TR:UPM-ETSIINF/DIA/2014-2

Regression models with MoPs Bayesian networks

Gherardo Varando GHERARDO.VARANDO@UPM.ES

Concha Bielza MCBIELZA@FI.UPM.ES

Pedro Larrañaga PEDRO.LARRANAGA@FI.UPM.ES

Departamento de Inteligencia Artificial
Universidad Politecnica de Madrid
Campus de Montegancedo, s/n
28660 Boadilla del Monte, Madrid, Spain

Abstract

We present a model of Bayesian network for continuous variables, where densities and conditional densities are estimated with B-spline MoPs. We use a novel approach to directly obtain conditional densities estimation using B-spline properties. In particular we implement naive Bayes and wrapper variables selection. Finally we apply our techniques to the problem of predicting neurons morphological variables from electrophysiological ones.

Keywords: Bayesian networks, regression, MoP, conditional density estimation, non parametric.

1. B-spline

B-splines or basis splines (Schoenberg, 1946) are polynomial curves that form a basis for the space of piecewise polynomial functions (Faux and Pratt, 1979) over a closed domain $\Omega_X = [\epsilon_X, \xi_X] \subset \mathbb{R}$. Given an increasing knot sequence (or split points) of $L_X + 1$ real numbers $\boldsymbol{\delta} = \{a_0, a_1, \ldots, a_{L_X}\}$ in the approximation domain $\Omega = [\epsilon, \xi]$ with $a_{i-1} < a_i$, $\epsilon = a_0$ and $\xi = a_L$, one can define $M = L + r - 1$ different B-splines with order r spanning the whole domain Ω. The jth B-spline $B^r_{X,j}(x), j = 1, \ldots, M$, is

$$B^r_j(x) = (a_j - a_{j-r}) H(x - a_{j-r}) \sum_{t=0}^{r} \frac{(a_{j-r+t} - x)^{r-1} H(a_{j-r+t} - x)}{w'_{j-r}(a_{j-r+t})}, \quad x \in \Omega, \tag{1}$$

where $w'_{j_X - r_X}(x)$ is the first derivative of $w_{j_X - r_X}(x) = \prod_{u=0}^{r_X} (x - a_{j_X - r_X + u})$, $a_k = a_0$ for every $k < 0$, $a_k = a_L$ for every $k > L$ and $H(x)$ is the Heaviside function

$$H(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

B-splines form a basis of piecewise polynomial function of order r and thus every piecewise polynomial ϕ over Ω, which is differentiable in (ϵ, ξ), could be expressed as,

$$\phi(x) = \sum_{i=1}^{L+r-1} \alpha_j B^r_j(x).$$

We have some useful properties,

1. ϕ is differentiable in (ϵ, ξ), and continue in $[\epsilon, \xi]$. (*regularity*)
2. If $\alpha_j >= 0$ for every $j = 1, \dots, L + r - 1$ then $\phi(x) >= 0$. (*positivity*)
3. $\phi(x) = 0$ for every $x \notin [\epsilon, \xi]$. (*compact support*)
4. $\phi(x) = 1$ if and only if $\alpha_j = 1$ for every $j = 1, \dots, L + r - 1$. (*partition of unity*)
5. $\int_{\mathbb{R}} B^r_j(x) = \frac{a_j - a_{j-r}}{r}$.

We can extend the above definitions for piecewise n-variate polynomial. In particular, for r_i the order of the B-splines, and $L_i + 1$ the number of knots for the i-th dimension, we have

$$\phi(\mathbf{x}) = \sum_{\substack{j_1=1,\dots,M_1 \\ \vdots \\ j_n=1,\dots,M_n}} \alpha_{j_1,\dots,j_n} \prod_{i=1}^{n} B^{r_i}_{X_i,j_i}(x_i), \quad \mathbf{x} \in \Omega_1 \times \cdots \times \Omega_n,$$

where $M_i = L_i + r_i - 1$, is a n-variate piecewise polynomial over hyper-rectangular pieces, defined by the Cartesian product of the knots sequences. As the univariate case, we have the follow properties,

1. ϕ is differentiable in $(\epsilon_1, \xi_1) \times \cdots \times (\epsilon_n, \xi_n)$, and continue in its closure.
2. If $\alpha_{j_1,\dots,j_n} >= 0$ for every $j_1, \dots, j_n$ then $\phi(\mathbf{x}) >= 0$.
3. $\phi(\mathbf{x}) = 0$ for every $\mathbf{x} \notin [\epsilon_1, \xi_1] \times \cdots \times [\epsilon_n, \xi_n]$.

2. Densities and Conditional Densities Estimation

In this section we expose methods for densities and conditional densities estimations, based on B-spline MoPs.

2.1 Densitiy

López-Cruz et al. (2013) developed a method, which is an extension of Zong (2006); Zong and Lam (1998), for the estimation of multivariate densities with B-spline, in particular, given knots sequences and orders for each dimension, the algorithm find the coefficients of the B-spline representation that maximise the likelihood (MLE MoP). We moreover observe that the given method outputs proper densities, that is, integrate to one and are non-negative. We present now a simple heuristic search strategy (Varando et al., 2014) to find knots sequences and orders for every dimension, for a given dataset $\mathcal{D}$ of observations for the random vectors $(X_1, \dots, X_n)$ with density $f(x_1, \dots, x_n)$. Algorithm 1 is a simple greedy search over the space of the parameters, we start from the B-spline MoP with order 2 and 2 knots in every dimension and we gradually increase the order or the number of knots in one of the dimensions, selecting at every step the solution that minimize the BIC score.

Algorithm 1: MoP search algorithm

```
Data: $\mathcal{D}$ a dataset of $(X_1, \dots, X_n)$ observations
Result: $\phi_f$ MoP B-spline approximation of $f$
$\epsilon = \min(\mathcal{D})$;
$\xi = \max(\mathcal{D})$;
$\phi$ =MLE MoP with orders 2 and 2 knots for every dimension over
$[\epsilon_1, \xi_1] \times \dots \times [\epsilon_n, \xi_n]$;
$score = AIC(\phi, \mathcal{D})$;
$score_{new} = score$;
while $score_{new} \leq score$ do
    $score_{new} = score$;
    $\phi = \phi_{new}$ ;
    for $i = 1, \dots, n$ do
        $\phi_{i,1}$ =MLE MoP with $i$-th order increased by one ;
        $\phi_{i,2}$ =MLE MoP with $i$-th number of knots increased by one ;
        $score_{i,j} = BIC(\phi_{i,j}, \mathcal{D})$ ;
    end
    $score_{new} = \min(score_{i,j})$;
    $\phi_{new}$ = MoP corresponding to $score_{new}$;
end
return $\phi_f = \phi$;
```

2.2 Conditional Density

We consider now a conditional density $g(x|\mathbf{y})$, that is a function of $(x, \mathbf{y}) = (x, y_1, \dots, y_m)$ such that,

$$\int_{\mathbb{R}} g(x|\mathbf{y}) dx = 1 \quad \forall \mathbf{y} \in \mathbb{R}^m. \tag{2}$$

In Varando et al. (2014) we present two algorithms for estimating conditional densities as g, those algorithm use the method of López-Cruz et al. (2013) for multivariate densities (with an heuristic search as in Algorithm 1) combined with a conditional sampling or a Lagrange basis interpolation technique to obtain a MoP approximation of $g(x|\mathbf{y})$. The problems of the algorithms in Varando et al. (2014) are that:

- They perform a lengthy two-steps procedure: estimation of the MoP of the joint probability, estimation of the MoP of the conditional density.
- One of the two methods (Lagrange interpolation) outputs general MoPs and not B-spline MoPs, they are not continuous, have an huge numbers of parameters, are not proper conditional densities (Equation 2) and do not have theoretical properties (as consistency, MLE, non-negative, integrate to one).
- The method that perform a conditional sampling and then learn a B-spline MoP is computationally costly and dose not provide a proper conditional density, that is a function that satisfies Equation 2.

Observing that the above problems could be a huge drawback in building Bayesian network regression models we present now a novel approach to directly find, given knots sequences, $\boldsymbol{\delta}_i$ and orders r_i for each dimension, the B-spline MoP $\phi(x|\mathbf{y})$ that maximise the conditional likelihood, that is

$$\mathcal{CL}(\phi, \mathcal{D}) = \sum_{(x,\mathbf{y}) \in \mathcal{D}} \phi(x|\mathbf{y}).$$

The algorithm is an adaptation of the original algorithm of Zong (2006) for finding the MLE of the B-spline coefficients. For the sake of simplicity we present now the algorithm in the bivariate case, that is for estimating a conditional density $f(x|y)$ with $x, y \in \mathbb{R}$, the extension to $f(x|\mathbf{y})$ $x \in \mathbb{R}$ and $\mathbf{y} \in \mathbf{R}^m$ is obvious.

Consider a dataset of N i.i.d. observations $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$ of (X, Y) with density $f(x, y)$, suppose that for every $y \in \Omega_Y = [\epsilon_Y, \xi_Y]$ there exist a conditional density of $X|Y$, that is $f(x|y)$ such that

$$\int_{\mathbb{R}} f(x|y) dx = 1.$$

Given $\Omega_X = [\epsilon_X, \xi_X]$, knots sequences $\boldsymbol{\delta}_X$, $\boldsymbol{\delta}_Y$ and orders r_X, r_Y we want to find a bivariate B-spline MoP $\phi_{\{\alpha_{i,j}\}}(x|y)$ such that,

1. $\phi_{\{\alpha_{i,j}\}}(x|y) = \sum_{i=1}^{L_X+r_X-1} \sum_{j=1}^{L_Y+r_Y-1} \alpha_{i,j} B_i^{r_X}(x) B_j^{r_Y}(y)$

2. $\int_{\Omega_X} \phi(x|y) dx = 1 \quad \forall y \in \Omega_Y$

3. $\phi(x|y) \geq 0$

4. $\phi = \arg\max \mathcal{CL}(\phi, \mathcal{D})$

As in Zong (2006) we convert conditions (2), (3) above for the coefficients $\alpha_{i,j}$

- $\sum_{i=1}^{L_X+r_X-1} \alpha_{i,j} \int B_i^{r_X} = 1 \quad \forall j = 1, \ldots, L_Y + r_Y - 1$

- $\alpha_{i,j} \geq 0 \quad \forall i = 1, \ldots, L_X + r_x - 1$ and $j = 1, \ldots, L_y + r_Y - 1$

And we obtain an iterative methods that converges to the set of coefficients that maximise the conditional likelihood.

```
Algorithm 2: Maximum Conditional Estimation
  Data: $\mathcal{D}$ a dataset of $(X, Y)$ observations, orders $r_X$, $r_Y$ and knots sequence $\boldsymbol{\delta}_X$, $\boldsymbol{\delta}_Y$
  Result: $\{\alpha_{i,j}\}$ the set of B-spline coefficients correspondent to the maximum
          conditional likelihood model
  $\alpha_{i,j} = \frac{1}{\sum_{i=1}^{L_X+r_X-1} \int B_i^{r_X}}$ for every $i, j$ ;
  $LCL = \log(\mathcal{CL}(\phi_{\{\alpha_{i,j}\}}, \mathcal{D})$
  while $(LCL_{new} - LCL > toll)$ do
      $\alpha_{i,j}^{new} = \sum_{(x,y)\in\mathcal{D}} \frac{B_i^{r_X}(x) B_j^{r_Y}(y)}{\phi_{\{\alpha_{i,j}\}}(x|y)}$ ;
      $\alpha_{i,j} = \frac{\alpha_{i,j}^{new}}{\sum_{i=1}^{L_X+r_X-1} \alpha_{i,j} \int B_i^{r_X}}$;
      $LCL_n ew = \log(\mathcal{CL}(\phi_{\{\alpha_{i,j}\}}, \mathcal{D})$ ;
  end
  return $\{\alpha_{i,j}\}$ ;
```

We observe that Algorithm 2 is exactly as the algorithm for multivariate MoP's in López-Cruz et al. (2013), with a different kind of normalization that ensures the conditional density properties.

We can then apply the same heuristic search as in Algorithm 1 to search for the set of knots and value for the orders parameters that maximize the BIC score.

To resume we now have two algorithm, implemented as two functions in R, that computes joint density and conditional density B-spline MoP estimation, maximising the BIC score (AIC score could be used too). We will denote those functions as *search.mop* and *condtionialSearch.mop*. We observe that the learned (conditional) B-spline MoP densities are actual (conditional) densities, i.e. they are non-negative and integrate to one (for every conditioning value).

3. B-spline Mops Bayesian network

We define a framework for define, computing inference and learning structure for non parametric Bayesian network based on B-spline MoPs densities and conditional densities.

A B-spline MoPs Bayesian network (BMoP-BN) is a Bayesian network such that densities and conditional densities are specified with B-spline MoPs.

So given a defined BN structure $\mathcal{N}$ over some variables $X_1, \ldots, X_n$, fitting the corresponding BMoP-BN to a dataset $\mathcal{D}$ is equivalent to computing with functions *searc.mop* or *conditionalSearch.mop* the densities ad conditional densities specified by the BN structure $\mathcal{N}$. We will refer to this operation as *fit.bmopbn*$(\mathcal{N}, \mathcal{D})$.

3.1 Inference

To compute posterior density we just multiply each densities and conditional densities substituting evidence when given, the result is a function of the variables with no evidence. We can now compute posterior mean by numerical integration or by sampling, and mode by numerical optimization.

Those methods for computing posterior density, mean and mode are very naive and do not take advantage of the B-spline MoP form of the densities. To make the most of the particular type of densities we are using we should use an algorithm that multiply B-spline MoPs and obtains the result as a B-spline MoP, moreover using an exact method for the integration and/or for deriving the posterior density in B-spline form would implies a gain in computation complexity and in accuracy of the results.

4. Regression Models

We use now BMoP-BN to perform regression, we present different models based on BMoP-BN. Given a target variable Y and features $X_1, \dots, X_n$ we consider the following methods.

BMoP-NB we consider a Naive Bayes (NB) structure, that is the BMoP-BN with arcs from Y to every X_i.

BMoP-NBWR The same as BMoP-NB but with features selection via a wrapper approach. We start with a NB without features and at each step we try every possible BMoP-NB with one more feature, the one that perform best in term of MSE (estimated with k-fold cross validation) is selected.

Where MSE is the mean square error defined as

$$MSE = \frac{1}{|\mathcal{D}|} \sum_{(y,\mathbf{x}) \in \mathcal{D}} (y - \hat{y}(\mathbf{x}))^2$$

and *rmse* its root.

We implemented the estimation of the predicted values both with posterior mean, and with posterior mode.

5. Electro-Morphological Regression

5.1 Problem Description

Connecting Electro physiological and Morphological variables is a common problem in Neuroscience (Connors and Regehr, 1996). In particular interested is been devoted to connect morphology and functionality of single neurons (Torben-Nielsen et al., 2007; Maturana et al., 2014) or circuit of neurons (Shepherd et al., 2005).

Recent advances in searching the relationship between morphological and electro physiological variables of single neurons focus on *models* of neurons (multicompartment models in Maturana et al. (2014) and L-system models in Torben-Nielsen et al. (2007)) and try to replicate given fire-patterns with different possible models of neurons, using for example NEURON software for simulating the electro-response.

We would like to study, instead, directly the relationship between electro physiological and morphological variables, moreover in contrast to the state of the art, we try to predict morphological variables starting from electro physiological ones.

5.2 Results on Neurological Datatset

Datasets The raw dataset consist in 48 observation of 32 electrophysiological variables (E) and 46 morphological variables. The dataset contains missing data and constant variables that we remove. We split the morphological variables in two subsets, axon's variables (A) and basal/soma variables (B). For every set of variables (A, B or E) we select a subset of observation avoiding the missed data, respectively O_A, O_B and O_E. We have that

$$|A| = 22$$
$$|B| = 24$$
$$|E| = 20$$

We then build two datasets, one with $E \cup A$ variables ($data1$) and one with $E \cup B$ variables ($data2$). Every dataset has no missing value that is, $data1$ for example contains observation in $O_A \cap O_B$. The dataset $data1$ has 39 observations and $data2$ has 44 observations. In this document we report results just for the axonal variables.

Numerical Results We list now some evaluations, we compare our method to some state of the art regression models as K-NN (our simple R implementation), and M5'(WEKA), DecisionStump (WEKA), Linear Regression with attributes selection (as in WEKA, default options). We also report results for the constant regression model (Const) that predict the mean of the training values. As the dataset has few observation we perform leave-one-out cross validation.

	Axon_Term_ Mean_Branch _Length	Axon_Max_ Path_Length	Axon_total _volume	Axon_total _Surface _Area
3-NN	66.34	568	841	8307
5-NN	64.46	552	825	8285
10-NN	57.38	509	801	8172
20-NN weighted	56.61	489	779	7800
M5'	58.17	527	833	8431
DecisionStump	47.97	548	972	9550
Linear Regression	47.92	643	839	8522
Const	55.58	**485**	766	**7555**
BMoP-NB	52.41	658	**745**	8087
BMoP-NBWR	**44.48**	554	818	8256

Table 1: Root mean squared error, computed with leave-one-out cross validation. For every column the best result is marked in boldface.

From the results in Table 1 we observe that BMoPs models are able to achieve slightly better results than the other regression models, at the cost of a large increase in computational complexity (times of model's learning are not even comparable).

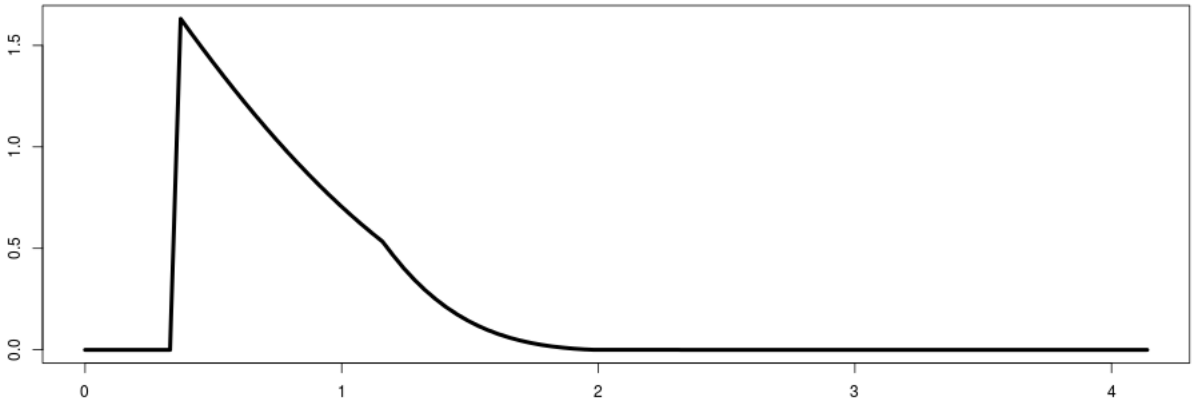

Figure 1: Posterior density for Axon_Term_ Mean_Branch_Length

5.3 Artificial Dataset

Consider a dataset with 50 observations of five independent variables $X_1, \ldots, X_5$ distributed as standardized Normal distributions ($mean = 0$ and $sd = 1$). We analyse our BMoP models over those variables, we perform the regression of $C = \sum_{i=1}^{5} X_i + \epsilon$ and of $Q = \sum_{i=1}^{5} X_1^2 + \epsilon$ where ϵ is some *noise*, Normal distributed with mean 0 and standard deviation 0.001. Results are reported in Table 2.

	C	Q
M5'	**0.001**	3.25
Linear Regression	**0.001**	3.1
Const	1.93	2.93
BMoP-NB	1.53	**2.66**

Table 2: Root mean squared error, computed with leave-one-out cross validation. For every column the best result is marked in boldface.

From the results on the artificial dataset we see how B-MoPs models seems able to deal with regression of non-linear functions. In the case of linear relations, we observe how M5' and Linear regression obviously obtain very good results. In particular M5' and Linear regression with just 50 observations are able to obtain the optimal errors (empirical error=theoretical errors). B-MoPs obtains results that are slightly better than the Constant regression both in non-linear as in linear case.

Acknowledgments

This research has been partially supported by the Spanish Ministry of Economy and Competitiveness through Cajal Blue Brain (C080020-09). The authors thankfully acknowledges the computer resources, technical expertise and assistance provided by the Supercomputing and Visualization Center of Madrid (CeSViMa).

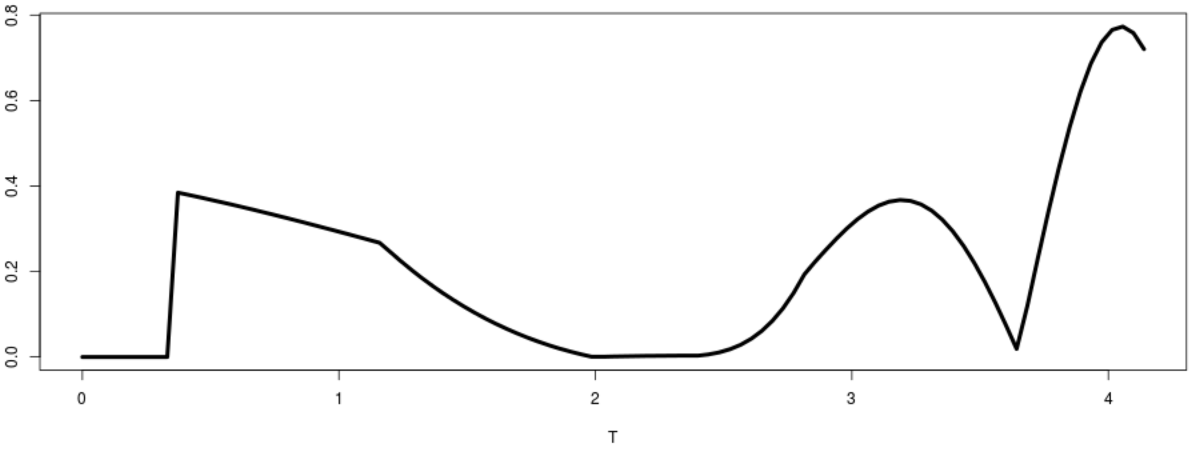

Figure 2: Posterior density for Axon_Term_ Mean_Branch_Length

References

Barry W Connors and Wade G Regehr. Neuronal firing: Does function follow form? *Current Biology*, 6(12):1560 – 1562, 1996. ISSN 0960-9822.

I.D. Faux and M.J. Pratt. *Computational Geometry for Design and Manufacture.* Wiley, 1979.

Pedro L. López-Cruz, Concha Bielza, and Pedro Larrañaga. Learning mixtures of polynomials of multidimensional probability densities from data using B-spline interpolation. *International Journal of Approximate Reasoning*, page submitted, 2013.

Matias I. Maturana, Tatiana Kameneva, AnthonyN. Burkitt, Hamish Meffin, and DavidB. Grayden. The effect of morphology upon electrophysiological responses of retinal ganglion cells: simulation results. *Journal of Computational Neuroscience*, 36(2):157–175, 2014. ISSN 0929-5313.

I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. Part A: On the problem of smoothing of graduation. A first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4:45–99, 1946.

Gordon Shepherd, Armen Stepanyants, Ingrid Bureau, Dmitri Chklovskii, and Karel Svoboda. Geometric and functional organization of cortical circuits. *Nat Neurosci*, 8:782 – 790, 2005.

Ben Torben-Nielsen, Karl Tuyls, and Eric O. Postma. On the neuronal morphology-function relationship: A synthetic approach. In Karl Tuyls, Ronald Westra, Yvan Saeys, and Ann Now, editors, *Knowledge Discovery and Emergent Complexity in Bioinformatics*, volume 4366 of *Lecture Notes in Computer Science*, pages 131–144. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-71036-3.

G. Varando, P.L. Lopez-Cruz, T. Nielsen, P. Larrañaga, and C. Bielza. Conditional density approximations with mixtures of polynomials. *International Journal of Intelligent Systems*, (accepted), 2014.

Z. Zong. *Information-Theoretic Methods for Estimating Complicated Probability Distributions*. Elsevier, 2006.

Z. Zong and K.Y. Lam. Estimation of complicated distributions using B-spline functions. *Structural Safety*, 20(4):341–355, 1998.