REGRESSION ANALYSIS THROUGH BAYESIAN CLUSTERING

Changku Kang\textsuperscript{1}, Subhashis Ghosal\textsuperscript{2}
Bank of Korea\textsuperscript{3}, North Carolina State University\textsuperscript{2}

Abstract

The article describes a method of estimating nonparametric regression function through Bayesian clustering. The basic working assumption in the underlying method is that the population is a union of several hidden subpopulations in each of which a different simple parametric regression regime is in force and the overall regression function arises as a result of superposition of these simple regression functions. A Bayesian clustering technique based on Dirichlet mixture process is used to identify clusters which correspond to samples from these hidden subpopulations. The clusters are formed automatically within a Markov chain Monte Carlo scheme arising from a Dirichlet mixture process prior for the density of the regressor variables. The number of components in the mixing distribution is thus treated as unknown allowing considerable flexibility in modeling. Within each cluster, we estimate model parameters by the standard least square method or some of its variations. Automatic model averaging takes care of the uncertainty in classifying a new observation to the obtained clusters. As opposed to most commonly used nonparametric regression estimates, which breaks up the sample locally, our method splits the sample into a number of subgroups which is free of the dimension of the regressor variable. Thus our method avoids the curse of dimensionality problem. Through extensive simulations, we compare the performance of our proposed method with that of commonly used nonparametric regression techniques. We conclude that when the model assumption holds and the subpopulation are not highly overlapping, our method has smaller estimation error particularly if the dimension is relatively large.

KEY WORDS: Clustering, curse of dimensionality, Dirichlet mixture, Markov chain Monte Carlo, regression.

1. Introduction

Consider the problem of estimating a regression function \( m(x) = \mathbb{E}(Y|X = x) \) based on sampled data \((X_1, Y_1), \ldots, (X_n, Y_n)\), where \( X \) is \( d \)-dimensional. Typically \( m(x) \) is estimated under the homoscedastic signal plus noise model

\[
Y_i = m(X_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d., } \mathbb{E}(\varepsilon) = 0, \quad \text{var} (\varepsilon) = \sigma^2, \quad (1)
\]

where \( m(x) \) has some parametric form or is estimated nonparametrically based only on smoothness assumptions. Common smoothing methods include the nearest neighbors, kernel, local linearity, orthogonal series or spline; see e.g., Efrosimovich (1999). These methods are extremely powerful in lower dimension. In higher dimension, because observations are relatively sparsely distributed over the space, one typically needs to assume additional structure in the regression function. Generalized additive models (GAM) (Hastie and Tibshirani (1990)), classification and regression trees (CART) (Breiman et al. (1984)) and multivariate adaptive regression splines (MARS) (Friedman (1991)) are among the popular estimation methods in higher dimension.

Bayesian methods for estimating a regression function has been developed by putting priors directly on the regression function. A popular method of assigning a prior on functions is by means of a Gaussian process. Wahba (1978) considered an integrated Wiener process prior for \( m(x) \). The resulting Bayes estimate, in the sense of a noninformative limit, is a smoothing spline. Priors based on series expansion with normally distributed coefficients, which are also Gaussian priors, have been investigated well for various choices of basis functions. Müller, Erkanli and West (1996) used Dirichlet mixture of normal on joint density of \((X, Y)\) to estimate the conditional mean. The method suffers heavily from curse of dimensionality in higher dimension.

Sometimes, however, \( X \) is sampled from a population which can be best described as a mixture of some hidden subpopulations. In this case, it is plausible that the regression function have different parameter values when the sample comes from different groups. This situation will typically arise if there is one unobservable categorical label variable which influences both \( X \) and \( Y \), but conditional on that missing label variable, the relationship between \( Y \) and \( X \) is very simple, such as linear, up to an additive error. However, as we do not observe the group labels, the overall regression function of \( Y \) on \( X \) is a weighted average of the individual regression functions, where the weights are given by the probabilities of the observation being classified into different groups given its value. In such a situation, an easy calculation (see the next section) shows that the model (1) fails. The mixture distribution arises naturally in biology, economics and other sciences. For example, in a marketing survey, suppose that consumers rate the quality of a product. Different consumers may give different weights to various factors depending on their background and mentality. In other words, the population actually consists of different subpopulations where different regression functions are in effect, but subpopulation membership is abstract and
is not observable. Had we observed the labels, regression analysis would have been straightforward. In the absence of the labels, we use the auxiliary measurements to impute the lost labels and use a simple regression analysis within each hypothetical group. The number of groups and group membership labels are nuisance parameters in the analysis, which should be integrated out with respect to their posterior distribution.

In this paper, we propose a method of estimating an unknown regression function by viewing it as a mixture of an unknown number of some simple regression functions associated with each subpopulation. These corresponding subgroups in the sample are estimated by identifying clusters in the data cloud. The clusters are automatically produced by a Dirichlet mixture model, a popular Bayesian nonparametric method. Standard regression methods are used in fitting a regression function within each cluster. Each iteration of the Markov chain Monte Carlo (MCMC) sampling based on the Dirichlet mixture process produces a partitioning of the data cloud and thus an estimate of the regression function. The final regression estimate is obtained by averaging out the estimate obtained in each iteration thus integrating out the nuisance label parameters with respect to their posterior distribution obtained from the Dirichlet mixture process prior. Thus our method may be viewed as an ensemble method like bagging [Breiman (1996)] and random forest [Breiman (2001)], but here the ensemble is produced by MCMC iterations rather than by bootstrap resamples.

The similarity and difference of our method with other nonparametric methods are interesting to note. Regression methods based on free-knot splines, kernels or nearest neighbors partition the space and fit some standard regression function in each part and combine these individual estimates. We also combine individual estimates based on different components of the sample. However, unlike other standard methods, we partition the observed samples rather than the space where they lie. Under the assumed condition that the samples come from a fixed (although unknown) number of groups, our method requires partitioning the sample in a number of groups that essentially remains fixed independently of the dimension of the space. In contrast, many standard nonparametric methods in higher dimension require splitting the space in a relatively high number of subregions fueling the demand for a larger sample size, thus leading to the curse of dimensionality problem. The simulations conducted in Section 3 confirm that our method performs better than standard nonparametric methods, particularly in higher dimension, provided that the mixture model representation is appropriate. In addition, we could get conservative pointwise credible bands very easily which are usually difficult with other methods. Another operational advantage of our method is that there is no need to choose any smoothing parameter.

There have been attempts to utilize clusters to partition data and estimate regression. The term clusterwise regression was first used by Spath (1979). DeSarbo and Cron (1988) proposed a conditional mixture maximum likelihood methodology to identify clusters and partition the sample before applying linear regression in each piece. The method estimates the parameters of the mixture distribution by maximum likelihood using the EM algorithm. The number of components \( k \) in the mixture is chosen by the Akaike information criterion. Their method is closest in spirit to ours. However, they do not take into account the uncertainty in the number of groups and each group membership. Geweke and Keane (2005) estimated the probability of an observation with value \( x \) belonging to different subpopulations by a multinomial probit model. The main reason for using this model is that it produces simple and tractable posterior distributions within the Gibbs sampling algorithm. However, there is no room for a data driven choice of \( k \) and the method suffers from the curse of dimensionality problem in higher dimension.

By extensive simulation studies, we shall compare the performance of our method with the kernel and spline methods in one dimension. In higher dimension, we shall compare with the estimates obtained from GAM and MARS.

The paper is organized as follows. Section 2 provides detailed description of our method, including the MCMC algorithms and the construction of credible bands. Simulation results are presented and discussed in Section 3.

2. Method description

Suppose that we have \( k \) regression models

\[
Y_i = m_j(X_i) + \varepsilon_i, \quad \varepsilon_i \text{i.i.d., } \mathbb{E}(\varepsilon_i) = 0, \quad \text{var}(\varepsilon_i) = \tau_i^2, \quad (1)
\]

\( i = 1, \ldots, n, \quad j = 1, \ldots, k, \) corresponding to \( k \) subgroups in the populations. Let \( J \) be an unobserved random variable indicating the subgroup label. The prior distribution of \( J \) is given by \( \pi_j = P(J = j) \) and thus the posterior distribution given \( X = x \) is given by \( \pi_j(x) = P(J = j | X = x) \). Then, after eliminating the unobserved \( J \), the regression of \( Y \) on \( X \) is given by

\[
m(x) = \sum_{j=1}^{k} \pi_j(x)m_j(x). \quad (2)
\]

Also observe that the conditional variance \( \text{var}(Y | X = x) \) is given by

\[
\sum_{j=1}^{k} \tau_j^2 \pi_j(x) + \sum_{j=1}^{k} m_j^2(x) \pi_j(x) - \left( \sum_{j=1}^{k} m_j(x) \pi_j(x) \right)^2
\]

which is not a constant in \( x \) even if all \( \tau_j \)'s are equal. Hence the commonly assumed nonparametric regression model (1) fails to take care of the simple situation of missing labels.

The regression function \( m(x) \) given by (2) may be estimated by its posterior expectation given the observed data. This can be evaluated in two stages — in the first
Section on Nonparametric Statistics

stage we can obtain the posterior expectation given the hidden subpopulation labels and the observed data, while in the second stage we average the resulting quantity with respect to the conditional distribution of the labels given the data. The second stage averaging may be taken conditional on the X values only. For a discussion on how the Y observations may also be used to detect the hidden labels, see Van Aelst et al. (2005). Thus we need to model the joint distribution of the X and the missing subpopulation labels, which is done below. The first stage analysis reduces to k independent parametric regression problems. Therefore the posterior expectation of $m_j(x)$, $j = 1, \ldots, k$, may essentially be replaced by the corresponding least square estimates. Alternatively, one may view $m_j$s as hyperparameters and the resulting substitution corresponds to an empirical Bayes approach. While a fully Bayesian analysis at this stage is quite possible, the minor difference between the posterior expectations and the least square estimates, particularly when the sample size is not very small, makes the additional computing requirement unappealing. Henceforth we shall consider this hybrid approach of least square estimation in the first stage coupled with a posterior analysis of missing subpopulation labels.

The joint distribution of X and the missing label J will be specified somewhat indirectly, through another set of latent variables. The technique is based on one of the most popular methods in Bayesian nonparametrics, namely that of Dirichlet mixture (DM). Let $X_1, \ldots, X_n$ be i.i.d. from a density

$$f(x; G, \Sigma) = \int \phi(x; \theta, \Sigma) dG(\theta),$$

where $\phi(\cdot; \theta, \Sigma)$ is the density function of the d-dimensional normal distribution with mean $\theta$ and dispersion matrix $\Sigma$. Let G follow $DP(M, G_0)$, the Dirichlet process with precision parameter M and center measure $G_0$ in the sense of Ferguson (1973). The model can be equivalently written in terms of latent variables $\theta_1, \ldots, \theta_n$ as

$$X_i|\theta_i \sim N(\cdot; \theta_i, \Sigma), \theta_i|G \sim G, \ G \sim DP(M, G_0).$$

Because Dirichlet samples are discrete distributions, the Dirichlet mixture prior has the ability to automatically produce clusters. More precisely, we may identify $X_1$ and $X_j$ as the member of a common cluster if $\theta_1 = \theta_j$. Note that the last event occurs because the posterior distribution of $(\theta_1, \ldots, \theta_n)$ given $(X_1, \ldots, X_n)$ is given by the Polya urn scheme described by

$$\theta_i|{(\theta_{-i}, X_1, \ldots, X_n)} \sim \frac{\phi_0 dG_i(\theta_i) + \sum_{j \neq i} q_{ij} \delta_{\theta_j}(\theta_i)}{\phi_0 + \sum_{j \neq i} q_{ij}},$$

$$G_i(\theta) \propto \phi(X_i; \theta, \Sigma) dG_0(\theta),$$

$$\phi_0 \propto \int \phi(x; \theta, \Sigma) dG_0(\theta),$$

$$q_{ij} \propto \phi(X_i; \theta_j, \Sigma),$$

where $\phi_0 + \sum_{j \neq i} q_{ij} = 1$. Here and throughout, the subscript “−i” stands for all but i. This way a probability distribution on partitions of $\{1, 2, \ldots, n\}$ according to subpopulation membership is obtained, which in effect describes a joint distribution of X observations and the corresponding labels. In particular, the distribution of the number of hidden subpopulations is also described by the process.

Taking advantage of the many ties in $\theta_i$’s, a more efficient algorithm can be constructed based on a reparameterization $(k, \varphi_1, \ldots, \varphi_k, s_1, \ldots, s_n)$, where $k$ is the number of distinct $\theta_i$ values, $\varphi_1, \ldots, \varphi_k$ are the set of distinct $\theta_i$'s and $s_i = j$ if and only if $\theta_i = \varphi_j$. In fact, to identify the clusters, we need not even know the values of $\varphi_1, \ldots, \varphi_k$; it suffices to know the configuration vector $s = (s_1, \ldots, s_n)$. This will allow substantial reduction of computational complexity at the MCMC step. The clusters are given by $I_j = \{i : s_i = j\}$, $j = 1, \ldots, k$, and $H = \{I_1, \ldots, I_k\}$ is a partition of $I = \{1, \ldots, n\}$. Integrating out $\varphi_1, \ldots, \varphi_k$ in the above step, we obtain the following algorithm to generate the clusters through a Gibbs sampler when $M$, $G_0$ and $\Sigma$ are given: $P(s_i = j|s_{-i}, X_1, \ldots, X_n)$ is proportional to

$$n_{-i,j} \int \phi(X_i; \theta, \Sigma) dH_{-i,j}(\theta), \text{ if } j = 1, \ldots, k_{-i},$$

$$M \int \phi(X_i; \theta, \Sigma) dG_0(\theta), \text{ if } j = k_{-i} + 1,$$

where $H_{-i,j}$ is the posterior distribution based on the prior $G_0$ and all observations $X_i$ for which $l \in I_j$ and $l \neq i$, that is,

$$dH_{-i,j}(\theta) \propto \prod_{l \in I_j, l \neq i} \phi(X_l; \theta, \Sigma) dG_0(\theta),$$

$$n_{-i,j} = \#I_j \setminus \{i\}, \ j = 1, \ldots, k_{-i}, \ k_{-i} \text{ is the number of distinct elements in } \{\theta_j : \ j \neq i\} \text{ and } s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n).$$

Letting $G_0$ have density $\phi(\cdot; \xi, \Phi)$, it follows that

$$P(s_i = j|s_{-i}, X_1, \ldots, X_n) \propto M \phi(X_i; \xi, \Sigma + \Phi),$$

for $j = k_{-i} + 1$. For $j = 1, \ldots, k_{-i}$, the corresponding expression can be simplified using

$$\int \phi(X_i; \theta, \Sigma) dH_{-i,j}(\theta)$$

$$= \frac{\int \phi(X_i; \theta, \Sigma) \prod_{l \in I_j, l \neq i} \phi(X_l; \theta, \Sigma) dG_0(\theta) \phi(\theta; \xi, \Phi) d\theta}{\int \prod_{l \in I_j, l \neq i} \phi(X_l; \theta, \Sigma) dG_0(\theta) \phi(\theta; \xi, \Phi) d\theta}$$

$$= \frac{n_{i-j}}{n_{i-j} + 1} \phi(X_i; \bar{X}_{i-j}, \frac{1}{n_{i-j} + 1} \Sigma)$$

$$\times \frac{\phi(\xi; \bar{X}_{-i,j}, \frac{1}{n_{-i,j} + 1} \Sigma + \Phi)}{\phi(\xi; \bar{X}_{-i,j}, \frac{1}{n_{-i,j} + 1} \Sigma + \Phi)}.$$

where

$$\bar{X}_{-i,j} = \frac{1}{n_{-i,j}} \sum_{l \in I_j, l \neq i} X_l$$

$$\bar{X}_{i-j} = \frac{1}{n_{i-j} + 1} (\sum_{l \in I_j, l \neq i} X_l + X_i).$$

The above expressions are controlled by the hyperparameters $M$, $\xi$, $\Phi$ and $\Sigma$. In a Bayesian setting, it
Section on Nonparametric Statistics

is natural to put priors on them too and update during the MCMC stage. However, as we have integrated out \(\varphi_1, \ldots, \varphi_k\), it is no longer possible to update \(\xi, \Phi \) and \(\Sigma\). We take an empirical Bayes approach and estimate these parameters within an MCMC step by taking advantage of the knowledge of the partition. As \(\xi\) is the marginal expectation of \(X\), we estimate it by \(\hat{\xi} = \bar{X}\). Note that \(\Sigma\) is the dispersion matrix within group, and hence can be estimated by

\[
\hat{\Sigma} = n^{-1} \sum_{j=1}^{k} \sum_{t \in I_j} (X_t - \bar{X}_j)(X_t - \bar{X}_j)^T,
\]

\[
\bar{X}_j = n_j^{-1} \sum_{t \in I_j} X_t.
\]

The matrix \(\Phi\) stands for the between variation and hence can be estimated by

\[
\hat{\Phi} = n^{-1} \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^T.
\]

Clearly both \(\hat{\Sigma}\) and \(\hat{\Phi}\) are positive definite provided that \(n_j > 1\) for all \(j\). The precision parameter \(M\) can be updated by Gibbs sampling using a clever data augmentation technique [Escobar and West (1995), Section 6]. However, to reduce computational time, we shall work with a fixed moderate value of \(M\) such as 1 or 2.

It may be noted that each step of MCMC changes the membership of at most one observation. Since sample means and dispersions can be calculated recursively when an observation is added to or deleted from a group, it is clear that one can avoid calculation from the scratch everytime (6).

Also it follows that we could form location-scale mixtures in (3) by allowing \(\Sigma\) to vary in each group as \(\Sigma_j, \ j = 1, \ldots, k\). The same methodology will work if \(\Sigma_j\)'s are estimated separately, for instance, by \(\Sigma_j = n_j^{-1} \sum_{t \in I_j} (X_t - \bar{X}_j)(X_t - \bar{X}_j)^T\).

Once the clusters have been identified, we can use a standard regression method, such as the method of least squares for linear regression, within each cluster giving estimates \(\hat{m}_j(x)\) of regression functions \(m_j(x), j = 1, \ldots, k\).

Polynomial regression may be used if we suspect the lack of linearity in the model. Variants of the least square estimate such as the ridge regression estimate [Hoerl and Kennard (1988)] or the LASSO [Tibshirani (1996)] may be used in place of ordinary least squares to lower the variation in the estimate if there is a lot of multicollinearity in the data. The latter is particularly very useful in higher dimension where typically a variable selection is desirable before estimation. Further, as the clusters produced by the Dirichlet mixture will invariably differ from the true clusters, the use of a robust alternative to the least square estimate is desirable to prevent misclassified observations altering the estimates significantly. These robust estimates require some tuning and are generally harder to compute compared to the least squares estimate. For the sake of simplicity, we do not pursue these modifications in the present paper.

To complete the estimation of \(m(x)\), it remains to estimate \(\pi_j(x), j = 1, \ldots, k\). In order to do that, we consider a Bayesian model based on the given observations and the clusters. The prior probability of a new observation coming from the \(j\)th subgroup is given by the empirical fraction of observations falling in the group, that is \(n_j/n, j = 1, \ldots, k\). Under the empirical model, the \(j\)th subpopulation is assumed to be normal with center at \(\bar{X}_j\) and dispersion matrix \(\Sigma\). If the value of the new observation is \(x\), then the likelihood of the \(j\)th subgroup is \(\phi(x; \bar{X}_j, \Sigma)\). Applying the Bayes theorem, the posterior probability of \(x\) coming from the \(j\)th subpopulation, based on the empirical model, is proportional to \(n_j \phi(x; \bar{X}_j, \Sigma)\). To estimate the regression function at \(x\) given the clusters, one may do a “model averaging” with respect to the empirical posterior probabilities to yield the estimate

\[
\hat{\mu}(x) = \frac{\sum_{j=1}^{k} n_j \phi(x; \bar{X}_j, \Sigma) \hat{m}_j(x)}{\sum_{j=1}^{k} n_j \phi(x; \bar{X}_j, \Sigma)}.
\]

Alternatively, one may do a “model selection” to select the most plausible submodel and then use the estimate within that submodel to yield the estimate

\[
\hat{\mu}(x) = \hat{m}_j(x), \quad j = \arg \max\{j : n_j \phi(x; \bar{X}_j, \Sigma)\}.
\]

Both of the above estimates (10) and (11) are based on a given value of \(k\) and the given break-up of the clusters. Of course, all of these are unknown. In order to remove the dependence on a particular cluster formation, we form an ensemble of estimates given by (10) or (11) — a fresh estimate in each MCMC iteration of the DM process. The final estimate may thus be formed by a simple averaging of these \(\hat{\mu}(x)\). We call these ensemble based estimates corresponding to (10) and (11) as DM-AVE (Dirichlet mixture — average) and DM-ML (Dirichlet mixture — most likely) respectively.

Piecing all the previous discussions together, the complete algorithm to calculate our estimate may be described as follows:

- Choose a sufficiently fine grid of \(x\)-values.
- Start with an initial configuration \((k, s_1, \ldots, s_n)\).
- Let \(s_{-i}, k_{-i}\) and \(n_{-i,j}\) be as defined above. Given \(s_{-i}\) and \(X_1, \ldots, X_n\), sample \(s_i\) from \(\{1, \ldots, k_{-i}, k_{-i} + 1\}\) with probabilities

\[
\begin{align*}
&\left\{ \begin{array}{ll}
\frac{n_{-i,j}}{n_{-i,j} + k_{-i} + 1} \phi(X_i; \bar{X}_{-i} + j, \frac{n_{-i,j}}{n_{-i,j} + k_{-i} + 1} \Sigma) \\
\frac{\phi(x; X_{j-1} - i, \Sigma + \Phi)}{\phi(X_{j-1} - i, \Sigma + \Phi)} \\
cM \phi(x; \bar{X}_j, \Sigma + \Phi) \\
\end{array} \right. \quad j = 1, \ldots, k_{-i}, \quad j = k_{-i} + 1,
\end{align*}
\]

where \(X_j, X_{j-1}, X_{-i,j}, \Sigma\) and \(\Phi\) are defined in equations (7), (8), and (9), and \(c\) is the appropriate constant of proportionality to make the sum to be one.
Section on Nonparametric Statistics

- Repeat the above procedure for \( i = 1, \ldots, n \), to complete one MCMC cycle.

- Compute DM-AVE and DM-ML as defined by (10) and (11) for each cycle.

- Repeat the cycle sufficiently large number of times to let the chain mix well.

- Compute the averages of the estimates DM-AVE and DM-ML for each cycle after a burn-in for each point on a chosen grid.

- Obtain the estimated function by joining the average values over the grid points.

In density estimation applications, one usually works with the initial configuration that \( k = n \), that is, all clusters are singleton. Here one cannot use that because \( \Sigma \) will then be the zero matrix, and thus the formula will break down. As an alternative, one may split the domain into small cubes and form the initial clusters by looking at the sample values falling into these regions, where singleton clusters will have to be merged with some neighboring ones.

Note that we substitute \( \xi, \Sigma, \Phi, \) and \( m_1(\cdot), \ldots, m_k(\cdot) \), as well as the group mean values \( \theta_1, \ldots, \theta_k \) within an MCMC iteration by the corresponding empirical estimates in the analysis, thus adopting an empirical Bayes approach. Implementation of a fully hierarchical Bayesian approach is possible provided that we do full updating of the Dirichlet mixture process. However, the fully Bayesian approach takes much longer time to run and does not seem to lead to more accurate estimates.

Our method can be expanded further to construct a pointwise band for the regression function. Let \( x \) be a fixed point in the space of the regressor. A confidence interval for \( m(x) = E(Y|X = x) \) is easily obtained from the method of least squares given the sub-grouping of the samples and the group to which the new observation \( X = x \) belongs. However, since both the group to which \( x \) belongs and the break-up of clusters are unknown, we can integrate them out according to their posterior distributions, leading to a “posterior expected pointwise confidence band” as described below. Treat \( X = x \) as a new observation and suppose that \( J \) stands for the cluster label to which \( x \) belongs. Given the clusters, let \( [L_j, U_j], j = 1, \ldots, k \), be 100(1 - \( \alpha \))\% confidence limits for \( m_j(x) \). This confidence interval may be thought of as a posterior credible interval for \( m_j(x) \). Especially, in large sample sizes the two notions will tend to agree. Let \( L \) be the \( \gamma/2 \)-quantile of \( L_1, \ldots, L_k \) and \( U \) be the \( 1 - \gamma/2 \)-quantile of \( U_1, \ldots, U_k \), respectively, where \( \gamma = \alpha - \alpha_0 \) and the associated probabilities are \( \pi_1(x), \ldots, \pi_k(x) \). Let

\[
P(m^*(x) \leq L) = \sum_{j:L_j \leq L} P(m_j(x) \leq L) \pi_j(x)
+ \sum_{j:L_j > L} P(m_j(x) \leq L) \pi_j(x)
\leq \frac{\gamma}{2} + \frac{\alpha_0}{2}.
\]

Similarly \( P(m^*(x) \geq U) \leq (\gamma + \alpha_0)/2 \), and so \( P(L \leq m^*(x) \leq U) \geq 1 - \alpha, \alpha = \gamma + \alpha_0 \). Thus by the convexity of \([L, U]\), \( m(x) \), being an average of \( m^*(x) \), has the same credible bound given the clusters. Finally, the bounds \( L \) and \( U \) may be averaged out over MCMC iterations leading the final posterior expected pointwise confidence band. The obtained band may not have frequentist validity though, and it more closely resembles a posterior credible band.

Since the success of our method is intimately linked with the ability to discover the correct patterns in the data as well as the number of clusters, in the simulation study, we shall monitor the distribution of the number of groups found by the posterior and the similarity of the obtained groupings with the actual ones.

Asymptotically, Dirichlet mixtures estimate densities consistently [cf. Ghosal et al. (1999)], which implies that the mixing distribution is estimated consistently in the weak topology. In view of a result of Donoho (1988), consistency essentially implies that the number of clusters in the pattern obtained by the posterior is asymptotically at least as much as the true number of groups. In general, the obtained number could be considerably bigger, which is also confirmed by the simulation study. However, not much harm is caused by unnecessary breaking up of groups, except for a loss of some precision of the estimated regression function because of smaller number of available samples. On the other hand, an inappropriate merging of two groups introduces serious bias in the estimates.

Since the success of our method is intimately linked with the ability to discover the correct patterns in the data as well as the number of clusters, in the simulation study, we shall monitor the similarity of the obtained groupings with the actual ones.

3. Simulation study

In this section we do several simulations to evaluate the performance of our method. First, data are generated from a mixture of normal distributions in the univariate and the multivariate case, and in different groups different regression functions are given with several parameters. We proceed to estimate the fitted regression function using our proposed method and also some other methods such as those based on kernel and spline smoothing in the univariate case and MARS and GAM in the multivariate case.
Section on Nonparametric Statistics

We compare our method with other methods by the empirical $L_1$-error and $L_2$-error between estimates and the true regression function, defined respectively by
\[ EL_1(f) = n^{-1} \sum_{i=1}^{n} |\tilde{f}(x_i) - f(x_i)| \quad \text{and} \quad EL_2(f) = \left( n^{-1} \sum_{i=1}^{n} (\tilde{f}(x_i) - f(x_i))^2 \right)^{1/2}. \]

In order to monitor how close is the partition induced by the Dirichlet mixture process to the true partition, we shall use Rand’s (1971) measure of similarity of clusterings. Given $n$ points, $\theta_1, \ldots, \theta_n$ and two clustering vectors, $s = \{ s_1, \ldots, s_k \}$ and $s' = \{ s'_1, \ldots, s'_{k'} \}$, Rand’s measure $d$ is defined by
\[ d(s, s') = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \gamma_{ij}}{\binom{n}{2}}, \tag{12} \]
where
\[ \gamma_{ij} = \begin{cases} 1, & \text{if there exist } k \text{ and } k' \text{ such that both } \theta_i \text{ and } \theta_j \text{ are in both } s_k \text{ and } s_k', \\ 0, & \text{otherwise}. \end{cases} \tag{13} \]

In practice, we use a simple computational formula for $d(s, s')$
\[
\left[ \binom{n}{2} - \frac{1}{2} \left( \sum_{i=1}^{n} \sum_{j=i+1}^{n} n_{i,j} \right)^2 + \sum_{i=1}^{n} \left( \sum_{j=i+1}^{n} n_{i,j}^2 \right) \right] \\
+ \sum_{i=1}^{n} \sum_{j=i+1}^{n} n_{i,j}^2 \Big/ \binom{n}{2}, \tag{14} \]
where $n_{i,j}$ is the number of points simultaneously classified in the $i$th cluster of $s$ and the $j$th cluster of $s'$. Note that $d$ ranges from 0 to 1. When $d = 0$, the two clusterings have no similarities, and when $d = 1$ the clusterings are identical. Rand’s measure will be computed in each MCMC step.

Simulation studies are conducted under several different combinations of the true model and different sample sizes. Throughout this section and in the next, we shall work with the choice $M = 1$ for the precision measure of associated Dirichlet process. To see how accurate the clustering is, we also monitor Rand’s measure. All of simulation work is done by the package R (version 2.0.1).

3.1 One dimension

To give a simple example we consider a univariate predictor $X$ having the following distribution.
\[ X \sim 0.3N(-0.9, \sigma^2) + 0.2N(-0.3, \sigma^2) + 0.4N(0.4, \sigma^2) + 0.1N(1.0, \sigma^2), \]
where $\sigma^2 = 0.01, 0.02, 0.03, 0.04$ are considered.

First generate 100 samples of $X$ from above mixture of normals and generate $Y$ using four linear functions with different coefficients are given within each group.
\[ f_1(x) = 2.3 + 1.1x, \quad f_2(x) = 1.5 - 0.6x, \]
\[ f_3(x) = 0.8 + 0.9x, \quad f_4(x) = 1.7 - 0.2x, \]
with an additive noise term having mean 0 and variance $\tau^2 = 0.03$.

In the MCMC sampling step, we generate 5000 samples and ignore 1000 as a burn-in period. For appropriate selection of burn-in period, the trace plot of the number of distinct clusters was used and 1000 steps were found to be sufficient. This simulation work is repeated 100 times. Each simulation took about 5 minutes.

First we compare our two methods DM-AVE and DM-ML. The boxplots, which are not displayed to save space, show almost identical performance. For this reason, we shall compare only DM-AVE with other methods in the subsequent studies.

For kernel estimation, we use a normal kernel with the help of the R function “ksmooth” and choose bandwidth by the default mechanism in that program. We also monitor the empirical $L_2$-error and $L_1$-error. It shows that our method has good performance for all the cases we investigated. Larger the value of $\sigma^2$, more overlapped the groups are, and smaller the value of Rand’s measure. This is expected because if some of clusters are severely overlapped, it is hard to separate them.

Comparing the errors by means of Wilcoxon rank test in the cases $\sigma^2 = 0.01$ and 0.02, we conclude that the median error of DM-AVE method is significantly smaller than that of other two methods at the 5% level of significance. For $\sigma^2 = 0.03$, the difference between DM-AVE and spline is significant but the difference with the kernel estimate is not significant. For $\sigma^2 = 0.04$, none of the differences are significant.

Now let us consider the case that the true regression function is not linear. Let $X$ be generated by the following distribution
\[ X \sim 0.3N(-1.9, \sigma^2) + 0.3N(-0.3, \sigma^2) + 0.4N(1.4, \sigma^2), \]
where $\sigma^2 = 0.01, 0.02, 0.04$ are considered. The regression functions in each cluster are given by
\[ f_1(x) = 0.3 + 0.3 \sin(2\pi x), \]
\[ f_2(x) = 1.2 - 0.6 \sin(2\pi x), \]
\[ f_3(x) = 0.5 + 0.9 \sin(2\pi x). \]

Note here that we have three clusters in this case. Again we generate 5000 MCMC samples after 1000 burn-in and repeated the experiment 100 times. In this case, we use the cubic regression as well as the linear regression. Figure 2 tells us that the cubic regression performs better than the linear regression method and other methods. Formal rank tests shows that cubic DM-AVE method is significantly better than the kernel method. The spline method here performs slightly better than cubic DM-AVE although the difference is not significant. It may be noted that the cubic is also not the correct functional.
Section on Nonparametric Statistics

Figure 1: (One dimension) : The boxplots of L2-error. Rand’s measures on average are 0.94, 0.87, 0.83 and 0.81, respectively.

form of the regression functions of the groups. Thus the DM-AVE method performs reasonably well even under misspecified models provided that a flexible function like a cubic is used.

3.2 Ten dimension

We consider the case when the predictor X has 10 variables and distributed as a mixture of four multivariate normal: \( X \sim \sum_{j=1}^{4} \omega_j N(\mu_j, \Sigma) \). We let \( \Sigma = \sigma^2 I_{10} \), \( \sigma^2 = 0.2, 0.3, 0.4, 0.5 \) and \( \tau^2 = 0.1 \). The four mean vectors are given by

\[
\begin{align*}
\mu_1 &= (1,1,1,1,2,4,2,4,4,0.5)^T, \\
\mu_2 &= (4,0.5,5,3,5,3,1,1,1,1)^T, \\
\mu_3 &= (2,5,2,5,4,0.5,4,0.5,5,3)^T, \\
\mu_4 &= (5,3,1,1,1,1,2,5,2,5)^T.
\end{align*}
\]

Let the weights be \( \omega = (0.3, 0.2, 0.4, 0.1) \) and the within subpopulation regression functions are given by

\[
f_j(x) = \alpha_j + \beta_j^T x, \quad j = 1, \ldots, 4,
\]

where \( \alpha_1 = 2.3, \alpha_2 = 1.5, \alpha_3 = 0.8, \alpha_4 = 1.7 \) and \( \beta_1, \beta_2, \beta_3, \beta_4 \) are the vectors of length 10 respectively given by

\[
\begin{align*}
(1,1,-2,1,1,-2,-0.6,1.2,-0.6,1.2,0.9,-1.1)^T, \\
(0.9,-1.1,-0.2,1.2,-0.2,1.2,1.1,-2.1,1.1,-1)^T, \\
(-0.6,1.2,-0.6,1.2,0.9,-1.1,0.9,-1.1,-0.2,1.2)^T, \\
(-0.2,1.2,1.1,-2,1.1,-2,-0.6,1.2,-0.6,1.2)^T.
\end{align*}
\]

We generated 200 sample data from the resulting population and obtain estimates of regression. The whole simulation is repeated 100 times. We consider four different \( \sigma^2 \) values 0.2, 0.3, 0.4 and 0.5. Our method DM-AVE is compared with GAM and MARS. In the MCMC sampling scheme, we generate 5000 samples after ignoring first 1000 as burn-in. Figure 3 shows that DM-AVE is better than GAM and MARS for all four choices of \( \sigma^2 \). The differences are significant at 5% significance level according to the Wilcoxon rank test.

REFERENCES


Figure 2: (One dimension) : The case where the true model is not linear. The plots of $L_1$-error are shown. Average Rand's measures are 0.97, 0.96, 0.95 and 0.94, respectively.

Figure 3: (Ten dimension): Boxplots of $L_1$-error with respect to $\sigma^2$ values. Average Rand's measures are 0.84, 0.77, 0.76 and 0.76, respectively.