Multiresolution dictionary learning for conditional distributions

Francesca Petralia, Joshua T. Vogelstein and David B. Dunson

Department of Statistical Science, Duke University, Durham, NC

Summary. Nonparametric estimation of the conditional distribution of a response given high-dimensional features is a challenging problem. In many settings it is important to allow not only the mean but also the variance and shape of the response density to change flexibly with features, which are massive-dimensional with a distribution concentrated near a lower-dimensional subspace or manifold. We propose a novel multiresolution approach which starts by learning a multiscale dictionary of densities, constructed as Gaussian within each set of multiscale partition tree for the features. We place a novel multiresolution stick-breaking process prior on the dictionary weights to construct a conditionally Bayesian approach. The algorithm scales efficiently to massive numbers of features, and can be implemented efficiently with slice sampling. State of the art predictive performance is demonstrated for toy examples and an important application to predicting drug response to anti-depressants.

Keywords: Density regression; Dictionary learning, Manifold learning; Mixture of experts; Multiresolution stick-breaking; Nonparametric Bayes

1. Introduction

2. Model

2.1 Model Structure

Let $x_i = (x_{i1}, \dots, x_{ip})' \in \mathcal{X}$ denote a vector of predictors, with p large, and let $y_i \in \mathcal{Y}$ denote a response variable. We are interested in estimating the conditional density $f(y|x)$ of the response given predictors, while taking into account that the predictors are high-dimensional but thought to be concentrated near a lower-dimensional subspace $\mathcal{M}$ embedded in $\Re$ that may or may not correspond to a Riemannian manifold. There is a literature on obtaining multiscale representations of such subspaces.

Suppose we define a multiscale partition of $\mathcal{X}$. Generation one corresponds to the entire $\mathcal{X}$. At generation two, $\mathcal{X}$ is split into two mutually exclusive partition sets, $\mathcal{X}_1$ and $\mathcal{X}_2$ with $\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2$ and $\mathcal{X}_1 \bigcap \mathcal{X}_2 = \emptyset$. At each subsequent generation, each set is further refined into two children. For example, in generation $j = 3$, $\mathcal{X}_1$ is split into $\mathcal{X}_{11}$ and $\mathcal{X}_{12}$ while $\mathcal{X}_2$ is split into $\mathcal{X}_{21}$ and $\mathcal{X}_{22}$. This proceeds for k generations.

Any predictor value $x \in \mathcal{X}$ has a corresponding ancestral history $A(x) \in \{1,2\}^k$ encoding their location in the partition tree, with $A_j(x) \in \{1,2\}^j$ denoting the elements up to generation j and the first element equal to 1 by definition. We characterize the conditional density $f(y|x)$ as a convex combination of multiscale *dictionary* densities. At level (generation) one, the global parent density is denoted f_1. In generation two, the dictionary densities corresponding to partition sets $\mathcal{X}_1$ and $\mathcal{X}_2$ are denoted f_{11} and f_{12}. For $j = 2, \ldots, k$, the dictionary density at generation j for ancestry $a_j \in \{1,2\}^j$ is f_{a_j}. The resulting conditional density is characterized as

$$f(y|x) = \sum_{j=1}^{k} \pi_{A_j(x)} f_{A_j(x)}(y), \tag{1}$$

where $0 \leq \pi_{a_j} \leq 1$ and $\sum_{j=1}^{k} \pi_{a_j} = 1$ for any $a \in \{1,2\}^k$ with a_j the first j elements of a.

Each $A_j(x)$ is a $j \times 1$ binary vector encoding the path through the partition tree up to generation j specific to predictor value x. For two predictor values x and x' located close together, it is expected that the paths will be similar, which leads to similar weights on the dictionary densities. In the extreme case in which x and x' belong to the same leaf partition set, we have $A(x) = A(x')$ and the path through the tree will be the same. In this case, we also will have $f(y|x) = f(y|x')$ so that up to k levels of resolution the densities $f(y|x)$ and $f(y|x')$ are identical. If the paths through the tree differ only in the final generation or two, the weights will typically be similar but the resulting conditional densities will not be identical.

2.2 Two-Stage Approach & Treed Stick-Breaking

Although we will consider fully Bayesian approaches, to start with we consider a simple two stage approach. In the first stage, apply some existing multilevel partitioning approach, such as those proposed by Mauro & gang, to obtain a multilevel partition that we will then consider as fixed. In addition, for $j = 1, \ldots, k$ and $a \in \mathcal{A}_j$ we separately estimate each of the dictionary densities by either just assuming a normal form and using maximum likelihood or Bayesian MAP estimation or using frequentist kernel smoothing. This is done by taking the data for all those subjects having predictors in that partition set and estimating the dictionary density based on only these data. To be specific, for estimating density $f_{a_j}(y)$, we use the data $\{y_i : x_i \in \mathcal{X}_{a_j}\}$, noting that $a_j \in \{1,2\}^j$ is a j-dimensional binary vector. This is certainly very fast and can be implemented in parallel for the different partition sets.

We then conduct the analysis treating the dictionary elements and partition sets as fixed and placing a prior on the weights $\pi_a^{(j)}$ for all $a \in \{1,2\}^k$. A natural choice corresponds to a multiscale stick-breaking process that is defined as follows. For each node in the binary partition tree including the root node, define a stick length, $V_{a_j} \sim \text{beta}(1, \alpha)$, for $a_1 = 1$ and $a_j \in \{1,2\}^j, j = 2, \ldots, k$. Potentially we can consider a tree that goes to infinite depth and then truncate as an approximation but first we will just consider finite depth. The parameter α encodes the complexity of the model, with $\alpha = 0$ corresponding to the case in which $f(y|x) = f_q(y)$. We relate the weights in (1) to the stick-breaking random variables as follows:

$$\pi_{a_j} = V_{a_j} \prod_{h=1}^{j-1} (1 - V_{a_h}), \quad j = 1, \ldots, k-1, \tag{2}$$

with $V_a = 1$ for the leaf stick-breaking random variables to ensure that $\sum_{j=1}^k \pi_{a_j} = 1$ for any $a \in \{1,2\}^k$ with $a_1 = 1$ and a_j denoting the first j elements of a.

3. Estimation

Can we do data augmentation Gibbs sampling very fast under this model? We introduce $S_i \in \{1, \ldots, k\}$ for $i = 1, \ldots, n$ denoting the level of the dictionary density that subject i uses. Then, we can run a Gibbs sampler that simply lets

1. Update S_i by sampling from the multinomial conditional posterior with

$$\Pr(S_i = j \,|\, -) = \frac{\pi_{A_j(x_i)} f_{A_j(x_i)}(y_i)}{\sum_{h=1}^k \pi_{A_h(x_i)} f_{A_h(x_i)}(y_i)}, \quad h = 1, \ldots, k. \tag{3}$$

2. Update stick-breaking random variable V_{a_j} from the beta full conditional

3. Update α from its gamma full conditional

4. In addition, we can allow the dictionary elements to be unknown by placing normal inverse-gamma priors on the mean and precision and updating these from the conditionally conjugate prior, though we initially fix the dictionary densities.

It seems there should be some simple analysis that can be done motivating the above specification in terms of mean square error and the bias-variance tradeoff. As we weight the rougher scale dictionary elements more heavily we reduce variance because there is much more data available about these densities, while as we weight the fine scale dictionary elements more we decrease bias at the expense of variance. Given the independence assumptions conditionally on the multiscale partition, it would seem some math analysis of MSE and the tradeoff is possible.

Also, intuitively the above structure should do well in providing a type of multiscale ensemble method with the Bayesian nonparametric prior for the weights providing a Bayesian way to obtain weights on the ensemble, though of course the method is certainly not fully Bayesian. It seems we can make it fast and apply to big problems involving tons of predictors and potentially obtain a decent gauge of uncertainty through random weights and possibly dictionary elements, while keeping the multiscale partition fixed at first for tractability. Perhaps mainly useful as a black box for prediction but I would suspect some inferences could be back out in terms of impact of specific predictors, etc.

References

Chung, Y, and Dunson, DB (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *JASA*, 104, 1646-1660.

Dunson, DB, and Park, J-H. (2007). Kernel stick-breaking processes. *Biometrika*, 95, 307-323.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, 138, 252-290.

Holmes, CC, Denison, DGT, Ray, S, and Mallick, BK (2005). Bayesian prediction via partitioning. *JCGS*, 14, 811-830.

Jara, A. and Hanson, T.E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika*, 98, 553-566.

Reich, BJ, Kalendra, E, Storlie, CB, Bondell, HD, Fuentes, M (2012). Variable selection for high dimensional Bayesian density estimation: Application to human exposure simulation. *Applied Statistics*, 61, 47-66.

Tokdar, S.T., Zhu, Y.M. and Ghosh, J.K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis*, 5, 319-344.