ON SELECTING THE HYPERPARAMETERS OF THE DPM MODELS FOR THE DENSITY ESTIMATION OF OBSERVATION ERRORS

Asma Rabaoui\textsuperscript{1}, Nicolas Viandier\textsuperscript{2}, Juliette Marais\textsuperscript{2}, Emmanuel Duflos\textsuperscript{1,3}

\textsuperscript{1}LAGIS UMR CNRS 8146, Ecole Centrale de Lille, Villeneuve d’Ascq, France
\textsuperscript{2}INRETS, LEOST, Villeneuve d’Ascq, France
\textsuperscript{3}Team-Project SequeL, INRIA Lille Nord-Europe, Villeneuve d’Ascq, France

\section*{ABSTRACT}

The Dirichlet Process Mixture (DPM) models represent an attractive approach to modeling latent distributions parametrically. In DPM models the Dirichlet process (DP) is applied especially when the distribution of latent parameters is to be considered as multimodal. DPMs allow for uncertainty in the choice of parametric forms and in the number of mixing components (clusters). The parameters of a DP include the precision $\alpha$ and the base probability measure $\mathbb{G}_0(\mu, \Sigma)$. In most applications, the choice of priors and posteriors computation for the hyperparameters $(\alpha, \mu, \Sigma)$ clearly influences inferences about the level of clustering in the mixture. This is the main focus of this paper. We consider the problem of density estimation of an observation noise distribution in a dynamic nonlinear model from a Bayesian nonparametric viewpoint. Our approach is illustrated in a real-world data analysis task dealing with the estimation of pseudorange errors in a GNSS based localization context.

\textbf{Index Terms}— Density estimation, Bayesian nonparametrics, Dirichlet Process, Mixture models, Hyperparameters

\section{1. INTRODUCTION}

Mixture modeling is a successful and widely used density estimation method, capable of representing the phenomena that underlie many real-world datasets. However, the main difficulty in mixture analysis is how to choose the number of mixture components. Model selection methods in general treat the number of components as an unknown constant and set its value based on the observed data. Such an approach lacks flexibility, since in practice we often need to model the possibility that new observations come from as yet unseen components. The DPM model, studied in Bayesian nonparametrics \cite{1, 2}, allows for exactly this possibility. The construction of probability measures on the space of distribution functions has received considerable attention in the Bayesian nonparametrics literature \cite{3}. A rich literature exists on its theoretical properties, and it has been used in a variety of problems \cite{4, 5, 6}.

Let consider a pdf $F$ and $y_1, \ldots, y_n$ a set of vectors statistically distributed according to $F$, $y_i \sim F(.)$ We will consider a nonparametric model allowing to estimate $F$ as follows:

$$ F(y) = \int_{\Theta} f(y|\theta) d\mathbb{G}(\theta) $$

where $\theta \in \Theta$ is called the latent variable (also called cluster), $f(\cdot|\theta)$ is the mixed pdf and $\mathbb{G}$ is the mixing distribution.

Under a Bayesian framework, it is assumed that $\mathbb{G}$ is a Random Probability Measure (RPM) distributed according to a prior distribution. In this paper, we have selected as a RPM the DP prior. In mixture models the DP is used to specify latent patterns of heterogeneity, particularly when the distribution of latent parameters is thought to be clustered (multimodal) \cite{1}.

Recall that in the finite mixture model, each data point is drawn from one of $k$ fixed, unknown distributions. For example, the simplest Gaussian mixture assumes that each observation has been drawn from one of $k$ Gaussians, parameterized by $k$ different means. To allow the number of mixture components to grow with the data, we move to the general mixture model setting. This is best understood with the hierarchical graphical model representation in \cite{1}. The core of the DPM model can basically be thought of as a simple Bayes model given by the likelihood $y_i \sim f(y_i|\theta_i)$ and prior $\theta_i \sim \mathbb{G}(\theta_i)$, with added uncertainty about the prior distribution $\mathbb{G}$:

\begin{align*}
  y_i & \sim f(y_i|\theta_i), \ i = 1, \ldots, n \\
  \theta_i & \sim \mathbb{G} \\
  \mathbb{G} & \sim DP(\mathbb{G}_0, \alpha)
\end{align*}

In \cite{7} the Dirichlet Process has been introduced as a measure on measures. The DP is parameterized by a base distribution $\mathbb{G}_0$ on a (measurable) space and a positive scaling parameter $\alpha$. Suppose we draw a random measure $\mathbb{G}$ from a DP, and independently draw $n$ random variables $\theta_i$ from $\mathbb{G}$. Marginalizing out the random measure $\mathbb{G}$, the joint distribution of $\{\theta_1, \ldots, \theta_n\}$ follows a Polya urn scheme \cite{2, 1}. The underlying random measure $\mathbb{G}$ is discrete with probability one. Moreover, it can be shown that the predictive distributions, computed by integrating out $\mathbb{G}$, admits the following Polya urn representation \cite{2}:
\[ \theta_{i+1}|\theta_{1:i} \sim \frac{1}{\alpha + i} \sum_{k=1}^{i} \delta_{\theta_k} + \frac{\alpha}{\alpha + i} G_0 \] (2)

Where \( \delta_{\theta_k} \) denotes a degenerate distribution of unit mass centered at \( \theta_k \). This factorization implies that \( \theta = (\theta_1, \ldots, \theta_n) \) is randomly partitioned into classes of distinct values such that the elements of \( \theta \) within a class share the same value. Therefore, a random draw from \( G \) may be computed as follows:

\[ \theta_{i+1}|\theta_{1:i}, G = \begin{cases} \theta_{c^*_k}, & \text{with probability } n_{i,c}/(\alpha + i) \\ G_0, & \text{with probability } \alpha/(\alpha + i) \end{cases} \] (3)

where \( \theta_{c^*_k} (c = 1, 2, \ldots, k) \) denotes the \( c^* \)th of \( k \) distinct values in \( \theta_{1:i} = (\theta_1, \ldots, \theta_k) \) and \( n_{i,c} \) denotes the number of elements in \( \theta_{1:i} \) that equals \( \theta_{c^*_k} \). In other words, a random draw from \( G \) either equals one of the previous draws or is drawn independently from the base probability measure \( G_0 \). This is due to the discreteness of the random measure \( G \). The parameter \( \alpha \) obviously plays a pivotal role in the distribution of \( \theta \). In (3) note that the probability that \( \theta_i \) differs from all previously drawn parameter values is proportional to \( \alpha \). The parameter \( G_0 \) is a distribution function and is the location parameter of the DP prior. The parameter \( \alpha \) is the strength of belief in the prior; it is a type of dispersion parameter. If \( \alpha \) is large, the distributions which are samples from the DP prior will concentrate near \( G_0 \). If \( \alpha \) is small, DP prior will be more diffuse.

The paper is organized as follows. Section 2 introduces the motivating problem. Our aim using this family of densities based on DPM is to be able to capture the right shape of the noise pdfs and hence statistical inference for the parameters of interest will be improved and reliable. Therefore, in section 3 we show how to tune the DPM hyperparameters in a flexible way. The posterior distribution of the precision \( \alpha \) will be estimated as part of the Gibbs sampler, and \( G_0 \) parameters will be chosen carefully in a data-adaptive way for a better fitting of the data distribution shape. In section 4 the efficiency of the proposed approaches is demonstrated by conducting validation experiments involving simulated GNSS signals.

2. THE MOTIVATING PROBLEM

In a dynamic nonlinear model with additive observation errors, it is usual to assume that errors are normally distributed or can be approximated by a finite Gaussian mixture. This can cause problems when there are for example outlying errors leading to many modes in the density distribution form whose variability over time induces poor inference for the parameters of interest. It is our intention therefore to model the additive error using a highly flexible family of density functions, which are based on a Bayesian nonparametric model involving the Dirichlet process.

In a hierarchical model, a hyperprior can be placed on the parameters of the DP by simply putting prior distributions on \( \alpha \) and the parameters of \( G_0 \). The data is then used to calculate the posterior distribution. We can say that the DP adds another layer in the hierarchical model, and allows if necessary to deviate from the shape of \( G \) after the data is seen.

Since \( \alpha \) is an important parameter is this problem, one may want to put a prior on \( \alpha \) and then use the data to calculate a posterior distribution. In understanding what values of \( \alpha \) to use, the relationship between alpha and the expected number of clusters in the data might be considered. Escobar and West [1] developed a clever Gibbs sampler wherein draws from the conditional posterior distribution of \( \alpha \) are computed by drawing successive samples from relatively familiar distributions (Beta and Gamma).

Concerning the centring distribution \( G_0 \), since little work has been conducted on how to tune the hyperparameters, in this paper priors on \( G_0 \) parameters will be specified using a data-adaptive way without applying a pure sampling approach. Most of the time, in real-world application contexts, the observed data can help in defining a prior for some of the hyperparameters in the considered model. Computing every unknown hyperparameter with pure sampling is not the most effective approach in all cases since this can be practically difficult and time consuming.

3. HYPERPARAMETERS ESTIMATION

3.1. Prior and posterior distributions for \( \alpha \)

In [8], the author shows how, with respect to a flexible class of prior distributions for parameter \( \alpha \), the posterior may be represented in a simple conditional form that is easily simulated. As a result, for a better fitting of DPM models, inference about this parameter may be developed with the existing Gibbs sampling algorithms.

We have mentioned in this paper that a key feature of the Dirichlet is its discreteness, which in our context implies that the pairs \((\mu_j, \Sigma_j), (j = 1, \ldots, n)\), concentrate on a set of some \( k \leq n \) distinct pairs. To sample the precision parameter \( \alpha \), we first determine the prior distribution for \( k \), the number of normal components in the mixture (clusters). At each stage of the simulation analysis, a specific value of \( k \) is simulated from the posterior for \( k \) (together with sampled values of the means and variances of the normal components) which also depends critically on this hyperparameter \( \alpha \). In [1, 8] it was shown how, based on a specific but flexible family of prior distributions for \( \alpha \), the parameter vector \( \theta \) may be augmented to allow for simulation of the full joint posterior now including \( \alpha \).

As in [8], a Gamma prior for \( \alpha \) will be used. Suppose \( \alpha \sim G(a, b) \), a gamma prior with shape \( a > 0 \) and scale \( b > 0 \). In this case, \( p(k|\alpha, n) \) may be expressed as a mixture of two gamma posteriors, and the conditional distribution of the mixing parameter given \( \alpha \) and \( k \) is a simple Beta.
\[(\alpha|x,k) \sim \pi_x G(a+k, b - \log(x)) + (1 - \pi_x) G(a+k-1, b - \log(x))\]  
(4)
with weights \(\pi_x\) defined by
\[
\frac{\pi_x}{1 - \pi_x} = \frac{(a + k - 1)}{n(b - \log(x))}. \tag{5}
\]
where \(0 < x < 1, k > 1\) and \((x|\alpha, k) \sim \beta(\alpha + 1, n)\). More details about this sampling can be found in [8]. It is important to mention that \(\alpha\) should be sampled at each Gibbs iteration stage in the simulation process. The current sampled values of \(k\) and \(\alpha\) permit to compute a new value of \(\alpha\).

3.2. The \(G_0\) parameters

DP mixtures have received a considerable interest in the bayesian nonparametric literature. Despite this activity, little has been written on choosing a centring distribution \(G_0\) for these models. In this work, the latent sources of heterogeneity to be specified through the \(G_0\) parameters are considered using a data-adaptive way. More details about this, will be developed in the experimental part.

Note that we considered the most common instance of the general model which is the normal-normal DPM model in which

\[
G_0(\mu, \sigma) = \mathcal{N}(\mu_0, \sigma/k_0), \mathcal{W}^{-1}(\Sigma_0, dof)
\]

where \(\mu_0\) is the mean of the Normal distribution \(\mathcal{N}\), \(k_0\) is a scale parameter, \(\Sigma_0\) and \(dof\) are respectively the covariance and the degree of freedom for the inverse Wishart distribution \(\mathcal{W}^{-1}\) (see [2] for more details about these distributions).

4. MODELING PSEUDORANGE ERRORS

4.1. Context

In multi-sensors based systems, each sensor transmits a signal (an information) to the receiver (or antenna). These sensors can work according to three different operation modes (failure mode, degraded mode and normal mode). In this paper, the GNSS (Global Navigation Satellite System) constellation [9] is considered as a sensor network. Consequently, each GNSS satellite is considered as a sensor [10]. The failure mode appears when the antenna can not receive a signal from a satellite because of local masks. In the normal mode, the satellite signal reaches the antenna in line of sight (LOS). Finally, the degraded mode occurs when the signal reaches the antenna after one or more reflections (non line of sight (NLOS)). To compute a position, the receiver needs to track signals from at least four different satellites. The distance between the satellite and the receiver deduced from the signal propagation time is called pseudorange.

Errors induced by atmospheric propagation and satellites clock bias can be corrected by correction models [9]. After corrections, the pseudorange error is only expressed in function of two different error sources as \(e_i^s = w_i + m_i^s\), where \(m_i^s\) is the potential error due to the signal reflections between the satellite \(s\) and the receiver at the time \(t\) and \(w_i\) is the receiver noise. According to the reception conditions, \(e_i^s\) can switch between different observation models as following:

\[
\begin{cases}
   LOS & : m_i^s = 0, e_i^s \sim \mathcal{N}(0, \sigma) \\
   NLOS & : m_i^s \neq 0, e_i^s \sim \mathcal{N}(0, \sigma)
\end{cases}
\]

In the case when a signal is received in LOS, the pseudorange error distribution is considered white-Gaussian. In constrained environments, the density form can change abruptly or during a long period of time according to the obstacle nature. Moreover, several obstacles (vehicles, pedestrians, etc) can induce random errors. Therefore, to estimate the pseudorange error density with accuracy, the DP Mixture was proposed.

4.2. Error Density Modeling using DPM

In our application of on-line pseudorange error density estimation, the latent variables \(\theta_i\) are the mean and the variance of each Gaussian distribution included in the infinite mixture, i.e. \(\theta_i = (\mu, \sigma)\). To compute a vehicle position at each step of the filtering process, \(N\) particles \(\theta_1^r, \ldots, \theta_N^r\) are computed. The problem with GNSS applications is that signal propagation can be considered in different modes. In LOS reception, the pseudorange error noise follows a zero-mean Gaussian distribution with a standard deviation \(\sigma\) inferior or equal to 1 (actually it depends of the receiver used). In NLOS reception, the pseudorange error can be higher than 100 m.

In the proposed DPM model, the \(\alpha\) will be estimated by sampling as detailed in section 3. The same base distribution \(G_0\) cannot be applied in all navigation situations. Here, we propose to adapt both each time using the distributions showed in section 3. To optimize the DPM, we propose to use an additional parameter \(r_i^s\) for each satellite \(s\) at each time \(t\). \(r_i^s\) is an estimation of the propagation state (LOS, NLOS or blocked). It updates the parameters of \(G_0\). In LOS reception the \(G_0\) hyperparameters are drawn as follows:

\[
\mu_i \sim \mathcal{N}(\mu_0, \sigma_i/k_1)
\]
\[
\sigma_i \sim \mathcal{W}^{-1}(\Sigma_1, dof_1)
\]

where \(\mu_0\) is the mean of the Normal distribution, \(k_0\) is a scale parameter, \(\Sigma\) and \(dof\) are respectively the covariance and the degree of freedom for the inverse Wishart distribution. However, in NLOS reception the distribution of the hyperparameters is adapted:

\[
\mu_i \sim \mathcal{N}(\mu_0 + res_i^s, \sigma_i/k_2)
\]
\[
\sigma_i \sim \mathcal{W}^{-1}(\Sigma_2, dof_2)
\]

with \(\Sigma_2 > \Sigma_1, dof_2 > dof_1, k_2 < k_1\) and \(res_i^s\) is the estimation of the pseudorange error at each time \(t\).
4.3. Results and Interpretations

The choice of DPM for the modeling of the observation errors is well justified and its suitability in the considered problem is proved. In our experiments, the pseudorange error density is tested with the DPM approach and is compared to a finite Gaussian Mixture (GM) based modeling. The GM and the DPM are used in a sequential Monte Carlo filtering approach. Table 1 shows the localization performances for the two models in terms of mean error and availability of the navigation solution (threshold of 3 meters).

Table 1. Comparison of positioning errors using DPM and GM

<table>
<thead>
<tr>
<th></th>
<th>Mean error</th>
<th>Min error</th>
<th>Max error</th>
<th>%error&lt;3m</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM</td>
<td>2.01</td>
<td>0.0014</td>
<td>8.284</td>
<td>89%</td>
</tr>
<tr>
<td>GM</td>
<td>6.14</td>
<td>0.0019</td>
<td>25.49</td>
<td>51%</td>
</tr>
</tbody>
</table>

In preliminary experiments, we considered a range of values in order to explore the extend to which our results are sensitive to different choices of $\alpha$. While the overall shape of the distribution remains the same across values of $\alpha$, higher values of $\alpha$ appear to produce an increased number of localized features and even indicate additional modes for a subset of the observations. The estimated means and variances of the model parameters are, however, quantitatively very similar. Furthermore, we find little gain from using large values of $\alpha$. In our DPM model, we used a gamma distribution as a prior for the $\alpha$. The major advantage of this choice is that with this prior the conditional distributions are easy to sample by applying the approach of section 3.1. Fig. 2 shows that estimating the $\alpha$ leads to a good positionning accuracy compared to the result obtained when we use a fixed $\alpha$. Practically, in most situations a considerable uncertainty exists about the probable value of the number of model clusters $k$. In our context, nor the prior mean neither the prior variance of the $k$ are known. Fig. 1 shows clusters and their locations for two satellites (signals from satellite 18 are mostly received in LOS and those of satellite 13 both in LOS and NLOS).

It seems to be extremely hard to have such informations on data in a real-world application. Therefore, the parameters a and b of the gamma prior were adjusted carefully. For example, we noticed that small values of a and b lead to nearly similar values of the $\alpha$ probability density, which means a lack of variability in the distribution of $\theta$.

![Fig. 1. Locations and numbers of clusters in the DPM for two different sensors (Satellite 13 and Satellite 18)](image)

![Fig. 2. Estimation of the pseudorange errors using DPM models with fixed and estimated values of $\alpha$)](image)

5. CONCLUSION

The DP prior is extremely flexible, it allows to model a huge variety of distributional forms. In real applications, its effectiveness is proved, since in most cases little is known about priors to be considered for the data distribution. In this paper, we illustrated through experiments that a good choice of the DPM hyperparameters leads to a better density estimation. The proposed density estimation solution was strongly motivated by the complexity of the considered data.

6. REFERENCES