An Empirical Bayesian Kernel Density Estimator

Soma Sekhar Dhavala Souporno Ghosh

Introduction

Often times there is a need to infer the true underlying probability based on the observations, such as in, including but not limited to, data-mining, optimizing the process control parameters etc., Histograms, very rudimentary empirical density estimators, divide the whole data range into either equal or unequal sub intervals (bins) and then obtain the frequency of occurrence of each bin. They could lead to completely different estimates if the bin-width, and their locations are chosen differently.

The kernel density estimators (KDEs) offer practical alternatives to histograms, providing smooth density estimates. KDEs belong to a class of non-parametric methods of estimation, which assume no fixed structure of the underlying density and completely estimate the true density based on the observations alone.

However, a fundamental challenge in KDEs is controlling the degree of smoothness that a method provides. As the degree of smoothness is very subjective, many methods were proposed which minimize certain cost functions or consider some ad hoc criteria to select the smoothening criteria.

In this report, we propose and investigate a type of non-parametric KDE from a Bayesian stand-point.We are inspired by wavelet-based KDE in the sense of analyzing the density at different scales and the scales are weighed probabilistically.

The report is organized as follows: We introduce a few types of KDEs. We introduce our model in the next Section and analyze the choice of our priors. Later, we apply our model to simulated data as well as real data. We discuss the various modeling issues therein. Finally, we conclude the report by summarizing the main results.

Review of KDEs

In KDE, a kernel function (essentially a smoothening function) is centered at each data point and the contribution of each data point over a local neighborhood is calculated. Thus the contribution of an observed data point $y(i)$ to the point at which we want to evaluate the density, say , at x depends on how far apart $y(i)$ and x are.

Formally speaking, let $y(i)$ be i.i.d random samples from a continuous density $f(x)$. Then, $\hat{f}(x)$ the KDE of $f(x)$, is given by:

$$\hat{f}(x) = \sum_{i=1}^{N} \Phi\left(\frac{x - y(i)}{h}\right) \tag{1}$$

where Φ is the smoothening kernel, $y(i)$ the observed data and h the bandwidth.

The optimal bandwidth that minimizes Average Integrated Mean Square Error (AIMSE) is given by [1]:

$$h_{\text{opt}} = 0.9\,\sigma\,N^{-0.2} \tag{2}$$

However, there may not a single bandwidth that offers the best smoothening at all locations. Consider a case, where te data is coming from a mixed distribution having spread-out means. Then, we might need different smoothness parameters for these two different distribution. A natural modification would be to vary the bandwidth at each location, leading to a adaptive bandwidth KDE, given by:

$$f(x) = \sum_{i=1}^{N} \Phi\left(\frac{x - y(i)}{h_i}\right) \tag{3}$$

where $h(i)$ is the local bandwidth. A particular choice for the bandwidth is:

$$h(i) = \frac{h_{\text{opt}}}{\hat{f}(y(i))} \tag{4}$$

The adaptive bandwidth KDE performs better than the constant bandwidth KDE in terms of AIMSE but it suffers from tail problems, i.e., it tends to produce bumpy tails. One way to account for tails, is by choosing a different kernel. Another approach is to use the asymptotic properties of based on the order statistics, which would naturally consider the location and thus the tails in the KDE. The rank adaptive KDE is given by [2]:

$$f(x) = \sum_{i=1}^{M} \Phi\left(\frac{x - x_{(i)}}{h_i}\right) \tag{5}$$

where $x_{(i)}$ is the ith order statistic and N is the number of ordered statistics. The local bandwidths are chosen as

$$h_i = h_{\text{opt}} \left(\frac{p_i(1 - p_i)}{f^2(x_i)}\right)^{0.5} \tag{6}$$

where p_i is the empirical c.d.f of the order statistics (several definitions exist):

$$p_i = \frac{i}{N + 1} \tag{7}$$

and the true density is replaced by some pilot estimate, in this case a constant BW KDE with optimal bandwidth.

We can view the above equation as a mixture of some parametric densities with different scale and location parameters. We immediately observe that, at each location determined by the order statistics, we employ a particular scale. For e.g, if we choose the standard normal density as the kernel, we can view the bandwidth as the scale parameter and the KDE is in fact a finite mixture model. This might work reasonably well in many situations, but may not model concave-shaped densities (like exponential densities which have a faster decay than Gaussians), one natural choice is to consider a hierarchy of mixture densities, i.e, we assume that at each *pivot*, we model the component of the mixture model, as another finite mixture model. The hierarchy of mixtures thence formed could model arbitrary distributions with a greater degree of flexibility.

The interpretation of hierarchy of mixtures is not new. For example, in the wavelet-shrinkage based KDEs, one is essentially using shifted and scaled versions of the mother wavelet (a wavelet can not be a pdf because its average value is zero) to approximate the underlying function [3]. However they produce non-bona fide densities because the underlying smoothening functions are not densities.

In the next section we will introduce our model.

Model

The proposed form of the KDE tries to approximate the density at each order statisitc as finite mixture of Gaussians with different scales, given by:

$$f(x) = \sum_{i=1}^{M} \beta_i \sum_{j=1}^{J} \alpha_i^j \Phi\left(\frac{x - x_{(i)}}{h_i^j}\right) \tag{8}$$

where x_i is the ith pivot and j corresponds to a particular scale. The scales are varied on a dyadic scale with their the median being the same as the local bandwidth in rank-adapted KDE. We can view β_is as the associated weights of the density at the ith pivot and α_i^js weigh the densities with different scales but at the same pivot. We constrain these weights so that the KDE is in fact a valid probability density function, given by:

$$\begin{gathered} \beta_i, \alpha_i^j \in [0,1] \\ \textstyle\sum_{i=1}^{M} \beta_i = 1 \text{ and } \sum_{j=1}^{J} \alpha_i^j = 1 \end{gathered} \tag{9}$$

We treat the weights as random quantities rather than deterministic but unknown. This philosophy allows us to incorporate subjective knowledge about the problem at hand, and more importantly, it makes drawing almost any inference about the parameter possible, like 95% posterior confidence intervals. Thanks to the advancements in MCMC methods and increased hardware efficiency. However, in the present problem, we have little knowledge about the distributional properties of the weights. Hence, we ascertain the ambiguity by placing hyper priors on them.

Following are the prior specifications for the parameters:

$$\begin{aligned} \boldsymbol{\beta}_{|\boldsymbol{\zeta}} &\sim \text{Dirichlet}(\boldsymbol{\zeta}) \\ \boldsymbol{\zeta} &\sim \text{Multinomial}(\boldsymbol{p}) \end{aligned} \tag{10}$$

$$\begin{aligned} \boldsymbol{\alpha}_{i|\boldsymbol{\eta}} &\sim \text{Dirichlet}(\boldsymbol{\eta}_i) \\ \boldsymbol{\eta}_i &\sim \text{Multinomial}(\boldsymbol{\lambda}_i) \end{aligned} \tag{11}$$

where $\sum_{j=1}^{J} \lambda_i^j = 1$ and $\sum_{i=1}^{N} w_i = 1$

We chose Dirichlet prior since βs are proportions (mixing weights) and consequently the Multinomial hyper prior. Of the several choices possible, we chose the empirical pmfs as the hyper-prior constants for $\boldsymbol{\beta}$, given by

$$w_i = \hat{f}(x_i) \tag{12}$$

We could have been non-informative by specifying equal proportions in the hyper-prior. We derived the priors based on the data as we are dealing with potentially hundreds of parameters and we are afraid that the chains may not converge Thus our approach is an *e*mpirical Bayesian analysis

Likewise for $\boldsymbol{\lambda}$, we choose:

$$\lambda_i^j = \frac{\exp(-|j - \left(\frac{J+1}{2}\right)|)}{\sum_{k=1}^{N} \exp(-|j - \left(\frac{J+1}{2}\right)|)} \tag{13}$$

We give maximum weight to the local bandwidth suggested by the rank-adapted KDE and taper-off the weights as we move-away from the nominal value. Thus over-smoothening and bumpy tails can be avoided. However, the Bayesian set-up would allow us to pick-up those scales if the data suggests otherwise.

The choice of scales and the way the scales are chosen is also subjective. In the present analysis, we have chosen the number of scales to be five (usually an odd number, with out loss of generality). We are inspired by the work in wavelet literature. Hence, we have decided in favor of varying the scales on a dyadic scale, given by:

$$h_i^j = h_i \, 2^{\frac{J+1}{2} - j} \tag{14}$$

MCMC computations

Having set-up the model, it remains to set-up the MCMC chains. We have chosen Gibbs sampler, not due to its simplicity but rather the complexity involved in specifying a nice proposal density for Metropolis-Hastings step, which is more of an art than of a science, has lead to choose the other way.

The MCMC chains can be generated using the Gibbs sampler. The algorithms is as follows:

- step-1: for t = 1,2,...(chain loop)
- step-2: choose i ∈ [1,N] without replacement (pivot loop)

- step-3: draw $\beta_i(t) \sim p(\beta_{i|\theta^c})$
- step-4: choose j ∈ [1,J] without replacement (scale loop)
- step-5: draw $\alpha^j_{i,|\theta^c}(t) \sim p(\alpha^j_{i|\theta^c})$
- step-6: go to step-1 until the desired number posterior samples are drawn

Note that in the steps two and three, we have chosen to visit the pivots and scales *randomly*, i.e, we jump from one pivot to another randomly and within a pivot, we visit different scales also randomly. We have decided to use this random jumps to avoid the the chains getting-stuck, a phenomena that we have observed when looped sequentially in a deterministic fashion, the usual Gibbs sampler way!

In order to implement the Gibbs sampler, we need to know the full conditionals. We first note that, the posterior is proportional to:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \ell(x)\pi(\boldsymbol{\beta}_{|\boldsymbol{\zeta}})\,\pi(\boldsymbol{\zeta})\,\pi(\boldsymbol{\alpha}_{|\boldsymbol{\eta}})\,\pi(\boldsymbol{\eta}) \tag{15}$$

where $\pi(.)$ representing a prior density and the likelihood is given as

$$\ell(x) = \prod_{m=1}^{M}\sum_{i=1}^{M}\sum_{j=1}^{J}\beta_i\alpha_i^j\Phi\left(\frac{x_m - x_{(i)}}{h_i^j}\right) \tag{16}$$

If we factor-out the terms that are only dependent on a particulat α or β, we would observe

$$\ell(\alpha^j_{i|\theta^c}) \propto \prod_{m=1}^{M}\left(\alpha_i^j\gamma^j_{i,0}(m) + \gamma^j_{i,1}(m)\right) \tag{17}$$

where

$$\gamma^j_{i,0}(m) = \beta_i\Phi\left(\frac{x_m - x_{(i)}}{h_i^j}\right) \tag{18}$$

and

$$\gamma^j_{i,1}(m) = \sum_{\substack{k,l \\ k,l\neq\{i,j\}}}\beta_i\,\alpha_i^j\,\Phi\left(\frac{x_m - x_{(i)}}{h_i^j}\right) \tag{19}$$

where θ^c is the set of all parameters excluding the one whose distribution we wanted to consider which should be evident from the context, for e.g., $p(\alpha^j_{i|\theta^c})$ is the full conditional density of α_i^j

Likewise, for β, we get,

$$\ell(\beta_{i|\theta}) \propto \prod_{m=1}^{M}\left(\beta_i\psi_{i,0}(m) + \psi_{i,1}(m)\right) \tag{20}$$

where

$$\psi_{i,0}(m) = \beta_i \sum_{j=1}^{J} \alpha_i^j \Phi\left(\frac{x_m - x_{(i)}}{h_i^j}\right) \tag{21}$$

$$\psi_{i,1}^j(m) = \sum_{\substack{k,l \\ k\neq i}} \beta_i \, \alpha_i^j \, \Phi\left(\frac{x_m - x_{(i)}}{h_i^j}\right) \tag{22}$$

We can use rejection sampling to draw samples from the conditional density with the conditional Dirichlet as the proposal density, i.e.,

$$p(\alpha_{i|\theta^c}^j) \propto \ell(\alpha_{i|\theta^c}^j)\pi(\alpha_{i|\theta^c}^j) \tag{23}$$

We note that the first term on the right-hand side is a polynomial of degree M in α_i^j and second term is a conditional Dirichlet density. In this case, the rejection sampling reduces to the following simple algorithm, since $\alpha_i^j \in [0, 1]$, to:

Let $Q(\alpha_i^j)$ be the Mth degree polynomial in α_i^j, with q_k being the coefficient associated with the term ${\alpha_i^j}^k$ (and $q_{-1} = 0$) then

- step-1: $u \sim$ U[0,1]
- step-2 : choose $k \in [0, M]$ such that $\sum_{l=-1}^{k-1} q_l < u \leq \sum_{l=-1}^{k} q_l$
- step-3: generate α_i^j from $\pi(\alpha_{i|\theta^c}^j)$
- step-4: $u \sim$ U[0,1]
- step-5: if $u \leq \alpha_i^{j,k}$, accept the proposal, else go to step-3

Now, we need to be able to generate draws from the conditional Dirichlet as follows (step-3 of the above algorithm):

At a particular iteration in the MCMC chain,

- $\boldsymbol{\zeta} \sim$ Multinomial($\boldsymbol{\pi}$)
- $\alpha_i^j \sim \Gamma(\zeta_i^j, 1)$ (notation followed from Gelman Text Book)
- for each i, $\alpha_i = \frac{\alpha_i^j}{\sum_{i=1}^{N} \alpha_i^j}$

We can generate β_is also in the same fashion. In the next Section, we present analysis of simulated as well as real data.

Simulations & Analysis

We have applied our model to the Galaxy data described in [4]. There are a total of 82 observations and the data were scaled so that x_1 is 0.05 and x_{82} is 0.95. We retained all the order statistics while carrying out the MCMC simulations. The weights of the prior distribution for α and β are shown in Figures 1 and 2 respectively. It should be noted that the hyper prior for β is almost uniform, indicating that each order statistics is visited very few times.

We run the Gibbs sampler as described in the earlier Sections. The MCMC chain for some βs is shown in Fig. 3. It should be noted that the chain stays in a particular state for a while before taking any jumps. This phenomena was even more predominant when we traversed the pivots in a sequential deterministic manner. We observed from the autocorrelation plots for the βs that indeed the correlations was significant upto ten lags. Hence, we have systematically sampled the posterior chain to make the samples appear to be from a independent chain. All the βs. the mixing weights of the distributions at the pivots, are shown in Fig. 4. We see that few components have near zero inclusion. On an average, all pivots have approximately the same probability of inclusion. However, that is not the case with *alpha*s which seem to mix nicely.

From our empirical pdf plots we observed that, most of the scaling weights have right-skewed posterior distributions which is reflected in Fig. 5 where we plotted the chains for some scaling weights at a particular pivot. Hence, we have chosen to represent the population statistics with the median instead of the mean. We obtained the posterior median for all the parameters. The posterior median for β are shown in Fig. 6 and the 95% confidance bands are shown in Fig. 7. It can be observed that the upper CIs are far-off from the median indicating a right-skewed distribition.Similarly, posterior median for α at different scales are shown in Fig. 8. It should be noted that, the middle scales,which corresponds to the local bandwidth of the adpative-bandwith approach has large posterior probabilty. We belive that we need to run more number of chains inorder to correctly asses the effect of prior on the posterior inference.

Using the medians as the parameter estimates, we compute the kernel density estimate. Various KDEs, namely, the constant bandwidth KDE, adaptive bandwidth KDE, our proposed method are depicted in Fig. 8.Our method prdces results in comparison to the rank-adaptive method. It picks up the vallies and peaks in the data. We believe that we need to assign different weights to obtain the desired smoothening. We would be investigating different weights for the prior for α.

To save computational complexity, we have considered significant order statistics which account for certain proportion of the observed samples. Since we have few observations, we considered all the ordered statistics in the computations. Also, as is the case usually, the order statistics are distinct and they almost occur with the same frequency. This is because the data is continuous and unless the the data is censored, it is vey improbable that we observe the same order statistics frequently. Hence in our next analysis, we considered uniform weights for βs treating them as deterministic quantities.

We have simulated the data from a mixture of normal densities. Similar to the earlier analysis, we have obtained the posterior samples using the Gibbs sampler. The only difference was that we assigned equal weights to the βs. The chains for different scaling weights are shown in Fig. 10. In Fig. 11, we show the posterior median estimates for different pivots at scales corresponding to over smoothening, median and under smoothening. We obtain the KDE by using the medians of the MCMC chains. The resultingKDE, along with other methods is shown in Fig. 12.

Conclusions

Our proposed method, generalizes the rank-adaptive KDE. The Bayesian set-up allows one to control the degree of smoothness through hyper priors. From our preliminary results, we discoved that the mixing weights at the pivots are almost equi-weighed. In the MCMC simulations, we have to traverse the chain stochastically to avoid the chain getting stuck temporarily before it moves by itself. Over all, the proposed method seems to be a viable candidate for further research warranting extensive simulations.

References

[1] S.J Sheather, "The performance of six popular bandwidth selection methods on some real data sets," *Comput. Statistics*, vol. 7, pp. 225–250, 1992.

[2] David B. Kim., "Rank adapted kernel density estimation," 2001.

[3] Peter Muller * and Brani Vidakovic, "Bayesian inference with wavelets: Density estimation," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 456468, 1998.

[4] K. Roeder and L. Wasserman, "Practical Bayesian density estimation using mixture of normals," *JASA*, vol. 92, pp. 894–902, 1997.

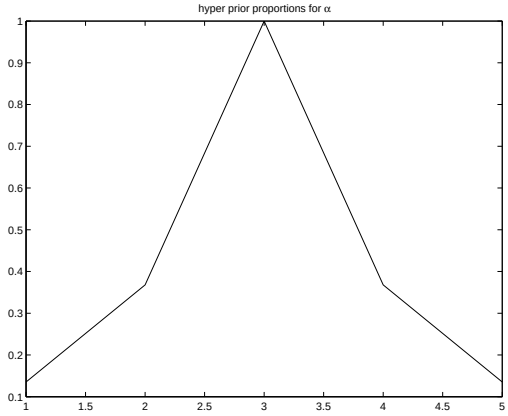

Figure 1: hyper prior weights for α at all pivots

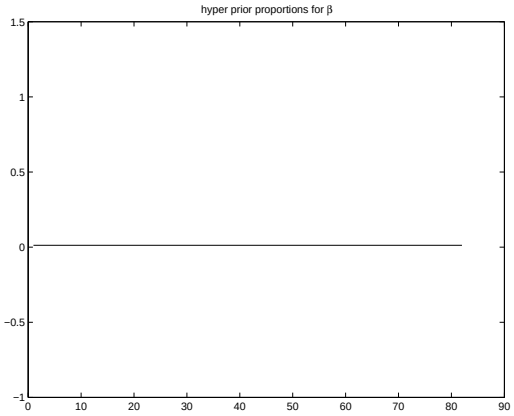

Figure 2: hyper prior weights for α at all pivots

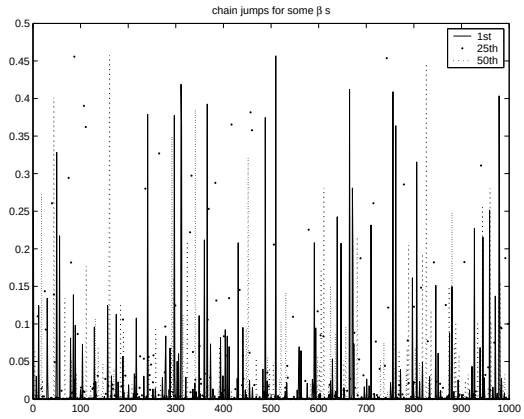

Figure 3: some β_is in the MCMC chain

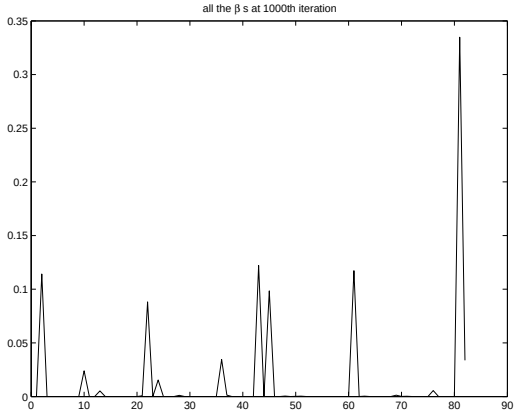

Figure 4: β_is at 1000th iteration of MCMC

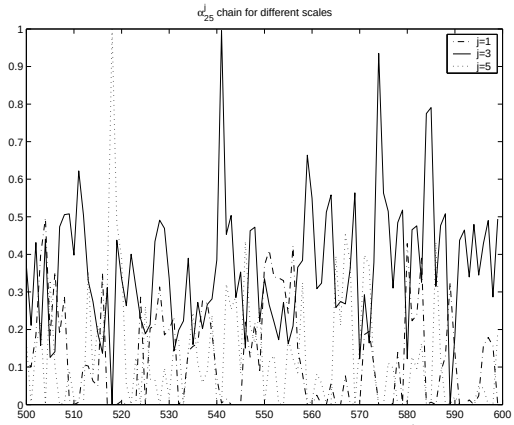

Figure 5: MCMC chain from 500 to 600 for α^j_{25}s at different scales

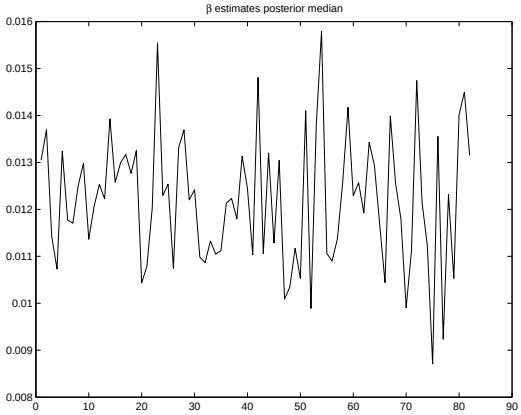

Figure 6: posterior median of the MCMC chains for β_is

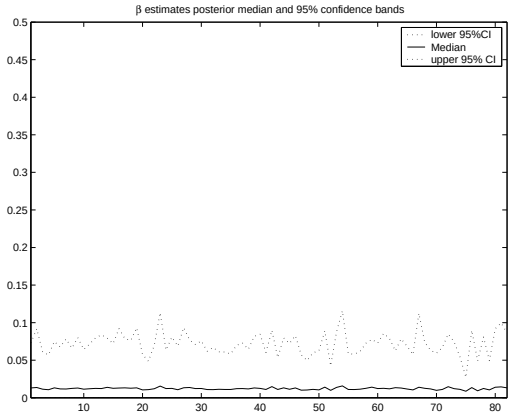

Figure 7: posterior 95% CIs for the β_is

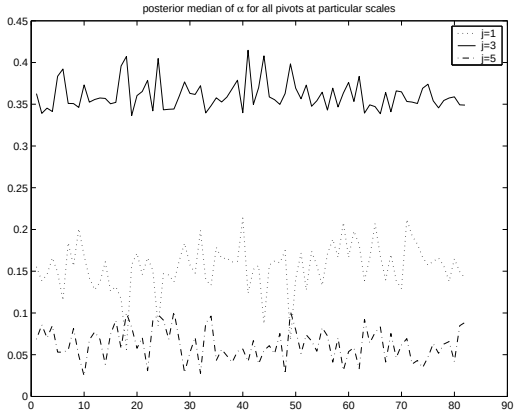

Figure 8: posterior median of α at particular scales

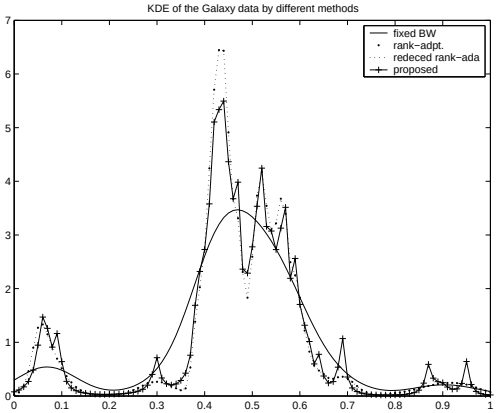

Figure 9: KDEs from methods for the Galaxy data

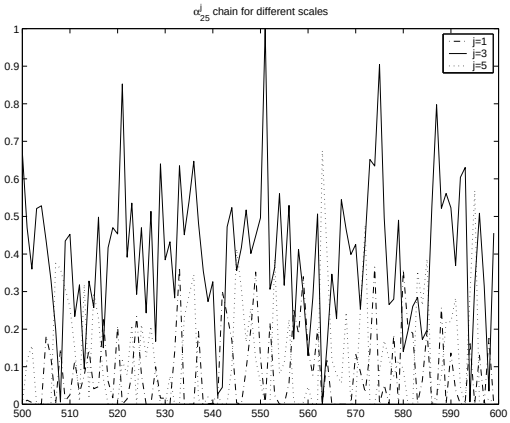

Figure 10: MCMC chain from 500 to 600 for α^j_{25}s at different scales (GMM)

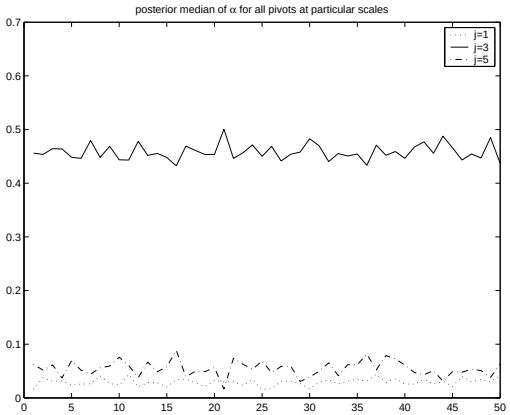

Figure 11: posterior median of α at particular scales (GMM)

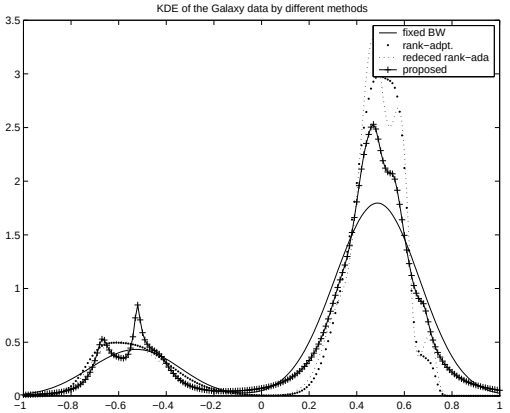

Figure 12: KDE estimates of the Gaussian Mixture density