CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 744

Bayesian Nonparametric Models for Conditional Densities Based on Orthogonal Polynomials

Andriy Norets and Marco S. Petterson

December 2024

CSEF - Centre for Studies in Economics and Finance
DEPARTMENT OF ECONOMICS AND STATISTICS – UNIVERSITY OF NAPLES FEDERICO II
80126 NAPLES - ITALY
Tel. and fax +39 081 675372 – e-mail: csef@unina.it

CSEF
Centre for Studies in Economics and Finance

WORKING PAPER NO. 744

Bayesian Nonparametric Models for Conditional Densities Based on Orthogonal Polynomials

Andriy Norets* **and Marco S. Petterson†**

Abstract

The paper considers a nonparametric Bayesian model for conditional densities. The model considered is a mixture of orthogonal polynomials with a prior on the number of components. The use of orthogonal polynomials allows for a great deal of flexibility in applications while maintaining useful approximation properties. We provide the posterior contraction rate in the case of Legendre polynomials. The algorithm proposed allows for cross-dimensional moves, allowing it to choose the optimal number of terms in the series expansion conditional on a penalty parameter. We also provide Monte Carlo simulations that show how well the model approximates known distributions also in finite sample situations.

JEL Classification: C11, C14, C13

Keywords: Bayesian nonparametrics, orthogonal polynomials, variable dimensions model

* Brown University

† University of Naples Federico II and CSEF. Email: marcostenborg.petterson@unina.it.

1 Introduction

Economics is often concerned with estimating the relationship between a dependent variable and a set of covariates. One tool to properly characterize this relationship is the conditional distribution of the dependent variable with respect to the covariates. In particular, characterizing the full conditional distribution, rather than only estimating the effect of the covariates on the average dependent variable, can give important insights into the mechanisms at play in the relationship and how they vary across different values of the covariates. Examples in which conditional distributions have been used include distribution of earnings in Geweke and Keane (2007), estimation of health expenditures in Keane and Stavrunova (2011), or the examples considered in Norets and Pelenis (2014).

There are, in fact, several studies on the estimation of conditional distributions both in classical nonparametrics and in Bayesian nonparametrics. The Bayesian approach to the estimation of conditional distributions has several attractive properties, one of which is the natural way to quantify the uncertainty of the conditional distribution through the posterior distribution, and it has also been shown to perform well in out-of-sample predictions.

In order to conduct nonparametric estimation, a suitable set of basis functions needs to be chosen. In this work, we choose to focus on orthogonal polynomials. In particular, every function can be represented as an infinite sum of weighted orthogonal polynomials once a set of orthogonal polynomials has been chosen. There are several reasons why orthogonal polynomials are appropriate in our setting.

First of all, they provide flexibility which is not always found in other approaches. In fact, a very common choice of basis is the B-spline. Such basis, though, forces the econometrician to fix a maximum level of smoothness of the function and then commit to that choice of basis. Orthogonal polynomials do not need these additional assumptions on the shape of the function to be approximated, they can easily be precomputed up to any degree and new degrees can be added to the expansion if the algorithm finds it optimal to do so, without changing any of the previous calculations. This implies that we can have a variable number of terms throughout the simulations, we take advantage of this possibility by adding a retrospective sampling step (Papaspiliopoulos and Roberts 2008). In particular, we will also use the optimality result obtained by Norets (2021) to improve the acceptance probability of the algorithm. Furthermore, the orthogonality properties of the polynomials imply that adding new terms does not change substantially the meaning of the terms included so far. This is not necessarily true in approaches like the mixture of normals.

Secondly, in the context of conditional densities, it is important that the estimator considered can be integrated analytically. Orthogonal polynomials have useful properties in this case for two reasons: being polynomials they are easy to integrate analytically and secondly, their orthogonality properties imply that such

calculations are even easier. Specifically, in the case of Legendre polynomials, the weight function $w(x) = 1$ further eases the calculations.

Something to keep in mind is that in order to apply a Bayesian framework we must ensure non-negativity of the likelihood function at all times. In order to do this, we will square the sum of polynomials, which allows us to maintain differentiability and ensures non-negativity. One of the reasons why we want differentiability of the model is because, when performing posterior simulations, we are going to use Hamiltonian Monte Carlo (HMC). HMC is a gradient-based algorithm which we describe in some more detail in Section 3, a more detailed analysis can be found in Betancourt (2017).

As of now, this paper places a specific focus on Legendre polynomials, but the framework we aim at developing is more general and encompasses a large set of orthogonal polynomials, e.g. Chebyshev polynomials. We choose Legendre polynomials because of their weight function, which makes several of the calculations easier from an analytical point of view and somewhat more intuitive as well.

In this paper, we provide both results concerning frequentist properties of Bayesian nonparametric procedures and an algorithm which allows for a variable number of terms in the considered series expansion. The frequentist properties of Bayesian nonparametric procedures have important consequences. In particular, notice that for nonparametric Bayesian functions the term prior refers to the prior put on a space of functions, often infinite-dimensional, and it is therefore crucial to understand the properties of such prior in order for it not to be dogmatic in its choice and furthermore they allow to understand the rate at which the model contracts to the distribution generating the data. Several studies have obtained adaptive posterior convergence rates for models concerning densities estimation, Huang (2004), Scricciolo (2006), Van Der Vaart and Van Zanten (2009), and some other have been concerned specifically with multivariate joint densities which then imply convergence rates for conditional densities, Van Der Vaart and Van Zanten (2008) and Shen, Tokdar, and Ghosal (2013), Norets and Pati (2017).

In our setting we do not model the marginal distributions of the covariates but we instead put a prior directly on the conditional distribution hence sidestepping the concerns raised by Wade et al. (2014) on the fact that, when using mixtures, some components may tend to provide a good fit for the marginal distributions but worsen the fit of the conditional.

The paper is structured as follows. Section 2 presents the main theoretical results of the paper. Section 3 describes the used algorithm, Section 4 contains some results of Monte Carlo simulations. Section 5 concludes. The Appendix includes some useful results on Legendre polynomials.

2 Theoretical Results

2.1 Notation

Let $\mathcal{Y} \subset [-1,1]$ be the response space, $\mathcal{X} \subset [-1,1]^{d_x}$ be the covariate space, and $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$. Let then $\mathcal{F}$ denote the space of conditional densities with respect to the Lebesgue measure

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

We then assume that we have a random sample (Y^n, X^n) from the joint density $f_0 g_0$ available to us. In particular, we have $f_0 \in \mathcal{F}$ and g_0 is a density on $\mathcal{X}$ with respect to the Lebesgue measure. Let us define then the Hellinger distance for conditional distributions, which will be useful for the derivation of the main results

$$d_h(f_1, f_2) = \left(\int \left(\sqrt{f_1(y|x)} - \sqrt{f_2(y|x)} \right)^2 g_0(x) dy dx \right)^{1/2}$$

The operator $\lesssim$ denotes less or equal up to a multiplicative constant. $N(\epsilon, B, \rho)$ denotes the $\epsilon-$packing number of the set B with respect to the metric ρ. The $\epsilon-$packing number is defined as the maximum cardinality of an ϵ- dispersed subset of B with respect to the distance ρ. In other words the maximum number of points in B such that they are at least ϵ apart using the distance ρ, intuitively this number gives a sense of how "big" the space that we are considering is.

We also define a class of locally Hölder functions $\mathcal{C}^{\beta, L, \tau_0}$ as functions $f : \mathbb{R}^d \to \mathbb{R}$ such that for $k = (k_1, \ldots, k_d)$ we have that $k_1 + \cdots + k_d \leq \lfloor \beta \rfloor$, the mixed partial order derivative of order k, $D^k f$ is finite and for $\Delta z \in \mathcal{Z}$

$$\left| D^k f(z + \Delta z) - D^k f(z) \right| \leq L(z) \|\Delta z\|^{\beta - \lfloor \beta \rfloor} e^{\tau_0 \|\Delta z\|^2}$$

Finally, we define the generalized Kullback-Leibler neighborhood as

$$\mathcal{K}(f_0, \epsilon) = \left\{ f : \int f_0 g_0 \log(f_0/f) < \epsilon^2, \int f_0 g_0 \left[\log(f_0/f) \right]^2 < \epsilon^2 \right\}$$

2.2 Assumptions on DGP

We assume that that $\mathcal{X} = [-1,1]^d$, or that the covariates have been rescaled to be such that this is true. This assumption comes naturally from the fact that we will be using Legendre polynomials for our calculations, Legendre polynomials are in fact orthogonal in this space. In the more general setting of orthogonal polynomials the covariate space can be chosen appropriately depending on the

space on which the orthogonal polynomials are defined.

We choose g_0 to be the uniform distribution over $[-1,1]$, which implies that the joint distribution has the same smoothness as f_0, the conditional distribution.

We assume $f_0 \in \mathcal{C}^{\beta,L,\tau_0}$ and also that $f_0(y|x) \in (0,\infty)$. The main reason for this assumption is that it allows for an equivalence between Hellinger distance and Kullback-Leibler divergence, such equivalence will be useful for the derivation of the main result.

It is known, from theKhasminskii (1979), the minimax rate for a $d+1$ dimensional density belonging to a $\beta-$Hölder class is $(n/log(n))^{-\beta/(2\beta+d+1)}$, with respect to the sup-norm.

2.3 Assumptions on the prior

We make some general assumptions and we verify them with Legendre polynomials.

We consider a prior Π on $\mathcal{F}$ defined by a mixture of orthogonal polynomials[1]. We model the joint density as

$$fg = \left(\sum_{j:||j||_\infty = m} a_j P_j(y,x) \right)^2$$

and consequently the conditional density is

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

like in Shen and Ghosal (2015) we make the following assumptions on the priors of the coefficients

- For some $b_1, b_2, b_3 > 0$ and $0 \leq t_2 \leq t_1 \leq 1$, we have $\Pi\left(\|a - a_0\|_2 \leq \epsilon\right) \geq \exp\left\{-b_1 J \log(1/\epsilon)\right\}$ and that $\Pi(m) \in \left[\exp\left\{-b_2 m \log^{t_1} m\right\}, \exp\left\{-b_3 m \log^{t_2} m\right\}\right]$

In the application we consider in the simulations we put independent Normal priors on the coefficients and and a discrete exponential prior on the number of terms used, both of these priors satisfy the conditions above.

If we where to consider univariate polynomials there would be no confusion on whether m represents the number of polynomials of the degree of the polynomials considered. When it comes to the multivariate setting some clarifications become necessary. We consider m to be the number of terms used also in the multivariate setting. In application it may be necessary to decide an order in which

[1]Notice that we will be using normalized orthogonal polynomials, such that $\forall j \int_{-1}^{1} P_j(x) P_j(x) dx = 1$.

to add the polynomials to the series expansion as it is not necessarily obvious in which dimension it is best to increase the degree of the polynomial considered. Alternatively, it can be left to be chosen at random among a set of polynomials, this is in fact done in the simulations results of Section 4.

In order to be able to use an equivalence result between the Kullback-Leibler divergence and the Hellinger distance we will assume that $\sum_{j=1}^{m} a_j P_j(y,x)$ is bounded away from 0. Future developments of this paper will look at a more explicit condition to define a set of parameters $a_j \;\; j = 1,\dots,m$, in which this is true. It is worth noting though that this set exists in the case of Legendre polynomials as the zeros of these polynomials are different for each degree and therefore there is no x nor y that can make the sum equal to 0 for any a_j .

Finally, we assume that

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

we know this sort of result to be true for several orthogonal polynomials, including Legendre polynomials both from Shen and Ghosal (2015) and Theorem 6.1 of Hesthaven, S. Gottlieb, and D. Gottlieb (2007) reported in the appendix. The generalization to multidimensional polynomials should be straightforward. Furthermore the fact that we are considering $\left(\sum_{j:||j||_\infty=m} a_j P_j(y,x)\right)^2$ rather than the classical Legendre series expansion,$\sum_{j:||j||_\infty=m} \hat{a}_j P_j(y,x)$, should also not constitute a problem as it is natural to view it as just the squared approximation of $\sqrt{f_0 g_0}$.

2.4 Results

We want to find the posterior contraction rate of our prior to the density f_0 that generated our sample. As said above, we are in a similar setting to the one considered by Shen and Ghosal (2015) and therefore we follow the lines of their Theorems 1 and 2 considering our specific choice of prior and the fact that we are in the setting of conditional densities. Since the proof of theorems similar to this is widespread in the literature we do not report the proof here.

Theorem 1. *Suppose we observe independent observations* $\{Y_i, X_i\} \;\; i = 1,\dots,n$, *following some joint density* $f_0 g_0$. *Let* ϵ_n *and* $\tilde{\epsilon}_n$ *bet positive sequences such that* $\tilde{\epsilon}_n \leq \epsilon_n$ *with* $\epsilon_n \to 0$ *and* $n\tilde{\epsilon}_n^2 \to \infty$. *Suppose* $\mathcal{F}_n \subset \mathcal{F}$ *is a sieve defined as*

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

for which the following conditions are satisfied

$$\log N(\epsilon_n, \mathcal{F}_n, d_h) \leq c_1 n\epsilon_n^2 \tag{1}$$

$$\Pi\left(\mathcal{F}_n^c\right) \leq c_3 exp\left\{-c_2 n\tilde{\epsilon}_n^2\right\} \tag{2}$$

$$\Pi\left(\mathcal{K}\left(f_0, \tilde{\epsilon}_n\right)\right) \geq c_4 exp\left\{-c_5 n\tilde{\epsilon}_n^2\right\} \tag{3}$$

Then there exist an $M > 0$ such that

$$\Pi\left(f : d_h\left(f, f_0\right) > M\epsilon_n | Y^n, X^n\right) \xrightarrow{P_0^n} 0$$

where P_0 is the probability measure corresponding to $f_0 g_0$.

Let us now discuss the intuition of the assumptions. The main idea behind this kind of results is that asymptotically we want our model to tend to the true DGP of the data. The question is then at what speed such approximation takes place and this is exactly the contraction rate ϵ_n we are looking for.

We focus on the two main conditions: Equation 3 which we call KL condition and Equation 1 which we call entropy bounds. We notice that in condition 1 the left hand side increases as the ϵ_n decreases and vice-versa for the right hand side. This condition ensures the existence of a test ϕ^n of $f = f_0$ against $\{f \in \mathcal{F}_n : \rho\left(f, f_0\right) > M\epsilon_n\}$ with decreasing error of both types.

The KL condition is concerned with prior thickness, we want the prior to put enough mass to the Kullback-Leibler neighborhoods of the true density, to do this we need to be able to define a measure of closeness between the true density and the estimated density which we do through the Kullback-Leibler neighborhoods. The proofs of the results will then rely on being able to state equivalently this condition as closeness of the estimated parameters.

These two conditions together then define the best contraction rate that can be obtained as proved in the seminal work of Ghosal, Ghosh, and Van Der Vaart (2000) and in several papers after for different models.

Theorem 2. *Let ϵ_n and $\tilde{\epsilon}_n$ bet positive sequences such that $\tilde{\epsilon}_n \leq \epsilon_n$ with $\epsilon_n \to 0$ and $n\tilde{\epsilon}_n^2 \to \infty$. Assume there exist sequences of positive numbers $\bar{M}_n, M_n$ and A_n. Let $c_1, \ldots, c_{10}$ be positive constants with $c_{10} > 1$ and let the following assumptions hold:*

$$M_n\left(\log M_n + \log A_n + 2C\log n\right) \leq n\epsilon_n^2$$
$$d_H^2(f_0 u, fg) \leq \|a_0 - a\|_2^2 \leq \sum_{j=m+1}^{\infty} a_{0j}^2 \leq C_1 m^{-2\beta/d} \leq \tilde{\epsilon}_n^2$$
$$c_9\overline{M}_n \log^{t_1}\overline{M}_n + c_8\overline{M}_n \log\left(2c_7\left(\overline{M}_n\right)/\tilde{\epsilon}_n\right) \leq c_6 n\tilde{\epsilon}_n^2$$
$$n\tilde{\epsilon}_n^2 \leq C_2 M_n \log^{t_2} M_n \text{ for any constant } C_2,\ M_n \exp\left\{-C_2 A_n^{t_3}\right\} \leq (c_{10} - 1)\exp\left\{-n\tilde{\epsilon}_n^2\right\}$$

Under the assumptions of Sections 2.2 and 2.3 the conditions of Theorem 1 hold with

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

We note that several of the assumptions rely on the same idea as Shen and Ghosal (2015). The proof has several steps that we construct in the following propositions

As a first step to obtain the results we bound the Hellinger distance between conditionals by the distance between joint, in the same spirit as Norets and Pati (2017) by using their Lemma 9.2. In particular

$$d_h^2(f_0, f) \lesssim d_H^2(f_0 u, f g)$$

for an arbitrary f and g.

We can then use the properties of orthogonal polynomials to bound the squared Hellinger distance between polynomials.

In particular for our choice of prior the following result holds

Proposition 1. *Let $f_0 u$ and $f g$ be as previously defined, we then have that*

$$d_H^2(f_0 u, f g) \leq \|a_0 - a\|_2^2 \tag{4}$$

Proof. Recall the definition of squared Hellinger distance

$$d_H^2(f_0 u, f g) = \int \int \left\{ \sqrt{\left(\sum_{j=0}^{\infty} a_{0j} P_j(y, x) \right)^2} - \sqrt{\left(\sum_{j:||j||_\infty = m} a_j P_j(y, x) \right)^2} \right\}^2 dy dx$$

Notice that we can write any function like this because the Legendre polynomial constitute and orthonormal basis. This can then be rewritten as

$$\begin{aligned}
&= 2 \left[1 - \int \sqrt{\left(\textstyle\sum_{j=0}^{\infty} a_{0j} P_j(y, x) \right)^2 \left(\sum_{j=0}^{m} a_j P_j(y, x) \right)^2} dy dx \right] \\
&= 2 \left[1 - \int | \left(\textstyle\sum_{j=0}^{\infty} a_{0j} P_j(y, x) \right) \left(\sum_{j=0}^{m} a_j P_j(y, x) \right) | dy dx \right] \\
&\leqslant 2 \left[1 - | \int \left(\textstyle\sum_{j=0}^{\infty} a_{0j} P_j(y, x) \right) \left(\sum_{j=0}^{m} a_j P_j(y, x) \right) dy dx | \right] \\
&\leqslant 2 \left[1 - \int \left(\textstyle\sum_{j=0}^{\infty} a_{0j} P_j(y, x) \right) \left(\sum_{j=0}^{m} a_j P_j(y, x) \right) dy dx \right] \\
&= 2 \left[1 - \left(\textstyle\sum_{j=0}^{\infty} a_{0j} a_j \right) \right]
\end{aligned}$$

where the last equality follows from the properties of normalized orthogonal polynomials. Note that we can write $2 \left[1 - \left(\sum_{j=0}^{\infty} a_{0j} a_j \right) \right]$ because we assume that $\forall j > m \; a_j = 0$.

We then note that $\sum_{j=0}^{\infty} a_{0j}^2 = \sum_{j=0}^{\infty} a_j^2 = 1$ as proven in the Appendix and hence we can rewrite the last expression as

$$= \|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2$$

as we wanted to show. Notice that $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2 = \sum_{j=m+1}^{\infty} a_{0j}^2$ as we wanted. □

Notice we focus on using the Hellinger distance even though the conditions use KL-divergence. We can do this as there exist an equivalence relation between the Hellinger distance and the Kullback-Leibler divergence as long as the ratio of the two distributions is bounded above and away from zero.[2] In this case, the relevant ratio between distributions that allows the equivalence between the Hellinger distance and the KL-divergence is

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

but we recall we assumed the that the denominator is bounded away from zero in section 2.3. Recall we also assumed the true joint density is also bounded.

We are now interested in computing the entropy bounds

$$\log N\left(\epsilon_n, \mathcal{F}_n, d\right) \leq n\epsilon_n^2$$

Using the definition of packing number and Equation 4 we have that

$$\begin{aligned}
\log N\left(\epsilon_n, \mathcal{F}_n, d\right) &\leq \log\left\{\sum_{j=1}^{M_n} N\left(n^{-2C}, \left\{a \in \mathbb{R}^j, \|\boldsymbol{a}\|_\infty \leq A_n\right\}, \|\cdot\|_2\right)\right\} \\
&\leq \log\left\{M_n\left\{\sqrt{M_n} A_n n^{2C}\right\}^{M_n}\right\} \\
&\leq M_n\left(\log M_n + \log A_n + 2C\log n\right) \\
&\leq n\epsilon_n^2
\end{aligned}$$

where the first inequality follows by the fact that ϵ_n is lower bounded by n^{-1} by assumption.

Next we verify Condition 2

$$\begin{aligned}
\Pi\left(\mathcal{F}_n^c\right) &\leq \Pi(m > M_n) + \sum_{j=1}^{M_n} \Pi\left(\boldsymbol{a} \notin [-A_n, A_n]^j\right) \Pi(m = j) \\
&\leq \exp\left(-b_2' M_n \log^{t_2} M_n\right) + M_n \exp\left\{-C A_n^{t_3}\right\} \\
&\leq a_1 \exp\left\{-n\bar{\epsilon}_n^2\right\}
\end{aligned}$$

this condition follows by just applying the assumptions stated above, which in fact mirror the idea of the KL-condition, we assume that the prior puts sufficiently small probability on the part of the space that is outside the sieve. This

[2] See also Lemma B.1 in Ghosal and Vaart (2017).

condition plays a more important role for results concerning consistency, which is not addressed in this paper as of now.

Finally, we verify the KL-condition, we notice though that binding the distance between the prior and the true distribution by the distance between the parameters allows us to state the condition directly on the parameters which are much easier to verify using our assumptions.

$$\begin{aligned}\Pi\left\{f : d_h\left(f_0, f\right) \leq 2\tilde{\epsilon}_n\right\} &\geq \Pi\left(m = \bar{M}_n\right)\Pi\left(\left\|\boldsymbol{a} - \boldsymbol{a}_0\right\|_2 \leq \tilde{\epsilon}_n\right) \\ &\geq \exp\left\{-c_1 \bar{M}_n \log^{t_1} \bar{M}_n\right\} \exp\left\{-c_3 \bar{M}_n \log\left(\frac{1}{\tilde{\epsilon}_n}\right)\right\}\end{aligned}$$

where d is the Hellinger distance.

Then we can apply the result of Theorems 1 and 2 of Shen and Ghosal (2015) follows and the contraction rate is the one stated in Theorem 2.

3 Algorithm

In this section we explain the steps of the algorithm we propose and use in the simulations of the next section.

Note that, in the applications, when specifying the degree of the polynomials used, we do not use the tensor product of polynomials in order to try to reduce the curse of dimensionality that often arises in these cases when using orthogonal polynomials, the degree of the polynomials instead is the maximum total degree of the polynomials considered. The use of total degree polynomials is not new to economics, in fact they were introduced by Gaspar and L. Judd (1997) under the name of complete polynomials. Another possible scheme adopted in the approximation theory literature is called the euclidean degree, in which the polynomials are chosen depending on the Euclidean norm of their degrees, this approach is shown to have interesting properties in Trefethen (2017). Notice that for a given maximum degree m using complete polynomials rather than the tensor product of the polynomials reduces the number of terms is just $\sum_{j=1}^{m} \binom{j+d-1}{j}$.

Furthermore within each of the iterations for which we allow a change in dimensions we apply Hamiltonian Monte Carlo method to extract the parameters of interest. We present below the main ideas of the algorithm.

3.1 Hamiltonian Monte Carlo

The HMC is a Monte Carlo method that makes use of the gradient of the distribution in order to simulate a chain of values from the distribution considered. Two of the aspects that make this algorithm very useful are: the fact that the distribution considered does not need to be normalized and the speed at which it is able

to extract samples. As the name suggests the algorithm relies on the Hamiltonian dynamics of the distribution and the gradient allows it to follow precisely the curvature of the function considered.

We describe the main functioning with a physical analogy, the main mathematical concepts underlying the HMC are grounded in differential geometry and for a more extensive analysis of the algorithm we refer the reader to Betancourt (2017). The physical analogy is to imagine a hockey puck which is kicked in a random direction on a surface, which is in our case the negative of the posterior distribution, and it is then stopped after a random amount of time. The sample is extracted in the point in which the puck stops. Clearly, in order to be able to calculate in which direction the puck will be going after the first kick the slope of the surface is needed as it will influence the direction in which it is moving and this is where the gradient of the function considered is used. Notice that the normalization constant of the posterior is not needed for this to work.

It is very easy to differentiate orthogonal polynomials and therefore an implementation of our model in the HMC can be achieved without major intricacies

3.2 Cross-dimensional move

We allow for a varying number of components in our Legendre series expansion and therefore include this step in the algorithm. Notice that this implies that the dimension of the parameter space varies from one iteration to the other. We therefore apply the techniques developed in Norets (2021) in order to obtain the maximum probability of acceptance for cross-dimensional moves.

In oder to apply such techniques we need to be able to approximate as well as possible a the posterior distribution of the additional parameter, be it to be added or removed, but we also need to be able to calculate the normalization constant of such distribution in order to extract a random sample from it. To do this we apply techniques similar to the adaptive rejection sampling introduced by Gilks, Best, and Tan (1995) The main idea is to consider a grid of possible values for the parameter to be extracted and for each of these points on the grid we calculate the exact value of the unnormalized log-posterior. We then linearly approximate the log-posterior within the grid and assume an exponential decay outside such grid. The linear approximation makes it easy to calculate the normalization constant of this piecewise function and we then use the inverse CDF method in order to extract the random sample. The acceptance probability for this draw, and therefore the change in number of components considered is then

$$\alpha\left(m^{*}, m\right) = \frac{p\left(Y|m^{*}, a_{1m^{*}}, X\right)\Pi\left(a_{1m^{*}}|m^{*}\right)\Pi\left(m^{*}\right)}{p\left(Y|m, a_{1m}, X\right)\Pi\left(a_{1m}|m\right)\Pi(m)} \cdot \left(\frac{1\left\{m^{*}=m+1\right\}}{\tilde{\pi}_{m}\left(a_{m+1}|a_{1m}, Y, X\right)} + 1\left\{m^{*}=m-1\right\}\tilde{\pi}_{m-1}\left(a_{m}|a_{1m-1}, Y, X\right)\right) \tag{5}$$

where$m^{*}$is the proposed new number of terms and $\tilde{\pi}_{m}$ is defined as

$$\begin{aligned}\tilde{\pi}_{m}\left(a_{m+1}|a_{1m}, Y, X\right) &= p\left(a_{m+1}|Y, m+1, a_{1m}, X\right) \\ &\propto p\left(Y|m+1, a_{1m+1}, X\right)\Pi\left(a_{1m+1}|m+1\right)\end{aligned}$$

3.3 Main steps of the algorithm

We report here the main steps of the algorithm. First of all we notice that the model we are considering is identified up to a constant. Hence we need to fix one parameter to be able to properly identify the other parameters of the series expansion. In application we do this by fixing the first parameter $a_{0,\ldots,0} = 0.5$, clearly any number can be chosen.

1. Precompute the multivariate orthogonal polynomials for every value of the random sample up to a given degree, decided by the econometrician. This may be an time-expensive operation but only needs to be done once.

2. Run one iteration of the Hamiltonian Monte Carlo sampler, possibly using as starting point the Maximum-a-Posteriori of the log probability density. The surface can become irregular, especially when a high degree of polynomial is considered or the dimension is large, hence the efficiency is improved when a good starting point is chosen

3. Propose a change in the number of terms of the used series expansion. We now have two cases

   (a) an increase in the number of terms is proposed

      i. Compute the corresponding orthogonal polynomial if not done so before the start of the iterations.

      ii. Construct a grid of possible values of the parameter to be extracted and evaluate the posterior at each of points of the grid, keeping constant the other parameters

      iii. Construct a linear approximation of the log density that using the values computed in the points of the grid and compute the corresponding normalizing constant

      iv. Extract a random sample from the approximation and compute the value it takes within the approximation

v. Compute the acceptance probability according to Equation 5 and consequently decide whether to accept the draw

(b) a decrease in the number of terms is proposed

i. Construct a grid around the parameter to be removed and evaluate the posterior at each of the point

ii. Calculate the linear approximation of the log density given the grid, the normalization constant and the corresponding value that the parameter takes in this approximation

iii. Compute the acceptance probability according to Equation 5 and consequently decide whether to accept the draw

4. Run an iteration of the HMC starting from the parameter values extracted from step 3 with a simulation length chosen by the econometrician. Go to step 3 until the desired number of iterations has been done.

4 Simulations

In this section we present some numerical results of the Monte Carlo simulations on finite samples that we ran following a known Data Generating Process.

At this stage of the paper we do not yet have a fully working version of the code which allows to change for a cross-dimensional move, we present some preliminary numerical results in a dedicated subsection below.

4.1 Fixed maximum total degree

We use different maximum total degrees $m = 5, 10, 15$ and 20 and also different sample sizes $n = 500, 1000, 2000$. In order to evaluate the performance of the estimator we report mean absolute error and root mean squared error

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \left(\hat{f}\left(y_i|x_j\right) - f_0\left(y_i|x_j\right) \right)^2}{N_y N_x}}$$

We report the MAE and RMSE for a set of covariates $x_i = [-0.9, -0.5, -0.1, 0.1, 0.5, 0.9]$ and a grid of equally spaced y_i and observe how the approximation changes for different combinations of maximum total degree and sample size.

We consider the following DGP

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

wherex_i and ϵ_i are $i.i.d$ random variables with density$1 - |x|$ on $[-1, 1]$, which is the same DGP as considered in Norets and Pelenis (2014), rescaled for our purposes. The coefficients were given a $Normal(0, 1)$ prior.

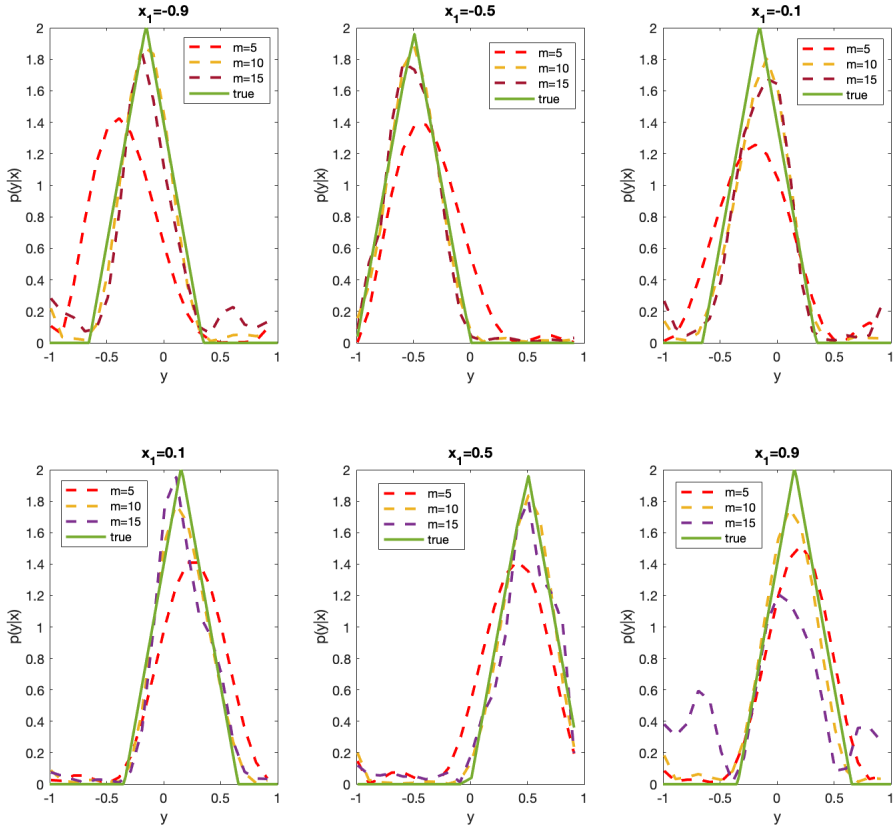
MAE	$m = 5$	$m = 10$	$m = 15$	$m = 20$
$n = 500$	0.2484	0.1018	0.1957	0.3384
$n = 1000$	0.2061	0.0748	0.1283	0.2287
$n = 2000$	0.1939	0.0538	0.0656	0.1194

RMSE	$m = 5$	$m = 10$	$m = 15$	$m = 20$
$n = 500$	0.3364	0.1376	0.2797	0.4568
$n = 1000$	0.2873	0.1057	0.1832	0.3513
$n = 2000$	0.2756	0.0776	0.0907	0.1723

We notice that the fit improves with the number of observations ant it also seem to improve with the maximum total degree of the polynomials used. When reaching though $m = 20$ and also somewhat with $m = 15$ we have a worsening of the fit, probably because of the increase in the number of parameters, which with total degree increase by a binomial coefficient factor $\binom{m+d-1}{m}$.

We also plot the mean of the approximations corresponding to the MCMC draws together with the true value of the distributions at a some specific values of the covariate x. We plot them for different values of $m = 5, 10, 15$ just to show how the approximation level changes increasing the degree of the polynomials. In particular, we plot the simulations for $n = 1000$ and we plot the mean conditional likelihood estimated from 1000 MCMC draws with a burn-in of 500. The solid line represents the true value of the conditional densities for the covariate specified, while the dotted line represent the degrees $m = 5, 10, 15$ in dark blue, orange and green respectively. We can see that the best fit is obtained by $m = 10$ while $m = 15$ includes too many fluctuations and is probably overfitting the distribution.

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



4.2 Flexible number of terms

In this subsection we present some preliminary, but promising, results of our algorithm when it is allowed to change the total number of terms in the series expansion, considering a penalty for the number of terms. The DGP used and the statistics reported are the same as the previous section. We report them for sample sizes $n = 1000$ and $n = 5000$.

The way the polynomials are chosen in this case is by pre-computing a number of polynomials up to a total degree of 15. The algorithm as described in Section 3 is run and which polynomial is added or removed is chosen at random between all the precomputed polynomials. The only polynomial which is not allowed to be removed is the first one, which is necessary for the normalization.

The prior on the number of components is

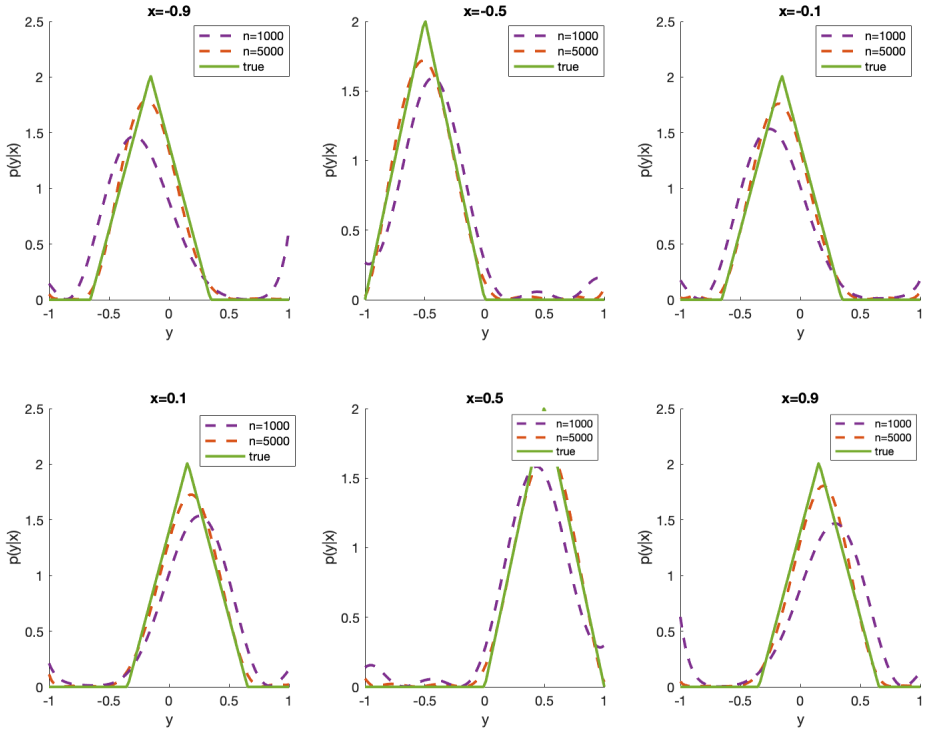
$$\Pi(m) = \frac{e^H - 1}{e^{-H}} e^{-Hm}$$

with a hyper-parameter $H = 1$, in order to make sure the algorithm works correctly we have also performed a Joint Distribution Test, as proposed by Geweke (2004). The t-statistics, after $20,000$ simulations were all not statistically significant.

	MAE	RMSE
$n = 1000$	0.1816	0.2459
$n = 5000$	0.0491	0.0728

In the figure the solid line represents the true value of the conditional densities for the specified covariate while the dotted purple and orange line represent the mean conditional density given 1000 MCMC draws for $n = 1000$ and $n = 5000$ respectively. As we can see the values are comparable to, or better of, some of the values for a fixed total degree of polynomials considered even though here we allow for much more flexibility and therefore also verifying which components have been selected can give some additional insights on the behavior of the true DGP.

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



5 Conclusions

We have showed that orthogonal polynomials series expansions have attractive properties when used as basis for the nonparametric estimation of conditional densities. In particular an adaptive contraction rate can be found given the conjectures on approximation error of multivariate orthogonal polynomials, which are likely to hold. This has been shown in the special case of Legendre polynomials. The flexibility of these series furthermore imply they are the optimal tool to approximate functions of unknown smoothness as they can be easily computed and no prior assumption on the maximum level of smoothness is needed. In application the structure of orthogonal polynomials should allow for a faster convergence to the optimal number of polynomials given the penalty on number of terms.

Future work will be on proving the conjecture on the approximation error, solidifying the numerical results for a varying number of terms in the expansion and for different families of orthogonal polynomials.

References

Betancourt, Michael (2017). *A Conceptual Introduction to Hamiltonian Monte Carlo*. DOI: 10.48550/ARXIV.1701.02434. URL: https://arxiv.org/abs/1701.02434.

Gaspar, Jess and Kenneth L. Judd (1997). "Solving Large-Scale Rational-Expectations Models". In: *Macroeconomic Dynamics* 1.1, 45â75. DOI: 10.1017/S1365100597002022.

Geweke, John (2004). "Getting it right: Joint distribution tests of posterior simulators". In: *Journal of the American Statistical Association* 99.467, pp. 799–804.

Geweke, John and Michael Keane (2007). "Smoothly mixing regressions". In: *Journal of Econometrics* 138.1, pp. 252–290.

Ghosal, Subhashis, Jayanta K Ghosh, and Aad W Van Der Vaart (2000). "Convergence rates of posterior distributions". In: *Annals of Statistics*, pp. 500–531.

Ghosal, Subhashis and Aad van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, pp. 1–646. ISBN: 9781139029834. DOI: 10.1017/9781139029834.

Gilks, Wally R, Nicky G Best, and Keith KC Tan (1995). "Adaptive rejection Metropolis sampling within Gibbs sampling". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 44.4, pp. 455–472.

Hesthaven, Jan S., Sigal Gottlieb, and David Gottlieb (2007). *Spectral Methods for Time-Dependent Problems*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press. DOI: 10.1017/CBO9780511618352.

Huang, Tzee-Ming (2004). "Convergence rates for posterior distributions and adaptive estimation". In: *The Annals of Statistics* 32.4, pp. 1556–1593.

Keane, Michael and Olena Stavrunova (2011). "A smooth mixture of tobits model for healthcare expenditure". In: *Health economics* 20.9, pp. 1126–1153.

Khasminskii, Rafail Z (1979). "A lower bound on the risks of non-parametric estimates of densities in the uniform metric". In: *Theory of Probability & Its Applications* 23.4, pp. 794–798.

Norets, Andriy (2021). "Optimal auxiliary priors and reversible jump proposals for a class of variable dimension models". In: *Econometric Theory* 37.1, pp. 49–81.

Norets, Andriy and Debdeep Pati (2017). "Adaptive Bayesian estimation of conditional densities". In: *Econometric Theory* 33.4, pp. 980–1012.

Norets, Andriy and Justinas Pelenis (2014). "Posterior consistency in conditional density estimation by covariate dependent mixtures". In: *Econometric Theory* 30.3, pp. 606–646. ISSN: 14694360. DOI: 10.1017/S026646661300042X.

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

Scricciolo, Catia (2006). "Convergence rates for bayesian density estimation of infinite-dimensional exponential families". In: *Annals of Statistics* 34.6, pp. 2897–2920. ISSN: 00905364. DOI: 10.1214/009053606000000911. arXiv: 0708.0175.

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

Shen, Weining, Surya T. Tokdar, and Subhashis Ghosal (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures". In: *Biometrika* 100.3, pp. 623–640. ISSN: 00063444. DOI: 10.1093/biomet/ast015.

Trefethen, Lloyd N. (2017). "Cubature, Approximation, and Isotropy in the Hypercube". In: *SIAM Review* 59.3, pp. 469–491. ISSN: 0036-1445.

Van Der Vaart, A. W. and J. H. Van Zanten (2008). "Rates of contraction of posterior distributions based on Gaussian process priors". In: *Annals of Statistics* 36.3, pp. 1435–1463. ISSN: 00905364. DOI: 10.1214/009053607000000613.

— (2009). "Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth". In: *Annals of Statistics* 37.5 B, pp. 2655–2675. ISSN: 00905364. DOI: 10.1214/08-AOS678. arXiv: arXiv:0908.3556v1.

Wade, Sara, David B Dunson, Sonia Petrone, and Lorenzo Trippa (2014). "Improving prediction from Dirichlet process mixtures via enrichment". In: *The Journal of Machine Learning Research* 15.1, pp. 1041–1071.

Appendix

Legendre Polynomials

Useful results

The following is a useful fact for the derivation of some of the results of the main text. If $f(x)$ is a pdf and $f = \left(\sum_{j=1}^{m} a_j P_j(x)\right)^2$ then $\sum_{j=1}^{m} a_j^2$. The result holds also for x being a vector but we prove it for the univariate case to avoid cumbersome notation that does not add intuition.

Note that if

$$f = \left(\sum_{j=1}^{m} a_j P_j(x)\right)^2$$

then

$$\begin{aligned} f &= \left(\sum_{j=1}^{m} a_j P_j(x)\right)^2 \\ &= \sum_{j=1}^{m} \left(a_j P_j(x)\right)^2 + 2\sum_{j<i}^{m} a_j P_j(x) a_i P_i(x) \end{aligned}$$

we can then take the integral on the left and right hand side and get

$$\begin{aligned} \int f(x)dx &= \int \left(\sum_{j=1}^{m} \left(a_j P_j(x)\right)^2 + 2\sum_{j<i}^{m} a_j P_j(x) a_i P_i(x)\right) dx \\ &= \sum_{j=1}^{m} \int \left(a_j P_j(x)\right)^2 dx + 2\sum_{j<i}^{m} \int a_j P_j(x) a_i P_i(x) dx \end{aligned}$$

and hence, by the definition of Legendre polynomials

$$\int f(x)dx = \sum_{j=1}^{m} \left(a_j\right)^2 = 1$$

as we wanted to show.

The following result can be found in Hesthaven, S. Gottlieb, and D. Gottlieb (2007) and proves that the condition we are interested in on the approximation rate of the orthogonal polynomials is true

$$\|u - \mathcal{P}_N u\|_{L^2_w[-1,1]} \leq C N^{-p} \|u\|_{H^p_w[-1,1]}$$

the result should be generalizable along the same lines to a multivariate setting without any particular complications.

Derivation of univariate Legendre series coefficients

We know we can represent any function $f(x)$ as

$$f(x) = \sum_{i=1}^{\infty} a_i P_i(x)$$

we are then interested in what form these coefficients a_i take and to do this we notice we can do the following.

We multiply each side by P_j and integrate between -1 and 1 getting hence

$$\int_{-1}^{1} P_j(x) f(x) dx = \int_{-1}^{1} P_j(x) \sum_{i=1}^{\infty} a_i P_i(x) dx$$

Thanks to the Fubini- Tonelli theorem we can the order of summation and integration. We then have

$$\int_{-1}^{1} P_j(x) f(x) dx = \sum_{i=1}^{\infty} a_i \int_{-1}^{1} P_j(x) P_i(x) dx$$

and hence

$$\int_{-1}^{1} P_j(x) f(x) dx = a_j \frac{2}{2j+1}$$

therefore we get that

$$a_j = \frac{2j+1}{2} \int_{-1}^{1} P_j(x) f(x) dx$$

Notice that if we assume that the polynomial is normalized

$$a_j = \int_{-1}^{1} P_j(x) f(x) dx$$

which implies that if $f(x)$ is a density $a_0 = 1$ and $a_1 = E[x]$.

Derivation of multivariate Legendre series coefficients

We know that we can represent any multivariate function $f(\mathbf{x})$ as

$$f(\mathbf{x}) = \sum_{k}^{\infty} a_k \prod_{j=1}^{d} P_k(x_j)$$

with $k = (k_1, \ldots, k_d)$. Let us now retrace the same steps as the univariate case. We multiply by $\prod_{j=1}^{d} P_l(x_j)$ with l being a specific combination of degrees $l = (l_1, ..., l_d)$ and integrate by $[-1, 1]^d$ and I believe I should get something like

$$\int_{-1}^{1}...\int_{-1}^{1}\prod_{j=1}^{d}P_l(x_j)f(\mathbf{x})dx_1...dx_d=\int_{-1}^{1}...\int_{-1}^{1}\prod_{j=1}^{d}P_l(x_j)\sum_{k}^{\infty}a_k\prod_{j=1}^{d}P_k(x_j)dx_1...dx_d$$

We apply Fubini-Tonelli and again get

$$\int_{-1}^{1}...\int_{-1}^{1}\prod_{j=1}^{d}P_l(x_j)f(\mathbf{x})dx_1...dx_d=\sum_{k}^{\infty}a_k\int_{-1}^{1}...\int_{-1}^{1}\prod_{j=1}^{d}P_l(x_j)\prod_{j=1}^{d}P_k(x_j)dx_1...dx_d$$

and therefore

$$a_k=\left(\prod_{j=1}^{d}\frac{2l_j+1}{2}\right)\int_{-1}^{1}...\int_{-1}^{1}\prod_{j=1}^{d}P_l(x_j)f(\mathbf{x})dx_1...dx_d$$

Clearly these coefficients are computationally intensive to calculate actually above a certain number of dimension but this expression should allow for bounding as in the results presented previously.