Hacettepe Journal of Mathematics and Statistics
Volume 46 (3) (2017), 525 – 545

Nonparametric Bayesian approach to the detection of change point in statistical process control

Issah N. Suleiman*† and M. Akif Bakır‡

Abstract

This paper gives an intensive overview of nonparametric Bayesian model relevant to the determination of change point in a process control. We first introduce statistical process control and develop on it describing Bayesian parametric methods followed by the nonparametric Bayesian modeling based on Dirichlet process. This research proposes a new nonparametric Bayesian change point detection approach which in contrast to the Markov approach of Chib [6] uses the Dirichlet process prior to allow an integrative transition of probability from the posterior distribution. Although the Bayesian nonparametric technique on the mixture does not serve as an automated tool for the selection of the number of components in the finite mixture. The Bayesian nonparametric mixture shows a misspecification model properly which has been explained further in the methodology. This research shows the principal step-bystep algorithm using nonparametric Bayesian technique with the Dirichlet process prior defined on the distribution to the detection of change point. This approach can be further extended in the multivariate change point detection which will be studied in the near future.

Keywords: Nonparametric, Bayesian, Change point, Clustering, Mixture model, Dirichlet process.

2000 AMS Classification: 62F15

Received : 23.06.2016 *Accepted* : 10.09.2016 *Doi* : 10.15672/HJMS.2017.419

*Gazi University, Department of Statistics, 06500 T.Okullar-Ankara-Turkey Email: jazeerabay@gmail.com

†Corresponding Author.

‡Gazi University, Department of Statistics, 06500 T.Okullar-Ankara-Turkey, Email: mabakir@gazi.edu.tr

1. Overview

Change point is the detection of distributional changes in a time ordered observation which is essential in the statistical variation analysis. There are various work done on the detection of change point in statistical data analysis. The general approaches to the detection of variation in a time-ordered observation can be categorized into classical and Bayesian approach. One of the earliest classical approach to the detection of change point is the CUSUM-procedure by Page [14] which was further investigated by Lorden [10] and Moustakides [12]. On the Bayesian framework Shiryeav [19] was one of the first to present a model in a continuous time where the change point is assumed to be a random variable with some prior distribution.

Change point detection methods can be further categorized into parametric and nonparametric approach. The parametric approach incorporate knowledge of the data into the detection of the variation whereas the nonparametric makes no distributional assumption about the data. This research will focus mainly on nonparametric Bayesian approach for the change point detection.

The earlier research on change point detection as described in the survey article of Zack [24] considers where only one change point exist or more. Bayesian parametric approach as in Broemeling [4], Smith [20], [21], [22] and Cobb [7] uses parametric hypotheses to the change point estimation. Pettit [16] used ranks to determine the (approximate) posterior distribution of the change point.

Other comprehensive reviews are given in Bhattacharya [1] and in the book of Brodsky and Darkhovsky [3] on nonparametric models. Muliere and Scarsini [13] studied change point detection using the nonparametric Bayesian approach by computing the posterior distribution for the change point using the Bayes estimate with Ferguson-Dirichlet prior. They assumed the simplest case where F_1 and F_2 are priori independent. However this is not a realistic in many cases as the distribution might not necessarily be independent and unbounded. In this article we use the nonparametric Bayesian approach to detect change point by not restricting the distributions to be just priori distributed and also we develop that into the multiple change point case using the Dirichlet prior.

This paper focuses on using the Bayesian technique to nonparametric approach in the detection of change point in any statistical process. Bayesian nonparametric approach provides a Bayesian framework for model selection and adaptation using nonparametric model which is essential in clustering and as well change point detection. Nonparametric Bayesian model is a famous and a powerful procedures to many difficult statistical problems which includes clustering and change point detection. The key assumption which underlines nonparametric Bayesian model is the claim that sets of random variables are drawn from some unknown distribution. The principle goes on further to elaborate that, this unknown distribution by itself is also drawn from some prior distribution. In this research we take into account that the unknown distribution is drawn from a prior distribution which is of Dirichlet process. Some existing research uses the Pitman-Yor [17] process as the prior. The Pitman and Yor (1997) process which recently was introduced is based on a two-parameter generalization of the Dirichlet process. Theoretically by Bayesian we imply that, a prior is needed over the mixing distribution say G , and in our case the most common prior to use is Dirichlet process, DP. The essence of the Dirichlet process DP is parametrized by a concentration parameter $\alpha > 0$ with a base distribution H which is a prior over distribution (which is probability measure) G which implies that, for any finite partition $A_1, ..., A_n$ of any given parametric space, the induced random vector $(G(A_1), ..., G(A_n))$ is a Dirichlet distributed with the parameters $(\alpha H(A), ..., \alpha H(A))$. Bayesian nonparametric mixture uses mixing distribution which

consist of countably infinite number of atoms which can be represented as:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

This will be explained further in the methodology. The resulting solution gives rise to a mixture model which has an infinite components. The DP induces a distribution over partitions of integers which ensures that, samples drawn from a DP becomes discrete distributions as represented in Figure 1. In section 2 we will be discussing the nonparametric Bayesian approach to change point and build on it the hierarchical Bayesian nonparametric. We will then explain the clustering approach to change point and discuss why we opted to use a modified Dirichlet process as the prior by performing a full Bayesian.

2. Nonparametric Bayesian approach and change point

Control chart appears to be an effective tool for the monitoring of the process in statistical process control. In some process, multiple correlated quality characteristics are interested. In such cases, multivariate control charts are applied for the monitoring process. Control chart therefore in short, serves as structural display that helps us in the detection of changes that occur in the process. It does this by issuing an out-of-control signal. Although the time in which the out-of-control signal is given is not the real time of the occurrence of the change. It still serves as a key notice to the researcher as we are aware that it gives the signal with a delay which actually depends on the size of the shift. The real time of the change in the statistical process is termed as the change point.

In general, change point analysis is the process of detecting distributional changes that occur within some observed time-ordered sequence. Talih and Hengartner [23] expressed this as a statistical change that occurs in a financial modeling such that, the correlated assets are traded and modeled based on the historical data represented as multivariate time series. Change point analysis is also used in the detection of credit card fraud (Bolton and Hand [2]) and other anomalies. In practical applications, the applications of change point also be found in signal processing: where change point analysis can be used to detect significant changes within a stream of images (Kim et al., [9]).

In Bayesian statistics, the observed data are considered to be constant with an unknown parameter which is a random variable. The basic principle of Bayesian statistics is that any forms of uncertainty are represented as randomness. Assuming we define a random variables θ within the parametric space T, the interest here is to define some assumptions on how θ is distributed. This is usually achieved by choosing a specific distribution $Q = L(\theta)$ The distribution of Q is referred to as the prior distribution (prior). Then we can finally define our Bayesian model M as an observational model such that Q represents our prior. Generally, data is generated in two stages under the Bayesian model such that

$$\theta \sim Q$$

$$X_1, X_2, ..., \backslash\Theta$$

where the observed data here are conditionally iid rather than iid. Our objective here as Bayesian approach has always been determining our posterior distribution, which can be defined as the conditional distribution of Θ given the data,

$$Q[\Theta \in \cdot \backslash X_1, , ..., X_n = x_n]$$

which gives a different form of parameter estimation compared to the classical approach such that, the value of the parameter remains uncertain with a given finite number of observations. This uncertainty is expressed in the Bayesian scheme by the posterior distributions.

Therefore a nonparametric Bayesian model is a Bayesian model with an infinite dimensional space. In order to define our nonparametric Bayesian model we need to define our prior which is the probability distribution defined on the infinite-dimensional space.

3. Hierarchical Bayesian non-parametric model

Nonparametric models are simply statistical model technique to model selection and adaptation whose model size depends directly on the data size. This implies that, the sizes of the models are allowed to increase as the data size increases. This characteristics of the non-parametric model opposes the parametric model technique which uses a fixed number of parameters. The non-parametric methods have been very much used in the classical approach to statistical data analysis. Although the theoretical results for the nonparametric models are typically harder to prove than in the case of the parametric models. There had been some theoretical appealing features that have been stipulated for a wide range of models.

Bayesian nonparametric approach introduces the Bayesian framework for the model selection and adaptation through the principles of nonparametric models. The basics of Bayesian framework defines as the prior and the posterior on a single fixed parametric space but in case of the nonparametric model, the model size increases as the data size increases. This give rise to the non-triviality of the nonparametric problems in the case of the Bayesian formulation. The Nontrivial solution which comes as a result of the Bayesian approach to the nonparametric problem implies, the use of an infinite-dimensional parameter space, and to invoke only a finite subset from the parameters on any given finite data set. Such that, the subsets from the parameters increases as the data size increases. Therefore, Bayesian nonparametric models can be interpreted as "of finite but unbounded" since they are form "infinite-dimensional parametric space".

Hence a Bayesian nonparametric model is any model that introduce Bayesian framework on an infinitedimensional parametric space and can be analyzed on a finite sample in a manner that uses only finite subset of the parameters to explain the sample data. The Bayesian nonparametric model is used to fit a single model that can explain explicitly the complexity of the data sample or process.

Traditional mixture models which are generally used for the model fitting, group data into a pre-specified number of latent clusters. The Bayesian nonparametric mixture model, which in our case Dirichlet process mixture infers the number of clusters from the data and allows the number of clusters to grow as new data points are observed.

The hierarchical Dirichlet process (HDP) is a prior for Bayesian nonparametric mixed membership modeling of data groups. Hierarchically, it can be defined as

$$G_d \sim DP(\alpha, G_0)$$

where, α represents a nonparametric or a semi-parametric prior distribution, and G_0 denotes the base measure which is often taken to be a parametric distribution with its parameters endowed with prior distributions as well.

The hierarchical model here takes its definition from the prior and hyper prior from the features of the Dirichlet distribution which can be represented as follows:

$$\pi \backslash \sim Dirichlet\, \alpha \sim (\alpha/k, ..., \alpha/k)$$
$$\theta_k^* \backslash H \sim H$$
$$z_i \sim \pi$$
$$x_i \sim z_i,\ \varphi \sim F(\theta_{zi})$$

where,

π : the mixing proportion,

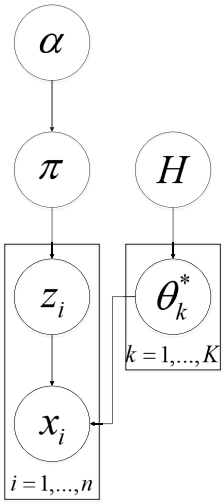


Figure 1. Dirichlet process mixture model

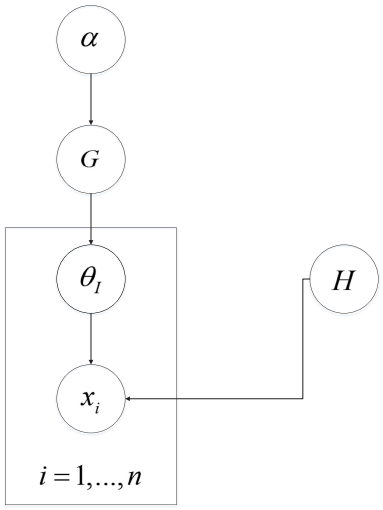


Figure 2. Dirichlet process mixture model

k : the number of clusters in each partition,
θ_k^* : the parameter in each of the partition of cluster k,
z_i : categorical (φ) or categorical distribution,
φ : the hyper-prior function with the parameter θ_{zi}.

Dirichlet process mixture model arises as infinite class cardinality limit uses: clustering and density estimation. This clusterial sequence of the prior can be represented diagrammatically in the Dirichlet mixture model as indicated in Figure 1.

The Figure 1 represents a Dirichlet process mixture model G with the parameters α and H.

$$G\backslash\alpha, H \sim DP(\alpha, H)$$

$$\theta_I\backslash G \sim G$$

$$x_I\backslash\theta_I \sim F(\theta_I)$$

denotes the sample drawn from the z_i with the parameter θ_I which is a hyper function (prior). Thus, the hierarchical distribution, $G_0 \sim DP(\gamma, H)$ where γ the parameter of the hyper- parametric function, can be defined from the property of the Dirichlet process as represented in Figure 2.

4. The clustering approach

The theoretical definition of clustering problem can be represented in the form: given a data set $X = \{X_1, X_2, ..., X_m\}$ and let k be any integer then we can define our clustering problem as a mapping function $f : X \to \{1, ..., k\}$. Using the Dirichlet process which is basically used in particularly data clustering. Assuming we observed the following observations in our systems $x_1, x_2, ..., x_n$ and our research objective here is to subdivide the samples into subsets (clusters). Which in our case will define the difference

in the observational reading which is as a result of the change detected in the reading? The observations within each cluster therefore should be mutually similar which in our change point detection case will imply within the specification limits. Recently there is a growing interest in the approaches to finding multiple clustering. These methods can be categorized as sequential (iterative) or simultaneous approach. The sequential approach finds an alternative clustering given that there exist one or more clusters in the process. However, the simultaneous approach tries to find multiple clusters simultaneously in the process. There are, however, a few recent work done on nonparametric Bayesian model regarding clustering and partitioning: the cross-categorization which utilizes a CRP-CRP and the Gibbs sampling (application of infinite models) for inference, and multiple cluster detection using the stick-breaking approach with a variational inference which allows the detection of the parameter of the model and the latent variable. Both methods assume that the feature in each clusters are disjoint and accordingly modelled as a partitioning problem with the constraint that $\sum_{\nu} y_{d,v} = 1$ (which implies the elements in each clusters belongs to ν). In our case we utilize the Dirichlet process for inference and multiple clustering (partitioning) using the sequential clustering approach. Suppose a process which has a known number of change points k in the sequence and the location of the change point is known. We can then define our change points $0 < \tau_1 < \tau_2 < ... < \tau_k < T$ which partitions the sequence into a $k+1$ clusters such that, the observed values within a particular cluster are have same characteristics (identically distributed) and observational values between different clusters are not identically distributed. Harchaoui and Cappe (2007), Rigaill [18] and Lung-Yut-Fong et al. [11] explained that, the simple and naive approach to the change point location estimation in this case will be quickly using computational intractable for $k \geq 3$.

Change point have been extensively been analyzed in the course of identifying any structural change in a sequence of time series observations mostly when the data are of known parametric form. In this report, we seek to use methods of exploration through the use of clustering techniques to detect change points in nonparametric Bayesian settings. Such that, no assumptions are made with regards to the distributional structure of the observed data. Many of the literatures written on Change points focuses primarily on the offline setting where inferences are made regarding the detection of changes by retrospectively study. But then the online setting which infers to making a sequential analysis with respect to every new observation received. The detection of distributional changes is essential as it aids in locating possible change points in previous observations. The nonparametric change point technique makes no distributional assumptions regarding the data. This research focuses entirely on using the nonparametric Bayesian approach to clustering in detecting the change point. An overview of parametric offline techniques can be found in Eckley et al. [8]. The figures below represent the output of a simulated data which shows the occurrence of change point as a result of various factors.

In Figure 3 the means of the simulated data was varied and simulated which clearly gives a signal of the variation in mean at the point 200 as can be seen from the output above. Also in Figure 4 the variance of the data samples was varied and the output from there equally shows the occurrence of change point but this time as a results of change in variance. A similar analysis was done when the data sample was simulated from different distributions and also varying regression pattern as in Figure 6 and Figure 5 respectively, in which case the variation as a results of change point occurring was equally observed. In summary what we want to explain is that, change point can occur as a result of many factors among which is explain using the Figures 3-6.

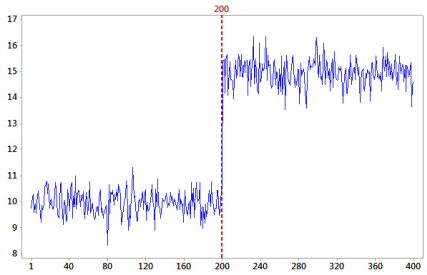


Figure 3. Change in the mean.

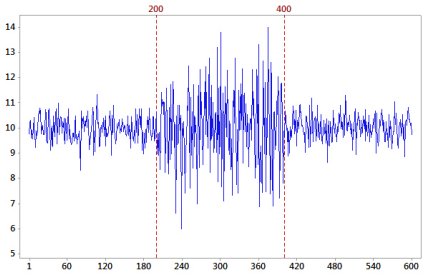


Figure 4. Change in the variance.

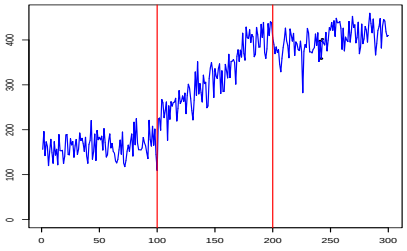


Figure 5. Change in the regression

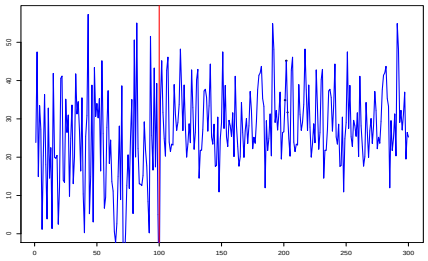


Figure 6. Change in the distribution

5. Dirichlet process mixture models

The nonparametric approach to clustering here is by performing a full Bayesian on the model by considering the prior distribution to be a Dirichlet process. To go about this method of clustering using nonparametric Bayesian approach, we need to define two basic characteristics:

- The likelihood term (how the data in hand is affected by the supposed parameters). We recall that in the nonparametric model we usually consider the parameter θ as a function.

  Assuming we consider a density estimation problem where the observed data $y_i \sim G, i = 1, ..., n$ and G is an infinite dimensional distribution. Inference under the Bayesian paradigm usually defines the completion by specifying a prior for the unknown distribution G. Then we define our Likelihood term as: $p(Y \backslash \theta)$
- With these we can define our prior distribution on the supposed parametric θ as: $p(\theta)$ where our prior here is defined in terms of Dirichlet process.

Therefore, we adapt the mixture model approach to clustering of the partition and performing a full Bayesian on each clusters. Mixture model by it definition is a probabilistic model use to represent the presence of subdivision of elements within an overall population without the need for any observed data set to identify the sub-division to which it belongs to. In mixture models, most of the algorithms approaches to clustering appears to require the need for the number of data clusters to be known in the aim

of selecting an approximate number. Nevertheless, the Dirichlet process mixture model (DPMM) happens to provide a new platform of using non-parametric Bayesian framework to explicitly describe distributions over mixture models with an infinite number of mixture components. By definition, a Dirichlet process (DP) is parameterized by a base distribution G_0 and a concentration parameter or semi-parameter α. Where α is used as a prior over the distribution G from the mixture components such that for any observed data points X_i we can define the DPMM as

$$G\backslash\alpha, G_0 \sim DP(\alpha, G_0)$$

$$\theta_i\backslash G \sim G$$

$$x_i\backslash\theta_i \sim F(\theta_i)$$

6. Posterior distribution

Let $G \sim DP(\alpha, H)$ be any distribution (random) over the space θ such that $\theta_1, \theta_2, ..., \theta_n$ represents an independent sequence drawn from the distribution G. Then, we can easily infer that the independent sequence θ_i takes values from the probability space Θ once G is a distribution over the space Θ. Our research interest here will be to determine the posterior distribution of G given the observed values $\theta_1, \theta_2, ..., \theta_n$. Letting $A_1, A_2, ..., A_n$ represents any finite measureable partitions on the space Θ such that $n_k = \#\{i : \theta_i \in A_k\}$ also represents the number of observed values of A_k. The essence of partitioning is to use the sequential approach to change point detection which allows a step-by-step finding of change point given an existing one in the space θ (this is explained further in the methodology).

In a very naive example assuming we have a sample observations which is observed from a control system, then we can easily simulate this using a sample data containing $n = 100$.

7. Methodology

Change point analysis is the process of detecting distributional changes that occur within some observed time-ordered sequence. We can start laying down our methodology by assuming that $Z_1, Z_2, ..., Z_\tau \overset{iid}{\sim} F_1$ and $Z_{\tau+1}, Z_{\tau+2}, ..., Z_T \overset{iid}{\sim} F_2$ represents any independent sequence of time-ordered observations. We would assume throughout this research that, the time between the observations are non-negative and non-null (i.e. is fixed or random) such that, our time index will imply time-order.

Now considering a simple (single change point) case, we can always hypothesize this system with a single change point location say τ. Letting $Z_1, Z_2, ..., Z_\tau \overset{iid}{\sim} F_1$ and $Z_{\tau+1}, Z_{\tau+2}, ..., Z_T \overset{iid}{\sim} F_2$ representing the unknown probability distributions of the two distributions. In a simple case, we can briefly describe the distributions of these two different observation. The hypothesis here will always be to test for the homogeneity in the two distributions ($H_0 : F_1 = F_2$ vs $H_0 : F_1 \neq F_2$). Now theoretically, since the distributions show a continuous nature in the univariate observational form, we can apply the Kolmogorov-Smirnov test. So using this test, if our H_0 is rejected then we can conclude by saying there is evidence of change point in the process at τ otherwise we conclude therefore that there is no distributional difference in F_1 and F_2. The above case is the setting assume for the case where the change point locaton is known, but we can modify this setting by assuming instead that the change point location is unknown but assumes that at most there exit one change point. In this case, the natural way to proceed to this is choosing as any possible location of the change point based on some

statistical criterion. We perform the test hypothesis on the homogeneity of the two distributions by defining our as a subset from the sets $\{1, 2, ..., T-1\}$. This should necessarily incorporate the fact that is unknown.

In a different case, assuming a process which has a known number of change points k in the sequence and the location of the change point is known. We can then define our change points $0 < \tau_1, < \tau_1, ... < \tau_k < T$ which partitions the sequence into a $k+1$ clusters such that, the observed values within a particular cluster are have same characteristics (identically distributed) and observational values between different clusters are not identically distributed. Cappe and Harchaoui [5], Rigaill [18] and Lung-Yut-Fong et al. [11] explained that, the simple and naive approach to the change point location estimation in this case will be quickly using computational intractable for $k \geq 3$. Another remedy is maximizing the objective function through the use of dynamic programming. In a more general case, assuming both the number of change point and as well their respective locations are unknown. In situation, a naive way to estimating will be infeasible. We therefore deploy the bisection and model selection techniques which obvious are the popular technique under these conditions.

Now, as the aim of this research is applying the principle of Nonparametric Bayesian technique using the Dirichlet process prior techniques to change point analysis. Although the Bayesian nonparametric technique on the mixture does not serve as an automated tool for the selection of the number of components in the finite mixture. The Bayesian nonparametric mixture shows a misspecification model property. To explain this better, we can perform the Dirichlet process on the data sample of a given size n, then for any clustering solution which is supported by the posterior (as Bayesian) and as well for any corresponding finite and random number of clusters in the process. We therefore can define the posterior distribution on the number of clusters though it does not imply model selection technique (since there is a single model involve). By this, the possible values of the clusters are assumed to be mutually exclusive such that we can simple assume a solution for the number of clusters. However, for a Dirichlet process we use a random measure on an infinite number of sequences. By this we mean, the model assumption implicit in a DP mixture is that as $n \to \infty$, *we will surely observe an infinite number of clusters.*

Bayesian non-parametric or semi parametric frameworks have always been useful in many statistical applications especially with the clustering techniques. We can start explain this further by looking at the Dirichlet process mixture models.

8. Dirichlet process application

An intuitive description of the Dirichlet process as in infinite dimensional generalization of the Dirichlet distribution can be made by first considering it in a form of Bayesian mixture model consisting of K components:

$$\pi \backslash \alpha \sim Dir\left(\frac{\alpha}{k}\right), ..., \left(\frac{\alpha}{k}\right) \qquad \theta_k^* \backslash H \sim H$$

$$z_i \backslash \pi \sim Mult(\pi) \qquad x_i \backslash z_i, \{\theta_k^*\} - F(\theta_{z_i}^*)$$

where π is the mixing proportion, α is nonparametric or a semi-parametric prior distribution and G_0 base measure.

We can define the Dirichlet process on a space measure in respect with our observed data with an infinitedimensional parametric space. Let (x, Ω) represent any measurable space with the measure $\mu = \alpha G_0$ (unnormalized density) on a finite, additive, nonnegative and non-null. Then we say that a random probability measure ρ^μ on (x, Ω) is Dirichlet process with parameter μ if the following conditions are satisfied: whenever $\{A_1, A_2, ..., A_K\}$ is a measureable partition on the space $\mathbf{\Omega}$ (i.e. each of the partitions

$\mu(B_k) > 0$ for $\forall k$), then the joint distribution of the random probabilities of the partitions can be expressed as $\rho^{\mu}(A_1), ..., \rho^{\mu}(A_k)$ distributed according to standard Dirichlet distribution $\mu(A_1), ..., \mu(A_k)$ Ferguson (1973-1974). That is to say, ρ^{μ} is a Dirichlet process takes it characteristics from the Dirichlet distribution on any finite partition of the original space.

Generally, most of the popularly known clustering algorithms require the number of the data clusters to be known *a priori* or used as heuristics in the selecting of an approximated number of clusters. Nevertheless, the Dirichlet process mixture models (DP-MMs) provides a new technique of using the non-parametric Bayesian framework to define the distribution over mixture models with an infinite number of mixture components.

The Dirichlet process has identified as a commonly-used prior distribution for a process with an unknown probability distribution.

$$F(.) \sim DP(\theta, F_0)$$

where F_0 is a probability measure which represents the prior belief in the distribution F with the weighted parameter θ (which equally represents the degree of the belief in the prior from F_0. Here the essence of the Dirichlet process is that, it induces a discretized posterior distribution which serves well in our Bayesian framework. Ferguson (1974), using a Dirichlet process, the distribution $DP(\theta, F_0)$ prior for $F(.)$ results in a posterior mixture of F_0 and point masses of unknown observations X_i:

$$F(.)\backslash X_1, X_2, ..., X_n \sim DP\left(\theta + n, F_0 + \sum_{i=1}^{n}\right)$$

Such that, density estimation is done with convolution with kernel functions to produce a continuous density estimate instead of discreteness though it is not a disadvantage in other statistical applications. We can therefore explain the detection of change point using the posterior distribution under the Dirichlet process prior in the next section.

9. Simple change point

The general idea of using the Bayesian framework in making inference for a simple change-point problem can be easily explained as:

$$\theta \sim \pi$$

$$(X_1, ..., X_n)\theta \sim f(x\backslash\theta)$$

Such that the posterior is used to compute the point estimator which represents the posterior mean θ. The posterior can then be plotted by drawing a large sample of $\theta_1, ..., \theta_n$ from the posterior $\pi(\theta\backslash X)$.

Assuming we have any randomly distributed data $(X_1, X_2, ..., X_n)$ rom any Dirichlet process which is conditionally *i.i.d.* according to d.f. F such that f is a DP whose parameter is also a measure and that $X_1, X_2, ..., X_n\backslash F$ are are from any nonparametric distribution which can be expressed in the form:

$$X_i \sim F_1 \qquad i = 1, ..., C$$

$$X_i \sim F_2, \qquad i = C + 1, ..., n$$

where C is the change point and unknown. Nonparametric Bayesian inference is carried out by using the Dirichlet process priors to F_1 and F_2 . We can therefore write the model in the form:

$$P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n\backslash C, F_1, F_2)$$

where C here is the unknown change point. This implies that, the observations $(X_1, X_2, ..., X_n)$ are conditionally *i.i.d* according to F_1 up to the time C and *i.i.d* according to F_1 from

time $(C+1)$. Now we can then make the inference that, if $C=0$ or $C=n$ then there is no change point in the distribution and hence the Data are conditionally $i.i.d$ as F_2 ya da F_1 accordingly.

Cifarelli and Regazzini (1978), the prior assumes that C and (F_1, F_2) are independently distributed with probability mass function $p(c)$ and (F_1, F_2) as a mixture of products of Dirichlet processes. The prior distribution can then be expressed in the form,

$$(F_1, F_2, C) \sim p(C) \int \mathfrak{Y}(\alpha_1(.;\theta_1))\mathfrak{Y}(\alpha_2(.;\theta_2)dH(\theta_1,\theta_2)$$

Muliere and Scarsini (1985) studied a special case of this model assuming the distributions F_1, F_2 to be independent. Mira and Petrone [15] also by the application of the Gibbs sampler algorithm approximated the posterior distribution in the above model. This can therefore be shown that, when $\alpha_1(.;\theta_1)$ and $\alpha_2(.;\theta_2)$ with densities $M_2 f_2 \alpha_1(.;\theta_1)$ and $M_1 f_1 \alpha_1(.;\theta_2)$ respectively, as explain above using the probability measure theory, the likelihood function is:

$$f(x_1, x_2, ..., x_n \backslash c, \theta_1, \theta_2) = \frac{1}{M_1^{[r]}} \prod_{i=1}^{c^*} M_1 f_1(.;\theta_1) \frac{1}{M_2^{[n-1]}} \prod_{i=c+1}^{n^*} M_2 f_2(.;\theta_2)$$

Here the sign $*$ implies taking the product over a distinct values only. We can therefore make an inference that, the posterior distribution of (C, θ_1, θ_2) can be computed using the Bayes theorem. In particular in the continuous case and if the observations $(x_1, x_2, ..., x_n)$ are distinct, we can therefore define the mass function condition on the distinct observations,

$$p(c \backslash x_1, ..., x_n) \propto k(c, m_1, M_2, n) I(c) p(c)$$

where, for $c = 0, 1, ..., n$ such that, we can define the indicator function of c is,

$$I(c) = \int \prod_{i=1}^{c} f_1(x_i;\theta_1) \prod_{i=c+1}^{n} f_2(x_i;\theta_2) dH(\theta_1,\theta_2)$$

such that

$$k(c, M_1, M_2, n) = \frac{M_1^c M_2^{[n-c]}}{M_1^{[c]} M_2^{[n-c]}}$$

where this expression represents the tie factor within the observations. We will then apply the discussed method in a case study using a simulated data.

10. Case study

In this section we will illustrate the algorithms for the detection of change point using the Bayesian nonparametric approach. We will be using command tool in our case study.

We will begin by considering the simple case for the detection of variation of changes in the system or distribution. For this we randomly generate 100 independent observed sample from 5 samples $(x_1, x_2, x_3, x_4, x_5)$. We will try to clarify by starting with the parametric approach then setup the procedure for the nonparametric Bayesian.

Commands

```
% Algorithm of the Nonparametric Bayesian approach to Change point
% estimation assuming (X1,... , Xn) are any randomly distributed data from
% nonparametric distribution. Such that;
%  X1 ,...,Xn each of the samples with size n=20
```

```
% Simple Case

%  H0: X1∼F1 i=1,...,c
%  H1: X2∼F2 i=c+1,...,n

% where c represents our change point.
% This implies, performing Bayesian on the two distribution to make
% inference on the distribution.
% Now;
% nonparametric Bayesian infers the use of Dirichlet process. This means
% that; for any measure space with a finite partition {A1,...,An} we can
% define our DP on the distribution such that;

%  G~DP(?,G0)

% (G(A1),...,G(An)) ; which means defining DP on the partitions.
% By Definition: Xi\theta~F(x\theta_i), theta~G, G~DP(alpha,G0)

% So here by DP we will define a distribution on the Prior as;
% P(theta)=Beta(a,b) such that for any uniform distribution a=b=1, P(theta)
    =Beta(1,1)

N = 100
n1=20
n2=20
n3=20
n4=20
n5=20
mu1=10
mu2=10
mu3=10
sigma = 1

Lambda= 4

x1=normrnd(mu1,sigma,1,n1)
x2=normrnd(mu2,sigma,1,n2)
x3=normrnd(mu3,sigma,1,n3)
x4=poissrnd(Lambda,1,n4)
x5=poissrnd(Lambda,1,n5)
```

We can then run a distribution fit tool command to calculate the Statistics about the data (e.g.parameter estimation, Likelihood and Log-Likelihood). Now considering a simple case of x_1 and x_2 we run a dfittool and find the descriptive statistics of x_1 and x_2 from the manage fit option. Since we will be interested in comparing our nonparametric inference with the parametric inference, we therefore use Bayes Factor to make inference about the parameter. Hence we can estimate our posterior from the distribution using the likelihood estimation from the "dfittool" for both x_1 and x_2 from the command

```
L1=Likelihood of x1
```

```
L2=Likelihood of x2
```

With similar approach let us conduct an application with the simulated data. We intend to apply the algorithm explained above in this report to detect whether or not the data set or samples which in this case $(x_1, x_2, x_3, x_4, x_5)$ are from the same distribution and use that as a basics to conclude that the variation that is between the samples are due to a change point which is resulted from the samples coming from different distribution hence different parametric values.

Using the practical examples from this report, assuming $(x_1, x_2, x_3, x_4, x_5)$ are drawn from any simulated values such that, we define a DP prior distribution on our parameter. We can determine our Posterior Distribution from the expression

$$f(D\backslash x) = f(x\backslash\theta)f(\theta)/f(x)$$

Such that the expression $f(x)$ defines our evidence probability which can be define in the form:

$$f(x) = Df(x\backslash\theta)f(\theta)d(\theta)$$

This can be determined from the expression below from the command:

```
g= Likelihood * Prior
```

which in this case our Likelihood function is calculated from the samples $(x_1, x_2, x_3, x_4, x_5)$

$$Likelihood = g = f(x\backslash\theta)f() = f(x_1, x_2\backslash\theta)f(\theta)$$

where $f(x\backslash\theta) = Likelihood$ and $f(\theta) = B[0.3, 0.3]$.

Since our observed data samples are from a multinomial distribution whose prior is defined on Dirichlet process then we expect our posterior distribution as well to be a DP from the Lemma of Dirichlet being a conjugate to multinomial observations. As in our case study, since our data sets are derived from a multinomial distribution such that,

$$D \sim c\backslash p$$

where $p(c = j\backslash\theta) = \theta_j$. Then the posterior is equally a Dirichlet which can be expressed in the form:

$$p(\theta/c = j, \alpha) = \frac{p(c = j/\theta)p(\theta/\alpha)}{p(c = j/\alpha)} = Dir(\alpha)$$

Hence the Posterior can be observed to also be a DP. As in our example, for which is the hyper parameter (prior parameter). Suppose our data set is represented by $\{1, 1, 1, 2, 2\}$ such that, we observed the samples $c = \{1, 2\}$ from the distribution with the prior $\alpha = \{0.3, 0.3\}$.

Then we can define our Posterior distribution as:

$$p(\theta/c = j, \alpha) = \{3.3, 2.3\}$$

Which is equally a Dirichlet process indicating the occurrence of clustering in the process at those particular points.

If we have two distributions A_1 and A_2, then we can compare the marginal likelihoods of each, i.e., compare $Pr(\{x_i\}|A_1)$ to $Pr(\{x_i\}|A_2)$ and ask which is better (larger) or even different. Or, if we have more than two distributions, we can compute the marginal likelihoods of each and ask which among the set is the largest. The whole point of marginalization is to eliminate the effect that different numbers of parameters have (that is, distributions with more parameters are more "complex" and thus more flexible, but, as a result, they assign lower likelihoods to all the data sets they can generate). In contrast, simpler distributions assign higher likelihoods to a smaller range of data sets,

and thus should win under this kind of distribution comparison. Thus, marginalization in the Bayesian framework is a kind of formalization of Occam's razor. Bayes factors are a slight variation on the general marginalization approach that makes the procedure look a lot like a likelihood ratio test. That is, a Bayes factor is the ratio of marginal likelihood of A_1 to that ofA_2:

```
% K=(Pr({xi} | A1))/(Pr({xi} | A2))
% The interpretation of Bayes factors is done by heuristic:
% if K > 1 then the result is interpreted as strong support for A1, while
% if
% K < 1, we rule in favor of A2.
% If K = 1,
```

Then we say that we cannot explicitly make a decision about the distribution or we cannot tell which of the distribution is better. This is one place where a true likelihood ratio test offers an advantage over the Bayesian approach: in a likelihood ratio test, by assuming that the data are random variables and thus that the likelihoods (or marginal likelihoods) are also random variables, the likelihood ratio test can quantitatively estimate just how close to 1 is "too close to call".

As in this research it was observed that, when the mean of the samples x_1, x_2 and x_3 the decision or hypothesis H_0 cannot be rejected, which implies that, the samples x_1, x_2 and x_3 are obtained from the same distribution while in the case of x_4 and x_5 when the theta was slightly varied, there was a clear observation of evidence rejecting H_0 implying that x_4 and x_5 unlike x_1, x_2 and x_3 are not from the same distribution and hence they are independently distributed. Below are the commands use for the hypothetical inference.

```
% For z2
n2=20;  %number of samples to collect
s=1;  % assume standard deviation s=1
mu2;  % assume the mean mu2=10
time=-2:0.1:5;   %the time series interval
L=zeros(1,length(time));  %Place holder for likelihoods
% Calculate Likelihoods for each parameter value in the range
L2 = exp(-sum((x2-mu1).^2)/(2*s^2))
% neglect the constant term (1/(sqrt(2*pi)*sigma))^N as it will pull down
% the likelihood value to zero for increasing value of N
[maxL,index]=max(L);  %Select the parameter value with Maximum Likelihood
display('Maximum Likelihood of A');
display(time(index));
% Bayes Factor;
K=(Pr({xi} | A1))/(Pr({xi} | A2))
B1= L1/L2
n3=20;    %Number of Samples to collect
% the process is repeated for B2, B3, B4 and B5
```

The result no doubt affirms the theoretical claims. Inferring that, the observed sample data which were from the same distribution shown a positive test for the null hypothesis which claims the observed sample data are from the same distribution. Hence no change point detected. Whiles in the other cases where the sample were randomly simulated from an unknown distribution, we observed a sudden change in the pattern showing clearly the presence of change point. The Table 1 shows the values of the simulated data

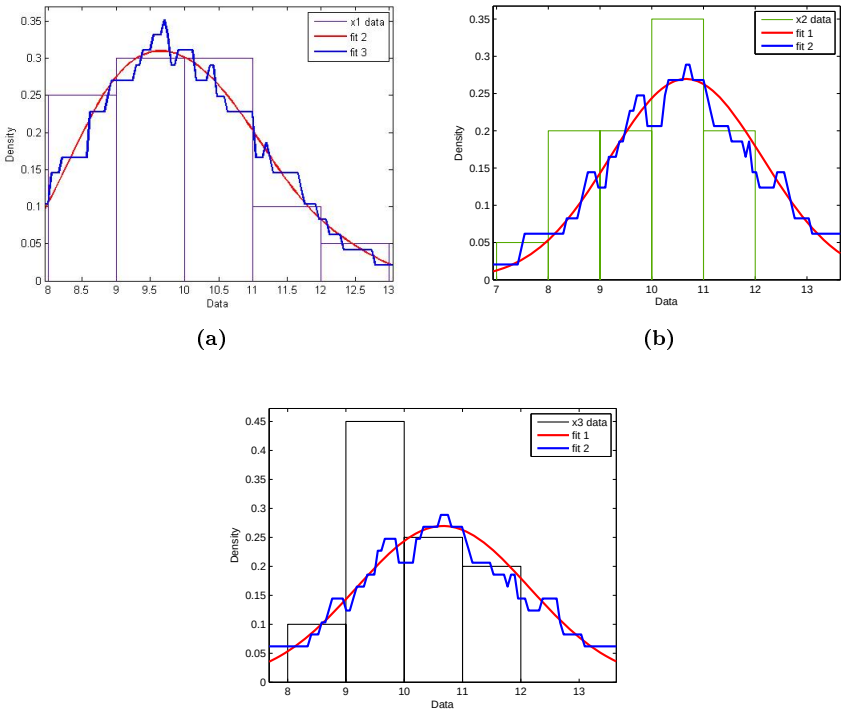
from the command from a normal distribution with $n = 20$ observed data. A cumulative hazard graph is plotted from the simulated data $(x_1, x_2, x_3, x_4, x_5)$. The essence of the cumulative hazard graph is to observe how the observed data points behaves with respect to the specification limits (cluster: whether it belongs to the same cluster) and the line of fit which is the targeted value (the targeted value here infer to the change point). The specification limits are the limits within which we define the quality of a process or observed data points (same cluster). In this case it shows whether or not the observed data points comes from similar distribution or not. A density function graph is equally plotted to observe the normality of the simulated data points. We observed that the simulated data points (x_1, x_2, x_3) were normally distributed whereas (x_4, x_5) showed a nonparametric behavior.

Table 1. The simulated sample of (x_1, x_2, x_3)

Sample no	x_1	x_2	x_3
1	8.9109	10.6715	9.1315
2	10.0326	8.7925	9.9699
3	10.5525	10.7172	9.8351
4	11.1006	11.6302	10.6277
5	11.5442	10.4889	11.0933
6	10.0859	11.0347	11.1093
7	8.5084	10.7269	9.1363
8	9.2577	9.6966	10.0774
9	8.9384	10.2939	8.7859
10	12.3505	9.2127	9.9932
11	9.3844	10.884	11.5326
12	10.7481	8..529	8.7859
13	9.8076	8.9311	8.8865
14	10.8886	9.1905	9.9932
15	9.2352	7.0557	11.5326
16	8.5977	11.4384	9.2303
17	8.5776	10.3252	10.3714
18	10.4882	9.2451	9.0744
19	9.8226	11.3703	9.7744
20	9.8039	8.885	11.1174

The Figure 7 represents the empirical density functions for the simulated sample (x_1, x_2, x_3), where (a), (b) and (c) are the density functions of (x_1, x_2, x_3) respectively. A density function graphs of (x_1, x_2, x_3) clearly show that they are all normally distributed about the mean.

The Figure 8 shows the cumulative hazard graph. The essence of the cumulative hazard graph is to observe how the observed data points (x_1, x_2, x_3) fair with the specification limit and the line of fit which is the targeted value (the targeted value here infer to the change point). The specification limits are the limit within which we define the quality of a process or observed data points. In this case it shows whether or not the observed data points comes from similar distribution or not. The Figure 8 shows a cumulative hazard of (x_1, x_2, x_3) which clearly with the red line representing the line of fit clearing showing that (x_1, x_2, x_3) falls within our specification limit. We can equally infer from this that (x_1, x_2, x_3) appears from the same cluster and hence from the same distribution.



(a) (b)

(c)

Figure 7. The empirical density functions of (x_1, x_2, x_3).

The Figure 9 represents the empirical density functions for the simulated sample (x_4, x_5), where (a), (b) and are the density functions of (x_4, x_5) respectively.

Density function of (x_4, x_5) indicating clearly nonparametric characteristics because the distribution is far from the targeted value (change point). The Figure 9 shows clearly that the sample is not normally distributed as in the case of Figure 7.

The Figure 10 shows a cumulative hazard of the samples, (x_4, x_5) which clearly can be observed to have deviated from the targeted value. The line of fit (change point) which can be seen in the brown line which show the deviation of the samples (x_4, x_5) from the confidence bounds (cluster).

Figure 11 shows the probability plot for the simulated samples $(x_1, x_2, x_3, x_4, x_5)$.

We can easily observed that, there are clearly two clusters which can be seen. These clusters are (x_1, x_2, x_3) and (x_4, x_5). Hence we can easily conclude that the data in each clusters are have the same characteristics and hence the same distribution but data across clusters are different and hence from distinct distributions.

11. Contribution

Statistically, there are some cases that, the data observed often do not conform to the normality assumption. In such cases, using the parametric model in estimation usually might not be the right method to the data analysis.

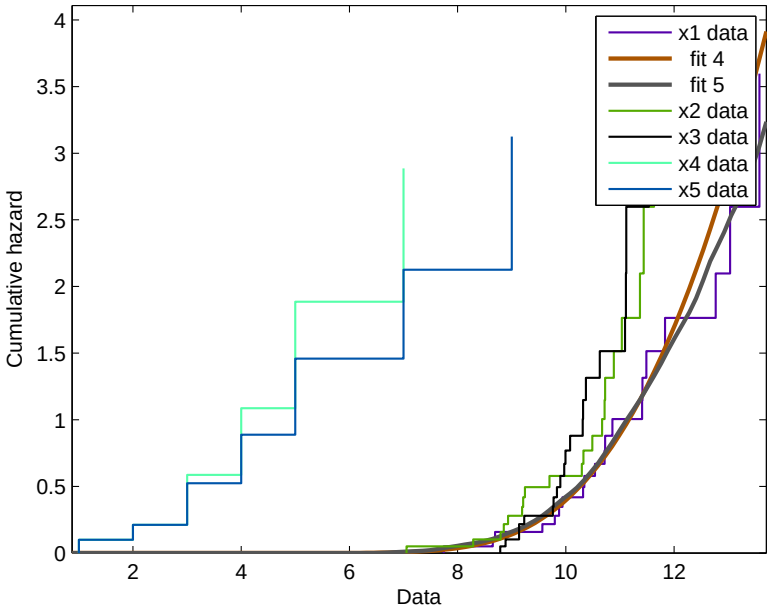
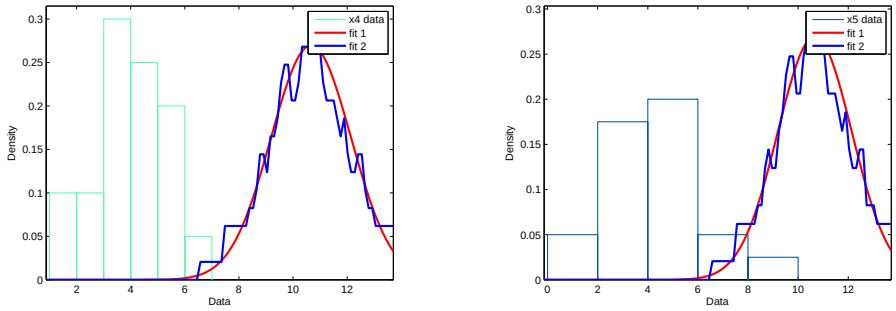


Figure 8. The hazard function of (x_1, x_2, x_3)

Table 2. The simulated sample of (x_4, x_5)

Sample no	x_4	x_5
1	5	2
2	4	2
3	2	3
4	7	3
5	6	5
6	4	4
7	4	5
8	4	6
9	3	4
10	3	6
11	1	7
12	3	5
13	5	4
14	3	3
15	3	5
16	5	5
17	0	8
18	4	3
19	4	8
20	4	4



(a) (b)

Figure 9. The empirical density functions of (x_4, x_5).

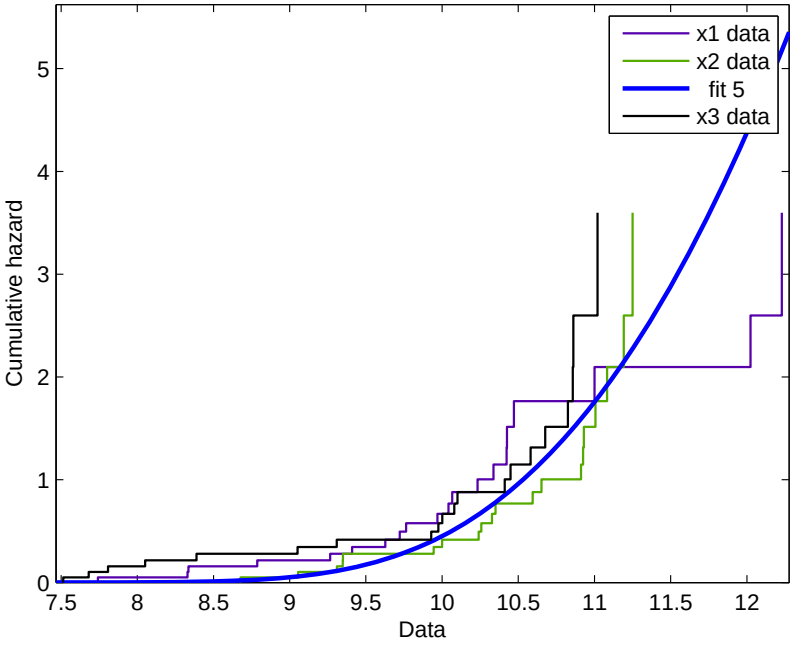


Figure 10. The cumulative hazard of (x_4, x_5)

Here the Dirichlet process distribution appears well in the model analysis. That is to say, in many cases, extreme values are more likely than would be dictated by a normality assumption. This is mostly observed when the data is from a financial source. The aim of this research is applying the principle of nonparametric Bayesian technique using the Dirichlet process prior techniques to detect the change point. Although the Bayesian nonparametric technique on the mixture does not serve as an automated tool for the

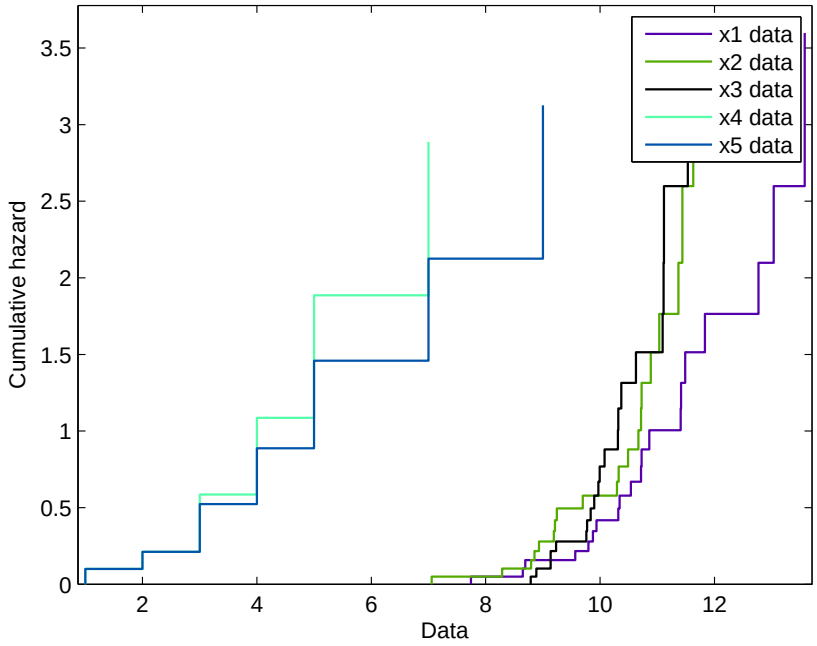


Figure 11. The probability plot for the simulated samples $(x_1, x_2, x_3, x_4, x_5)$

selection of the number of components in the finite mixture. The Bayesian nonparametric mixture shows a misspecification model properly which has been explained further in the methodology.

Considerable research has been done relating to statistical process control aiming to enhance the quality of product and also reduce the variability in the statistical results or outcome in the form of industrial, clinical or economic produce. A lot of these researches adopt the assumption that the observations from a multivariate process are independent. But then, there is no doubt about the increase in the rate of control chart giving false alarm in the presence of an autocorrelation.

This research aims at using the Bayesian nonparametric approach to statistical analysis in detecting any variation that occur in control chart and making interpretations using the outcomes with some practical examples. The researcher achieves this aim by taking a critical consideration of theories about change point detection and extending it to the Bayesian nonparametric concept where the observed data are an unknown distribution or violates the normality assumption. The researcher updates the posterior distribution using the likelihood function putting in mind the conjugate Dirichlet priors. Such that if the distribution and the prior distribution is from a particular distribution then the posterior distribution will equally be from same distribution (example if the distribution comes from a multinomial distribution with the prior distribution equally defined from the multinomial distribution then the Posterior will equally be a multinomial distribution).

Researchers for many years have worked on finding more efficient techniques from the use of these paradigms in designing models that can be used analytically in developing

a more scheme for a better inference and decision making. A lot of these research focus on how to define the limits of the control chart using variety methods. The use of the control chart serve well for both statisticians and companies as it give a much better and more effective way for statisticians to make a good inference and equally serves as an effective ways for companies to maintain and improve product quality and standard.

12. Conclusion

Parametric model operates on restrictive assumptions that does not confine with many real world process and applications that exhibit behaviours that are not well defined. It is often difficult to justify the use of parametrric models in changees detection in a process or system due this reasons. This paper deals with procedures that can be adapted to detect change points without making any restrictive assumptions. We have therefore been able to affirm that the nonparametirc Bayesian approach gives a statistical approach to the detection of any variation that occur in any statistical process by making an interpretations using the outcomes with some practical examples from a simulated values. We achieved this aim by taking a critical consideration of change point theories and extended that in the Bayesian nonparametric concept. The posterior distribution was also updated using the likelihood function putting in mind the property of conjugacy in Dirichlet priors. Giving the step-by-step Dirichlet process approach using clustering in model mixture but allowing the data to determine the complexity of the model, we clearly oberved that, the nonparametric Bayesian approach can be used to detect the occurance of change point in statistical process. I will suggest a further study and expansion of this field in the multivariate case as this study is carried out in the univariate case.

References

[1] Bhattacharya *Some aspects of change-point analysis*,In Change Point Problems E. Carlstein, H.G. Muller and D. Siegmund (eds.), (IMS Lecture Notes-Monograph Series, 1994) **23**, 28-56.

[2] Bolton, R. and Hand, D. *Statistical Fraud Detection: A Review*, Statistical Science, **17**, 235-225, 2002.

[3] Brodsky, B.E. and Darkhovsky B.S. *Nonparametric methods in change-point problems*, (Kluwer Academic Publ., The Netherlands, 1993).

[4] Broemeling, L.D. *Bayesian procedures for detecting a change in a sequence of random variables*, Metron , **30**, 214-227, 1972.

[5] Cappe, O. and Harchaoui, Z. *Retrospective multiple change-point estimation with kernels*, IEEE Computer Society Statistical Signal Processing, IEEE/SP Workshop on, 768-772, 2007.

[6] Chib, S. *Estimation and comparison of multiple change-point models*, Journal of Econometrics, **86**, 221-241, 1998.

[7] Cobb, G.W. *The problem of the Nile*, Biometrika, **65**, 243-251, 1978.

[8] Eckley, I.A., Fearnhead, P. and Killick, R. *Analysis of Changepoint Models*, (Cambridge University Press, In D Barber, AT Cemgil, S Chiappa (eds.), Bayesian Time Series Models, 2011).

[9] Kim, A.Y. et al. *Using labeled data to evaluate change detectors in a multivariate streaming environment*, Signal Process, **89**, 2529-2536, 2009.

[10] Lorden, G. *Procedures for Reacting to a Change in Distribution*, The Annals of Mathematical Statistics, **42**(6), 1897-1908, 1971.

[11] Lung-Yut-Fong, A., Levy-Leduc, C., and Cappe, *Homogeneity and Change-Point Detection Tests for Multivariate Data Using Rank Statistics*, arXiv, 1107.1971, 2011.

[12] Moustakides G.V. *Optimal stopping times for detecting changes in distributions*, Ann. Statist. **14**(4), 1379-1387, 1986.

[13] Muliere P. and Scarsini M. *Change-point problems: A Bayesian nonparametric approach*, Aplikace Matematiky, **30**, 397-402, 1985.

[14] Page, E. *Continuous Inspection Schemes*, Biometrika, **14**, 100-115, 1954.

[15] Petrone S. and Raftery, A.E. *A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability*, Statistics & Probability Letters, **36**, 69-83, 1997.

[16] Pettit, A.N. *Posterior probabilities for a change-point using ranks*, Biometrika, **68**, 443-450, 1981.

[17] Pitman, J. and Yor, M. *The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator*, Ann. Probab., **25**(2), 855-900, 1997.

[18] Rigaill, G. *Pruned Dynamic Programming for Optimal Multiple Change-Point Detection*, arXiv:1004.0887, 2010.

[19] Shiryaev, A.N. *On optimum methods in quickest detection problems*, Theory of Probability and Its Applications, **8**, 22-46, 1963.

[20] Smith, A.F.M. *A Bayesian approach to inference about a change point in a sequence of random variables*, Biometrika , **62**, 407-416, 1975.

[21] Smith, A.F.M. *A Bayesian analysis of some time-varying models*, (In Recent Developments in Statistics, eds. Barra, J.R. et. al., North-Holland, Amsterdam, 1977), 257-267.

[22] Smith, A.F.M *Change-point problems: approaches and applications*, Trab. Estadist., **31**, 83-98, 1980.

[23] Talih M. and Hengartner, N. *Structural Learning With Time-Varying Components: Tracking the Cross-Section of Financial Time Series*, Journal of the Royal Statistical Society, **67**, 321-341, 2005.

[24] Zacks, S. *Classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures*, Stat. Anal. Donnees, **7**, 48-81, 1982.