Is your prior better than mine?

A small-sample comparison in density regression through mixtures.

Michele Peruzzi
michele.peruzzi@phd.unibocconi.it

Sonia Petrone
sonia.petrone@unibocconi.it

Sara Wade
s.wade@warwick.ac.uk

INTRODUCTION

Comparing prior distributions and models is a debated issue, and possibly even more so in nonparametrics, where all models are rich and flexible. Some priors may be better suited for some problems in terms of their properties, but this becomes less evident when priors and models have complex features. One could look at frequentist properties such as consistency and rates of convergence, but even then, it may still be hard to draw a line.

We address the problem in the context of nonparametric regression via mixture models. The literature is rich, but fragmented, and we contribute by providing a unifying framework and a faily flexible computational strategy that can be easily adapted to different models.

BAYESIAN DENSITY REGRESSION

Mixture models are used to model complex relationships between response and covariates, so that both the mean and the error distribution of *y|x* may evolve flexibly with the covariates, as their attractive balance between smoothness and flexibility make them suitable to model local features. If we are modeling the response jointly with the covariates, then the mixture model has the form

$$f_P(y,x) = \int K(y,x;\theta)dP_x(\theta)$$

In a Bayesian setting, a prior distribution is placed on the mixing measure, with the Dirichlet Process (DP) being the most common choice. Particularly useful in what follows is the stick-breaking representation of the DP:

$$\mathrm{P} = \sum_{j=1}^{\infty} w_j \delta_{\tilde{\theta}_j},$$

where

$$w_1 = v_1,$$
$$w_j = v_j \prod_{j'<j}(1-v_{j'}) \quad \text{for } j>1 \qquad v_j \overset{iid}{\sim} \mathrm{Beta}(1,\alpha)$$

and, independent of the weights, $\tilde{\theta}_j \overset{iid}{\sim} P_0$.

In this case, and whenever the mixing measure is a.s. discrete, the mixture can be written as

with weights and atoms determined by a random probability measure such that

$$f_P(y,x) = \sum_{j=1}^{\infty} w_j K(y,x;\tilde{\theta}_j),$$

$$\mathrm{P} = \sum_{j=1}^{\infty} w_j \delta_{\tilde{\theta}_j} \qquad \sum_{j=1}^{\infty} w_j = 1 \ a.s. \qquad w_j > 0$$

and the DP is the most common choice in this case. The joint model is useful in stochastic regression, that is when both *y* and *X* are random variables. Inference on the conditional density is possible as a byproduct from the joint density.

CONDITIONAL MODEL

If the focus is on the conditional density, then a more direct approach is desirable. We then model *y|x* directly:

$$f_{P_x}(y|x) = \int K(y;x,\theta)dP_x(\theta)$$

The task is then to define the covariate-dependent mixing distribution. So we generalize the model above:

$$\mathrm{P}_x = \sum_{j=1}^{\infty} w_j(x)\delta_{\tilde{\theta}_j(x)}$$

DEPENDENT DIRICHLET PROCESS

MacEachern's (1999, 2000) Dependent Dirichlet Process takes either the weights or the atoms in a DP to be stochastic processes on *X* to introduce covariate-dependence. Using the stick-breaking representation:

$$w_j(x) = v_j(x)\prod_{j'<j}(1-v_{j'}(x)) \quad \text{for } j>1,$$
$$w_1(x) = v_1(x), \quad v_j(x) \sim \mathrm{Beta}\,(1,\alpha(x))$$

where v_j are independent across *j* and independent of each of the atoms, also stochastic processes on *X*.

This general model can be restricted to two main cases: covariate-dependent weights, or covariate-dependent atoms. An example of the latter is the single-p DDP.

KERNEL STICK-BREAKING

In the first case, Dunson and Park (2008) define the stick-breaking weights via a bounded kernel K:

$$v_j(x) = v_j K(x;\tilde{\psi}_j)$$

This way of constructing the weights gives the model its name. Examples of kernels include

$$v_j(x) = v_j \exp(-\tilde{\tau}_j||x-\tilde{\mu}_j||^2)$$

which we label **KSB-E** (i.e. exponential), and

$$v_j(x) = v_j \prod_{h=1}^{p} \mathbf{1}(|x_h - \tilde{\mu}_{j,h}| < \tilde{\tau}_j^{-1})$$

which we label **KSB-I** (i.e. indicator).

NORMALIZED WEIGHTS (JOINT-LIKE)

An alternative approach is to obtain covariate-dependent weights by suitably rewriting the conditional density from the joint model. This approach is outlined in Antoniano-Villalobos et al. (2014). We label this approach **JLW** (as in joint-like weights).

In this case, we have

$$w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'})}$$

where

$$0 \le w_j \le 1 \qquad \sum_{j=1}^{\infty} w_j = 1$$

and the kernel should be appropriate for the covariates (e.g. a Normal kernel for continuous covariates).

INTERPRETATION

The three models considered here are only a small selection of what is available in the literature, but this should constitute a starting point to try to understand what implications the modeling choices have in actual, real-world, finite sample data analysis. In terms of interpretation, the **JLW** model splits the covariates into regions of applicability. For every mixture component *j*, there is some central location μ_j where regression model *j* applies best, and a scale parameter describing the rate at which the applicability of the model decays around μ_j. A different kernel may model the varying applicability differently. The **KSB-E** looks similar, but since the exponential kernels are added to the stick-breaking construction, it is actually not as straightforward. The same applies for the **KSB-I**. In both cases, however, component mixtures will be best applicable to observations that are closer to their centers.

COMPUTATIONAL STRATEGY

Since our goal is to compare different models, we develop a single algorithm that is general enough to be applicable to these, and other models as well. We adopt the adaptive truncation algorithm of Griffin (2016) for BNP mixtures. The starting point is a finite mixture; more components are added via SMC steps, stopping at a level that provides a good approximation of the infinite mixture. The advantages are twofold: on one hand, we avoid running MCMC on a large number of components, and on the other, we have an appealing stopping rule. Our implementation of the algorithm involves an estimation of the finite mixtures via an adaptive RW Metropolis (Algorithm 6 in Griffin and Stephens, 2013).

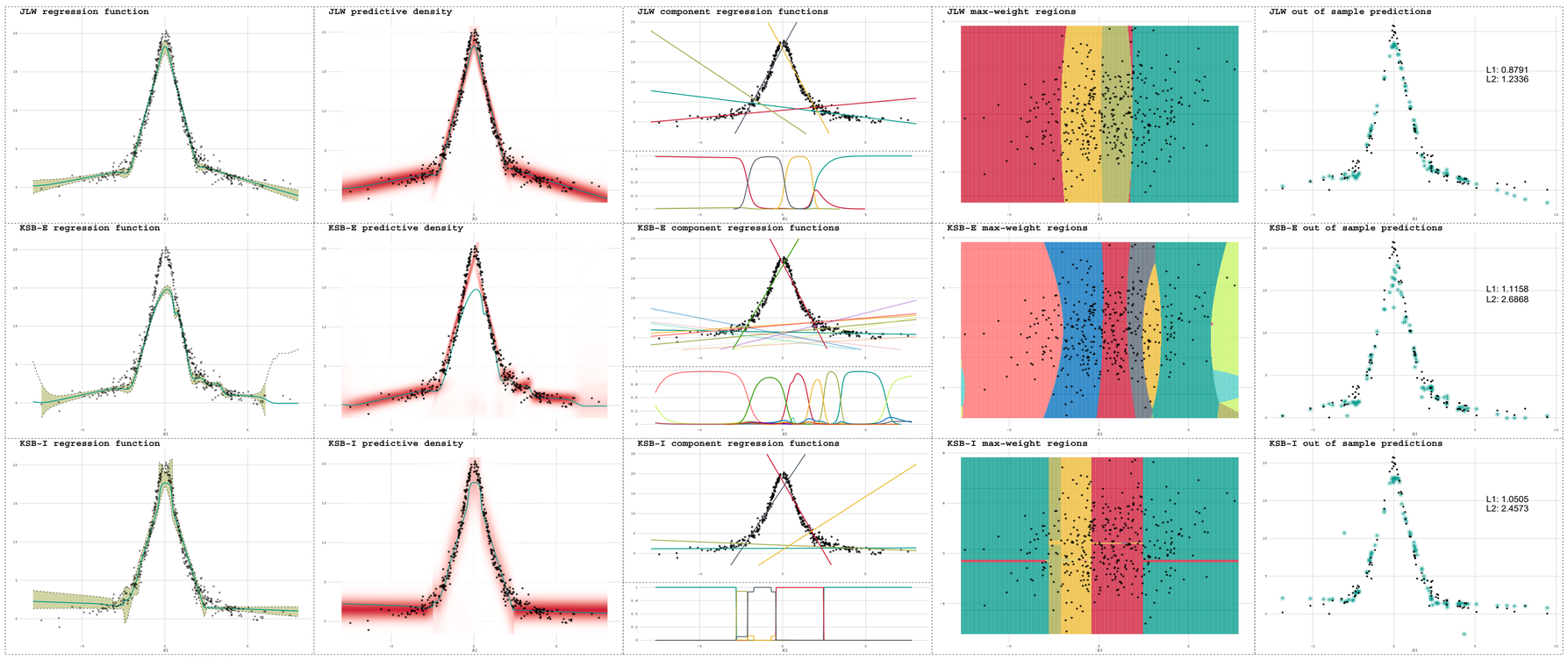
APPLICATIONS AND RESULTS

We test and compare the three models on two simulated datasets where the regression mean and the error distributions are non-standard. The SMC for the two KSB models is still undergoing testing, so we fit a finite-mixture with K=30 components on all models to level the playing field. The SMC steps for the **JLW** algorithm allowed us to determine that K<20 would be good enough for this model and data, and we assume K=30 will be good enough for the other two models, as well. Given the somewhat similar interpretation of the parameters, we set the same prior on the location and scale parameters of the mixture components. The sample size was n=400 and n=200 in the two simulations, respectively. Relatively fast computation times were achieved via usage of Julia and efficient memory management during MCMC.

From our testing, the **JLW** model looks the most smooth and stable out-of-sample, with few components receiving non-negligible weights. The **KSB-E** instead shows a tendency to keep non-negligible weights on a larger number of components, somehow hindering its performance. The **KSB-I** features a clear-cut definition of the dependence of the component weights on the covariates, and this may be desirable in some cases. The predictive densities look overall smoother and more concentrated around the regression function in the **JLW** model. The **KSB-I** and **KSB-E** models, instead, give contrasting results in the two simulations. Also different are the shapes one gets by looking at the components with maximum weight for every covariate value (x_1, x_2). We also note that the overlap of components in the **KSB-I** model is minimal.

FUTURE RESEARCH

This work provides a unifying framework and a shared platform for computations of bayesian nonparametric models for regression. Future work will involve extending the algorithm to other covariate-dependent weights models, and models with covariate-dependent atoms like the the single-p DDP with GP means.



Simulation 1. The **KSB-E** shows a concentrated density, which however does not decay smoothly and still shows weight being placed far from the regression mean. This is probably due to its tendency to keep non-negligible weights on multiple components. Overall, the **JLW** model seems more balanced. The covariate-dependent weighting shows abrupt changes for the **KSB-I** model due to the indicator kernels.

REFERENCES

I. Antoniano-Villalobos, S. Wade, and S.G. Walker. A bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in alzheimer's disease. *Journal of the American Statistical Association*, 109(506):477–490, 2014.

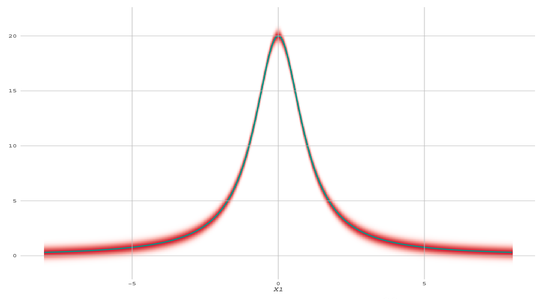
D.B. Dunson and J.H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008.

J.E. Griffin. An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing*, 26:423–441, 2016.

J.E. Griffin and D.A. Stephens. Advances in Markov chain Monte Carlo. In *Bayesian Theory and Applications*. OUP, 2013.

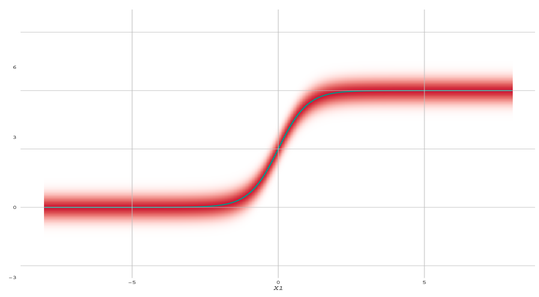
S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, Alexandria, VA, 1999. American Statistical Association.

S.N. MacEachern. Dependent Dirichlet processes. *Technical Report, Department of Statistics, Ohio State University*, 2000.



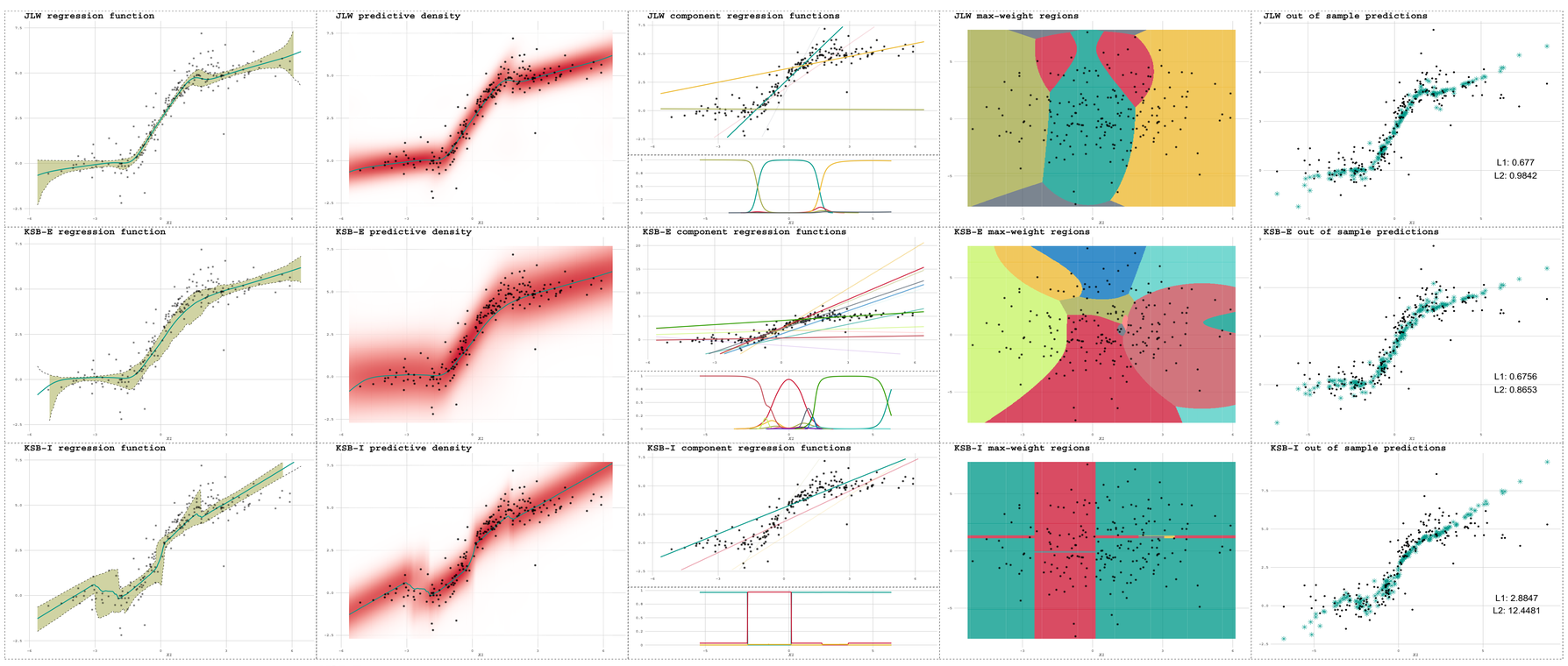
The true density and regression function in Simulation 1.

$$y_i = \frac{20}{1+x_{i,1}^2} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \frac{x_{i,2}^2}{32} + e^{-1})$$



The true density and regression function in Simulation 2.

$$y_i = \frac{5}{1+\exp(-2x_{i,1})} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \frac{x_{i,2}^2}{32} + e^{-1})$$



Simulation 2. Of the three models, the **KSB-E** has best out of sample predictions, but again, it shows a tendency to give somewhat larger weights to a larger number of mixture components compared to the other two models. This is especially true in the transition areas, where the model appears more unstable than the other two. Of the three models, the one with the more concentrated density is **JLW**.