Nonparametric Bayesian Conditional Density Models based on Orthogonal Polynomials

Andriy Norets and Marco Stenborg Petterson

Department of Economics

BROWN

Model

The paper considers a nonparametric Bayesian model for conditional densities

- We consider the space of conditional distributions

$$\mathcal{F} = \{f : \mathcal{Y} \times \mathcal{X} \to (0, \infty) - \text{Borel measurable}, \int f(y|x)dy = 1 \quad \forall x \in \mathcal{X}\}$$

  and we are interested in estimating nonparametrically $f_0 \in \mathcal{F}$.

- The model for conditional densities is based on an orthogonal series expansion with a prior on the coefficients and on the number of terms in the expansion,

$$f(y|x, \boldsymbol{a}, m) = \frac{\left(\sum_{j:||j||_\infty = m} a_j P_j(y, x)\right)^2}{\int \left(\sum_{j:||j||_\infty = m} a_j P_j(y, x)\right)^2 dy}$$

  where $j = (j_1, \ldots, j_d)$ indicates the degrees of the dimensions considered and $P_j(y, x)$ is a multivariate orthogonal polynomial

- The structure allows to naturally consider cross dimensional moves
- Orthogonality should imply stability of existing parameters when changing the number of components
- This model allows to analytically calculate the conditional distribution
- The series expansion is squared in order to ensure non-negativity

Literature Review

We refer to two main models: B-spline (Shen and Ghosal, 2016) and mixture of experts (Norets and Pati, 2017). We hope to improve on both:

- B-spline $\to$ fix a priori an upperbound on smoothness and interdependent when changing number of nodes
  - Orthogonal polynomials do not vary according to the level of smoothness
- Mixture of experts $\to$ difficult to simulate from posterior for variable number of experts, which is required for optimal posterior contraction rate
  - Orthogonality of polynomials should imply more stable parameters $\to$ higher acceptance rate

Assumptions

We assume a prior on m and $\boldsymbol{a}$ such that for some $b_1, b_2, b_3 > 0$ and $0 \le t_2 \le t_1 \le 1$

$$\Pi\left(\|a - a_0\|_2 \le \epsilon\right) \ge \exp\left\{-b_1 J \log(1/\epsilon)\right\}$$

and that

$$\Pi(m) \in \left[\exp\left\{-b_2 m \log^{t_1} m\right\}, \exp\left\{-b_3 m \log^{t_2} m\right\}\right]$$

We also assume

$$\frac{\sum_{j=0}^{\infty} a_{0j} P_j(y, x)}{\sum_{j:||j||_\infty = m} a_j P_j(y, x)}$$

to be bounded away from zero and infinity.
Finally, we assume that the approximation error is such that

$$\sum_{j=m+1}^{\infty} a_{0j}^2 \lesssim m^{-2\beta/d}$$

Convergence of Posterior

Under the assumptions above and following Shen and Ghosal (2015) we have that the posterior contraction rate is

$$\epsilon_n = n^{-\frac{\beta/d}{2\beta/d+1}} (\log n)^{\frac{\beta/d}{2\beta/d+1}}$$

MCMC: Cross-dimensional Move

- We allow for a flexible number of terms $\to$ RJMCMC in which we add or delete one polynomial term at a time
- An optimal RJMCMC proposal distribution for the coefficient of the new polynomial term is the conditional posterior distribution of this coefficient (Norets, 2018)
- The conditional posterior can only be computed up to a normalization constant; hence, we approximate the conditional posterior with a piecewise exponential function on a grid, for which the normalization constant is available in closed form and simulation from which is straightforward
  - The method is similar to Adaptive Rejection Sampling (Gilks, 1992)

MCMC: Within-dimension

- In application we use complete polynomials (Judd and Gaspar, 1997), rather than tensor products, to reduce the curse of dimensionality
- Large number of dimensions can make the posterior difficult to explore. We therefore use Hamiltonian Monte Carlo
  - No proposal function is needed, only log-posterior and gradient

Simulations

- We simulate observations from

$$y_i = \frac{\sin(\pi x_i) + \epsilon_i}{2}$$

  where x_i and ϵ_i are $i.i.d$ random variables with density $1 - |z|$ on $[-1, 1]$
- We assume $N(0, 1)$ as prior on the coefficients
- We fix the first coefficient to 0.5 since our model is identified up to a multiplicative constant
- We calculate the Mean Absolute Error

$$\text{MAE} = \frac{\sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \left|\hat{f}\left(y_i|x_j\right) - f_0\left(y_i|x_j\right)\right|}{N_y N_x}$$

Numerical Results

- We start by analyzing performance for a fixed maximum polynomial degree

MAE	$m=5$	$m=10$	$m=15$	$m=20$
$n=500$	0.2484	0.1018	0.1957	0.3384
$n=1000$	0.2061	0.0748	0.1283	0.2287
$n=2000$	0.1939	0.0538	0.0656	0.1194

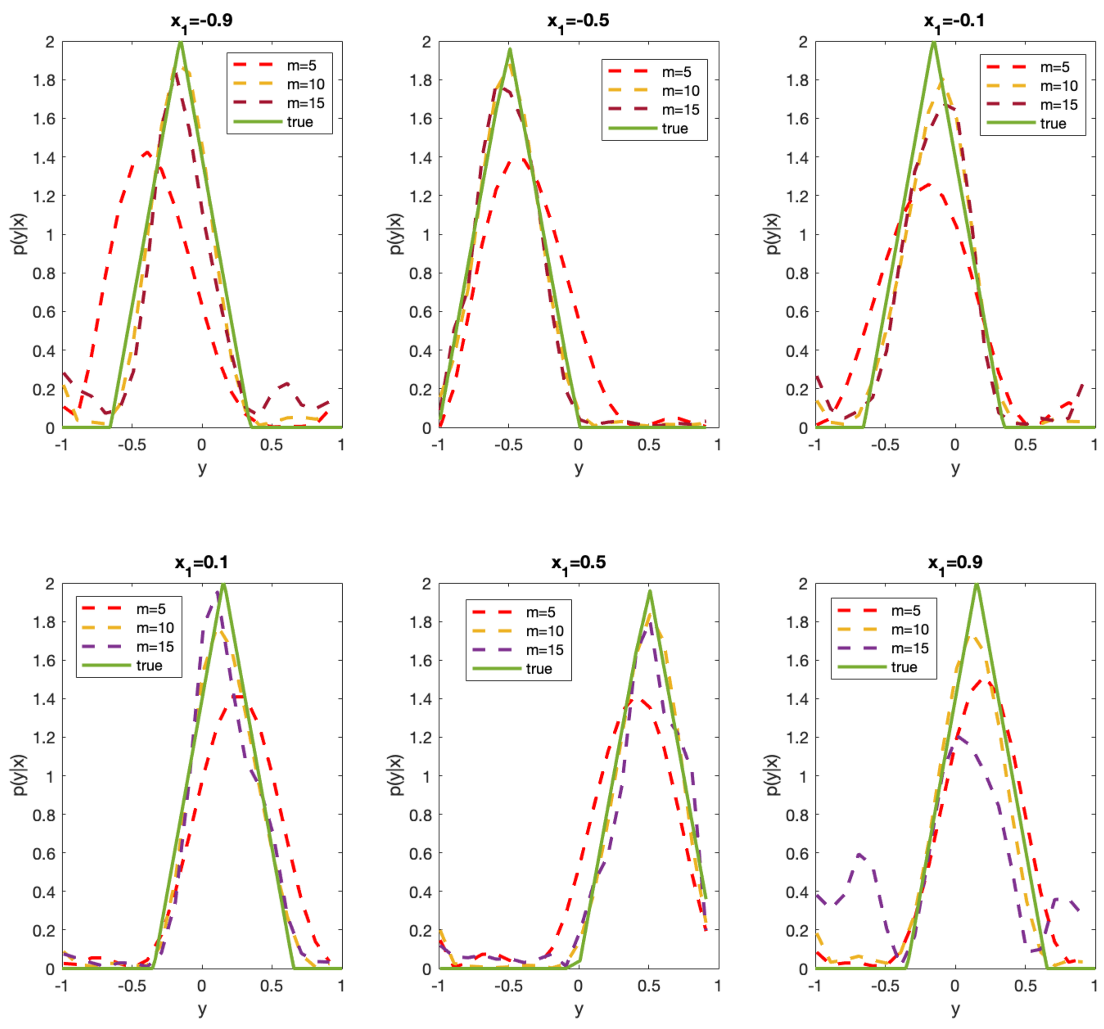

Figure 1: Simulations for $n = 1000$ and $m = 5, 10, 15$

- We now allow the degree to vary choosing at random from some precomputed polynomials

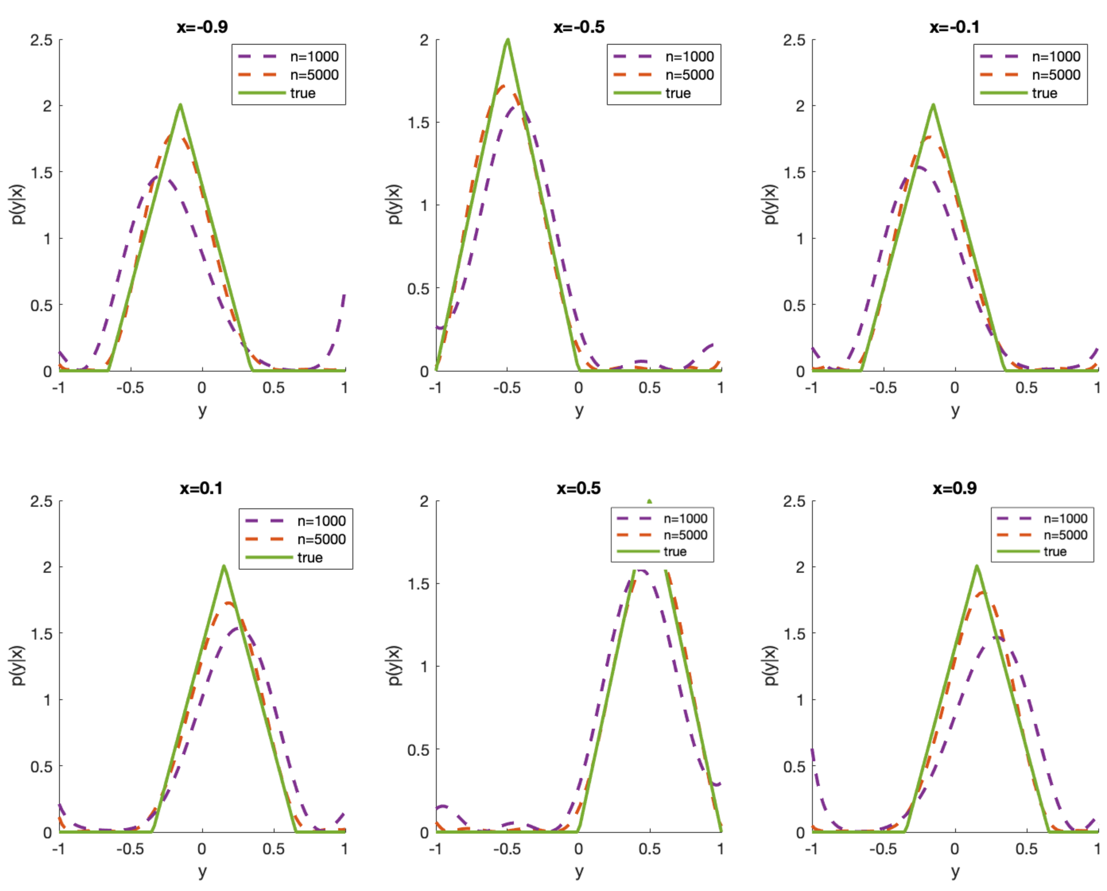

Figure 2: Simulations for $n = 1000$ and $n = 5000$

Future Work

- Complete vs tensor product polynomials
- Other families of orthogonal polynomials
- Relax boundedness away from zero assumption
- More extensive simulation experiments and applications

References

[1] Bradley P Carlin and Siddhartha Chib. “Bayesian model choice via Markov chain Monte Carlo methods”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.3 (1995), pp. 473–484.

[2] Peter J Green. “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”. In: *Biometrika* 82.4 (1995), pp. 711–732.

[3] Kenneth L. Judd and Gaspar. *Solving Large Scale RE models*. 1997. DOI: 10.1017/S1365100597002022.

[4] Andriy Norets. “Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models”. In: *Mimeo* (2018).

[5] Omiros Papaspiliopoulos and Gareth Roberts. “Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical model”. In: *Biometrika* (2008), pp. 1–32. arXiv: 0710.4228. URL: http://arxiv.org/abs/0710.4228.

[6] Weining Shen and Subhashis Ghosal. “Adaptive Bayesian Procedures Using Random Series Priors”. In: *Scandinavian Journal of Statistics* 42.4 (2015), pp. 1194–1213. ISSN: 14679469. DOI: 10.1111/sjos.12159.