The imprecise Dirichlet model for predicting consumer behaviour

Rosephine Georgina RAKOTONIRAINY (georgina@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Doctor Ian Durbach
University of Cape Town, South Africa

22 May 2014

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Consumer behaviour or purchase behaviour is the study of individuals and their processes to select and buy products. In this project, we use the imprecise Dirichlet model (IDM) to predict future purchase from purchase history. The IDM is a model for making inferences from multinomial data where the sample space is not precisely known; inferences are thus expressed in terms of posterior lower and upper probabilities. We apply this model to "panel data" which consists of a stream of purchases, indicating which product was bought at each point in time. The focus of this research project is first, to evaluate whether the IDM performs well for such prediction; then, to study the sensitivity of IDM to non-stationarity in the underlying purchase behaviour. For this, we study the change-point model which aims to identify the change in a sequence of purchases. We assess the relationship between the output of the change-point model and the accuracy of the IDM.

Ny atao hoe "consumer behaviour" dia fianarana mikasika ny toetra sy fomba fanaon'ny mpanjifa rehefa hividy zavatra. Ity asa ity dia mamakafaka ny fampiasana modely iray atao hoe IDM raha mety hanaovana faminavinana ny zavatra vidian'ny mpanjifa amin'ny ho avy amin'ny alalan'ny zavatra novidiany taloha. Ampiharina amin'ny "panel data" io modely io ahafahana manantontosa izany. Ny tanjon'ity asa ity izany dia mijery hoe mety ve io modely io hanaovana izany faminavinana izany ary koa mijery ny toetoetr'io modely io raha toa ka miovaova ny safidin'ny mpanjifa. Amin'io farany io dia iresaka momba ny "change-point model" isika izay modely afaka mamantatra ny fiovana tahak'izany ka ampifandray ny valiny azo avy amin'io sy ilay IDM.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Rosephine Georgina RAKOTONIRAINY, 22 May 2014.

Contents

1. Introduction

Market research aims to discover what consumers want or need. Once that research is completed, it can be used to determine how to market the product. One type of study within this area is the "consumer behaviour". It aims to study how and why consumers buy products over time. It is also a study which looks for patterns or regularities in "time series" representing purchase history. There exists various types of purchase behaviour. An example is the $brand\text{-}loyalty$. Consumers become committed to one brand of product and make repeated purchases over time. An alternative is the so-called $variety\text{-}seeking$ behaviour, by which consumers switch from one product to another product.

Such variety-seeking behaviour represents disadvantages and vulnerabilities for marketers. An understanding of the different consumer behaviour is important in defending and expanding market share. In addition, the knowledge of consumer behaviour enables the marketers to understand and predict buying behaviour of consumers in the marketplace. One important question is: "'can we use purchase history to predict future purchases?"

In this project, we try to answer this question. We will study the "Imprecise Dirichlet Model" (IDM) proposed by Walley (1996) and use it as a model to predict future purchases. Our data consists of "panel data" reporting purchases of one product category "washing powder" by 8389 panelists in Australia over 6 years. Our aim is not to adopt panel data models which describe the individual behaviour both across time and across individuals. Rather, we aim to evaluate the performance of IDM when predicting future purchases using panel data. Furthermore, we will study the sensitivity of the IDM to non-stationarity in the underlying purchase behaviour. In fact, a sequence of purchases may undergo changes at unknown time. So, in the second part of the present work, we present the notion of a "Change-Point model" which aims to identify and estimate such changes. Finally, we try to correlate the results from the change-point model and the IDM.

The remainder of this work is organized as follows: the next chapter covers the details of the methodology used. It includes some backgrounds and description of all desirable principles in order to make inferences. Then we will discuss the IDM and the change-point model. In chapter 3, we will describe the data and some details about the implementation. We, finally, summarize the results obtained in chapter 4.

1.1 Definition

We define here some terms that we use throughout this work.

- **Panel data**: it provides information on individual behaviour, both across individuals and over time. It is also called longitudinal data or cross-sectional time series data.
- **Time series**: it represents a sequence of observations which are ordered in time. In our context, this refers to the sequence of purchases of an individual at multiple point in time.
- **Statistical inference**: it makes use of information from a sample to draw conclusions (inferences) about the population from which the sample was taken.
- The **prior**: it expresses a state before any observation is taken into account, i.e. in the absence of observed data or evidence. In our context, we often define a prior probability distribution for the parameter in the model to express uncertainty about it before data is taken into account.

- The **posterior**: it defines a position after observing new information or some events. In a Bayesian framework, the posterior probability of the parameter is normally calculated by updating the prior probability by using "Bayes' theorem".
- **Bayes' theorem**: it is a direct application of conditional probabilities.
- **Conjugate distributions**: if the prior and posterior distributions of the parameter belong to the same family, then they are called conjugate. We will see an example, the Dirichlet distribution, in the Bayesian approach and IDM.
- **Hyperparameters**: this term refers to the parameters of prior distributions in order to distinguish them from the parameters of the model of the underlying data.
- **p-value**: it is defined as the probability of obtaining a result equal to or more extreme than what was actually observed, assuming that the null hypothesis is true. The null hypothesis, often denoted by H_0, refers to a general or default position and is assumed true until evidence indicates otherwise. Smaller p-values indicate greater statistical significance since it tells us that the hypothesis under consideration may not adequately explain the observation.

2. Methodology

2.1 Preliminaries

The imprecise Dirichlet model was introduced by Walley (1996). The main problem that he considered is the following: "I have ... a closed bag of coloured marbles. I intend to shake the bag, to reach into it and to draw out one marble. What is the probability that I will draw a red marble?" ... "Suppose that we draw a sequence of marbles whose colours are (in order) blue, green, blue, blue, green, red. What conclusions can you reach about the probability of drawing a red marble on a future trial?".

Before the first draw we notice that there is no information about what colours of marbles are in the bag. We do not know, whether there is no red marble inside or all of the marbles are red. If we try to answer our initial question, one naive answer is to say that, because there are two possible outcomes (red or non-red) and no information to favour either, the probability of red is $\frac{1}{2}$. But, another set of 'equipossible' outcomes is {red, blue, green, other colours}, which yields probability $\frac{1}{4}$ for red. There are many other ways to define the set of possible outcomes of the drawing and consequently many different probabilities for the event that red is drawn. Thus, we cannot rely on this idea.

Another approach can be justified, however. As we are completely ignorant about the contents of the bag, that recommends the $vacuous$ probability which is to choose the upper probability of red to be 1 and the lower probability to be 0. This vacuous probability models the prior ignorance and does not depend on the sample space. The IDM is a model that generates such vacuous prior lower and upper probabilities for the outcome of future trials. Furthermore, this prior ignorance can be updated after observing repeated trials. In other words, as we draw marbles many times, we gain information from the observed marbles about the colour of the marbles inside the bag. Consequently, the vacuous probability becomes precise and in the IDM, the prior lower and upper probabilities become posterior lower and upper probabilities.

This problem of a bag of marbles shares similarities with the problem of predicting consumer behaviour. In fact, there exists two issues in modelling purchase behaviour. We are often given a list of brands but we are never sure that this list contains all brands. Moreover, the list of brands is often long but each consumer will only think of a small proportion of these brands. So even if we know most of all possible brands, the problem is that we do not know, $a\ priori$, what brands are relevant to a consumer before we observe them. And as we observe more purchase sequences, we gain information about the buying behaviour of consumers. These issues motivate the use of IDM. Before giving details about the IDM, let us first define some tools that will be needed.

2.1.1 Multinomial sampling. We consider a sample of consumers making purchases in a market with K different brands [1]. For each consumer, we observed N successive purchases which is the sum of all counts of purchases of each brand. That is, $N = \sum_{i=1}^{K} n_i$, $i = 1, \ldots, K$, where n_i is the number of purchases of the brand i. As explained above, the sample space is not defined $a\ priori$, this implies that K is arbitrary. In addition, we associate each brand i with a parameter θ_i which is the probability of selecting the brand i at each purchase occasion. The probabilities $\theta_1, \ldots, \theta_K$ are unknown parameters and may vary among consumers. Thus, the probability of observing $\mathbf{n} = (n_1, \ldots, n_K)$, conditioned on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, can be modelled by a multinomial distribution:

[1] We consider one product category in this problem.

$$P(\mathbf{n}|\boldsymbol{\theta}) = \frac{N!}{n_1! \cdots n_K!} \theta_1^{n_1} \cdots \theta_K^{n_K}, \tag{2.1.1}$$

where $\frac{N!}{n_1! \cdots n_K!}$ is the multinomial coefficient so that the probabilities sum to one across K.

If we consider this probability as a function of $\boldsymbol{\theta}$ with $\mathbf{n}$ considered as fixed, we get the observed likelihood function

$$L(\boldsymbol{\theta}|\mathbf{n}) \propto \prod_{j=1}^{K} \theta_j^{n_j}. \tag{2.1.2}$$

2.1.2 Inference problems. The basic idea of the IDM is to make inference from multinomial data i.e. a sample of N observations yielding the counts $\mathbf{n} = (n_1, \ldots, n_K)$ over K categories with probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. One question that might arise is "inference about what?". In our context, we consider two types of inference

- *Parametric inference*: making inferences about the unknown parameter $\boldsymbol{\theta}$.
- *Predictive inference*: predicting the counts $\mathbf{n}' = (n'_1, \ldots, n'_K)$ in M future observations.

For our purpose, we are more interested in the second type of inference (the predictive one) as it consists of making predictions about future observations.

Whether we consider parametric or predictive inferences, two questions may arise. The first one is about assessing whether one event B is probable or not. Under the IDM, the answer is given by the *lower* and *upper probabilities* $\underline{P}(B)$ and $\bar{P}(B)$. The second question consists of making inferences about some real-valued derived parameter $\delta = D(\boldsymbol{\theta})$. The answer, for IDM case, is obtained by taking the expectation or by using the cumulative distribution function of δ.

2.2 Desirable principles

In Walley (1996), several principles were proposed to be considered for making inferences, especially for *objective inference* as it aims to learn about the parameter $\boldsymbol{\theta}$ or about future data from prior ignorance. In general, two approaches are used to provide objective inference methods [2], namely the *frequentist* and the *objective Bayesian approaches*. In our context however, both approaches suffer from shortcomings, in that they do not satisfy some of the following important principles.

2.2.1 Symmetry Principle (SP). This principle states that prior uncertainty about any event relative to $\boldsymbol{\theta}$ or $\mathbf{n}'$ (the count of future trial), should be invariant with respect to permutation of categories. That is, we have no reason to favour one possible outcome to another and therefore the probability model should be symmetric.

2.2.2 Embedding Principle (EP). For each event A, the probability assigned to A should not depend on the probability space in which A is embedded. In particular, the probability assigned *a priori* to the event A should be invariant with respect to refinements and coarsenings of categories.

[2] Objective inference methods: "let the data speak for themselves". These methods are used to make inferences from the observed data only, assuming no or little prior knowledge about the parameters is available.

2.2.3 Representation Invariance Principle (RIP). This principle and the EP were proposed by Walley (1996). It asserts that posterior inferences of any event A should not depend on refinements or coarsenings of categories. The satisfaction of RIP is the crucial property of IDM, which we will see in the next section. This motivates the use of IDM. However, objective Bayesian approaches do not satisfy this principle. Two different sample spaces give two different inferences.

2.2.3.1 Example. To illustrate that objective Bayesian approaches violate the RIP, let us consider the following example. Suppose that we observe a sequence of purchases (in order) as follows A,B,A,A,B,C, where A,B and C denote the brands. We set two possible sample spaces, $\Omega_1 = \{$ A,B,C, others $\}$ and $\Omega_2 = \{$ C, others $\}$. Recall that we do not know $a\ priori$ the relevant sample space. We wish to calculate the probability $P(\text{C}|\mathbf{n})$ which is the probability of observing the brand C in the next purchase.

In Bayesian framework, prior uncertainty about the parameter $\boldsymbol{\theta}$ in (2.1.1) is described by a single Dirichlet distribution whose density is given by

$$p(\boldsymbol{\theta}) \propto \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}, \tag{2.2.1}$$

which we write as

$$\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha}),$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \alpha_j > 0$ for any j. We call α_j's the prior strengths.

There is also an alternative parametrization of the Dirichlet distribution, which is

$$\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha}) = Dir(s, \mathbf{t}),$$

where $s = \sum_{j=1}^{K} \alpha_j$ and $\mathbf{t} = \frac{\boldsymbol{\alpha}}{s}$, such that $\mathbf{t} = (t_1, \ldots, t_K)$ with $t_j > 0$ and $\sum_{j=1}^{K} t_j = 1$.

We call s [3] the $total\ prior\ strength$ and $t_1, \ldots, t_K$ are the $prior\ frequencies$.

In addition, from previous studies, e.g. (Bernard, 2008), four Dirichlet priors have been proposed as models for describing prior ignorance about $\boldsymbol{\theta}$. All are symmetric Dirichlet, i.e. $t_j = \frac{1}{K}$ for $j = 1, \ldots, K$, and with different total prior strength s. They are

- Haldane's improper prior (1948) $\alpha_j = 0$ $(s = 0)$;
- Perks (1947) $\alpha_j = \frac{1}{K}$ $(s = 1)$;
- Jeffreys (1946) $\alpha_j = \frac{1}{2}$ $(s = \frac{K}{2})$;
- Bayes-Laplace's uniform prior $\alpha_j = 1$ $(s = K)$.

Now, the posterior distribution of $\boldsymbol{\theta}$ is obtained from Bayes' theorem by updating the prior distribution of $\boldsymbol{\theta}$ (2.2.1) with the observed likelihood (2.1.1). This is expressed by

$$\boldsymbol{\theta}|\mathbf{n} \sim Dir(\mathbf{n} + \boldsymbol{\alpha}) = Dir(N + s, \boldsymbol{t}^*),$$

[3] There is considerable disagreement over what value of s should be used among Bayesians especially those who adopt symmetric Dirichlet. In fact, smaller values of s produce stronger conclusions, whereas larger values of s produce more cautious inferences. There are many arguments that suggest $s = 2$ or $s = 1$ for consistency between the different Bayesian and frequentist inferences. In this project, we use $s = 2$ or $s = 1$.

where $\boldsymbol{t}^* = (t_1^*, \ldots, t_K^*)$ and $t_j^* = \frac{n_j + st_j}{N+s}$, $j = 1, \ldots, K$. The values $t_1^*, \ldots, t_K^*$ determine the posterior frequencies.

Hence, the posterior probability $P(\mathsf{C}|\mathbf{n})$ is given by

$$P(\mathsf{C}|\mathbf{n}) = t_c^* = \frac{n_c + st_c}{N+s} = \frac{n_c + sK^{-1}}{N+s}, \tag{2.2.2}$$

where n_c is the number of purchases of the brand C in the N observed sequences, where $N = 6$ in our example. The table below shows the probability $P(\mathsf{C}|\mathbf{n})$ calculated from (2.2.2) for the two different sample spaces set above.

	$s = 1$	$s = 2$	$s = \frac{K}{2}$	$s = K$
Ω_1	0.179	0.188	0.188	0.2
Ω_2	0.214	0.25	0.214	0.25

Thus, this result confirms that objective Bayesian approaches do not obey the RIP.

2.2.4 Stopping Rule Principle (SRP). This principle asserts that inferences should not depend on the stopping rule, which means they should not depend on data that might have occurred but have actually not. The frequentist approach does not satisfy this principle.

To see that the frequentist approach violates the SRP, let us again consider the example (2.2.3.1). In this case, we wish to test whether the probability of observing the brand C in the next purchase, denoted by θ_c, is less than $\frac{1}{2}$. So the null hypothesis H_0 is $\theta_c \geq \frac{1}{2}$ against the alternative hypothesis $H_1 : \theta_c < \frac{1}{2}$. Consider two possible stopping rules.

- First, suppose that we decide to stop the experiment after observing 6 purchases. The test statistic is the number of purchases of the brand C. The observed data A,B,A,A,B,C shows that the brand C was bought one time out of six purchases. So, the p-value, when $\theta_c = \frac{1}{2}$, is the probability that no more than 1 brand C is purchased in six purchases. This probability can be computed from the binomial coefficient as follows

$$p_1 = \frac{1}{2^6}\left[\binom{6}{0} + \binom{6}{1}\right] = \frac{7}{64} = 0.109.$$

  Thus, since p_1 exceeds 0.1, the conventional frequentist conclusion is that there exists no evidence against the null hypothesis.

- Consider a second stopping rule: we decide to stop observing the sequences when the brand C is purchased for the first time. Thus, the test statistic is the number of purchases to get the first brand C which is 6 here. Therefore, the p-value, when $\theta_c = \frac{1}{2}$, is the probability that six purchases are needed so that the brand C is purchased for the first time. This value is

$$p_2 = \frac{2}{2^6} = 0.031.$$

  As $p_2 < 0.05$, there is enough evidence against H_0.

Hence, we get two different conclusions for two different choices of stopping rules.

2.2.5 Likelihood Principle (LP). Bernard (2005) suggested that "posterior inferences should depend on the data through the likelihood only". Precisely all inferences on the parameter $\boldsymbol{\theta}$ brought by an observation $\mathbf{n}$ are contained in the likelihood function. Moreover, in Robert (2007), if $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are two observations depending on the same parameter, such that there exists a constant c satisfying $L(\boldsymbol{\theta}|\boldsymbol{n}_1) = cL(\boldsymbol{\theta}|\boldsymbol{n}_2)$ for every $\boldsymbol{\theta}$, they then must lead to identical inferences.

Since the Bayesian approach is entirely based on the posterior distribution which depends on observed data $\boldsymbol{n}$ only through the likelihood, this principle is automatically satisfied in a Bayesian setting.

2.2.6 Coherence Principle (CP). This principle is put forward in a Bayesian context, in which uncertainty is described by personal probabilistic assessments and interpreted as defining an acceptable betting rate, (Bernard, 2005). In fact, it states that all our assignment of probabilities to all possible events should be such that it is not possible to make a definite gain. It is a self-consistency requirement which ensures that combinations of several bets cannot lead to a sure loss.

In contrast of the common objective inference approaches (objective Bayesian approaches and frequentist), the IDM satisfies all of the above principles. This fact motivates the use of IDM in our context. In the following section, we will study the IDM itself and look at the application of it for predicting future purchases.

2.3 The imprecise Dirichlet model

For this model, prior uncertainty about the parameter $\boldsymbol{\theta}$ is described by a set of prior distribution. Indeed, in Bayesian models, prior uncertainty about $\boldsymbol{\theta}$ is described by a single Dirichlet distribution. However, under the IDM, it is expressed by a set of Dirichlet distributions.

Concretely, this set is expressed by

$$M_0 = \left\{ Dir(s, \mathbf{t}); \quad 0 < t_j < 1, \sum_{j=1}^{K} t_j = 1 \right\},$$

After any observations or counts $\mathbf{n}$ are taken into accounts, each Dirichlet distribution of $\boldsymbol{\theta}$ in the set M_0 is updated by Bayes' theorem into another Dirichlet on $\boldsymbol{\theta}|\mathbf{n}$. Hence, posterior uncertainty about $\boldsymbol{\theta}$ is described by the set M_N that consists of all Dirichlet distributions $Dir(N+s, \boldsymbol{t}^*)$ where $\boldsymbol{t}^* = (t_1^*, \ldots, t_K^*)$ and $t_j^* = \frac{n_j + s t_j}{N+s}$. This set is expressed by

$$M_N = \left\{ Dir(N+s, \boldsymbol{t}^*); \quad 0 < t_j^* < 1, \sum_{j=1}^{K} t_j^* = 1 \right\}.$$

From now on, we denote by $P_{s,\mathbf{t}}(.)$ a prior probability provided by a particular $Dir(s, \mathbf{t})$ in the set M_0, and by $P_{s,\mathbf{t}}(.|\mathbf{n})$ a posterior probability from a particular $Dir(N+s, \boldsymbol{t}^*)$ in the set M_N.

In the next 3 subsections, we will discuss first how to use the IDM to answer the questions that we arose in (2.1.2). Then, we will see different methods used, under the IDM, for making parametric inference. Lastly, we will describe how to make predictive inference using IDM. This last part is the one that we need for our analysis.

2.3.1 Answering inference questions with the IDM. We stated two types of inference questions in (2.1.2). Each question can be answered before (prior inference using M_0) or after (posterior inference using M_N) taking data into account.

For the first question, let B_j denote the event that the brand j will be purchased in the next purchase. Prior uncertainty about B_j is expressed by the prior lower and upper probabilities, $\underline{P}(B_j)$ and $\bar{P}(B_j)$, which are obtained by minimizing and maximizing $P_{s,\mathbf{t}}(B_j)$ with respect to t_j. Under the Dirichlet prior distribution, the probability $P_{s,\mathbf{t}}(B_j)$ is equal to the mean of θ_j, which is t_j. So by optimizing with respect to t_j, we obtain the vacuous probability $\underline{P}(B_j) = 0$ and $\bar{P}(B_j) = 1$.

Similarly, posterior uncertainty about B_j is expressed by the posterior lower and upper probabilities $\underline{P}(B_j|\mathbf{n})$ and $\bar{P}(B_j|\mathbf{n})$, obtained by minimizing and maximizing $P_{s,\mathbf{t}}(B_j|\mathbf{n})$ with respect to t_j. Under the posterior Dirichlet distribution, the predictive probability $P_{s,\mathbf{t}}(B_j|\mathbf{n})$ is equal to the posterior mean of θ_j which is $t_j^* = \frac{n_j + s t_j}{N+s}$. So by minimizing and maximizing t_j^* with respect to t_j, we obtain

$$\underline{P}(B_j|\mathbf{n}) = \frac{n_j}{N+s} \quad \text{(achieved in the limit } t_j \to 0)$$

and

$$\bar{P}(B_j|\mathbf{n}) = \frac{n_j + s}{N+s} \quad \text{(achieved in the limit } t_j \to 1).$$

If we consider the example (2.2.3.1), the posterior lower and upper probabilities of observing the brand C in the next purchase, for $s = 2$ and $N = 6$, are

$$\underline{P}(\mathsf{C}|\mathbf{n}) = \frac{1}{6+2} = 0.125 \quad \text{and} \quad \bar{P}(\mathsf{C}|\mathbf{n}) = \frac{1+2}{6+2} = 0.375.$$

For the second question, consider a real-valued derived parameter $\delta = D(\boldsymbol{\theta})$. Prior inferences about δ can be summarized by the prior lower and upper cumulative distribution functions (CDFs) $\underline{F}_\delta(d) = \underline{P}(\delta < d)$ and $\bar{F}_\delta(d) = \bar{P}(\delta < d)$. Furthermore, posterior lower and upper CDFs are defined by $\underline{F}_\delta(d|\mathbf{n}) = \underline{P}(\delta < d|\mathbf{n})$ and $\bar{F}_\delta(d|\mathbf{n}) = \bar{P}(\delta < d|\mathbf{n})$. Notice that these lower and upper probabilities, either prior or posterior, are obtained by minimizing and maximizing the appropiate $P_{s,\mathbf{t}}(.)$ and $P_{s,\mathbf{t}}(.|\mathbf{n})$ with respect to $\mathbf{t}$.

2.3.2 Inferences about the parameter $\boldsymbol{\theta}$. We know that the underlying parameter $\boldsymbol{\theta}$ in the model is unknown. In this subsection, we describe two ways for making inferences about this parameter. Under the IDM, inferences about $\boldsymbol{\theta}$ can be summarized by stating the posterior lower and upper means and variances for $\boldsymbol{\theta}$, or by quoting a credible interval for it:

- The posterior lower and upper expected values of θ_j, $j = 1, \ldots, K$, are simply the posterior lower and upper means of the posterior Dirichlet distributions after minimizing and maximizing with respect to t_j. That is,

  $$\underline{E}(\theta_j|n) = \frac{n_j + s}{N+s} \quad \text{and} \quad \bar{E}(\theta_j|n) = \frac{n_j}{N+s}.$$

  If we consider the example (2.2.3.1), with $s = 2$, we have

  $$\underline{E}(\theta_c|\mathbf{n}) = 0.125 \quad \text{and} \quad \bar{E}(\theta_c|\mathbf{n}) = 0.375.$$

  The posterior lower and upper variances of θ_j are difficult to compute, but they can be shown, as in Walley (1996), to be

$$\underline{V}(\theta_j|\mathbf{n}) = \frac{n_j(N-n_j) + s\ \min\{n_j, N-n_j\}}{(N+s)^2(N+s+1)}$$

and

$$\bar{V}(\theta_j|\mathbf{n}) = \frac{n_j(N-n_j) + \frac{1}{4}(N+s)(s+1)^2}{(N+s)(N+s+1)^2}.$$

- A credible interval for θ_j, $j = 1, \ldots, K$, can be expressed by $I = [\theta_*, \theta^*]$ where θ^* and θ_* are chosen so that respectively the posterior lower probability $\underline{P}(I|\mathbf{n})$ reaches a credibility γ and the posterior upper probability $\bar{P}(I|\mathbf{n})$ reaches a credibility $1-\gamma$. In other words, θ^* can be obtained from $\underline{P}(I|\mathbf{n}) = \gamma = G(\theta^*)$, where G is the CDF of $Beta(s+n_j, N-n_j)$ (i.e $\theta^* = G^{-1}(\gamma)$) and $\theta_* = H^{-1}(1-\gamma)$ where H is the CDF of $Beta(n_j, s+N-n_j)$.

2.3.2.1 Remark.

  - This interval can be interpreted as the usual confidence interval by which the expected value of $\boldsymbol{\theta}$ should lie within.
  - Notice that $\underline{P}(I|\mathbf{n})$ is achieved when $t_j \to 1$ and $\bar{P}(I|\mathbf{n})$ is achieved as $t_j \to 0$.
  - Usually the value of γ is 0.95.
  - We use a Beta distribution to compute the values of the endpoints of the credible interval. Indeed, we know that the posterior probability of $\boldsymbol{\theta}$ is described by a Dirichlet distribution $Dir(N+s, \boldsymbol{t}^*)$. But we saw in (2.2.3) that IDM satisfies the RIP which implies that it suffices, when making inferences about single θ_j $(j = 1, \ldots, K)$, to consider only two categories c_j and its complement. Hence, the IDM can be reduced to the imprecise Beta model. That is, prior and posterior inferences about θ_j can be derived by considering a Beta distribution on θ_j of the form $Beta(n_j + st_j, N - n_j + s - st_j)$ with $0 < t_j < 1$, and by setting $\mathbf{n} = n_j = 0$ for prior inferences. The general scheme of this process is illustrated below.

    For prior inferences

$$\begin{cases} \boldsymbol{\theta} \sim Dir(s, \mathbf{t}) \\ \downarrow \\ P(\boldsymbol{\theta}) \propto \Pi_{j=1}^{K} \theta_j^{st_j - 1} \end{cases} \iff \begin{cases} \theta_j \sim Beta(st_j, s - st_j) \\ \downarrow \\ P(\theta_j) \propto \theta_j^{st_j - 1}(1-\theta_j)^{s - st_j - 1}. \end{cases}$$

    For posterior inferences

$$\begin{cases} \boldsymbol{\theta}|\mathbf{n} \sim Dir(N+s, \boldsymbol{t}^*) \\ \downarrow \\ P(\boldsymbol{\theta}|\mathbf{n}) \propto \Pi_{j=1}^{K} \theta_j^{n_j + st_j - 1} \end{cases} \iff \begin{cases} \theta_j|\mathbf{n} \sim Beta(st_j + n_j, N + s - st_j - n_j) \\ \downarrow \\ P(\theta_j|\mathbf{n}) \propto \theta_j^{n_j + st_j - 1}(1-\theta_j)^{N + s - st_j - n_j - 1}. \end{cases}$$

  - A 95% credible interval for θ_c of the example (2.2.3.1) is given by

$$I = [\theta_*, \theta^*] = [0.0036, 0.7095].$$

    Indeed, $\theta^* = G^{-1}(\delta)$ and $\theta_* = H^{-1}(1-\delta)$, where $G \sim Beta(3,5)$, $H \sim Beta(1,7)$ and $\delta = \frac{1+0.95}{2}$.

2.3.3 Prediction about the future number of successes. In this part, we describe how to predict the number of purchases of a particular brand in any number of future purchases as well as obtain confidence interval around the prediction.

Let us denote by Y the number of purchases of the brand j in M future purchases. Probabilistic predictions about the future purchase can be made by computing the lower and upper CDFs for Y as follows.

Under the Dirichlet prior distribution for $\boldsymbol{\theta}$, the posterior distribution of Y is a beta-binomial distribution defined by

$$P(Y = i|\mathbf{n}) = \binom{M}{i} \frac{B(\alpha + x + i, \beta + N - x + M - i)}{B(\alpha + x, \beta + N - x)}, \tag{2.3.1}$$

for $i = 0, ..., M$; where B denotes the beta function, x is the number of purchases of the brand j, $\alpha = st_j$, $\beta = s - st_j$ and t_j is the prior probability of j.

Indeed, Y can be interpreted as the number of successes in M future purchases and according to (2.3.2.1), $\boldsymbol{\theta} \sim Dir(N + s, \boldsymbol{t}^*) \Longleftrightarrow \theta_j \sim Beta(st_j + n_j, N + s - st_j - n_j)$. Thus, Y follows a binomial distribution $Bin(M, \theta_j)$, i.e.

$$P(Y = i; M, \theta_j) = \binom{M}{i} \theta_j^i (1 - \theta_j)^{M-i}. \tag{2.3.2}$$

and

$$\pi(\theta_j) = \frac{\theta_j^{\alpha+x-1}(1 - \theta_j)^{\beta+N-x-1}}{B(\alpha + x, \beta + N - x)}. \tag{2.3.3}$$

Hence, the compound distribution (2.3.1) is obtained from (2.3.2) and (2.3.3) by

$$\begin{aligned} P(Y = i|\mathbf{n}) &= \binom{M}{i} \frac{1}{B(\alpha + x, \beta + N - x)} \int_0^1 \theta_j^{\alpha+x+i-1}(1 - \theta_j)^{\beta+N-x+M-i-1} d\theta_j \\ &= \binom{M}{i} \frac{B(\alpha + x + i, \beta + N - x + M - i)}{B(\alpha + x, \beta + N - x)}. \end{aligned}$$

Now, it follows that the posterior lower and upper CDFs for Y are

$$\underline{P}(Y \leqslant y|\mathbf{n}) = \sum_{i=0}^{y} \binom{M}{i} \frac{B(s + x + i, N - x + M - i)}{B(s + x, N - x)} \tag{2.3.4}$$

and

$$\bar{P}(Y \leqslant y|\mathbf{n}) = \sum_{i=0}^{y} \binom{M}{i} \frac{B(x + i, s + N - x + M - i)}{B(x, s + N - x)}, \tag{2.3.5}$$

for $y = 0, ..., M$. Notice that the CDF $P(Y \leqslant |\mathbf{n})$ is maximized with respect to t_j as $t_j \to 0$ and minimized as $t_j \to 1$.

From (2.3.4) and (2.3.5), we can construct a prediction set for Y. The one-sided γ prediction for Y is the smallest set $\{0, 1, \ldots, y^*\}$ such that $\underline{P}(Y \leqslant y^*|\mathbf{n}) \geqslant \gamma$. A conservative two-sided γ prediction set for Y of the form $\{y_*, \ldots, y^*\}$ can be constructed by finding the one-sided $\frac{1+\gamma}{2}$ prediction set of the form $\{0, \ldots, y^*\}$ (i.e. $\underline{P}(Y \leqslant y^*|\mathbf{n}) \geqslant \frac{1+\gamma}{2}$) and the other one-sided set of the form $\{y_*, \ldots, M\}$ (i.e. $\bar{P}(Y \leqslant y_*|\mathbf{n}) \geqslant \frac{1-\gamma}{2}$).

2.3.4 Application. In this part, we apply (2.3.3) in two sequences from our data. The goal is to evaluate if the prediction set contains the true outcome.

Consider the first sequence of purchases

$$S_1 = [A, A, B, A, C, A, D, D, E, F, G, A, F, A, F, A, G, H, F, H, H, E, A, E, A, A, G, A, F, F, A, I, G],$$

where A, B, C, C, E, F, G, H and I denote the different brands. We take the first $N = 20$ of this sequence, and denote by Y the number of occurrence of the brand A in the next $M = 10$ after N purchases. Notice that we can change the values of N and M as we want and also do this process for each brand within the sequence. The lower and upper CDFs for Y can be computed by using (2.3.4) and (2.3.5). For $s = 2$, the CDFs are

y	0	1	2	3	4	5	6	7	8	9	10
$\bar{P}(Y \leqslant y\|\mathbf{n})$	0.044	0.173	0.375	0.595	0.778	0.899	0.963	0.990	0.998	0.999	1
$\underline{P}(Y \leqslant y\|\mathbf{n})$	0.014	0.074	0.202	0.389	0.596	0.776	0.899	0.965	0.991	0.999	1

Thus, according to this table, in the next ten purchases it is highly probable that the brand A will be purchased no more than 7 times (lower probability 0.965) and at least once (upper probability 0.173). So, $\{1, 2, 3, 4, 5, 6, 7\}$ is a 95% prediction set. Similarly, a 60% prediction set is $\{3, 4, 5\}$. In fact, the upper probability of $Y = 3$ is 0.595 which is greater than 0.4 and the lower probability of $Y = 5$ is 0.776 which is greater than 0.6. These two prediction sets contain exactly the true number of purchases of the brand A in the next ten purchases. Actually, the number of occurrence of the brand A in S_1 for the next $M = 10$ purchases is 4 and this lies in the prediction sets for both either 95% or 60%.

2.3.4.1 Remark. If over many sequences, the two sided γ prediction set contains the true value in $100\gamma\%$ of all sequences, we say that the IDM is "well-calibrated".

Now, let us look at another sequence

$$S_2 = [A, A, A, A, A, A, A, A, A, B, A, A, A, A, A, A, A, A, A, A, C, C, C, C, C, C, C, C, C, C, C, C, C, C].$$

Similarly, let Y' be the number of occurrence of A in the next $M = 10$ purchases. For $N = 20$ the CDFs are

y	0	1	2	3	4	5	6	7	8	9	10
$\bar{P}(Y' \leqslant y\|\mathbf{n})$	1.5e-06	2.5e-05	2.2e-04	1.3e-03	5.9e-03	2.1e-02	0.06	0.17	0.38	0.7	1
$\underline{P}(Y' \leqslant y\|\mathbf{n})$	2.2e-08	4.9e-07	5.7e-06	4.5e-05	2.8e-04	1.5e-03	6.7e-03	0.02	0.09	0.32	1

Hence, a 95% prediction set for Y' is given by $\{6, 7, 8, 9, 10\}$. Indeed, 6 is the smallest value of Y' such that $\bar{P}(Y' \leqslant 6) \geqslant 0.05$ and for $Y' = 10$ the lower probability is 1 which is greater than 0.95. In addition, a 50% prediction set for Y' is $\{9, 10\}$. This is obtained by the same reasoning when replacing 0.95 and 0.05 by 0.5. Here, both of the prediction sets do not contain the exact number of purchases of A in M future purchases. For the 95% prediction set, it predicts with high probability that the number of occurrence of the brand A in ten purchases is within the set $\{6, 7, 8, 9, 10\}$. However, according to the sequence S_2, A does not occur for the next $M = 10$ purchases.

Thus, from these two examples and the study of the other sequences, we can state in conclusion that the model may fit well or poorly depending on the form of the sequence. We hypotheses that the model will be inaccurate for two types of sequence:

- first, a sequence like S_2, i.e. a long sequence of purchases of one brand then change to one long sequence of purchases of another brand.
- the second type is a sequence which consists of different brands, starting with a long sequence of purchases of one brand which does not occur later.

  For example, S = [A,B,B,A,A,A,A,A,A,A,A,B,C,D,D,C,B,E,F,B,B,D,D,C].

We can say, to explain the above fact, that the consumer may change his brand choice at an unknown time. Furthermore, we note that the underlying brand choice probability may vary. What can we say about this change? Can we model this change and assess its influence on the accuracy of IDM?

In the following section we try to answer these questions. We assess the change in the sequence of purchases by developing a multiple change-point model. We associate the observed sequence to an underlying sequence of choice probabilities which can be partitioned into blocks. The choice probabilities are equal within blocks but different between them. Our aim is to detect changes in consumer choice probabilities in order to assess the influence of these changes on the accuracy of the IDM.

2.4 Change-point model

The identification of change points are important in many data analysis problems. Many studies were developed and several authors presented Bayesian approaches for the change-point problem. In our context, we consider the *Product Partition Model* (PPM) proposed by Barry and Hartigan (1993). We will discuss below what PPM is, as shown in Loschi et al. (2003), but in a multivariate probability case and describe a practical way to detect the change points by using Gibbs sampling.

2.4.1 The PPM. Let $X_1, \ldots, X_d$ be an observed time series. In our case this time series represents the sequence of purchases. We associate with $X_1, \ldots, X_d$ a sequence of unknown parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$ which are the underlying choice probabilities in our context. Conditioned on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$, the sequence $X_1, \ldots, X_d$ has conditional marginal densities $f_1(X_1|\boldsymbol{\theta}_1), \ldots, f_d(X_d|\boldsymbol{\theta}_d)$ respectively. Consider a random partition ρ of the set $I = \{1, \ldots, d\} \cup \{0\}$ and a random variable B that represents the number of blocks in ρ. Assume that each partition $\rho = \{i_0, i_1, \ldots, i_b\}$, $0 = i_0 < i_1 < \ldots < i_b = d$ divides the sequence $X_1, \ldots, X_d$ into $B = b$ ($b \in I$) subsequences, which will be denoted by $\boldsymbol{X}_{[i_{r-1}i_r]} = (X_{i_{r-1}+1}, \ldots, X_{i_r})$, where $r = 1, \ldots, b$. Let $c_{[i_{r-1}i_r]}$ be the prior cohesion associated to the block $[i_{r-1}i_r]$ which represents the degree of similarity among the observations in $\boldsymbol{X}_{[i_{r-1}i_r]}$.

It is said that the random quantity $(X_1, \ldots, X_d; \rho)$ follows a PPM, denoted by $(X_1, \ldots, X_d; \rho) \sim PPM$, if

i) the prior distribution of ρ is

$$P(\rho = \{i_0, i_1, \ldots, i_b\}) = \frac{\prod_{r=1}^{b} c_{[i_{r-1}i_r]}}{\sum_{\mathfrak{C}} \prod_{r=1}^{b} c_{[i_{r-1}i_r]}}, \quad \text{for all } b \in I, \tag{2.4.1}$$

where $\mathfrak{C}$ is the set of all possible partitions of the set I into b blocks with endpoints $i_0, \ldots, i_b$ satisfying the condition $0 = i_0 < i_1 < \ldots < i_b = d$;

ii) conditioned on ρ, the sequence $X_1, \ldots, X_d$ has the joint density given by

$$f(X_1, \ldots, X_d|\rho = \{i_0, i_1, \ldots, i_b\}) = \prod_{r=1}^{b} f_{[i_{r-1}i_r]}(\boldsymbol{X}_{[i_{r-1}i_r]}), \tag{2.4.2}$$

where $f_{[i_{r-1}i_r]}(\boldsymbol{X}_{[i_{r-1}i_r]})$ is the density of the random vector $\boldsymbol{X}_{[i_{r-1}i_r]}$. In the parametric approach, this density is obtained as follows

$$f_{[i_{r-1}i_r]}(\boldsymbol{X}_{[i_{r-1}i_r]}) = \int_{\Theta_{[i_{r-1}i_r]}} f_{[i_{r-1}i_r]}(\boldsymbol{X}_{[i_{r-1}i_r]}|\boldsymbol{\theta})\pi_{[i_{r-1}i_r]}(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.4.3}$$

where $\Theta_{[i_{r-1}i_r]}$ is the parameter space corresponding to the common parameter, say $\boldsymbol{\theta}_{[i_{r-1}i_r]} = \boldsymbol{\theta}_{i_{r-1}} = \ldots = \boldsymbol{\theta}_{i_r}$, and $\pi_{[i_{r-1}i_r]}(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} \in \Theta_{[i_{r-1}i_r]}$. Notice that $\boldsymbol{\theta}_{[i_0i_1]}, \ldots, \boldsymbol{\theta}_{[i_{b-1}i_b]}$ are independent.

Also note that the number of blocks B in ρ has a prior distribution

$$P(B = b) \propto \sum_{\mathfrak{C}_1} \prod_{r=1}^{b} c_{[i_{r-1}i_r]}, \quad b \in I, \tag{2.4.4}$$

where $\mathfrak{C}_1$ is the set of all partitions of $I \cup \{0\}$ in b (fixed) blocks.

The goal is to get the posterior distributions of the parameters ρ, B, and $\boldsymbol{\theta}_k (k = 1, \ldots, d)$ in order to estimate the change in the time series. The three points below explain the computation of these posterior probabilities.

- In Barry and Hartigan (1993), the posterior distribution of $\boldsymbol{\theta}_k$, $k = 1, \ldots, d$ is given by

$$\pi(\boldsymbol{\theta}_k|X_1, \ldots, X_d) = \sum_{i=0}^{k-1} \sum_{j=k}^{d} r^*_{[ij]} \pi_{[ij]}(\boldsymbol{\theta}_k|\boldsymbol{X}_{[ij]}), \tag{2.4.5}$$

  where $r^*_{[ij]}$ denote the posterior relevance for the block $[ij]$, which is

$$r^*_{[ij]} = P([ij] \in \rho|X_1, \ldots, X_d).$$

  In Loschi et al. (2003), this posterior relevance is expressed by

$$r^*_{[ij]} = \frac{\lambda_{[0i]} c^*_{[ij]} \lambda_{[jn]}}{\lambda_{[0d]}},$$

  with $\lambda_{[ij]} = \sum_{\mathfrak{C}} \prod_{r=1}^{b} c^*_{[i_{r-1}i_r]}$ and $c^*_{[i_{r-1}i_r]} = c_{[i_{r-1}i_r]} f_{[i_{r-1}i_r]}(\boldsymbol{X}_{[i_{r-1}i_r]})$ is the posterior cohesion of the block $[i_{r-1}i_r]$.

- Moreover, in Loschi et al. (2003), the prior cohesion for the block $[i_{r-1}i_r]$ is interpreted as the probability that a new change takes place after $i_r - i_{r-1}$ instants given that a change has taken place at the instant i_{r-1}. That is

$$c_{[i_{r-1}i_r]} = \begin{cases} p(1-p)^{i_r - i_{r-1} - 1}, & \text{if } i_r < d \\ (1-p)^{i_r - i_{r-1} - 1}, & \text{if } i_r = d, \end{cases} \tag{2.4.6}$$

  where p is the probability that a change occurs at any instant in the sequence.

Hence, the prior distribution of ρ (2.4.1) conditioned on p can be expressed by

$$P(\rho = \{i_0, i_1, \ldots, i_b\}; p) \propto p^{b-1}(1-p)^{d-b}, \quad b \in I. \tag{2.4.7}$$

In addition, if we assume that p has a Beta prior distribution with parameters α and β denoted by $p \sim Beta(\alpha, \beta)$, then by using Bayes' theorem, the posterior probability of the random partition ρ is obtained from the prior probability updated with the likelihood (2.4.2). That is,

$$P(\rho = \{i_0, i_1, \ldots, i_b\} | X_1, \ldots, X_d) = P(\rho = \{i_0, i_1, \ldots, i_b\}) * \frac{\prod_{r=1}^{b} f_{[i_{r-1} i_r]}(\boldsymbol{X}_{[i_{r-1} i_r]})}{\sum_{\mathfrak{C}} \prod_{r=1}^{b} f_{[i_{r-1} i_r]}(\boldsymbol{X}_{[i_{r-1} i_r]})}, \tag{2.4.8}$$

where

$$P(\rho = \{i_0, i_1, \ldots, i_b\}) = \int_0^1 P(\rho = \{i_0, i_1, \ldots, i_b\}; p)\pi(p)dp,$$

with $\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$.

Explicitly, from (2.4.7),

$$\begin{aligned} P(\rho = \{i_0, i_1, \ldots, i_b\}) &\propto \int_0^1 P(\rho = \{i_0, i_1, \ldots, i_b\}; p)\pi(p)dp \\ &\propto \int_0^1 p^{b+\alpha-2}(1-p)^{d-b+\beta-1}dp \\ &\propto \Gamma(b+\alpha-1)\Gamma(d+\beta-b)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}. \end{aligned}$$

- For the random variable B, its prior distribution is given by

$$P(B = b) = \binom{d-1}{b-1} p^{b-1}(1-p)^{d-b}, \quad \forall b \in I,$$

where $\binom{d-1}{b-1}$ determines the number of different partitions of I into b blocks.

Besides, the posterior distribution of the event $B = b$, $b \in I$ is obtained by multiplying the posterior probability of ρ (2.4.8) by $\binom{d-1}{b-1}$.

Unfortunately the price paid for the ability of the PPM to exactly identify the multiple change points is that the calculation of the posterior distributions of ρ and B requires a high computational efforts. We need to consider all possible partitions that split the observed sequence into b blocks which is 2^{d-1}. Barry and Hartigan (1993) suggested the Gibbs sampling algorithm to compute these posterior distributions.

2.4.2 Gibbs sampling. Gibbs sampling was introduced by Geman and Geman (1984). It is an approach for generating random variables from a distribution without having to calculate the density, using Monte Carlo Markov Chain (MCMC) methods. It is often used in Bayesian statistics to sample from the posterior probability density function, e.g. (Gelfand and Smith, 1990). Here we apply the following Gibbs sampling scheme proposed by Barry and Hartigan (1993), in a multivariate probability case, in order to compute an approximation of the required posterior distributions.

Let U_t be an auxiliary random quantity which reflects the occurrence of a change-point at time-point t. That is,

$$U_t = \begin{cases} 1, & \text{if } \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t+1} \\ 0, & \text{if } \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t+1}, \end{cases}$$

for $t = 1, \ldots, d-1$. The random partition ρ is immediately identified by this random variable U_t. In fact, the posterior probability of each $\rho = \{i_0, \ldots, i_b\}$ can be estimated from the number of $\boldsymbol{U} = (U_1, \ldots, U_{d-1})$ for which the value of ρ is found. For the posterior distribution of B, it is given by $B = 1 + \sum_{t=1}^{d-1}(1 - U_t)$.

The algorithm is an iterative procedure which starts with an initial value $\boldsymbol{U}^{(0)} = (U_1^{(0)}, \ldots, U_{d-1}^{(0)})$ and generates the vector $\boldsymbol{U}^{(k)} = (U_1^{(k)}, \ldots, U_{d-1}^{(k)})$ at the k-th step. The t-th element $U_t^{(k)}$ is obtained from the conditional distribution

$$U_t^{(k)} | U_1^{(k)}, \ldots, U_{t-1}^{(k)}, U_{t+1}^{(k-1)}, \ldots, U_{d-1}^{(k-1)}; X_1, \ldots, X_d,$$

for $t = 1, \ldots, d-1$.

In practice, to compute the $\boldsymbol{U}^{(k)}$'s, it suffices to consider the following ratio

$$R_t = \frac{P(U_t = 1 | A_t^{(k)}; X_1, \ldots, X_d, p)}{P(U_t = 0 | A_t^{(k)}; X_1, \ldots, X_d, p)},$$

where $A_t^{(k)} = \{U_1^{(k)} = u_1, \ldots, U_{t-1}^{(k)} = u_{t-1}, U_{t+1}^{(k-1)} = u_{t+1}, \ldots, U_{d-1}^{(k-1)} = u_{d-1}\}$.

Furthermore, denote by b the number of blocks obtained if $U_t = 0$, and assume $p \sim Beta(\alpha, \beta)$. Hence, by considering the prior cohesion (2.4.6), each value $U_t^{(k)}$ for $t = 1, \ldots, d-1$ can be generated by using

$$R_t = \frac{f_{[xy]}(X_{[xy]}) \int_0^1 p^{b-2}(1-p)^{d-b+1} d\pi(p)}{f_{[xt]}(X_{[xt]}) f_{[ty]}(X_{[ty]}) \int_0^1 p^{b-1}(1-p)^{d-b} d\pi(p)} \tag{2.4.9}$$

$$= \frac{f_{[xy]}(X_{[xy]}) \Gamma(d+\beta-b+1)\Gamma(b+\alpha-2)}{f_{[xt]}(X_{[xt]}) f_{[ty]}(X_{[ty]}) \Gamma(d+\beta-b)\Gamma(b+\alpha-1)}, \tag{2.4.10}$$

where

$$x = \begin{cases} \max\{i, \text{ s.t. } 0 < i < t, U_i^{(k)} = 0\} & \text{if } U_i^{(k)} = 0, \text{ for some } i \in \{1, \ldots, t-1\}, \\ 0 & \text{otherwise} \end{cases}$$

and

$$y = \begin{cases} \min\{i, \text{ s.t. } t < i < d, U_i^{(k-1)} = 0\} & \text{if } U_i^{(k-1)} = 0, \text{ for some } i \in \{t+1, \ldots, d-1\}, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the criterion for choosing the values $U_t^{(k)}$ for $t = 1, \ldots, d-1$ becomes

$$U_t^{(k)} = \begin{cases} 1, & \text{if } R_t \geq \frac{1-u}{u} \\ 0, & \text{otherwise,} \end{cases} \tag{2.4.11}$$

where $u \sim Unif(0,1)$ (Uniform distribution).

In their work, Clark and Durbach (2014) considered a change-point analysis on respondent-level binary sequences of purchase data i.e. a sequence of 1's (if a brand of interest was purchased) and 0's (if another brand in the product category was purchased). In our context, we consider a categorical sequences of purchase data, that is we extend the analysis in Clark and Durbach (2014) by constructing a multinomial change-point model, which allows the purchase data to be modelled simultaneously across all brands, rather than from the point-of-view of a single brand. The details of procedure and example applied in one sequence of our data will be shown in the following section.

2.4.3 Details of procedure and worked example. We consider K (fixed) different brands, and a random variable X_τ, $\tau = 1, \ldots, n$ where the subscript τ represents the time of purchase made by the consumer and n denotes the last time of purchase. We associate X_τ with parameter $\boldsymbol{\theta}_\tau = (\theta_{\tau 1}, \ldots, \theta_{\tau K})$, where $\theta_{\tau i}$ corresponds to the probability that the individual purchases a brand i at time τ, for $i = 1, \ldots, K$. The vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a time series which represents an individual's purchase history. In order to compute the probability of change estimated from the Gibbs sampling, we need to compute (2.4.9) as follows

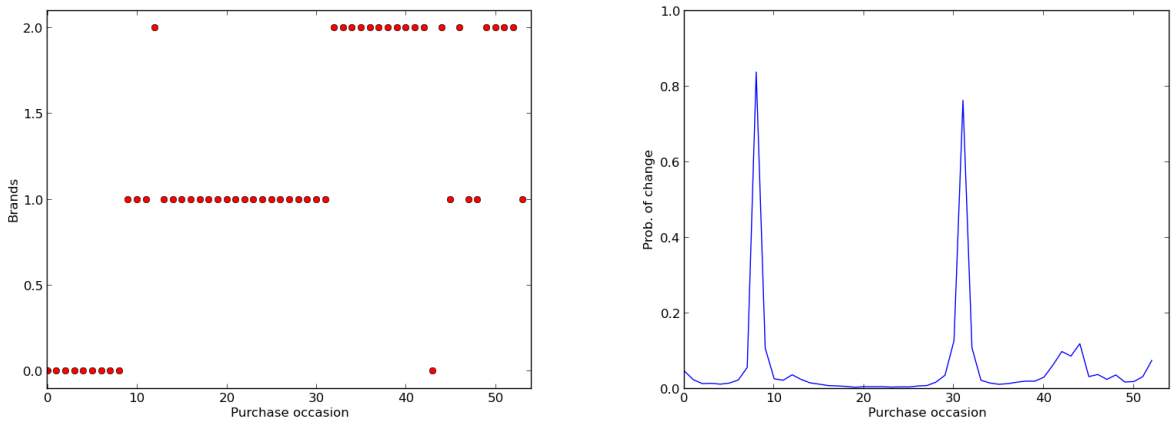
- First, notice that $X_\tau \sim Cat(\boldsymbol{\theta}_\tau)$ (categorical distribution), with $P(X_\tau = i|\boldsymbol{\theta}_\tau) = \theta_{\tau i}$, for $i = 1, \ldots, K$.
- Within a block $[ab]$, $\boldsymbol{\theta} = \boldsymbol{\theta}_\tau = \boldsymbol{\theta}_{[ab]}$ for every $a < \tau < b$ and $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. That is, $\pi(\boldsymbol{\theta}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$, where $B(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^K \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^K \alpha_j)}$.
- The likelihood $f_{[ab]}(\boldsymbol{X}_{[ab]})$ can be computed using (2.4.3). We have $f_{[ab]}(\boldsymbol{X}_{[ab]}|\boldsymbol{\theta}) = \prod_{j=1}^K \theta_j^{n_j}$, where n_j denote the number of brand j in the block $[ab]$ with $\sum n_j = s_{[ab]}$ (size of $[ab]$). So,

$$\begin{aligned} f_{[ab]}(\boldsymbol{X}_{[ab]}) &= \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K \theta_j^{n_j + \alpha_j - 1} \\ &= \frac{\Gamma(A)}{\Gamma(A+N)} \prod_{j=1}^K \frac{\Gamma(n_j + \alpha_j)}{\Gamma(\alpha_j)}, \end{aligned}$$

  where $A = \sum_{j=1}^K \alpha_j$ and $N = \sum_{j=1}^K n_j$.

From that, we can compute (2.4.9) easily, and do the iteration in Gibbs sampling.

Figure (2.1) demonstrates an example of the outcome from the Gibbs sampling. Figure (2.1a) indicates the observed sequence. It consists of 3 different brands denoted by 0,1,2 and has length 53. Figure (2.1b) shows the probability of change in the observed sequence of purchase (2.1a) obtained from the Gibbs sampling.



(a) Observed sequence

(b) Probability of change

Figure 2.1: Purchase history with posterior probability of change estimated by the change-point model.

Figure (2.1b) shows an apparent change in behaviour at time τ=9 and τ=32. The posterior probability of change is significantly high, close to one, which is reasonable for such obvious change.

2.4.3.1 Remark. For the computation, we placed a prior distribution of $Beta(3, 50)$ on p, and used a prior $Dir(1, 1, \ldots, 1)$ to represent the underlying choice probability $\boldsymbol{\theta}_\tau$. The prior $Beta(3, 50)$ was chosen so that it gives approximately a 5% prior probability that any time-point is a change-point. That is, we assign a small probability that a sequence contains a change-point at the beginning as we do not have information to the contrary. In addition, we considered a uniform Dirichlet prior distribution for $\boldsymbol{\theta}_\tau$ because this is consistent with prior ignorance and also the SP principle. We generated 6000 samples starting with a random partition ρ and discarded the initial 1000 iterations.

For the analysis of our data, we use the above procedure to generate the probability of change in each sequence.

3. Data

3.1 Description of purchase data

Our data consists of panel data reporting purchases of one product category (washing powder) with 95 different brands by 8389 panelists in Australia. Each panelist is governed by one sequence of purchases consisting of different brands. The three plots on the right show the summary of the data.

Figure (3.1) shows the distribution of the existing brands. It indicates that some brands are often purchased and some infrequently. For our analysis, we will consider the 7 leading brands (the 7 most purchased).

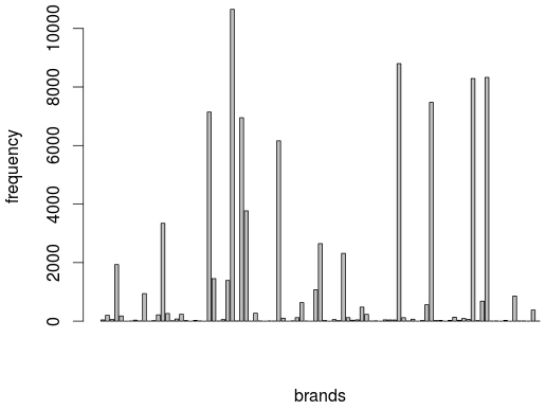


Figure 3.1: Brands frequency.

These top 7 brands were chosen because they represent 71% of the total purchases.

Figure (3.2) displays the distribution of the length of all sequences. There are many consumers with sequence length less than 40, but for our purpose we consider the sequences with length greater than 40 in order to get enough information about the consumer purchases. We call them "sequences of interest".

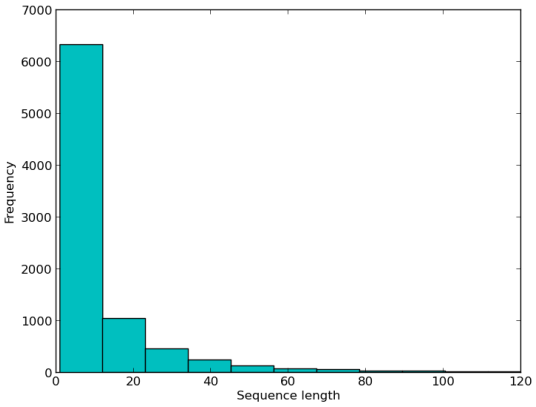


Figure 3.2: Sequence length frequency.

Furthermore, we deal with unbalanced panel data, which means customers are not observed in all time periods or at the same time. Figure (3.3) shows the distribution of the time lengths of all sequences with length 40 (in weeks).

This figure illustrates that some customers buy regularly and some of them infrequently. Around 58% of all consumers with sequence of length 40 buy the product every month and about 20% of them take more than 4 years to buy the product.

But, the inter-purchase time or the average time between 40 purchases of all sequences with length 40 is about 3.5 weeks. So, on average it will take approximately 3 years to collect data on 40 purchases.

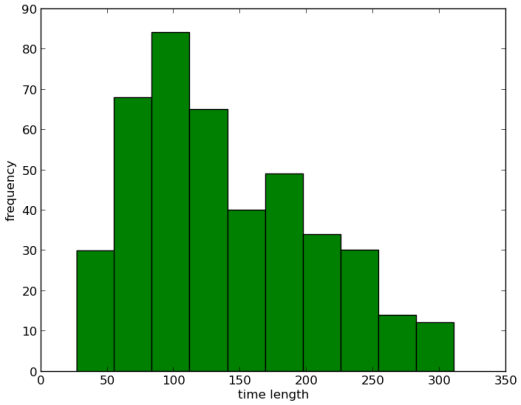


Figure 3.3: Distribution of time lengths.

Recall that our main aim is to analyse the performance of IDM possessing well-calibrated prediction intervals as a model to predict future purchase from purchase history. The procedure of our analysis is described below and we present our results in the next chapter.

3.2 Implementation

For each sequence of interest, which represents an individual's purchase history, we fit the IDM model as in (2.3.4).

First, we divided each sequence into 2 blocks. The first block is considered as the "training data" or the purchase history of the individual and the second block is referred to as the "test data" or the future purchase sequence. We denote by N the length of the training data and by M the test data. We took $N = 20$ and $M = 20$ during the analysis and for sequences with length greater than 40 we selected the first 40 observations only. The value of N and M can be changed, but these values were chosen to motivate the use of sequences with length 40 and more. Then, we run the IDM in the training data, for each of the 7 leading brands within each sequences [1]. Notice that within a sequence, it may happen that only one of the 7 leading brands occur, or two, or all. That is, the number of brands may differ between sequences. We then build a 95% prediction set, as in (2.3.4), for each leading brand in the sequence and compare the outcomes of the prediction to the test data. For this comparison, we set a new variable containing 1 if the prediction set contains the true value and 0 if the prediction is not correct. In this case if we average over all sequences we should get the coverage probability i.e. it should be 95%.

For the change-point model we use Gibbs sampling as described in (2.4.2) to generate a probability of change in the underlying purchase sequence. We placed a prior distribution of $Beta(3, 50)$ on the probability that any time-period is a change-point i.e. $Pr[U_\tau = 0]$ and used a prior $Dir(1, \cdots, 1)$ to represent the underlying choice probability $\boldsymbol{\theta}_\tau$.

For the computation of the Gibbs sampling, we generated 6000 samples starting with a random partition ρ. At each iteration i, we constructed a vector U_i of 0 or 1 indicating whether a change exists or not. We removed the first 1000 iterations to reduce autocorrelation between vectors. The output is a matrix by which the rows consist of the vector U_i at each iteration, $i = 1, \ldots, 6000$. Then, we took the average of each column. This gave us the probability that there is a change at time j, $j = 1, \ldots, 40$. Hence, for each sequence, we get a vector containing a probability that there is a change at each time-point. We recorded the maximum probability and used it as a feature to assess whether there is a relationship between the maximum and the coverage probability obtained from the IDM.

3.3 Outcomes

Given our aim is evaluating whether the IDM is an appropriate model to predict future purchase from purchase history, and also detecting changes in underlying purchase probabilities so that we can relate

[1] Notice that we do not need to run the IDM for all brands since we are interested only in these 7 leading brands as explained in (3.1).

the existence of change point in the sequence with the accuracy of the IDM; we summarize each sequence of interest in the following ways:

1. Proba = max($Pr[U_\tau = 0, \forall\tau]$) : the maximum probability that the sequence of purchases contains a change-point.

2. Perc = proportion of correct predictions made by the IDM : it represents the proportion of prediction intervals that contained the observed number of purchases of the brand during the test period.

3. Err = Perc - 0.95 : represents the error in the prediction outcomes.

4. Results

Our aim is to test whether the IDM is well-calibrated. Recall that we constructed a 95% prediction set for each leading brand in each sequence using the IDM and assigned in the variable Perc explained in (3.3) the proportion of correct prediction for each sequence. Therefore, a relevant test is to compare the average proportion of correct predictions with 0.95 because we expect that the 95% prediction set to contain the true value for 95% of sequences. The average of Perc is 0.82 which is obviously different from 0.95. But, a hypothesis test was performed to test this difference for statistical significance. We set the null hypothesis to be H_0: there is no significant difference between the true coverage proportion and 0.95. The t-test statistic is given by

$$t = \frac{\text{mean(Perc)} - 0.95}{\text{sd(Perc)}\sqrt{352}} = -6.704,$$

where sd stands for standard deviation. Hence, the p-value is determined by

$$p = 2 * P(T_{351} < -6.704) = 8.1e - 11.$$

As p is too small (strictly less than 0.05), the usual frequentist conclusion is that there is strong evidence to reject the null hypothesis in favour of the alternative one. Consequently, there is enough evidence to say that the IDM is not at all well-calibrated.

However, we described the change-point model in section (2.4.1). We saw that the underlying purchase sequence may change at unknown time. We adopted Gibbs sampling to detect this change and assigned in variable Proba as described in (3.3) the maximum probability that the purchase sequence contains a change.

Figure (4.1) shows the distribution of the maximum probability of change for all 352 sequences of interest.

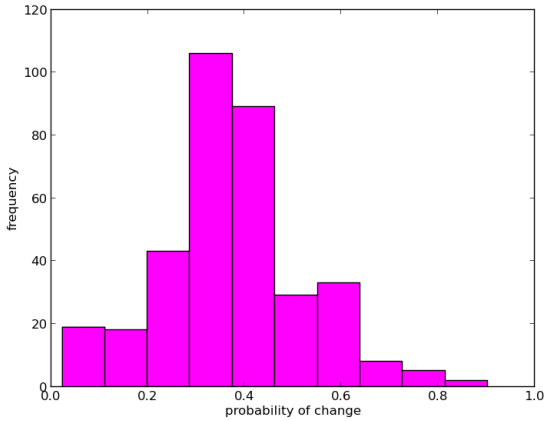


Figure 4.1: Distribution of the probability of change.

From the figure it follows that a clear change in purchase behaviour occurs for a minority of consumers. The observed proportion of sequences with a greater than 70% chance of containing a change-point is 2% (8 of 352 sequences). In addition, roughly 15% of consumers show medium evidence of change ($0.5 \leq$ Proba < 0.7), 61% moderate evidence of change ($0.2 <$ Proba < 0.5) and 9% very little evidence of change (Proba ≤ 0.2).

Basically, this result agrees with previous research. Some studies (e.g. Kahn (1995); Givon (1985)) indicate that the purchase sequence of a majority of stationary consumers are consistent with the "zero-order assumption". That is, a consumer is governed by a fixed probability of purchase. Givon (1985) found that laundry detergents had the 5th highest proportion of zero-order sequences, which means that consumers of this category of product do not often present changes in their purchase sequence.

So what can we say about the IDM and the change in the underlying purchase probabilities. Does the accuracy of the IDM depend on whether there has been a change in the purchase sequence? That is, does the IDM perorm significantly better in series where there is little evidence of a change?

Two different tests were considered in order to answer these questions.

- The first test concerns the correlation (if any) between the probability of change and the accuracy of the IDM from Err in (3.3). A correlation test was performed with null hypothesis H_0 : they are not correlated. The test gave a p-value equal to 0.0058, which tell us that there is enough evidence to reject H_0. That is, there is correlation between the change-point and the accuracy of the IDM. But how do they relate? This question motivates the construction of the second test.
- For the second test, we tried to assess whether the IDM does better in sequences with little or strong evidence of change. To be more specific, we divided the sequences of interest into three groups according to the probability of change. Denote by Group1 all sequences with little evidence of change (i.e. $Proba < 0.3$), by Group2 the sequences with medium evidence of change ($0.3 \leq Proba < 0.6$) and by Group3 all series with strong evidence of change ($Proba \geq 0.6$). We recorded the average of the variable Perc in (3.3) for each group. We tested each value to see whether it is significantly different from 0.95. The results are shown in Table (4.1).

Group	Proba intervals	mean(Perc)	t-statistic	df	p-value
Group1	[0-0.3[	0.957	0.4896	90	0.625
Group2	[0.3-0.6[	0.783	-6.866	237	5.73e-11
Group3	[0.6-1]	0.8	-2.125	22	0.044

Table 4.1: Test statistic results for each group.

As shown in Table (4.1), the average of the variable Perc is significantly different from 0.95 for Group2 and Group3 and this is confirmed by the common frequentist conclusion which states that there exists enough evidence to reject the null hypothesis for Group2 and Group3 (p-values strictly less than 0.05). However, the average of Perc is not significantly different to 0.95 for Group1. Consequently, there is enough evidence to say that the accuracy of the IDM depends on how much change occurs in the sequence of purchase. We can conclude that the IDM is well-calibrated for series which present a little or some evidence of change but performs relatively poorly for those which present quite high or strong evidence of change.

This result is not surprising, since the IDM assumes a stationary behaviour in the underlying purchase sequence. Indeed, the IDM models the consumer sequence as a multinomial random variable with a vector parameter $\boldsymbol{p}$ (probability) which is constant over time. However, customers at some time change their mind about how much they prefer the dominant brand over the others which reflects the change of the parameter $\boldsymbol{p}$ at the change time-point.

5. Conclusion

In a market, consumers are offered a great deal of choice. There are often a large number of brands within one product category. The problem of predicting consumer behaviour is then similar to the problem of bag of marbles considered by Walley (1996), where there is no prior information about what brands consumers will purchase.

In this project, we applied the IDM to predict future consumer behaviour given a purchase history. We first looked at the properties and principles that this model satisfies. We saw that the IDM does satisfy RIP which is the crucial property that motivates the use of it. Under this model, inferences are expressed in terms of posterior lower and upper probabilities. The probabilities are initially vacuous, reflecting prior ignorance, but they become more precise as the number of observations increases. Then, we assessed the extent of change in the underlying brand choice probability of consumers by developing the multiple change-point model. We developed a Gibbs sampling algorithm for the multinomial case. The code can be found on https://sites.google.com/a/aims.ac.za/georgina/project.

At the outset, we posed a question: "can we use purchase history to predict future purchase?" Our aim was then to evaluate if the IDM is a good model for predicting consumer behaviour. The answer is described in (4).

- After establishing the appropriate proportion of correct predictions and performing the relevant tests, it was shown that the IDM is not at all well-calibrated.
- But, we found that the accuracy of the IDM depends on the evidence of change existing in the sequence. The results indicated that IDM performs well for series or purchase sequences which present little evidence of change but performs poorly for those which present high or strong evidence of change.
- We, finally, found that a minority of consumers show relatively clear evidence of change which in basic agreement with previous research.

5.1 Discussion

The weakness of our analysis is that we worked on a quite small dataset. We considered 352 sequences which may not be sufficiently large to obtain reliable estimates of coverage probabilities. The above results are not sufficient to say definitively that the IDM is not well-calibrated or not working well to make such prediction. First, we fixed a threshold of 95% for the IDM. We need to evaluate how well the IDM is for the other credibility like 60% or 75%. In addition, we fixed the values of N and M which are respectively the number of training data and test data. We should evaluate how sensitive the IDM is when changing the values of N and M.

Moreover, the common model for describing consumer behaviour is the "Dirichlet model" or the "Negative-Binomial distribution model" as explained in Bassi (2011). It simultaneously models purchase frequency and brand choice. That is, it models the brand choice combined with consumer's inter-purchase rate. In this project, we just focused on the first part, which is the modelling of the brand choice.

Finally, several studies focus on estimating the change in the purchase sequence, especially by using the product partition model, e.g. (Clark and Durbach, 2014). They often consider different products and all

respondents or consumers. In our case, we considered only one product, and consumers with sequence length greater than 40.

5.2 Future work

Ideas for future work include the analysis of the sensitivity of the IDM to the amount of history i.e. the sequence length before prediction and the number of future forecasts. In addition, as we only looked at one type of product (washing powder), so a further study is to extend the result for different types of product. Lastly, the change-point model that we used is for a precise Dirichlet model, while the model that we used for predicting purchases is an imprecise Dirichlet model. So, one idea is to extent the change-point model to the imprecise approach, so that both models employ the same basic approach to uncertainty and ignorance.

Acknowledgements

Foremost, I thank Almighty God for his love. He said “Never will I leave you, never will I forsake you”. Hebrews 13:5

I would like to offer my special thanks to Dr. Ian Durbach for his assistance during the development of this project. I am grateful for his patience and guidance.

I also thank all AIMS staff, lecturers, tutors and students for their help and encouragement. In particular, I thank all Malagasy friends for their support.

Atolotro ho an’i Neny, Tonton ary Tantine ity asa ity izay tsy nitandro hasasarana nikarakara sy nivavaka ka naha toy izao ahy.

References

D. Barry and J. A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

F. Bassi. The dirichlet model: Analysis of a market and comparison of estimation procedures. *Marketing Bulletin*, 22, 2011.

J.-M. Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2):123–150, 2005.

J.-M. Bernard. Predictive inference: From bayesian inference to imprecise probability. 2008.

A. Clark and I. Durbach. Using bayesian change-point models to assess changes in customer loyalty overtime. *Management Dynamics*, 23(2):21–32, 2014.

E. M. Crowley. Product partition models for normal means. *Journal of the American Statistical Association*, 92(437):192–198, 1997.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

M. Givon. Variety seeking, market partitioning and segmentation. *International Journal of Research in Marketing*, 2(2):117–127, 1985.

B. E. Kahn. Consumer variety-seeking among goods and services: An integrative review. *Journal of Retailing and Consumer Services*, 2(3):139–148, 1995.

R. Loschi and F. Cruz. An analysis of the influence of some prior specifications in the identification of change points via product partition model. *Computational Statistics & Data Analysis*, 39(4):477–501, 2002.

R. H. Loschi, F. R. Cruz, P. L. Iglesias, and R. B. Arellano-Valle. A gibbs sampling scheme to the product partition model: an application to change-point problems. *Computers & Operations Research*, 30(3): 463–482, 2003.

C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, 2007. ISBN 9780387715995.

P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.