ is α-Hölder for some $\alpha \in (0, 2]$
- $\log p_0$ depends on $(y, \mathbf{x})$ only through a set of d_0 coordinates.
- Conditions on growth of $\|f_0\|_\infty, d_0, p$.

Theorem

There exists a constant $M > 0$ such that:

$$\Pi\left(h(p_0, p_{f,\sigma}) \geq M\epsilon_n \mid \{\mathbf{x}_i, y_i\}\right) \to 0$$

where

$$\epsilon_n = (n/\log n)^{-\frac{\alpha}{\alpha+1} \times \frac{\beta}{2\beta+d_0}} + \sqrt{d_0 \log(p+1)/n}$$

Simulations

Comparison Methods

- SBART-DS (Li et al. 2022+): Model conditional densities by taking a base model $h(y|\mathbf{x}, \theta)$ and modulating it via a link function $\Phi(\mu)$:

$$p(y \mid \mathbf{x}, \theta) \propto h(y \mid \mathbf{x}, \theta)\Phi\{r(y, \mathbf{x})\}$$

  Here, h is a normal linear regression and r is a (Soft) BART.

Comparison Methods

- SBART-DS (Li et al. 2022+): Model conditional densities by taking a base model $h(y|\mathbf{x}, \theta)$ and modulating it via a link function $\Phi(\mu)$:

$$p(y \mid \mathbf{x}, \theta) \propto h(y \mid \mathbf{x}, \theta)\Phi\{r(y, \mathbf{x})\}$$

  Here, h is a normal linear regression and r is a (Soft) BART.
- DPMM (Alejandro Jara et al. 2011): Fit a normal Dirichlet Process Mixture Model for $p(\mathbf{x}, y)$ and look at the implied conditional $p(y \mid \mathbf{x})$.

Simulation Setup

True model:

$$Y_i = f_0(X_i) + \epsilon_i(X_i)$$

where $f_0(X_i)$ is deterministic and $\epsilon_i(X_i)$ is a mixture of a normal and a log-gamma distribution.

Simulation Setup

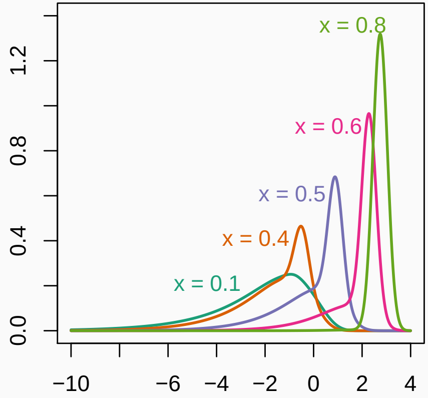
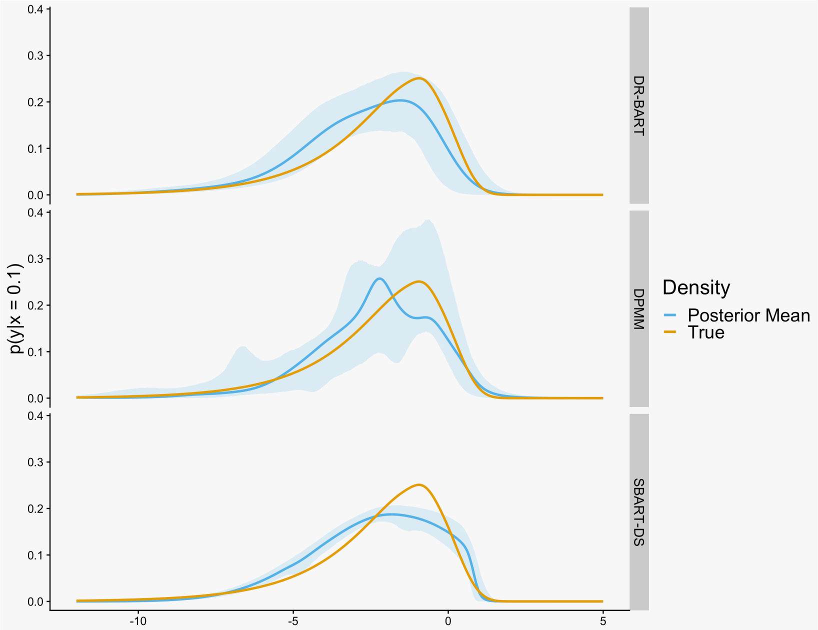


Figure 1: The conditional pdf for selected x values.

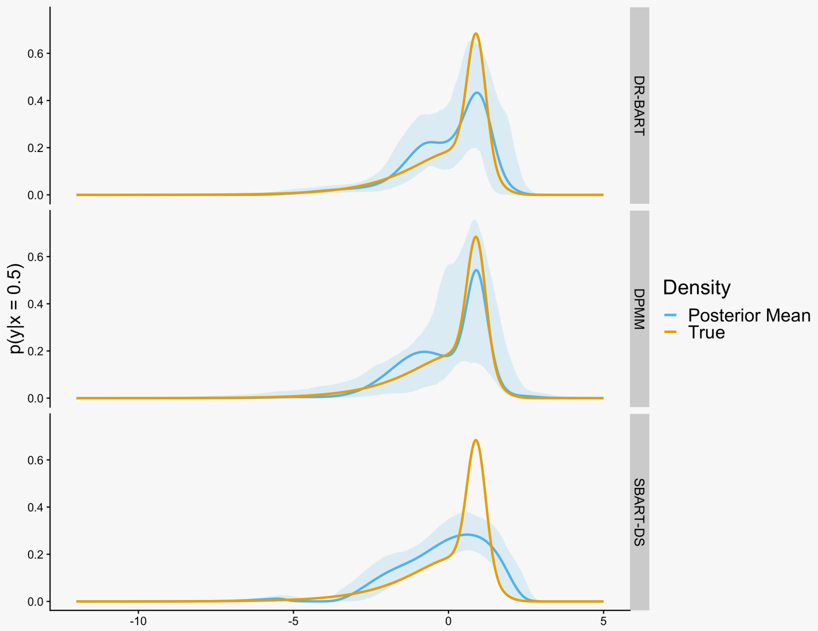
Simulation 1

Basic simulation: $X_1 \overset{ind}{\sim} U(0, 1)$.

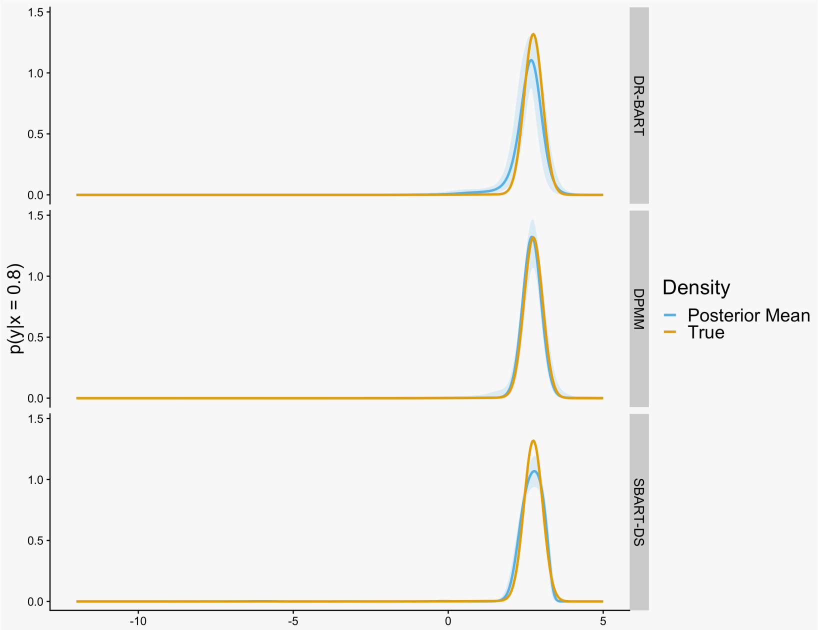
Simulation 1: $x = 0.1$



Simulation 1: $x = 0.5$



Simulation 1: $x = 0.8$



Simulation 2

Further comparison of BART-based models.

Simulation 2

Further comparison of BART-based models.

Mean function of true density: $f_0(x) = a(x - 0.5)^2$.

Simulation 2

Further comparison of BART-based models.

Mean function of true density: $f_0(x) = a(x - 0.5)^2$.

Recall: SBART-DS centered around linear base model.

Simulation 2

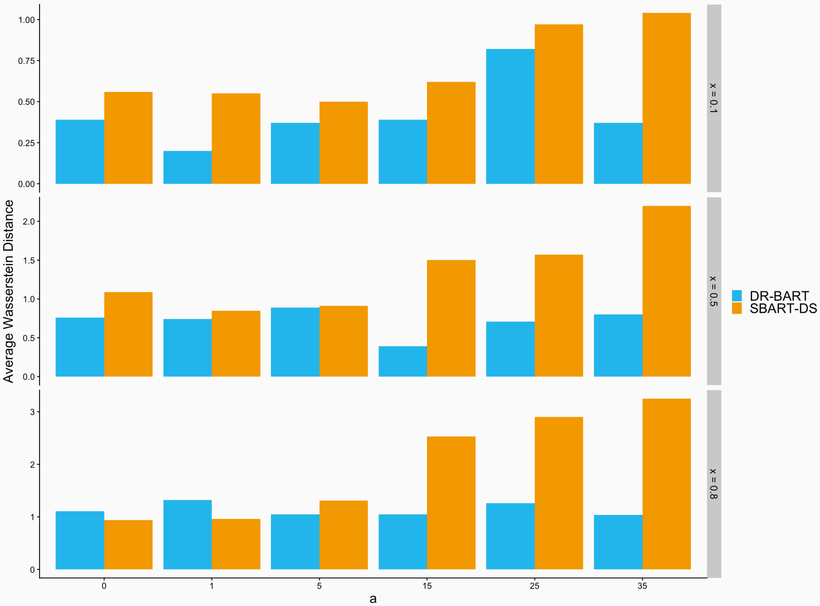
Further comparison of BART-based models.

Mean function of true density: $f_0(x) = a(x - 0.5)^2$.

Recall: SBART-DS centered around linear base model.

As a increases, this base model is increasingly misspecified.

Simulation 2



Summary

Adapted BART to perform density regression via a continuous latent variable model

DR-BART:

- Concentrates quickly about the true density
- Accurately point estimates densities and expresses appropriate uncertainty about these estimates
- Is more computationally efficient than competitors
- Has good default priors and doesn't rely on a base model

Moving Forward: Sensitivity Analysis

Dorie et al. 2016:

$$\begin{aligned} Y|U, \mu_{xz}, \sigma^2 &\sim N(\mu_{xz} + \zeta U, \sigma^2) \\ \mu_{xz}, \sigma^2 &\sim \text{BART}(X, Z) \end{aligned}$$

Moving Forward: Sensitivity Analysis

Dorie et al. 2016:

$$\begin{aligned} Y|U, \mu_{xz}, \sigma^2 &\sim N(\mu_{xz} + \zeta U, \sigma^2) \\ \mu_{xz}, \sigma^2 &\sim \text{BART}(X, Z) \end{aligned}$$

Proposal:

$$\begin{aligned} Y|U, \mu_{xzu}, \sigma^2 &\sim N(\mu_{xzu}, \sigma^2) \\ \mu_{xzu}, \sigma^2 &\sim \text{BART}(X, Z, U) \\ \text{s.t. } \text{BART}(X, Z, U) &\leq \zeta \end{aligned}$$

Moving Forward: Hierarchical Models

Linero et al. 2019:

$$
\begin{aligned}
(Y_i | \mathsf{X} = \mathsf{x}, \mathsf{h}, \boldsymbol{\omega}) &\sim f\{y | \mathsf{h}(\mathsf{x}), \boldsymbol{\omega}\} \\
h_m(\mathsf{x}) &= \sum_{t=1}^{T} g(\mathsf{x}; \mathcal{T}_t, \mathcal{M}_t^{(m)})
\end{aligned}
$$

Moving Forward: Hierarchical Models

Linero et al. 2019:

$$(Y_i | \mathsf{X} = \mathsf{x}, \mathsf{h}, \boldsymbol{\omega}) \sim f\{y | \mathsf{h}(\mathsf{x}), \boldsymbol{\omega}\}$$

$$h_m(\mathsf{x}) = \sum_{t=1}^{T} g(\mathsf{x}; \mathcal{T}_t, \mathcal{M}_t^{(m)})$$

Random Intercept BART:

$$Y_{ij} \sim N(\alpha_j + \mu_x, \sigma^2)$$

$$\mu_x \sim \text{BART}, \quad \alpha_j \sim F$$

Moving Forward: Hierarchical Models

Linero et al. 2019:

$$(Y_i | \mathsf{X} = \mathsf{x}, \mathsf{h}, \boldsymbol{\omega}) \sim f\{y | \mathsf{h}(\mathsf{x}), \boldsymbol{\omega}\}$$
$$h_m(\mathsf{x}) = \sum_{t=1}^{T} g(\mathsf{x}; \mathcal{T}_t, \mathcal{M}_t^{(m)})$$

Random Intercept BART:

$$Y_{ij} \sim N(\alpha_j + \mu_x, \sigma^2)$$
$$\mu_x \sim \text{BART}, \quad \alpha_j \sim F$$

Proposal:

$$Y_{ij} \sim N(\mu_{x\alpha}, \sigma^2)$$
$$\mu_{x\alpha} \sim \text{BART}, \quad \alpha_j \sim F$$

Moving Forward: Time Series

Stochastic volatility model

$$
\begin{aligned}
y_t &\sim N(0, \sigma_t^2) \\
\sigma_t &= \exp(\mu + u_t) \\
u_t &\leftarrow AR(1 \mid \theta)
\end{aligned}
$$

Moving Forward: Time Series

Proposal

$$y_t \sim N(0, \sigma_t^2)$$
$$\sigma_t = \exp(g(\mu, u_t))$$
$$u_t \leftarrow AR(1 \mid \theta), \quad g \sim \text{BART}$$

Moving Forward

Also network data, community detection, etc.

Thank you!

References i

- Angrist, Joshua et al. (2006). "Quantile regression under misspecification, with an application to the US wage structure". In: *Econometrica* 74.2, pp. 539–563.
- Daly, Mary C. et al. (2006). "Inequality and Poverty in United States: The Effects of Rising Dispersion of Men's Earnings and Changing Family Behaviour". In: *Economica* 73 (289), pp. 75–98.
- De Iorio, Maria et al. (2004). "An ANOVA model for dependent random measures". In: *Journal of the American Statistical Association* 99.465, pp. 205–215.
- De Iorio, Maria et al. (2009). "Bayesian nonparametric nonproportional hazards survival modeling". In: *Biometrics* 65.3, pp. 762–771.

Dorie, Vincent et al. (2016). “A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding”. In: *Statistics in Medicine*.

Dunson, D and Abhishek Bhattacharya (2011). “Nonparametric Bayes Regression and Classification Through Mixtures of Product Kernels”. In: *Proceedings of 9th Valencia International Conference on Bayesian Statistics*. Ed. by Jose M. Bernardo et al. Oxford University Press.

Dunson, D and Shyamal Peddada (2008). “Bayesian nonparametric inference on stochastic ordering”. In: *Biometrika* 95.4, pp. 859–874.

Fan, Jianqing et al. (2004). “A Crossvalidation Method for Estimating Conditional Densities”. In: *Biometrika* 91.4, pp. 819–834. ISSN: 00063444. URL: `http://www.jstor.org/stable/20441146` (visited on 06/27/2022).

References iii

- Gerfin, Michael (1994). “Income Distribution, Income Inequality and Life Cycle Effects - A Nonparametric Analysis for Switzerland”. In: *Swiss Journal of Economics and Statistics* 130, pp. 509–522.
- Geweke, John et al. (2007). “Smoothly mixing regressions”. In: *Journal of Econometrics* 138.1, pp. 252–290.
- Ghosal, S et al. (2007a). “Convergence Rates of Posterior Distributions for Noniid Observations”. In: *Annals of Statistics* 35, pp. 192–223.
- – (Apr. 2007b). “Posterior convergence rates of Dirichlet mixtures at smooth densities”. In: *The Annals of Statistics* 35.2. DOI: *10.1214/009053606000001271*. URL: *https://doi.org/10.1214/009053606000001271*.
- Griffin, Jim E et al. (2006). “Order-based dependent Dirichlet processes”. In: *Journal of the American Statistical Association* 101.473, pp. 179–194.

- Hannah, Lauren A. et al. (July 2011). “Dirichlet Process Mixtures of Generalized Linear Models”. In: *The Journal of Machine Learning Research*, pp. 1923–1953. ISSN: 1532-4435.
- Jacobs, Robert A et al. (1991). “Adaptive mixtures of local experts”. In: *Neural Computation* 3.1, pp. 79–87.
- Jara, A et al. (Sept. 2011). “A class of mixtures of dependent tail-free processes.”. In: *Biometrika* 98.3, pp. 553–566. ISSN: 0006-3444. DOI: 10.1093/biomet/asq082.
- Jara, Alejandro et al. (2011). “DPpackage: Bayesian Semi- and Nonparametric Modeling in R”. In: *Journal of Statistical Software* 40.5.
- Jeong, Seonghyun et al. (2020). *The art of BART: On flexibility of Bayesian forests*. eprint: arXiv:2008.06620.
- Jordan, Michael I et al. (1994). “Hierarchical mixtures of experts and the EM algorithm”. In: *Neural computation* 6.2, pp. 181–214.

References v

- Li, Yinpu et al. (2022+). “Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles”. In: *Journal of the American Statistical Association*.
- Linero, Antonio et al. (2019). “Semiparametric Mixed-Scale Models Using Shared Bayesian Forests”. In: *Biometrics*.
- Ma, L (2012). “Recursive partitioning and Bayesian inference on conditional distributions”. URL: `http://d7.stat.duke.edu/sites/default/files/papers/2012-03.pdf`.
- MacEachern, Steven N (1999). “Dependent nonparametric processes”. In: *ASA proceedings of the section on Bayesian statistical science*, pp. 50–55.
- – (2000). “Dependent Dirichlet Processes”. In: *Unpublished manuscript, Department of Statistics, The Ohio State University*.

References vi

- Molitor, John et al. (July 2010). “Bayesian profile regression with an application to the National survey of children’s health”. In: *Biostatistics* 11.3, pp. 484–98. ISSN: 1468-4357. DOI: *10.1093/biostatistics/kxq013*.
- Muller, P. et al. (Mar. 1996). “Bayesian curve fitting using multivariate normal mixtures”. In: *Biometrika* 83.1, pp. 67–79. ISSN: 0006-3444. DOI: *10.1093/biomet/83.1.67*.
- Park, Ju-Hyun et al. (2010). “Bayesian generalized product partition model”. In: *Statistica Sinica* 20.20, pp. 1203–1226.
- Pati, D et al. (2011). *Posterior Convergence Rates in Non-linear Latent Variable Models*. eprint: *arXiv:1109.5000*.
- Shahbaba, Babak et al. (Dec. 2009). “Nonlinear Models Using Dirichlet Process Mixtures”. In: *The Journal of Machine Learning Research* 10, pp. 1829–1850. ISSN: 1532-4435.

- Shen, W and S Ghosal (2014). "Adaptive Bayesian density regression for high-dimensional data". In: *arXiv preprint arXiv:1403.2695*.
- Shen, W, S Tokdar, et al. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures". In: *Biometrika* 100.3.
- Taddy, Matthew A. et al. (July 2010). "A Bayesian Nonparametric Approach to Inference for Quantile Regression". en. In: *Journal of Business & Economic Statistics* 28.3, pp. 357–369. ISSN: 0735-0015. DOI: 10.1198/jbes.2009.07331.
- Tokdar, S et al. (2010). "Bayesian density regression with logistic Gaussian process and subspace projection". In: *Bayesian Analysis* 5.2, pp. 319–344.
- Trippa, Lorenzo et al. (2011). "The multivariate beta process and an extension of the Polya tree model". In: *Biometrika* 98.1, pp. 17–34.

- Villani, Mattias et al. (2009). "Regression density estimation using smooth adaptive Gaussian mixtures". In: *Journal of Econometrics* 153.2, pp. 155–173.
- Wade, Sara, D Dunson, et al. (2014). "Improving Prediction from Dirichlet Process Mixtures via Enrichment". In: *Journal of Machine Learning Research* 15, pp. 1041–1071.
- Wade, Sara, Silvia Mongelluzzo, et al. (2011). "An Enriched Conjugate Prior for Bayesian Nonparametric Inference". EN. In: *Bayesian Analysis* 6.3, pp. 359–385.
- Wang, Lianming et al. (2011). "Bayesian isotonic density regression". In: *Biometrika* 98.3, pp. 537–551.
- West, Mike et al. (1993). *Hierarchical priors and mixture models, with applications in regression and density estimation*. Duke University.

References ix

Wittman, D. (July 2009). “What Lies Beneath: Using $p(z)$ To Reduce Systematic Photometric Redshift Errors”. In: *The Astrophysical Journal* 700.2, pp. L174–L177.

demo.bib