

# Chapter 2

## Inference Based on Complete Data

### 2.1 Introduction

In the statistical problem of nonparametric Bayesian analysis we have a random probability  $P$  belonging to  $\Pi$  and having a particular prior distribution. Given  $P = P$ , we also have a random sample  $X_1, \dots, X_n$ , which are iid  $P$  taking values in  $\chi$ . Based on the sample, our objective is to estimate a function  $\phi(P)$  of  $P$ , with respect to a certain loss function. Most of the applications presented in this chapter use the Dirichlet process prior or its variants—Dirichlet Invariant process and mixtures of Dirichlet processes. In Chap. 3 while dealing with censored data, we use other priors such as the neutral to the right processes which are more suited to such data, but obviously they are applicable in the uncensored data case as well.

In this chapter, first we will deal primarily with estimation problems and thereafter we will present hypothesis testing and other applications briefly. We will consider the distribution function (CDF) or its functionals. Since the Dirichlet process prior is conjugate, the strategy will be to obtain first the Bayes estimator of  $\phi$  for the no sample problem and then to update the parameter(s) of the prior to obtain Bayes estimator for any sample size  $n$ . Through out this chapter we assume that we have a random sample  $X_1, \dots, X_n$  from an unknown distribution function  $F$  (corresponding to  $P$ ) defined on the real line. In the case of two sample problem, we will have a second sample  $Y_1, \dots, Y_n$  from another distribution function, say,  $G$ . Both samples will be assumed to be independent. The loss functions used are a weighted (integral) squared error loss  $L_1$  for the distribution function and a squared error loss  $L_2$  for its functionals, where

$$L_1(F, \hat{F}) = \int (F(t) - \hat{F}(t))^2 dW(t); \quad L_2(\varphi, \hat{\varphi}) = (\varphi - \hat{\varphi})^2, \quad (2.1)$$

with  $W$  being a given weight function or a finite measure on  $(R, \mathcal{B})$ .

Through out this and the next chapter, we will denote the samples by bold letters, such as  $\mathbf{X} = (X_1, \dots, X_n)$ , the sample distribution by  $\hat{F}_n(t)$  and the Bayes estimator with respect to the Dirichlet prior  $\mathcal{D}(\alpha)$ , by  $\hat{F}_\alpha$ . Additionally, we let

$\bar{\alpha}(\cdot) = \alpha(\cdot)/\alpha(R)$ ,  $M = \alpha(R)$  and  $p_n = \alpha(R)/(\alpha(R) + n)$ . In some applications, we use  $\mathfrak{X}$  instead of  $R$ .

The topics in this chapter are organized as follows:

1. Estimation of a distribution function.
2. Tolerance region and Confidence bands.
3. Estimation of functionals of a distribution function.
4. Other applications
5. Bivariate distribution function.
6. Estimation of a function of  $P$ .
7. Two sample problems
8. Hypothesis Testing

Under these headings, applications to sequential estimation, empirical Bayes, linear Bayes, minimax estimation, bioassay, and other applications will be presented. The beauty of the Dirichlet process is that most of the results are in closed forms. Also, once the no-sample problem is solved, all that is needed to solve the problem for any sample size is to update the parameter of the Dirichlet process. This strategy is used repeatedly. Needless to say that many of the problems discussed here could also be solved by using other priors, such as processes neutral to the right, Polya trees, beta-Stacy, etc., although closed form of the results may not be guaranteed. Since the Dirichlet process is inadequate in handling problems such as density estimation, there has been recently intense activity in using Polya trees and mixtures of Polya trees in such problems. A brief discussion on these efforts is included in Sect. 2.5.4.

## 2.2 Estimation of a Distribution Function

In this section the Bayesian estimation of a distribution function with respect to the Dirichlet process and related prior processes is presented. Also included are the empirical Bayes, sequential and minimax estimation procedures.

### 2.2.1 Estimation of a CDF

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} F$ ,  $F$  defined on  $(R, \mathcal{B})$ . The objective is to estimate  $F$  based on  $\mathbf{X}$  under the loss function  $L_1$  and prior  $\mathcal{D}(\alpha)$ . For each  $t$ ,  $F(t) \sim Be(\alpha(-\infty, t], \alpha(t, \infty))$ . The risk is given by  $\mathcal{E}(L(F, \hat{F})) = \int \mathcal{E}(F(t) - \hat{F}(t))^2 dW(t)$ . The Bayes estimate of  $F$  for the no-sample problem is the posterior mean  $\hat{F}(t) = \mathcal{E}(F(t)) = F_0(t) = \alpha(-\infty, t]/\alpha(R)$ , where the expectation is taken with respect to  $\mathcal{D}(\alpha)$ . By Theorem 1.1 of Sect. 1.2, we have  $F|\mathbf{X} \in \mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$ . Therefore, for a sample of size  $n$ , the expectation is taken with respect to  $\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})$ , and the Bayes

estimator is  $\widehat{F}(t) = \mathcal{E}(F(t) \mid X_1, \dots, X_n)$  obtained as (Ferguson 1973)

$$\begin{aligned}\widehat{F}_\alpha(t) &= \widehat{F}(t \mid X_1, \dots, X_n) = \frac{\alpha(-\infty, t] + \sum_{i=1}^n \delta_{X_i}(-\infty, t]}{\alpha(R) + n} \\ &= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_n(t), \quad \text{say,}\end{aligned}\tag{2.2}$$

where  $\widehat{F}_n(t) = \frac{1}{n} \cdot \sum_{i=1}^n \delta_{X_i}(-\infty, t]$ , the empirical distribution function of the sample. Thus the Bayes rule  $\widehat{F}_\alpha$  may be interpreted as a mixture of the prior guess  $F_0$  and the empirical distribution function with respective weights,  $p_n$  and  $1 - p_n$ . At the same time,  $F_0$  can be interpreted as the ‘center’ around which the Bayes estimate resides. Robustness of this estimator is discussed in Hannum and Hollander (1983a, 1983b).

*Remark 2.1*  $M = \alpha(R)$  may be interpreted as a precision parameter or the prior sample size (Ferguson 1973). As  $\alpha(R) \rightarrow \infty$ ,  $\widehat{F}_\alpha$  reduces to the prior guess  $F_0$  at  $F$ . On the other hand, if  $\alpha(R) \rightarrow 0$ , the Bayes estimator reduces to the sample distribution function and hence it could be said that it corresponds to providing no information. However, Sethuraman and Tiwari (1982) take issue with this interpretation. For finite  $\alpha_0, \alpha_r, r \geq 1$  on  $R$ , they show that if, along with the sequence  $\alpha_r(R) \rightarrow 0$ , we have  $\sup_A |\bar{\alpha}_r(A) - \bar{\alpha}_0(A)| \rightarrow 0$  as  $r \rightarrow \infty$ ,  $A \in \mathcal{B}$ , then  $\mathcal{D}(\alpha_r) \rightarrow \delta_{Y_0}$ , where  $Y_0$  has the distribution  $\bar{\alpha}_0$ . This means that the information about  $P$  is that it is a probability measure concentrated at a particular point in  $R$ , and the point is selected according to  $\bar{\alpha}_0$ . This is a definite information about  $P$  and its discreteness.

### 2.2.2 Estimation of a Symmetric CDF

In the previous section,  $F$  was an arbitrary distribution function. Suppose now we wish to estimate  $F$  which is symmetric about a known point  $\eta$ . This suggests that the space of all distribution functions be restricted to the symmetric distributions only. So assume  $F$  to be distributed according to the Dirichlet Invariant process (Sect. 1.3), that is,  $F \in \mathcal{DGI}(\alpha)$ ,  $\mathcal{G} = \{e, g\}$  with  $e(x) = x$ ,  $g(x) = 2\eta - x$ . Then the Bayes estimate of  $F$  under the loss function  $L_1$  is (Dalal 1979a)

$$\begin{aligned}\widehat{F}_{\alpha\eta}(t \mid X_1, \dots, X_n) &= \frac{\alpha(-\infty, t] + (1/2) \sum_{i=1}^n (\delta_{X_i}(-\infty, t] + \delta_{2\eta - X_i}(-\infty, t])}{\alpha(R) + n} \\ &= p_n \cdot F_0(t) + (1 - p_n) \cdot \widehat{F}_{sn}(t),\end{aligned}\tag{2.3}$$

where  $\widehat{F}_{sn}(t)$  is  $\eta$ -symmetrized version of the empirical distribution. This is an analog of the Bayes estimator  $\widehat{F}_\alpha$ .

The Bayes estimator of  $F$  with respect to certain other prior processes will be presented when dealing with censored data in Chap. 3.

### 2.2.3 Estimation of a CDF with MDP Prior

Let  $G(\theta)$  stand for a random distribution function selected from a mixture of Dirichlet processes (Sect. 1.4) with index space  $U = R$ , parameter space  $\Theta = R$ , observation space also  $R$  and mixing distribution  $H$ . That is  $G \in \int_R D(\alpha_u) dH(u)$  and let  $\theta_1, \dots, \theta_n \stackrel{iid}{\sim} G$ , and given  $\theta_i$  let  $X_{i1}, \dots, X_{im_i}$  be a sample of size  $m_i$  from  $F_{\theta_i}(x)$ ,  $i = 1, \dots, n$ .

The Bayes estimate of  $G$  under the  $L_1$  loss function is given by  $\hat{G} = \mathcal{E}(G|\theta_1, \dots, \theta_n)$  if  $\theta_i$ 's are observed directly, and  $\hat{G} = \mathcal{E}(G|X_{i1}, \dots, X_{im_i})$  if  $X_{ij}$ 's are observed. One can use the formula given under property 4 of Sect. 1.4 to evaluate the former, and the formula given under property 5 to evaluate the latter. Antoniak (1974) has illustrated computational procedures by taking examples of small sample of size  $n = 2$ . As an example, he takes the transition measure  $\alpha_u(\cdot)/\alpha(R) = N(u, \sigma^2)$ , mixing distribution  $H = N(0, \rho^2)$  and sampling distribution  $F_\theta = N(\theta, \tau^2)$ , and obtains the expression for  $\hat{G}$  in a closed form. For larger sample size, the evaluations are difficult. However, computational algorithms are developed for handling such problems in Kuo (1986b), and are described in many publications that have appeared since her paper. See for example books by Dey et al. (1998) and Ibrahim et al. (2001).

In an effort to compromise between parametric and purely nonparametric models, Doss (1994) investigates prior distributions for  $F$  which give most of their mass to a “small neighborhood” of an “entire” parametric family. In other words, he considers the situation where a parametric family  $H_\theta$ ,  $\theta \in \Theta \subset R^p$  is specified. Thus, a prior on  $F$  is placed as follows. First choose  $\theta$  according to some prior  $\nu$ , then choose  $F$  from  $\mathcal{D}(\alpha(R)H_\theta)$  with specified  $\alpha(R) > 0$ . This leads to the mixture of Dirichlet processes priors,  $F \in \int \mathcal{D}_{\alpha(R)H_\theta} \nu(d\theta)$ . While this formulation encounters the same computational difficulties, it allows him to consider a more general set up when instead of exact values of  $X_i$  ( $\sim F$ ), it is only known that  $X_i \in A_i \subset R$ . Thus we may have  $A_i = \{x_i\}$  if  $X_i$  is an exact observation, and  $A_i = (c_i, \infty)$  if  $X_i$  is censored on the right by  $c_i$ . The task is to obtain the posterior distribution of  $F$  given the data. Doss develops an algorithm for generating a random distribution function from this conditional posterior distribution. Further details may be found in his paper.

### 2.2.4 Empirical Bayes Estimation

In the Sect. 2.2.2 we derived the Bayesian estimator of  $F$  assuming a Dirichlet process prior with parameter  $\alpha$ . It was assumed there that  $\alpha$  is known via a known prior guess  $F_0$  of  $F$ , and the total mass  $M$ . If this is not the case, we need to estimate  $F_0$  or  $M$  or both. This can be done via the *empirical Bayes* (EB) approach which is described now. The efficacy of the empirical Bayes estimator is judged by a criterion called ‘*asymptotic optimality*’: An empirical Bayes estimator is said to be *asymptotically optimal* relative to a class of Dirichlet process priors if the Bayes risk of the

EB estimator given  $\alpha$  converges to the Bayes risk of the Bayes estimator for all  $\alpha$ . This being a weak criterion, generally, the rate of convergence is also indicated.

Since  $\mathcal{E}[F(t)] = F_0(t)$  and  $\text{var}(F(t)) = F_0(t)(1 - F_0(t))/(M + 1)$ , the parameter  $\alpha$  is then expressed as  $\alpha(\cdot) = MF_0(\cdot)$ , which provides interpretation of  $M$  as a ‘precision’ or ‘accuracy’ or ‘uncertainty’ parameter and specification of  $F_0$  implies the random distribution function is centered around  $F_0$ . For this reason, it is felt that the empirical Bayes method, where the sample data is used for identifying  $F_0$ , is better rather than specifying some arbitrary  $F_0$ , whose validation may or may not be ascertained.

In the empirical Bayes framework, we are currently at the  $(n + 1)$ -th stage of an experiment, and information is available not only from the current stage, but also from the  $n$  previous stages. Thus we have a sequence of pairs  $(P_i, \mathbf{X}_i)$ ,  $i = 1, 2, \dots, n + 1$  of independent random elements, where  $P_i$ ’s are probability measures on  $(R, \mathcal{B})$  having a common Dirichlet process prior  $\mathcal{D}(\alpha)$ . Given  $P_i = P$ ,  $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{im_i})$  is a random sample of size  $m_i$  from  $P$ . The task is to estimate the distribution function corresponding to  $P$  at the  $(n + 1)$ -th stage or its functional. The strategy is to use the information provided by the previous  $n$  stages in estimating the parameters of the prior at the  $(n + 1)$ -th stage. This approach will be used in estimating the distribution function, the mean, and in general, any estimable parameters of degree 2 or 3.

*Remark 2.2* In many hierarchical modeling, the parameters at intermediate stages are assumed to have certain distributions with some hyper parameters. It is fine if there are valid justifications for such assignments. However, in absence of such information it is judged that the empirical Bayes methods may offer better solution since here the observed data itself is used to provide information on unknown parameters.

**Empirical Bayes Estimation of a CDF** Let  $F_1, F_2, \dots, F_{n+1}$  be  $n + 1$  distribution functions on the real line, and for  $i = 1, 2, \dots, n + 1$ , let  $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{im_i})$  be a sample of size  $m_i$  from  $F_i$ . We assume each  $F_i$  to have a common Dirichlet process prior,  $\mathcal{D}(\alpha)$ . Our prior information is incorporated through  $F_0$  and  $M$ . As before  $\hat{F}_j(t)$  is the sample distribution function of  $\mathbf{X}_j$  and  $p_j = \alpha(R)/(\alpha(R) + m_j)$ ,  $j = 1, \dots, n + 1$ . Consider the estimation of  $F_{n+1}(t)$  based on  $\mathbf{X}_1, \dots, \mathbf{X}_{n+1}$ . The Bayes estimator of  $F_{n+1}$  at the  $(n + 1)$ -th stage with respect to the prior  $\mathcal{D}(\alpha)$  and the loss function

$$L(F_{n+1}, \tilde{F}_{n+1}) = \int (F_{n+1}(t) - \tilde{F}_{n+1}(t))^2 dW(t) \quad (2.4)$$

is from Sect. 2.2.1 (suppressing the dependence on  $\alpha$ ),

$$\tilde{F}_{n+1}(t) = p_{n+1}F_0 + (1 - p_{n+1})\hat{F}_{n+1}(t). \quad (2.5)$$

The Bayes risk of  $\tilde{F}_{n+1}(t)$  with respect to  $D(\alpha)$ , denoted by

$$r(\alpha) = \mathcal{E}_{\mathcal{D}(\alpha)} \mathcal{E}_{F_{n+1}} [L(F_{n+1}, \tilde{F}_{n+1})],$$

is (Korwar and Hollander 1976)

$$\begin{aligned} r(\alpha) &= r(\tilde{F}_{n+1}, \alpha) = \frac{p_{n+1}}{\alpha(R) + 1} \int F_0(t)(1 - F_0(t))dW(t) \\ &= \frac{p_{n+1}}{\alpha(R) + 1} \sigma^2, \end{aligned} \quad (2.6)$$

where  $\sigma^2 = \int x^2 dF_0(x) - (\int x dF_0(x))^2$  is the variance of  $F_0$ .

If  $F_0$  and  $M$  are known,  $r(\alpha)$  can be evaluated completely. In the EB approach, we are able to estimate these parameters from the previous data and adjust this estimator, resulting in what is known as an *empirical Bayes* estimator.

Korwar and Hollander (1976), considered the case of  $\alpha$  when  $M$  is known but  $F_0$  is unknown. They estimated  $F_0(t)$  by the average of first  $n$  sample distributions,  $\frac{1}{n} \sum_{j=1}^n \hat{F}_j(t)$ , substituted this in (2.5) and proposed the following empirical Bayes estimator of  $F_{n+1}$ :

$$\bar{F}_{n+1}(t) = p_{n+1} \frac{1}{n} \sum_{j=1}^n \hat{F}_j(t) + (1 - p_{n+1}) \hat{F}_{n+1}(t). \quad (2.7)$$

They evaluated the Bayes risk of  $\bar{F}_{n+1}(t)$  as

$$r(\bar{F}_{n+1}, \alpha) = \mathcal{E}[L(F_{n+1}, \bar{F}_{n+1})] = r(\alpha) \left[ 1 + \frac{p_{n+1}}{n^2} \sum_{j=1}^n \frac{1}{1 - p_j} \right], \quad (2.8)$$

where the expectation is taken with respect to  $\mathcal{D}(\alpha)$  as well as  $\mathbf{X}_1, \dots, \mathbf{X}_{n+1}$ . When the samples are of same size as  $m$ ,  $r(\bar{F}_{n+1}, \alpha)$  reduces to  $r(\alpha)[1 + \alpha(R)/mn]$ . Clearly, as  $n \rightarrow \infty$ ,  $r(\bar{F}_{n+1}, \alpha) \rightarrow r(\alpha)$  for any  $\alpha$ . Thus they concluded that the estimator is asymptotically optimal and established the rate of convergence  $O(n^{-1})$ . Zehnwirth (1981) relaxed the assumption of  $M$  known in the case of equal sample size and estimated  $M$  in a clever way (to be described below) by the  $F$ -ratio statistic  $F_n$  in one-way analysis of variance based on  $\mathbf{X}_1, \dots, \mathbf{X}_n$  and showed that the resulting estimator of  $F_{n+1}$  is also asymptotically optimal with the same rate of convergence,  $O(n^{-1})$ .

Note that the estimators of  $F_0$  proposed by Korwar and Hollander and Zehnwirth were based only on the past data, but not the current data. Ghosh et al. (1989) modified these estimators to include the current data as well. Thus, it gives greater weight to the current data in estimating  $F_{n+1}(t)$  than that in the Hollander and Korwar and Zehnwirth estimators, and yields smaller risk than those estimators.

When  $\alpha(R)$  is known, their proposed empirical Bayes estimator of  $F_{n+1}(t)$  turns out to be

$$\tilde{F}_{n+1}^*(t) = p_{n+1} \hat{F}_0(t) + (1 - p_{n+1}) \hat{F}_{n+1}(t), \quad (2.9)$$

where  $\widehat{F}_0(t) = \sum_{j=1}^{n+1} (1 - p_j) \widehat{F}_j(t) / \sum_{j=1}^{n+1} (1 - p_j)$ , and the Bayes risk is

$$r(\widetilde{F}_{n+1}^*, \alpha) = r(\alpha) \left[ 1 + \frac{p_{n+1}}{\sum_{j=1}^{n+1} (1 - p_j)} \right], \quad (2.10)$$

which converges to  $r(\alpha)$  as  $n \rightarrow \infty$ , and hence it is asymptotically optimal. Comparing the risks of estimators with and without the use of the current data, it can be verified that  $r(\widetilde{F}_{n+1}^*, \alpha) - r(\alpha) \leq r(\overline{F}_{n+1}, \alpha) - r(\alpha)$  and hence an improvement is achieved by using the estimator  $\widetilde{F}_{n+1}^*(t)$  over  $\overline{F}_{n+1}$ . In fact if the sample sizes are equal,  $r(\widetilde{F}_{n+1}^*, \alpha) - r(\alpha) = (n/(n+1))[r(\overline{F}_{n+1}, \alpha) - r(\alpha)]$ .

Observe that  $\widetilde{F}_{n+1}^*(t)$  is a linear combination  $\sum_{j=1}^{n+1} w_j^* \widehat{F}_j(t)$ , with  $w_j^* = p_{n+1}(1 - p_j) / (\sum_{j=1}^{n+1} (1 - p_j))$ ,  $j = 1, \dots, n$ , and  $w_{n+1}^* = p_{n+1}(1 - p_{n+1}) / (\sum_{j=1}^{n+1} (1 - p_j)) + (1 - p_{n+1})$ . Clearly  $\sum_{j=1}^{n+1} w_j^* = 1$ . This gives a clue for them to show that indeed the Bayes risk of  $\widetilde{F}_{n+1}^*$  is smaller than the Bayes risk of any other estimator of the form  $\sum_{j=1}^{n+1} w_j \widehat{F}_j$  with  $\sum_{j=1}^{n+1} w_j = 1$ . By taking different choices of  $w_j$  we can see that this class includes the following estimators. The choice of  $w_j = p_m/n$  ( $j = 1, \dots, n$ ) and  $w_{n+1} = 1 - p_m$ , with  $m_1 = \dots = m_n = m$  and  $p_m = \alpha(R)/(\alpha(R) + m)$ , leads to Korwar and Hollander (1976) estimator  $\overline{F}_{n+1}(t)$ . Another possible choice of  $w_j = 1/(n+1)$  for  $j = 1, \dots, n+1$  which leads to the estimator  $\sum_{j=1}^{n+1} \widehat{F}_j/(n+1)$  of  $F_{n+1}$ . Also, the usual MLE estimator of  $F_{n+1}$  is  $\widehat{F}_{n+1}$  which is obtained when  $w_{n+1} = 1$ , and  $w_1 = \dots = w_n = 0$ .

When  $\alpha(R)$  is unknown, Zehnwirth (1981) proposed an estimator for  $\alpha(R)$  based on a one-way ANOVA table using the past data  $\mathbf{X}_1, \dots, \mathbf{X}_n$  for equal sample size  $m$  at each stage and proved it's consistency

$$m/(1 - F_n) \rightarrow \alpha(R) \quad \text{in probability as } n \rightarrow \infty. \quad (2.11)$$

Ghosh et al. (1989) provide an improvement over his estimator by including  $\mathbf{X}_{n+1}$  as well in the  $F_n$  statistic. Let  $\overline{X}_j = \sum_{i=1}^{m_j} X_{ji}/m_j$  be the mean of the sample values at  $j$ -th stage and  $\overline{X} = \sum_{j=1}^{n+1} m_j \overline{X}_j / \sum_{j=1}^{n+1} m_j$  denote the overall mean. Define

$$MSW = \sum_{j=1}^{n+1} \sum_{i=1}^{m_j} (X_{ji} - \overline{X}_j)^2 / \sum_{j=1}^{n+1} (m_j - 1) \quad \text{and} \quad (2.12)$$

$$MSB = \sum_{j=1}^{n+1} (\overline{X}_j - \overline{X})^2 / n,$$

the usual within and between mean squares, respectively. Simple evaluations involving the Dirichlet process yield (see Ghosh et al. 1989)

$$E(MSW) = (\alpha(R)/(\alpha(R) + 1))\sigma^2 \quad (2.13)$$

$$E(MSB) = (\alpha(R)/(\alpha(R) + 1))\sigma^2 + \xi.\sigma^2/(n(\alpha(R) + 1)), \quad (2.14)$$

where  $\xi = \sum_{j=1}^{n+1} m_j - \sum_{j=1}^{n+1} m_j^2 / \sum_{j=1}^{n+1} m_j$ . They proposed the following estimator of  $\alpha(R)$ .

$$\hat{\alpha}^{-1}(R) = \max\{0, (MSB/MSW - 1)n/\xi\}, \quad (2.15)$$

which is shown to be strongly consistent under some mild conditions. (Note that the Zehnwirth's (1981) estimator of  $\alpha(R)$  is based only on  $\mathbf{X}_1, \dots, \mathbf{X}_n$  and had assumed  $m_1 = \dots = m_{n+1}$ .) Substituting this estimator of  $\alpha(R)$  in  $p_j = (1 + m_j \alpha^{-1}(R))^{-1}$  they revise the estimate for  $F_0$  as

$$\tilde{F}_0(t) = \begin{cases} \sum_{j=1}^{n+1} (1 - \hat{p}_j) \hat{F}_j(t) / \sum_{j=1}^{n+1} (1 - \hat{p}_j), & \text{if } \hat{\alpha}^{-1}(R) \neq 0 \\ \sum_{j=1}^{n+1} \hat{F}_j(t) / (n+1), & \text{if } \hat{\alpha}^{-1}(R) = 0. \end{cases} \quad (2.16)$$

Finally for the case  $\alpha(R)$  unknown, Ghosh et al. (1989) utilizing these estimators proposed  $\hat{\hat{F}}_{n+1}$  as an improved empirical Bayes estimator of  $F_{n+1}$ , where

$$\hat{\hat{F}}_{n+1}(t) = \hat{p}_{n+1} \tilde{F}_0(t) + (1 - \hat{p}_{n+1}) \hat{F}_{n+1}(t), \quad (2.17)$$

and proved the asymptotic optimality of this estimator.

### 2.2.5 Sequential Estimation of a CDF

Ferguson (1982) derives the sequential estimator of  $F$  under the loss function  $L_1$  and prior  $\mathcal{D}(\alpha)$ , with  $F_0 = \bar{\alpha}$  as a specified distribution function on  $R$ . Then as noted earlier,  $\mathcal{E}(F) = F_0$  and the posterior distribution of  $F$ , given the sample  $X_1, \dots, X_n$  from  $F$ , is  $\mathcal{D}((M+n)\hat{F}_\alpha)$ , where  $\hat{F}_\alpha$ , as before, is the Bayes estimator of  $F$  under  $L_1$ , and the minimum Bayes risk is

$$\int \text{Var}(F(x)|\mathbf{X}) dW(x) = (1/(M+n+1)) \int_{\alpha} \hat{F}_\alpha(x) (1 - \hat{F}_\alpha(x)) dW(x). \quad (2.18)$$

In sequential estimation we need a stopping rule and a terminal estimator. It is enough to find a stopping rule, since once we have the stopping rule, the terminal Bayes estimator is  $\hat{F}_\alpha$  itself. Ferguson discusses the  $k$ -stage look ahead rule which, at each stage stops or continues according to whether the rule is optimal among those taking at most  $k$  more observations stops or continues. There is a positive cost  $c > 0$  to look for each additional observation. After observing  $X_1, \dots, X_n$ , the 1-stage look ahead rule that he develops calls for stopping after the first  $n$  observations for which

$$\int_{\alpha} \hat{F}_\alpha(1 - \hat{F}_\alpha) dW \leq c(M+n+1)^2. \quad (2.19)$$

Clearly the left hand side is bounded above by  $W(R)/4$  and the right hand side increases with  $n$ . Ferguson argues that the 1-stage look-ahead rule eventually calls

for stopping and bounds on the maximum sample size can be found. He provides justification for the optimality of this rule.

Ferguson also discusses the sequential estimation of the mean  $\mu = \int x dF(x)$  under the squared error loss function  $L_2$ . Let  $\mu_0 = \int x dF_0(x)$ , the prior estimate of the mean and  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . The minimum conditional Bayes risk is given by  $\text{Var}(\mu|X_1, \dots, X_n) = \sigma_n^2/(M+n+1)$ , where  $\sigma_n^2$  is the variance of the distribution  $\hat{F}_n$ , which can be expressed as

$$\begin{aligned} \sigma_n^2 &= \int (x - \mu_n)^2 d\hat{F}_n(x) \\ &= \left( M\sigma_0^2 + ns_n^2 + \frac{Mn}{M+n}(\bar{x}_n - \mu_0)^2 \right) / (M+n), \end{aligned} \quad (2.20)$$

with  $\sigma_0^2$  as the variance of  $F_0$  and  $ns_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2$ . Then his 1-stage rule is to stop after the first  $n$  observations for which  $\sigma_n^2 \leq c(M+n+1)^2$ . He also provides further some modified stopping rules and discusses their usage. His paper may be referred to for more details.

Sequential approach from the Bayesian point of view is also used by Hall (1976, 1977) in treating search problems with random overlook probabilities having a Dirichlet or a mixture of Dirichlet processes. Clayton and Berry (1985) treat one-armed bandit problem and Clayton (1985) a sequential testing problem for the mean of a population.

### 2.2.6 Minimax Estimation of a CDF

One of the first non-Bayesian application of the Dirichlet process was contained in Phadia (1971), where a sequence of Dirichlet process priors was used in deriving the minimax estimator of an unknown  $F$  based on a sample of size  $n$ . The technique was first to find an equalizer rule given by

$$\hat{F}_{mx}(t) = \frac{\sqrt{n}/2 + \sum_{i=1}^n \delta_{X_i}(-\infty, t]}{\sqrt{n} + n}.$$

Then a sequence of least favorable priors  $\mathcal{D}(\alpha_k)$  was defined, where  $\alpha_k$  was taken to be a finite measure giving equal weight  $\sqrt{n}/2$  to points  $\pm k$ ,  $k$  a nonnegative real number. Then it was shown that the Bayes risk of the Bayes estimator with respect to  $\mathcal{D}(\alpha_k)$  converges to the risk of the above equalizer rule as  $k \rightarrow \infty$ . Thus it was established that the above estimator was minimax under the  $L_1$  loss (minimax estimators for other loss functions were also obtained and in particular it was shown that the sample distribution function was minimax under a weighted quadratic loss function). However, at Ferguson's suggestion the results were simplified by taking a sequence of beta distribution priors (Phadia 1973).

## 2.3 Tolerance Region and Confidence Bands

In this section we first present the Tolerance region discussed by Ferguson in his 1973 paper, and then the construction of a confidence band as proposed by Breth (1978).

### 2.3.1 Tolerance Region

Ferguson (1973) treats the problem of deriving a tolerance region from a decision theoretic approach. Suppose we want to estimate the  $q$ -th quantile  $t_q$  of an unknown distribution  $F$  on the real line by an upper tolerance point  $a$  under the loss function

$$L(p, a) = pP((-\infty, a]) + qI_{(a, \infty)}(t_q), \quad (2.21)$$

where  $p$  is a constant,  $0 < p < 1$ . If  $t_q$  is known exactly,  $L$  is minimized by choosing  $a = t_q$ . But if  $t_q$  is not known precisely, then we need to minimize the Bayes risk with respect to the  $\mathcal{D}(\alpha)$  given by

$$\mathcal{E}(L(p, a)) = pu + q \int_0^q \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1} (1-z)^{(1-u)M-1} dz, \quad (2.22)$$

where  $u$  represents  $F_0(a)$ , and  $M = \alpha(R)$  as before. Let  $u = f(p, q, M)$  denote the point at which the minimum occurs. Then the Bayes rule for the no sample problem is given by  $a = f(p, q, \alpha(R))$ -th quantile of  $F_0$ . For a sample  $X_1, \dots, X_n$  of size  $n$ , the Bayes rule therefore is given by

$$\hat{a}_n(\mathbf{X}) = f(p, q, \alpha(R) + n)\text{-th quantile of } \hat{F}_n. \quad (2.23)$$

### 2.3.2 Confidence Bands

In the classical theory, confidence bands  $(F_L, F_U)$  for an unknown distribution function  $F$  are constructed for a given confidence level  $1 - \nu$ , such that  $\mathcal{P}(F_L \leq F \leq F_U) = 1 - \nu$ . Here  $F$  is considered to be fixed while  $F_L$  and  $F_U$  are random, they being functions of ordered sample values. In the Bayesian context, it is the other way around— $F$  is considered to be random and  $F_L$  and  $F_U$  are fixed and determined in terms of the prior and posterior probabilities. Breth (1978) treats this problem.

**Definition 2.3** (Breth) Suppose  $F \in \mathcal{D}(\alpha)$ . Then if  $\mathcal{P}\{F_L(t) \leq F(t) \leq F_U(t) \text{ for all } t\} = \nu_1$  ( $\nu_2$ ) is a prior (posterior) probability, the functions  $F_L(t)$  and  $F_U(t)$  constitute the boundaries for a fixed region within which the random distribution function lies with prior (posterior) probability  $\nu_1$  ( $\nu_2$ ).  $(F_L(t), F_U(t))$  are defined to be a pair of Bayesian confidence bands for the random distribution function  $F$  with prior (posterior) probability  $\nu_1$  ( $\nu_2$ ).

Let  $m$  be a fixed positive integer and for  $i = 1, 2, \dots, m$  define  $u_i$  and  $v_i$  such that  $u_i < v_i$  for all  $i$  and  $0 = u_0 \leq u_1 \leq \dots \leq u_m < 1$ ,  $0 < v_1 \leq v_2 \leq \dots \leq v_{m+1} = 1$ . Further, let  $I(x) = 1$  if  $x \geq 0$  and 0 otherwise, and  $J(x) = 1$  if  $x > 0$  and 0 otherwise. For  $-\infty = t_0 < t_1 < t_2 < \dots < t_m < t_{m+1} = \infty$ , define  $F_L(x) = \sum_{i=1}^m (u_i - u_{i-1})I(x - t_i)$  and  $F_U(x) = v_1 + \sum_{i=1}^m (v_{i+1} - v_i)J(x - t_i)$ . Also, for  $a > 0$ , let  $\alpha(R) = a + 1 > 0$  and  $\alpha(t)/\alpha(R)$  be a distribution function.

It is clear that  $\mathcal{P}\{F_L(t) \leq F(t) \leq F_U(t) \text{ for all } t\} = \mathcal{P}\{u_j \leq F(t_j) \leq v_j \text{ for } j = 1, \dots, m\}$ . Therefore, to be able to calculate the probabilities of this type, it suffices to be able to calculate general rectangular probabilities (Steck 1971) over the ordered Dirichlet distribution, since  $(F(t_1), \dots, F(t_m)) \sim D(a_1, \dots, a_m; a_{m+1})$  with  $a_j = \alpha(t_j) - \alpha(t_{j-1})$ ,  $j = 1, \dots, m$ . To calculate the boundaries with respect to the posterior probability, replace  $\alpha$  by  $\alpha^* = \alpha + n\hat{F}_n$ , where  $\hat{F}_n$  is the sample distribution function for the sample of size  $n$ .

It should be noted that as in the classical theory, there are many pairs of Bayesian confidence bands for  $F$  with the same probability content  $1 - \nu$ , say. In practice, a particular pair must be chosen to express quantitative confidence in  $F$ .

Breth (1978) uses recursive methods for computing  $\mathcal{P}\{u_j \leq F(t_j) \leq v_j \text{ for all } j\}$  for fixed numbers  $\{u_j\}$ ,  $\{v_j\}$  and  $\{t_j\}$  when  $F$  is a Dirichlet process. For details on calculations that are needed in practical applications, one can refer to his paper. In a follow up paper he (Breth 1979) discusses construction of Bayesian confidence intervals for quantiles and the mean, and also treats Bayesian tolerance intervals. The complexity in numerical calculation is evident. If  $\alpha$  is not stipulated a priori, it can be estimated (see Korwar and Hollander 1973).

In this connection it is worth mentioning that in non-Bayesian context, Phadia (1974) constructed the best invariant one and two-sided confidence bands for an unknown continuous distribution function. They were invariant under the group  $\mathcal{G}$  of transformations  $g_\phi(y_1, \dots, y_n) = (\phi(y_1), \dots, \phi(y_n))$ , where  $\phi$  is a continuous, strictly increasing function from  $R$  onto  $R$ . The confidence bands were step functions taking jumps at the ordered sample values. For a given confidence level, the values of jumps were calculated as a minimization problem using Steck's result.

**Simulation Method** Neath and Bodden (1997) also constructed  $(1 - \gamma)100\%$  Bayesian confidence bands  $F_L$  and  $F_U$  by using a simulation method. Let  $P$  be a random probability measure having a mixture of Dirichlet processes as prior distribution. In other words,  $\theta \sim G$ ,  $P|\theta \in \mathcal{D}(\alpha_\theta)$ . Let  $F$  be a distribution function corresponding to  $P$ . Given a random sample from  $F$ , first a value of  $\theta$  is obtained from the posterior distribution  $G_{\mathbf{X}}$  of  $\theta$ , given  $\mathbf{X}$ . The posterior distribution of  $F$  given the data and  $\theta$  is  $\mathcal{D}(\alpha_\theta + \sum_{i=1}^n \delta_{x_i})$ . However, this distribution is analytically intractable. Therefore, in constructing the confidence bands, they treat simulated sample of distribution functions  $F_1, \dots, F_N$  as the actual distributions and choose  $F_L$  and  $F_U$  such that

$$\frac{1}{N} \sum_{i=1}^N I\{F_L(t) \leq F_i(t) \leq F_U(t), t \in R\} \geq 1 - \gamma. \quad (2.24)$$

In choosing the ‘best’ bounds, the following two criteria are used.

$$\begin{aligned} \text{(i)} \quad & \min \left\{ \max_t [F_U(t) - F_L(t)] \right\} \quad \text{and} \\ \text{(ii)} \quad & \min \left\{ \int [F_U(t) - F_L(t)] dW(t) \right\}. \end{aligned} \quad (2.25)$$

The minimum is taken over all functions  $F_L$  and  $F_U$  such that (2.24) is satisfied. For the process of choosing  $F_L$  and  $F_U$ , they give two algorithms and discuss their implementation. They also provide a numerical example to illustrate the procedure.

**Bayesian Bootstrap Method** Hjort (1985) uses a Bayesian bootstrap method to construct confidence intervals for a function  $\theta(F)$  of an unknown  $F$ . Let  $F \in \mathcal{D}(\alpha)$  and  $X_1, \dots, X_n \stackrel{iid}{\sim} F$ . Then  $F|\mathbf{X} \in \mathcal{D}(\alpha + \sum_{i=1}^n \delta_{x_i})$  which can be written as  $\mathcal{D}(MF_0 + n\hat{F}_n)$  with  $F_0 = \alpha/M$ ,  $M = \alpha(R)$ . Let  $G(t) = P(\theta(F) \leq t | \hat{F}_n)$ . We need to find  $\theta_L$  and  $\theta_U$  such that  $P(\theta_L \leq \theta(F) \leq \theta_U) = 1 - 2\nu$ , say. Thus  $G^{-1}(\nu)$  and  $G^{-1}(1 - \nu)$  are the natural choices for  $\theta_L$  and  $\theta_U$ , respectively with  $G^{-1}(p) = \inf\{t : G(t) \geq p\}$ .

## 2.4 Estimation of Functionals of a CDF

In this section we discuss various applications in which Bayesian estimators of certain functionals such as the mean, median, variance, etc. are derived using the Dirichlet process priors.

### 2.4.1 Estimation of the Mean

Ferguson (1973) considered the Bayesian estimation of the mean  $\mu = \int x dP(x)$  with respect to the Dirichlet process prior and under the squared error loss  $L_2$ . It is assumed that  $\alpha$  has a finite first moment. The Bayes rule for the no-sample problem is the mean of  $\mu$ , say,  $\mu_0$  which, by property 1 of Sect. 1.2.2, is  $\hat{\mu} = \mathcal{E}_{\mathcal{D}(\alpha)} \int x dP(x) = \int x d\alpha(x)/\alpha(R) = \mu_0$ . The Bayes rule for a sample of size  $n$  therefore is obtained by updating the parameter  $\alpha$  to  $\alpha + \sum_{i=1}^n \delta_{x_i}$  and is given by

$$\begin{aligned} \hat{\mu}_{\alpha n}(\mathbf{X}) &= (\alpha(R) + n)^{-1} \int x d \left( \alpha(x) + \sum_{i=1}^n \delta_{x_i}(x) \right) \\ &= p_n \mu_0 + (1 - p_n) \bar{X}_n, \end{aligned} \quad (2.26)$$

where  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  is the sample mean. The Bayes estimator thus is, like  $\hat{F}_\alpha$ , a convex combination of the prior guess at  $\mu$ , namely  $\mu_0$ , and the sample mean.

As  $\alpha(R) \rightarrow 0$ ,  $\hat{\mu}_{\alpha n} \rightarrow \bar{X}_n$ , and as  $\alpha(R) \rightarrow \infty$ ,  $\hat{\mu}_n \rightarrow \mu_0$ . The Bayes risk of  $\hat{\mu}_n$  is  $r(\alpha) = p_n \sigma^2 / (\alpha(R) + n)$ . Alternatively, the estimator can also be obtained by taking  $g(x) = x$  in property 9 of Sect. 1.2.2. More generally, let  $Z$  be a measurable real valued function defined on  $(R, \mathcal{B})$  and  $\theta = \int Z dP$ . If  $P \in \mathcal{D}(\alpha)$  and  $\theta_0 = \int Z d\alpha / \alpha(R) < \infty$ , then the Bayes estimator of  $\theta$  under the loss  $L_2$  is given by

$$\hat{\theta}_{\alpha n}(\mathbf{X}) = p_n \theta_0 + (1 - p_n) \frac{1}{n} \sum_{i=1}^n Z(X_i). \quad (2.27)$$

Yamato (1984) showed that the mean  $\mu$  is distributed symmetrically about a constant  $\theta$  if the measure  $\alpha$  is symmetric about  $\theta$  and  $\int |x| d\alpha(x) < \infty$ .

If  $\alpha$  and  $\alpha(R)$  are unknown, the empirical Bayes method can be used.

**Empirical Bayes Estimation of the Mean** This can be dealt with in the same manner as the distribution function in Sect. 2.2.4 and the same notations will be used here as well. The Bayes estimator with respect to  $\mathcal{D}(\alpha)$  at the  $(n+1)$ -th stage is given by

$$\hat{\mu}_\alpha = p_{n+1} \mu_0 + (1 - p_{n+1}) \sum_{i=1}^{m_{n+1}} X_{n+1,i} / m_{n+1}. \quad (2.28)$$

The Bayes risk of  $\hat{\mu}_\alpha$  is given by  $r(\alpha) = p_{n+1} \sigma^2 / (\alpha(R) + 1)$ . For the empirical Bayes approach,  $\mu_0$  is estimated from the first  $n$  samples by Korwar and Hollander (1976) and the resulting estimator  $\hat{\mu}_n$  has the Bayes risk as  $r(\hat{\mu}_n, \alpha) = (1 + \alpha(R) / \sum_{i=1}^{n+1} m_i) r(\alpha)$ . Ghosh et al. (1989) estimates  $\mu_0$  from the past as well as current sample data as  $\hat{\mu}_0 = \sum_{j=1}^{n+1} (1 - p_j) \bar{X}_j / \sum_{j=1}^{n+1} (1 - p_j)$  and plugs in  $\hat{\mu}_\alpha$ . The resulting estimator is  $\hat{\mu}_{n+1}$  and its Bayes risk is

$$r(\hat{\mu}_{n+1}, \alpha) = r(\alpha) + p_{n+1}^2 \sigma^2 / \sum_{j=1}^{n+1} (1 - p_j). \quad (2.29)$$

They have shown that  $\hat{\mu}_{n+1}$  is asymptotically optimal and has a smaller Bayes risk than the estimator proposed by Korwar and Hollander (1976). Again, if  $\alpha(R)$  is unknown, it can be estimated as indicated in Sect. 2.2.4.

In the context of a finite population of size  $N$ , Binder (1982) considered the task of Bayes estimation of the population mean  $\sum_{i=1}^N X_i / N$ , where  $X_1, \dots, X_N$  are population values, by assuming that there is a super population  $P$  with prior  $\mathcal{D}(\alpha)$ , and given  $P$  these values are iid  $P$ .

Ghosh et al. (1989) have also considered the empirical Bayes estimation of the finite population distribution function. Tiwari and Lahiri (1989) have treated the Bayes and empirical Bayes estimation of variances from stratified samples and studied the risk performance of the empirical Bayes estimators.

### 2.4.2 Estimation of a Variance

Consider now the task of estimating the variance of an unknown probability distribution  $P$ . If  $\alpha$  has a finite second moment, then the variance of  $P$  is defined by

$$\text{Var } P = \int x^2 dP(x) - \left( \int x dP(x) \right)^2, \quad (2.30)$$

which is a random variable. Ferguson (1973) obtained the Bayes estimator under the squared error loss  $L_2$  assuming the Dirichlet process prior. The Bayes estimator of  $\text{Var } P$  for the no-sample problem is the posterior mean

$$\begin{aligned} \mathcal{E} \text{Var } P &= \mathcal{E} \int x^2 dP(x) - \mathcal{E} \left( \int x dP(x) \right)^2 = (\sigma_0^2 + \mu_0^2) - \left( \frac{\sigma_0^2}{\alpha(R) + 1} + \mu_0^2 \right) \\ &= \frac{\alpha(R)}{\alpha(R) + 1} \sigma_0^2, \end{aligned} \quad (2.31)$$

where  $\mu_0$  is as defined above in Sect. 2.4.1 and  $\sigma_0^2 = \int x^2 d\alpha(x)/\alpha(R) - \mu_0^2$  is the variance of  $F_0$ .

For a sample of size  $n$ , the Bayes rule is therefore obtained by replacing the parameter  $\alpha$  by  $\alpha + \sum_{i=1}^n \delta_{X_i}$ . After some simplification and rearrangement, we get the Bayes estimator of  $\text{Var } P$  as

$$\begin{aligned} \hat{\sigma}_n^2(\mathbf{X}) &= \frac{\alpha(R) + n}{\alpha(R) + n + 1} \text{Var}(\hat{F}_\alpha) \\ &= \frac{\alpha(R) + n}{\alpha(R) + n + 1} (p_n \sigma_0^2 + (1 - p_n) s_n^2 + p_n (1 - p_n) (\mu_0 - \bar{X}_n)^2) \\ &= \frac{\alpha(R) + n}{\alpha(R) + n + 1} \\ &\quad \times \left( p_n \sigma_0^2 + (1 - p_n) \left( p_n \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2 + (1 - p_n) s_n^2 \right) \right), \end{aligned} \quad (2.32)$$

where  $s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . The last equality expresses  $\hat{\sigma}_n^2$  as a mixture of three different estimates of the variance, as noted by Ferguson.

If the prior sample size  $\alpha(R) \rightarrow 0$ , keeping  $F_0$  fixed,  $\hat{\sigma}_n^2$  converges to the estimate  $\frac{1}{n+1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . This estimate is the best invariant or minimax estimator of the variance of a normal distribution under the loss  $(\text{Var } P - \hat{\sigma}^2)^2 / (\text{Var } P)^2$ .

### 2.4.3 Estimation of the Median

Next consider the problem of estimation of the median  $\theta$  defined as  $\theta = \text{med } P$ . Ferguson (1973) derived the Bayes estimator under the absolute error loss,  $L(\theta, \hat{\theta}) =$

$|\theta - \hat{\theta}|$ .  $\theta$  is unique with probability one and thus a well defined random variable. Under this loss function, any median of the distribution of  $\theta$  is a Bayes estimator of  $\theta$ . For the Dirichlet process prior with parameter  $\alpha$ , Ferguson points out that any median of the distribution of  $\theta$  is a median of the expectation of  $P$ , and conversely,

$$\text{med}(\text{dist. med } P) = \text{med } \mathcal{E}P. \quad (2.33)$$

Thus any number  $t$  satisfying

$$\frac{\alpha((-\infty, t))}{\alpha(R)} \leq \frac{1}{2} \leq \frac{\alpha((-\infty, t])}{\alpha(R)} \quad (2.34)$$

is a Bayes estimate of  $\theta$  with respect to the prior  $\mathcal{D}(\alpha)$  and absolute error loss. With  $F_0(t) = \alpha((-\infty, t])/\alpha(R)$ , the Bayes estimate for the no-sample problem is  $\hat{\theta} = \text{median of } F_0$  and for the sample of size  $n$ , it is

$$\hat{\theta}_{an} = \text{median of } \hat{F}_\alpha, \quad (2.35)$$

where  $\hat{F}_\alpha$  is the Bayes estimate of  $F$  derived in Sect. 2.2.1.

Doss (1985a, 1985b) also considers the problem of estimating the median but in a different nonparametric Bayesian framework. Let  $X_1, \dots, X_n$  be a random sample with distribution  $F_\theta$ , where  $F_\theta(x) = F(x - \theta)$  for some  $F$  that has median 0.  $F$  is assumed to be unknown and the problem is to estimate  $\theta$ . Rather than placing a prior on  $F$ , he chooses  $F_-$  and  $F_+$  from  $\mathcal{D}(\alpha_-)$  and  $\mathcal{D}(\alpha_+)$ , respectively, and defines  $F(t) = (F_-(t) + F_+(t))/2$ , where  $\alpha_-$  and  $\alpha_+$  are the restriction of  $\alpha$  to  $(-\infty, 0)$  and  $(0, \infty)$ , respectively. Then,  $F$  is a random distribution function such that  $F(0) = \frac{1}{2}$  (but not symmetric, although  $E(F(t)) = F_0$ ). In fact  $F$  so defined is a mixture of two Dirichlet processes. Let  $\mathcal{D}^*(\alpha)$  denote its distribution.

Let  $\alpha = MF_0$ , where  $F_0$  is a distribution function with median zero and for simplicity, no mass at zero. He places a prior on the pair  $(F, \theta)$  by assuming  $F$  and  $\theta$  independent,  $F \in \mathcal{D}^*(\alpha)$  and  $\theta$  having an arbitrary distribution  $\nu$ . Given  $\theta$  and  $F$ , let  $\mathbf{X}$  be a sample from  $F(x - \theta)$ . Assume that  $F_0$  has continuous density  $f_0$ . Then, Doss obtains the marginal posterior distribution of  $\theta$  given  $\mathbf{X}$  as

$$d\nu(\theta|\mathbf{X}) = \kappa(\mathbf{X}) \prod_{i=1}^n [f_0(X_i - \theta)] \Psi(\mathbf{X}, \theta) d\nu(\theta), \quad (2.36)$$

where  $\Psi^{-1}(\mathbf{X}, \theta) = \Gamma(M/2 + n\hat{F}_n(\theta))\Gamma(M/2 + n(1 - \hat{F}_n(\theta)))$ ,  $\hat{F}_n$  the sample distribution function,  $\prod^*$  represents the product taken over the distinct  $X_i$  and  $\kappa(\mathbf{X})$  is a normalizing constant.

Using the posterior distribution one can find the Bayes estimate of  $\theta$ . Doss states that the estimator is essentially a convex combination of the maximum likelihood estimator with respect to  $F_0$  and the sample median, with mixing weights depending on the sample values. He also shows that the Bayes estimator is consistent only if the true distribution of  $X_j$  is discrete. He also derives the posterior distribution of  $\theta$  in the case of  $F$  being a ‘neutral to the right type’ distribution discussed in Sect. 1.5.

### 2.4.4 Estimation of the $q$ -th Quantile

Ferguson (1973) extends the estimation of the median to the  $q^{\text{th}}$  quantile of  $P$ , denoted by  $t_q$ :  $P((-\infty, t_q)) \leq q \leq P((-\infty, t_q])$ , for  $0 < q < 1$ . The  $q^{\text{th}}$  quantile of  $P \in \mathcal{D}(\alpha)$  is unique with probability 1, so that  $t_q$  is a well defined random variable. He considers the following loss function,

$$\begin{aligned} L(t_q, \hat{t}_q) &= p(t_q - \hat{t}_q) \quad \text{if } t_q \geq \hat{t}_q \\ &= (1 - p)(t_q - \hat{t}_q) \quad \text{if } t_q < \hat{t}_q, \end{aligned} \quad (2.37)$$

for some  $p$ ,  $0 < p < 1$ . For this loss, any  $p^{\text{th}}$  quantile of the distribution of  $t_q$  is a Bayes estimator of  $t_q$ . The distribution of  $t_q$  is

$$\begin{aligned} \mathcal{P}\{t_q \leq t\} &= \mathcal{P}\{F(t) > q\} \\ &= \int_q^1 \frac{\Gamma(M)}{\Gamma(uM)\Gamma((1-u)M)} z^{uM-1} (1-z)^{(1-u)M-1} dz, \end{aligned} \quad (2.38)$$

where  $M = \alpha(R)$  and  $u = \alpha((-\infty, t])/\alpha(R) = F_0(t)$ . Setting this expression equal to  $p$  and solving the resulting equation for  $t$ , Ferguson obtains the  $p^{\text{th}}$  quantile of  $t_q$ . For fixed  $p$ ,  $q$ , and  $M$ , let this equation define a function  $u(p, q, M)$ . The Bayes estimate of  $t_q$  for the no-sample problem is the  $u^{\text{th}}$  quantile of  $F_0$ ,

$$\hat{t}_q = u(p, q, \alpha(R))\text{-th quantile of } F_0, \quad (2.39)$$

and for the sample of size  $n$ , it is

$$\hat{t}_q(\mathbf{X}) = u(p, q, (\alpha(R) + n))\text{-th quantile of } \hat{F}_\alpha. \quad (2.40)$$

If  $p$  and  $q$  are both  $\frac{1}{2}$ , this reduces to the estimate of the median, since  $u(\frac{1}{2}, \frac{1}{2}, M) = \frac{1}{2}$  for all  $M$ .

Doss (1985a, 1985b) extends his results of estimating the median to the estimation of quantiles as well, and discusses their properties.

### 2.4.5 Estimation of a Location Parameter

Dalal (1979b) considered the following model for sample observations. Let  $X = \eta + \varepsilon$ , where  $\eta$  is the location parameter and  $\varepsilon$ , the error term. Assume that  $\eta$  and  $\varepsilon$  are independent. The objective is to estimate  $\eta$  based on a random sample  $Y_1, \dots, Y_n$  from a  $\eta$ -symmetric distribution function  $F_\eta$ . That is  $F_\eta$  is assumed to be symmetric about  $\eta$ , but otherwise  $\eta$  and  $F_\eta$  are unknown. If  $\varepsilon \sim G$  and  $G \in \mathcal{D}(MF_0)$ , where  $F_0$  could be a standard normal distribution, then  $\mathcal{E}(G) = F_0$  and hence the errors are generated by a distribution in the neighborhood of  $F_0$ . With  $M$  large the neigh-

borhood becomes concentrated around  $F_0$ . Thus Dalal argues that the model can be interpreted from a robustness perspective as well. Let  $\eta$  be distributed according to a prior distribution  $\nu$ , the group of transformations  $\mathcal{G} = \{e, g\}$  with  $e(x) = x$ ,  $g(x) = 2\eta - x$ , and  $\alpha$  be a  $\eta$ -symmetric non-null finite measure on  $(R, \mathcal{B})$ . Given  $\eta$ , Dalal (1979b) assumes  $F_\eta$  to be distributed according to the Dirichlet Invariant process,  $\mathcal{DGI}(\alpha)$ , and obtains a Bayes estimate  $\hat{\eta}(\mathbf{y}) = \mathcal{E}_{\eta|\mathbf{y}}(\eta)$  of  $\eta$ , where the expectation is taken with respect to the conditional distribution  $\nu(\cdot|\mathbf{y})$  of  $\eta$  given  $\mathbf{y}$  averaged over  $F_\eta$ . However,  $\hat{\eta}(\mathbf{y})$  is not in a closed form and he encounters computational difficulties which is illustrated by an example consisting of 2 observations.

Let  $\alpha = MF_0$ , and assume that  $F_0$  has a density  $f_0$ , and that we have a sample of size one,  $Y_1 \sim F_\eta$  with  $F_\eta \in \mathcal{DGI}(\alpha)$ . Then  $\mathcal{E}(F_\eta) = F_0$  and the marginal conditional distribution of  $Y_1$  given  $\eta$  is  $F_0$ . Since  $\nu$  is prior distribution of  $\eta$ , the conditional density of  $\eta|Y_1 = y_1 \sim f_0(y_1)/\int f_0(x)d\nu(x)$ .

If we have a second observation  $y_2$ , then we run into difficulty since the distribution of  $Y_2|y_1, \eta, F_\eta \sim F_\eta$ . But  $F_\eta|y_1, \eta \in \mathcal{DGI}(\alpha + \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{2\eta-y_1})$  which results in a distribution of  $Y_2|y_1, \eta$  as a combination of continuous and discrete parts with point discrete masses at  $y_1$  and  $2\eta - y_1$ . Thus the evaluation of the posterior distribution of  $\eta|y_2, y_1$  gets complicated (see Dalal 1979b). The above argument is extended to the case of  $n$  observations and shown that if  $\nu$  is absolutely continuous, the posterior distribution of  $\eta|\mathbf{y}$  is a mixture of absolutely continuous and discrete probabilities. The mixing weights depend upon not only the distinct observations but also on their multiplicities. The discrete component concentrates its mass on the points  $(y_i + y_j)/2, i \neq j$ . However, the computational techniques lately developed should make it easy to compute the posterior distribution.

This and other aspects of Bayesian estimation of a location parameter are discussed in his paper in detail.

Doss (1985a) also discusses this model, but instead of errors drawn from a symmetric distribution, he takes them to be drawn from an  $F$  which has median 0, but otherwise unknown, and it is desired to estimate  $\eta$ . He places priors on the pair  $(F, \eta)$  and computes the marginal posterior distribution of  $\eta$  and takes the mean of the distribution as the estimate of  $\eta$ . In a follow up paper (Doss 1985b) he discusses consistency issues and shows that the Bayes estimates are consistent if the distribution of errors is discrete, otherwise they can be inconsistent.

### 2.4.6 Estimation of $P(Z > X + Y)$

Zalkikar et al. (1986) considered the problem of estimation of the parameter

$$\Delta(F) = P(Z > X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(x + y) dF(x) dF(y), \quad (2.41)$$

where it is assumed that the random variables  $X$ ,  $Y$  and  $Z$  are independent from the distribution function  $F$ , and  $S = 1 - F$ . This problem is encountered in reliability theory where it is desired to test whether new is better than used. Assume  $F \in \mathcal{D}(\alpha)$  and the squared error loss  $L_2$ . Based on a random sample  $\mathbf{X} = (X_1, \dots, X_m)$  from  $F$ , they derived the Bayes estimator as follows.

$$\begin{aligned} \widehat{\Delta}(F) = \frac{M+m}{(M+m)^{(3)}} & \left[ 2(M+m+1) + \widehat{F}_\alpha(0-) \right. \\ & \left. + (M+m) \int_{-\infty}^{\infty} S_\alpha(2y) d\widehat{F}_\alpha(y) + (M+m)^2 \Delta(\widehat{F}_\alpha) \right], \end{aligned} \quad (2.42)$$

where  $a^{(k)} = a(a+1) \dots (a+k-1)$ ,  $S_\alpha = 1 - \widehat{F}_\alpha$ .

When  $M \rightarrow 0$ , the estimator reduces to an estimator which is asymptotically equivalent to a U-statistic,

$$U_m = \frac{1}{m(m-1)(m-2)} \sum I[X_i > X_j + X_k], \quad (2.43)$$

where the summation is taken over all  $m(m-1)(m-2)$  distinct triplets  $(i, j, k)$ ,  $1 \leq i, j, k \leq m$ .

By using the earlier mentioned technique, they obtain an empirical Bayes estimate at the  $(n+1)$ -th stage utilizing the past as well as current data, which is also shown to be asymptotically optimal with the rate of convergence  $O(n^{-1})$ .

## 2.5 Other Applications

There are many other applications that have not been covered here. For example, Lo (1987) has studied the Bayesian bootstrap estimation of a functional  $\theta(P; X_1, \dots, X_n)$ , where the variables  $X_1, \dots, X_n$  are iid  $P$ , with  $P$  having the prior  $\mathcal{D}(\alpha)$ . He has provided large sample Bayesian bootstrap probability intervals for the mean, variance and confidence bands for the distribution function, the smoothed density and smoothed rate function. In a second paper (Lo 1988), he also considered a Bayesian bootstrap for a finite population.

Dirichlet process priors have also been used for bandits problem by Clayton and Berry (1985).

We do however present a few interesting applications.

### 2.5.1 Bayes Empirical Bayes Estimation

In a general empirical Bayes setting, we have  $n$  unobservable independent random variables  $\theta_i$ ,  $i = 1, 2, \dots, n$  from an unknown distribution  $G$ , and associated with

each  $\theta_i$ , we have a random variable  $X_i$  chosen independently from a distribution with density function  $f_i(x|\theta_i)$ ,  $i = 1, 2, \dots, n$ . The problem is to estimate  $\theta_i$ 's or  $G$  itself. A common procedure is to obtain first an estimator  $G_n$  of  $G$  from the data  $X_1, \dots, X_n$ , and then estimate  $\theta_i$  as the Bayes estimate with respect to the prior  $G_n$ . In the Bayesian approach to the empirical Bayes problem,  $G$  itself is to be considered random with a prior distribution. Berry and Christensen (1979) followed this route assuming the Dirichlet prior  $\mathcal{D}(\alpha)$  for  $G$ . Antoniak (1974) had shown that the posterior distribution of  $G$  is a mixture of Dirichlet processes with parameter  $\alpha + \sum_{i=1}^n \delta_{\theta_i}$  and mixing distribution  $H(\boldsymbol{\theta}|\mathbf{X})$ . Thus, the posterior distribution of  $G$  given  $\mathbf{X}$  in symbols is

$$G|\mathbf{X} \in \int \mathcal{D}\left(\alpha + \sum_{i=1}^n \delta_{\theta_i}\right) dH(\boldsymbol{\theta}|\mathbf{X}). \quad (2.44)$$

If we have an unconditional marginal distribution of  $\boldsymbol{\theta}$ , then  $dH(\boldsymbol{\theta}|\mathbf{X})$  can be expressed as

$$dH(\boldsymbol{\theta}|\mathbf{X}) = \prod_{j=1}^n f_j(x_j|\theta_j) dH(\boldsymbol{\theta}) / \left[ \int \prod_{j=1}^n f_j(x_j|\theta_j) dH(\boldsymbol{\theta}) \right]. \quad (2.45)$$

However, even in the simple case where  $f_i(x|\theta)$  is a binomial distribution with parameter  $\theta$ , Berry and Christensen (1979) found it difficult to evaluate and recommended some approximations. By using a lemma of Lo (1984), Kuo (1986a, 1986b) was able to express the Bayes estimator of  $\theta_i$  under the loss  $\sum_{i=1}^n (\theta_i - \widehat{\theta})^2$  in a concise form as a ratio of two  $n$ -dimensional integrals as follows.

$$\widehat{\theta}_i = \mathcal{E}(\theta_i|\mathbf{X}) = \frac{\int \dots \int_{R^n} (\theta_i \prod_{i=1}^n f_i(x_i|\theta_i)) \prod_{i=1}^n (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j}) (d\theta_i)}{\int \dots \int_{R^n} (\prod_{i=1}^n f_i(x_i|\theta_i)) \prod_{i=1}^n (\alpha + \sum_{j=1}^{i-1} \delta_{\theta_j}) (d\theta_i)} \quad (2.46)$$

for all  $i = 1, 2, \dots, n$ . Still it is hard to evaluate these integrals. She overcomes this problem by decomposing each of the multi-dimensional integrals as a weighted average of products of one dimensional integrals and approximating each of the weighted averages by an importance sampling Monte Carlo method. She illustrates the computation in detail with a numerical example.

This model has been discussed in Escobar and West (1995) and Escobar (1994). Lavine (1994) generalizes the approach by using a Polya tree prior for  $G$  and shows how the posterior distribution can be computed via the Gibbs sampler and demonstrates the advantages of using mixtures of Polya trees over mixtures of Dirichlet processes.

## 2.5.2 Bioassay Problem

The goal of the bioassay problem is to assess the dose-response relationship in a population. In particular, one is interested in the estimation of the distribution of

tolerance level to a drug administered to subjects at various dose levels. In order to determine an effective dose, one needs to collect data at different dose levels and their effect on the subject in mitigating the condition for which the drug is administered. The impact of the drug on subjects is represented by a CDF,  $F(t)$ , defined on  $[0, \infty)$  and represents the proportion of the population that would respond to dose  $t$ . This distribution is often known as dose-response curve in the field of bioassay.

Suppose a stimulus is administered to  $n_j$  subjects at dose level  $t_j$  with positive response in  $r_j$  subjects,  $j = 1, 2, \dots, L$ . Let  $F(t)$  represent the probability of getting positive response at dose level  $t$ . Thus  $r_j$ ,  $j = 1, 2, \dots, L$  are independent, each being a binomial random variable with parameters  $n_j$  and  $F(t_j)$ . Based on such quantal response data, the task is to estimate the response curve  $F$  nonparametrically from a Bayesian approach. This problem was first considered by Kraft and van Eeden (1964) who use a dyadic tailfree process as prior. The computations are difficult and were illustrated in the case of only three dose levels in their paper. Ramsey (1972) uses a Dirichlet process prior and obtains the modal estimates of  $F$  by maximizing the finite dimensional joint density of the posterior distribution which is not a Dirichlet.

Ferguson and Phadia (1979) noted that the bioassay problem may be considered as a censored sampling problem in which bioassay positive responses are observations censored on the left (since they could have responded to the drug at  $t_i^-$  but were observed at  $t_i$  only), and non-responses (failures) are observations censored on the right. Thus if all positive responses were considered as the real observations, they can be taken care of by updating the parameter of the Dirichlet prior. They showed (see Sect. 3.2.8) that the application of Ramsey's formulas when all observations are failures and Dirichlet process updated for real observations, yield the modal estimate of  $F$  (expression (3.22)). They also noted that in the case of all failures, Ramsey's modal estimate has a simple closed form (Ramsey's estimator was not) and is essentially given by the Susarla-Van Ryzin estimator of the survival function  $S = 1 - F$ .

Antoniak (1974) also assumes the Dirichlet process prior and worked out an exact solution in the case of two dose levels and showed that the posterior distribution leads to a mixture of Dirichlet processes. For example if there is only one dose at  $t_1$ ,  $F(t_1)$  has a beta distribution,  $Be(\alpha(0, t_1], \alpha(t_1, \infty))$  and the posterior distribution would be  $Be(\alpha(0, t_1] + r_1, \alpha(t_1, \infty) + n_1 - r_1)$ , and therefore, the Bayes estimator under the integrated squared error loss will be the mean of this distribution,  $\hat{F}(t_1) = (\alpha(0, t_1] + r_1)/(\alpha(0, \infty) + n_1)$ . This is deceptively simple. For two dose levels at  $t_1 < t_2$  it starts to get complicated. Antoniak worked out the details and produced the following estimator.

$$\hat{F}(t_1) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} a_{ij} \frac{\beta_1 + r_1 + i}{M + n_1 + n_2} \quad (2.47)$$

$$\hat{F}(t_2) = \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} b_{ij} \frac{\beta_1 + \beta_2 + n_1 + r_2 - j}{M + n_1 + n_2} \quad (2.48)$$

and for other values of  $t$ ,  $\widehat{F}(t)$  is obtained by the linear interpolation. Here

$$a_{ij} = b_{ij} / \sum_{i=0}^{r_2} \sum_{j=0}^{n_1-r_1} b_{ij} \quad \text{and} \quad (2.49)$$

$$\begin{aligned} b_{ij} = & \binom{n_1-r_1}{j} \binom{r_2}{i} \\ & \times \Gamma(\beta_1 + r_1 + i) \Gamma(\beta_2 + n_1 - r_1 + r_2 - i - j) \Gamma(\beta_3 + n_2 - r_2 + j) \\ & / (\Gamma(\beta_1) \Gamma(\beta_2) \Gamma(\beta_3)), \end{aligned} \quad (2.50)$$

with  $\beta_1 = \alpha(0, t_1]$ ,  $\beta_2 = \alpha(0, t_2]$  and  $\beta_3 = \alpha(t_2, \infty)$ . For the general case, the expressions are complicated and involve multiple integrals.

Bhattacharya (1981) develops procedures to compute finite dimensional distributions of the posterior distribution of a Dirichlet prior. Taking a lead from Ferguson and Phadia (1979), Ammann (1984) writes  $F(t) = 1 - \exp(-H(t))$  and assumes  $H$  to be a process with independent increments with no deterministic component. He then derives the posterior distribution of  $H(t)$  in terms of Laplace transforms. However, the expressions are no simpler.

In view of these difficulties, Kuo (1988) proposed a linear Bayes estimate of  $F$  which is a Bayes rule in the space generated by  $r_1, \dots, r_L$  and 1. She derives the estimator by point-wise minimization of the loss function  $\int (F - \widehat{F})^2 dW$  at each dose level. At any point  $t$  which is not a dose level, the estimate is defined by the linear interpolation of estimates at the two adjacent dose levels. Her result is as follows.

Let  $\text{cov}(\mathbf{r})$  denote the covariance matrix of  $r_1, \dots, r_L$ , and let  $D(i, t_j)$  denote the covariance matrix with the  $i$ -th column replaced by the column  $(\text{cov}(r_1, F(t_j)), \dots, \text{cov}(r_L, F(t_j)))^T$ . Also, let  $M = \alpha[0, \infty)$ ,  $F_0(t) = \alpha(t)/M$  and  $C$  be a class of decision rules which are linear combinations of  $r_1, \dots, r_L$  and 1. Then with  $F \in \mathcal{D}(\alpha)$ , the Bayes rule in this class at each dose level  $t_j$ ,  $j = 1, 2, \dots, L$  is given by

$$\widehat{F}(t_j) = F_0(t_j) + \sum_{i=1}^L n_i \widehat{\lambda}_i(j) [r_i/n_i - F_0(t_i)], \quad (2.51)$$

and at  $t$ ,  $t_j < t < t_{j+1}$

$$\widehat{F}(t) = \frac{F_0(t_{j+1}) - F_0(t)}{F_0(t_{j+1}) - F_0(t_j)} \widehat{F}(t_j) + \frac{F_0(t) - F_0(t_j)}{F_0(t_{j+1}) - F_0(t_j)} \widehat{F}(t_{j+1}) \quad (2.52)$$

where  $\widehat{\lambda}_i(j) = |D(i, t_j)| / |\text{cov}(r)|$ .

Kuo also shows that  $\widehat{F}(t_j)$  is an asymptotically unbiased and consistent estimator of  $F(t_j)$ . As  $M \rightarrow 0$ ,  $\widehat{F}(t_j) \rightarrow r_j/n_j$  and as  $M \rightarrow \infty$ ,  $\widehat{F}(t_j) \rightarrow F_0(t_j)$ . She points out that at times the estimator may not be monotone and if monotonicity is essential, one can use the pool-adjacent-violators algorithm (pp. 13–18 in Barlow et al. 1972) for obtaining the desired result.

In the case of  $M$  and  $F_0$  unknown, empirical Bayes method of Sect. 2.2.4 may be used.

### 2.5.3 A Regression Problem

In the bioassay problem, the objective was to estimate the dose-response curve. Antoniak (1974) points out that a similar problem that arise in regression problems can also be handled in the same way. Let  $G$  be a distribution function on  $[0, 1]$  and assume that  $G \in \mathcal{D}(\alpha)$ . At chosen points  $0 = t_0 < t_1 < \dots < t_k \leq 1$ , assume that we have samples  $\mathbf{X}_l = (X_{l1}, \dots, X_{lm_l})$  from  $F(x|G(t_l))$ ,  $l = 1, 2, \dots, k$ , and based on these samples, our aim is to make inferences about the parameters  $G(t_l)$ . Since  $G$  has a Dirichlet process prior, the joint distribution of  $(G(t_1), G(t_2) - G(t_1), \dots, 1 - G(t_k))$  is a Dirichlet distribution with parameters  $(\alpha(t_1), \alpha(t_2) - \alpha(t_1), \dots, \alpha(1) - \alpha(t_k))$ . He points out that the observations for different values of  $l$  will not be generally independent and thus the calculations become complex. He illustrates them by taking an example with  $k = 2$ . Note that in bioassay problems, at each value of  $l$ , the observations available were from a binomial distribution, where as in the regression problem, they arise from some known distribution.

Consider the general linear model  $Z = \mathbf{X}\beta + \epsilon$ , where  $\mathbf{X}$  is a vector of covariates,  $\beta$  is a vector of regression coefficients, and  $\epsilon$  is the error term. Traditionally the practice is to assume the error term to be distributed as a parametric distribution, typically normal distribution with mean zero. The nonparametric Bayesian approach is to assume the error term having an unknown distribution, and a prior is placed on the unknown distribution (see Antoniak 1974 for example) centered around a base distribution which may be taken as normal with mean zero. There are several papers along this line using different priors. Since the base of a Polya tree prior includes absolutely continuous distributions, it is found to be favorable over the Dirichlet process.

Lavine (1994) considers the model  $Y_i = \varphi(X_i, \beta) + \epsilon_i$ , where  $\varphi$  is a known function,  $X_i$  is a known vector of covariates,  $\beta$  is an unknown vector of regression parameters with prior density  $f$  and the  $\epsilon_i$  are independent with unknown distribution  $P$ . Assuming  $P|\beta \sim PT(\Pi_\beta, \mathcal{A}_\beta)$ , he derives the posterior distribution of  $\beta$  and shows that the posterior distribution of  $P|\beta$  is  $PT(\Pi_\beta, \mathcal{A}_\beta | Y_1 - \varphi(X_1, \beta), \dots, Y_n - \varphi(X_n, \beta))$ .

Walker and Mallick (1997b) use a finite Polya tree prior for the error distribution in a hierarchical generalized linear model centered around a known base probability measure (by taking partitions to coincide with the percentiles of the corresponding distribution function) and find this approach to be more appropriate than a parametric approach. They extend this approach to an accelerated failure time model (Walker and Mallick 1999) where the error distribution is assumed to have a Polya tree prior and show how to implement MCMC procedure with application to survival data. Procedure to simulate a random probability measure  $P$  from  $\mathcal{PT}(\Pi, A)$  is also indicated in their paper. This is done by first generating a finite set of beta

random variables and defining the random measure  $P_M$  by  $P(B_{\epsilon_1 \dots \epsilon_M})$  for each  $\epsilon_1 \dots \epsilon_M$  according to (1.108). Then one of the  $2^M$  sets is picked according to the random weights  $P(B_{\epsilon_1 \dots \epsilon_M})$  and then a uniform random variate is taken from this set. If one of the set chosen happens to be an extreme set, then the random variate is chosen according to the base measure  $G_0$  restricted to this set.  $\alpha$ 's are chosen such that they increase rapidly down towards level  $M$ . See their paper for details.

Hanson and Johnson (2002) argue that in practice it may be difficult to specify a single centering/base distribution  $G_0$ . Therefore, they recommend modeling the error distribution in a linear model as a mixture of Polya trees. A mixture of Polya tree distribution  $G$  is specified by allowing parameters of the centering distribution  $G_0$  and/or the family of real numbers  $\alpha$ 's to be random. That is,  $G|U, C \sim PT(\Pi_u, A_c)$ ,  $U \sim f_u(u)$ ,  $C \sim f_c(c)$ . They consider mixtures of Polya trees in which the partition is constructed by a parametric family of probability distributions with variance  $U$ . The effect of taking mixtures is to smooth out the partitions of a simple Polya tree. Hanson (2006) further justify the efficiency of using mixtures of Polya trees alternative to using parametric models and provide computational strategies to carry out the analysis and illustrate them by discussing several examples.

Kalbfleisch (1978), Wild and Kalbfleisch (1981) and Hjort (1990) cast the regression problem in terms of the Cox model to accommodate covariates in survival data analysis. Kalbfleisch (1978) used a gamma process as prior for the unknown distribution function, Wild and Kalbfleisch (1981) extended the work of Ferguson and Phadia (1979) in which the neutral to the right process was used, and Hjort (1990) uses a beta process as prior for the cumulative hazard function. Their work is summarized in Sect. 3.7.

### 2.5.4 Estimation of a Density Function

The nonparametric Bayesian density function estimation may be viewed as an application of the mixtures of Dirichlet processes.

Let  $X_1, \dots, X_n$  be a sample of size  $n$  from a density function  $f(x)$  with respect to some finite measure on  $R$ . Based on  $\mathbf{X} = (X_1, \dots, X_n)$ , consider the problem of estimating  $f(x)$  at some fixed point  $x$ , or some functional of  $f(x)$ , such as the mean  $\int x f(x) dx$ . For the Bayesian treatment, we need to assign a prior on the space of all density functions and be able to handle the posterior distribution analytically. In order that the posterior distribution is manageable, it would be preferable to find a conjugate family of priors. This is known to be difficult. Lo (1984, 1986) approaches this problem by using a kernel representation of the density function, and assigning a Dirichlet prior to  $G$ . His results are presented here.

Let  $G$  be a distribution function on  $R$  and  $\alpha$  a finite measure on  $(R, \mathcal{B})$ . Let  $K(x, u)$  represent a kernel defined on  $(\mathcal{X} \times R)$  into  $R^+$  such that for each  $u \in R$ ,  $\int_{\mathcal{X}} K(x, u) dx = 1$  and for each  $x \in \mathcal{X}$ ,  $\int_R K(x, u) \alpha(du) < \infty$ . (Lo takes  $\mathcal{X}$  and  $R$  to be Borel subsets of Euclidean spaces.) The posterior distribution of  $G|\mathbf{X}$  has been obtained by Antoniak (1974) as indicated earlier. For each  $G \in \mathcal{F}$ , define

$f(x|G) = \int_R K(x, u)G(du)$ , then  $f(\cdot|G)$  is a kernel representation of the density function  $f$  and  $G$  is known as a mixing distribution. Lo defines a prior distribution for random  $f$  by letting  $G$  to be a random distribution with Dirichlet process prior  $\mathcal{D}(\alpha)$ . This way the broad support for the prior on the space of  $G$  is extended to the broad support for the prior on the space of all density functions. Since  $G \in \mathcal{D}(\alpha)$ , it can be seen that for each  $x \in \mathcal{X}$ , the marginal density of  $X$  is  $f_0(x) = \int_{\mathcal{F}} f(x|G)\mathcal{D}_{\alpha}(dG) = \int_R K(x, u)\alpha(du)/\alpha(R)$ . Now the posterior distribution of  $G$  given the data  $\mathbf{X}$  can be seen to be

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_B \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)\mathcal{D}_{\alpha}(dG)}{\int_{\mathcal{F}} \prod_{i=1}^n \int_R K(x_i, u_i)G(du_i)\mathcal{D}_{\alpha}(dG)}, \quad (2.53)$$

for all  $B \in \mathcal{F}$ . By repeated application of his lemma (interchanging the order of integration),

$$\int_{\mathcal{F}} \int_R h(u, G)G(du)\mathcal{D}_{\alpha}(dG) = \int_R \int_{\mathcal{F}} h(u, G)\mathcal{D}_{\alpha+\delta_u}(dG)\alpha(du)/\alpha(R), \quad (2.54)$$

he shows that

$$\mathcal{P}(G \in B|\mathbf{X}) = \frac{\int_{R^n} \mathcal{D}_{\alpha+\sum \delta_{u_i}}(B)\mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}, \quad (2.55)$$

where

$$\mu_{n,k,\alpha}(C) = \int_C \prod_{i=1}^n K(x_i, u_i) \prod_{i=1}^n \left( \alpha + \sum_{j=1}^{i-1} \delta_{u_j} \right) (du_i) \quad (2.56)$$

for  $C \in \mathcal{B}^n$ ,  $d\mathbf{u} = \prod_{i=1}^n du_i$  and  $\mathbf{u} \in R^n$ . For any measurable function  $g$ , this leads to

$$\mathcal{E}(g(G)|\mathbf{X}) = \frac{\int_{R^n} g(G)\mathcal{D}_{\alpha+\sum \delta_{u_i}}(dG)\mu_{n,k,\alpha}(d\mathbf{u})}{\int_{R^n} \mu_{n,k,\alpha}(d\mathbf{u})}. \quad (2.57)$$

Now, by taking  $g(G) = f(x|G)$  and simplifying, the posterior expectation  $\hat{f}(x|G)$  of  $f(x|G)$  is derived as

$$\hat{f}_{\alpha}(x|G) = \mathcal{E}(f(x|G)|\mathbf{X}) = p_n f_0(x) + (1 - p_n)\hat{f}_n(x), \quad (2.58)$$

which is a convex combination of prior guess  $f_0(x)$  defined above, and a quantity  $\hat{f}_n(x)$ , to be defined below, which mirrors the sample distribution function, but is complicated.

Let  $N(\underline{P})$  denote the number of cells in the partition  $\underline{P}$  of  $\{1, 2, \dots, m\}$ ;  $C_i$  the  $i$ -th cell of  $\underline{P}$  with  $m_i$  elements in it,  $i = 1, \dots, N(\underline{P})$ ;  $g_i(u)$ ,  $i = 1, \dots, m$  are  $m$  positive or  $\alpha$ -integrable functions;

$$\varphi(\underline{P}) = \prod_{i=1}^{N(\underline{P})} \left\{ (m_i - 1)! \int_R \prod_{l \in C_i} g_l(u) \alpha(du) \right\} \quad (2.59)$$

and finally,  $w(\underline{P}) = \varphi(\underline{P}) / \sum_{\underline{P}} \varphi(\underline{P})$ . Then  $\hat{f}_n(x)$  is given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{\underline{P}} w(\underline{P}) \sum_{i=1}^{N(\underline{P})} m_i \left\{ \frac{\int_R K(x, u) \prod_{l \in C_i} K(x_l, u) \alpha(du)}{\int_R \prod_{l \in C_i} K(x_l, u) \alpha(du)} \right\}, \quad (2.60)$$

where the summation is taken over all partitions  $\underline{P}$  of  $\{1, 2, \dots, m\}$ .  $\hat{f}$  serves as a Bayes estimate under the loss function  $L(f, \hat{f}) = \int |f(x|G) - \hat{f}(x|G)|^2 W(dx)$ , where  $W$  is a weight function.

Lo discusses the choice of the kernel  $K$  and the parameter  $\alpha$  of the prior, and gives several examples of  $K(x, u)$  and  $\alpha$  and computes the Bayes estimators. His examples of kernels include histogram, normal with location and/or scale parameters, symmetric and unimodal densities, decreasing densities, etc. For example, if  $K$  is chosen to reflect the histogram model, the estimator reduces to the usual Bayes estimates of cell probabilities. Kuo's (1986a, 1986b) Monte Carlo method may be adapted to carry out the calculations. Details may be found in his paper. Lavine (1992) uses mixtures of Polya trees in density estimation.

Ghorai and Susarla (1982) considered an empirical Bayes approach to the above problem. Assuming  $\alpha(R)$  to be known, they obtained an estimator of  $f_0(x) = \int_R K(x, u) \alpha(du) / \alpha(R)$  based on previous  $n$  copies and substituted in the Bayesian estimator  $\hat{f}(x|G)$  at the  $(n+1)$ -th stage. Under certain conditions, they prove the asymptotic optimality of the resulting estimator.

Ferguson (1983) considered a different formulation of the density function. He modeled it as a countable mixtures of normal densities:  $f(x) = \sum_{i=1}^{\infty} p_i h(x|\mu_i, \sigma_i)$  where  $h(x|\mu, \sigma)$  is the normal density with mean  $\mu$  and variance  $\sigma^2$ . This formulation has countably infinite number of parameters,  $(p_1, p_2, \dots, \mu_1, \mu_2, \dots, \sigma_1, \sigma_2, \dots)$ . Since the interest is in estimating  $f(x)$  at a point  $x$ , and not in estimating the parameters themselves, it can be written as  $f(x) = \int h(x|\mu, \sigma) dG(\mu, \sigma)$ , where  $G$  is the probability measure on the half plane  $\{(\mu, \sigma) : \sigma > 0\}$  that gives weight  $p_i$  to the point  $(\mu_i, \sigma_i)$ ,  $i = 1, 2, \dots$ . While Lo assumes a Dirichlet process prior for the unknown  $G$ , Ferguson defines a prior via the Sethuraman representation of  $G$ . He defines the prior distribution for the parameter vector  $(p_1, p_2, \dots, \mu_1, \mu_2, \dots, \sigma_1, \sigma_2, \dots)$  as follows: vectors  $(p_1, p_2, \dots)$  and  $(\mu_1, \mu_2, \dots, \sigma_1, \sigma_2, \dots)$  are independent;  $p_1, p_2, \dots$  are the weights with parameter  $M$  in Sethuraman representation; and  $\xi_i = (\mu_i, \sigma_i)$  are iid with common gamma-normal conjugate prior for the two-parameter normal distribution. This shows that  $G$  is a Dirichlet process with parameter  $\alpha = MG_0$ , where  $G_0 = \mathcal{E}(G)$  is the conjugate prior for  $(\mu, \sigma^2)$ , and its infinite sum representation is  $G = \sum_{i=1}^{\infty} p_i \delta_{\xi_i}$  where as usual  $(p_1, p_2, \dots)$  and  $(\xi_1, \xi_2, \dots)$  are independent and  $\xi_i \stackrel{iid}{\sim} G_0$ . Now given a sample  $x_1, \dots, x_n$  of size  $n$  from a distribution with density  $f(x) = \int h(x|\xi) dG(\xi)$ , the posterior distribution of  $G$  given  $x_1, \dots, x_n$  has been obtained by Antoniak (1974) as mixture of Dirichlet processes

$$G|x_1, \dots, x_n \sim \int \dots \int \mathcal{D}(\alpha + nG_n) dH(\xi_1, \dots, \xi_n | x_1, \dots, x_n)$$

with  $nG_n = \sum_{i=1}^n \delta_{\xi_i}$ .  $H(\xi_1, \dots, \xi_n | x_1, \dots, x_n)$  is the posterior distribution of  $\xi_1, \dots, \xi_n$  given  $x_1, \dots, x_n$ . Since  $\mathcal{E}(\mathcal{D}(\alpha + nG_n)) = (MG_0 + nG_n)/(M + n)$ ,

$$\begin{aligned} \mathcal{E}(G(\xi) | x_1, \dots, x_n) \\ = p_n G_0(\xi) + (1 - p_n) \int \dots \int G_n(\xi) dH(\xi_1, \dots, \xi_n | x_1, \dots, x_n) \end{aligned} \quad (2.61)$$

and

$$\widehat{f}(x) = \mathcal{E}(f(x) | x_1, \dots, x_n) = p_n f_0(x) + (1 - p_n) \widehat{f}_n(x), \quad (2.62)$$

where:  $p_n = M/(M + n)$  as before,  $f_0(x) = \mathcal{E}(f(x)) = \sum_{i=1}^{\infty} \mathcal{E}(p_i) \mathcal{E}h(x | (\mu_i, \sigma_i)) = \mathcal{E}h(x | \mu, \sigma)$  and  $\widehat{f}_n(x)$  is given by

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \int \dots \int h(x | \xi_i) dH(\xi_1, \dots, \xi_n | x_1, \dots, x_n). \quad (2.63)$$

Following Lo,  $\widehat{f}_n(x)$  can be written as a ratio  $h(x, x_1, \dots, x_n)/h(x_1, \dots, x_n)$ , where

$$h(x_1, \dots, x_n) = \frac{1}{M^{(n)}} \int \dots \int \left( \prod_{i=1}^n h(x_i | \xi_i) \right) \prod_{n=1}^n d \left( MG_0 + \sum_{j=1}^{i-1} \delta_{\xi_j} \right) (\xi_i), \quad (2.64)$$

and computations are carried out by Kuo's (1986a, 1986b) Monte Carlo method.

Normal mixtures also turn up in Escobar (1994) and Escobar and West (1995). Escobar's set up is as follows. Let  $Y_i | \mu_i \sim N(\mu_i, 1)$ ,  $\mu_i | G \stackrel{iid}{\sim} G$ ,  $\mu_i$  and  $G$  are unknown. In contrast to Ferguson's and Lo's objectives, his objective is to estimate  $\mu_i$ 's (with the variance being known to be 1) based on observed  $Y_i$ 's using the nonparametric Bayesian approach. When  $G$  is known the Bayesian estimator is the posterior mean

$$\mathcal{E}(\mu_i | Y_i) = \frac{\int \mu_i \phi(Y_i - \mu_i) dG(\mu_i)}{\int \phi(Y_i - \mu_i) dG(\mu_i)}, \quad (2.65)$$

where  $\phi$  is the density of the standard normal distribution function. When  $G$  is unknown, empirical Bayes methods are typically used. Instead Escobar uses a Dirichlet process prior for  $G$ . Antoniak has shown that if Dirichlet process prior is used for  $G$ , then the posterior distribution of  $\mu_i$  is a mixture of Dirichlet process. Thus it was computationally difficult. Kuo (1986b) and Lo (1984) developed Monte Carlo integration algorithms, but Escobar points out that they are inefficient since they do not sample values conditionally based on the data. He introduces a new Gibbs sampler like method that remedied this problem.

Escobar and West (1995) describe a normal mixture model, similar to Ferguson's (1983), in terms of the predictive distribution of a future observation. For their model, given  $(\mu_i, \sigma_i^2)$ , we have a random sample, say  $Y_1, \dots, Y_n$ , such that  $Y_i | (\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, \dots, n$  and the objective is to find the predictive distribution of next observation  $Y_{n+1}$  which is a mixture of normals,

$Y_{n+1}|Y_1, \dots, Y_n \sim N(\mu_{n+1}, \sigma_{n+1}^2)$ . A usual practice is to put a parametric prior on vector  $\mathbf{v} = (\mu_1, \dots, \mu_n, \sigma_1^2, \dots, \sigma_n^2)$ . Ferguson models the common prior for  $v_i = (\mu_i, \sigma_i^2)$  as a Dirichlet process prior. Thus the data is considered as coming from a Dirichlet mixture of normals in contrast to Antoniak where the Dirichlet process processes were mixed with respect to a parametric distribution  $H(\theta)$ ,  $\alpha_\theta \sim H(\theta)$ . A particular case of  $(\mu_i, \sigma_i^2) = (\mu_i, \sigma^2)$  has been studied (see West 1990, 1992) in which the distribution of  $\mu_i$ 's is modeled as Dirichlet process with a normal base measure.

In view of the discreteness of Dirichlet process prior which induces multiplicities of observations,  $v_{n+1}|v_1, \dots, v_n$  will have distribution of the form given in property 19 of Sect. 1.2. Then they proceed on the line of Ferguson, derive the conditional distribution of  $Y_{n+1}|v_1, \dots, v_n$  which is a mixture of a Student's t-distribution and  $n$  normals  $N(\mu_i, \sigma_i^2)$ , and then it is shown that the unconditional predictive distribution is given by  $Y_{n+1}|Y_1, \dots, Y_n \sim \int \mathcal{P}(Y_{n+1}|\mathbf{v})d\mathcal{P}(\mathbf{v}|Y_1, \dots, Y_n)$ . Since the evaluation of  $\mathcal{P}(\mathbf{v}|Y_1, \dots, Y_n)$  is difficult even in small samples, they use Monte Carlo approximation using extensions of the iterative technique developed by Escobar (1994).

### 2.5.5 Estimation of the Rank of $X_1$ Among $X_1, \dots, X_n$

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} F$ . The problem of estimating the rank order  $G$  of  $X_1$  among  $X_1, \dots, X_n$  based on the knowledge of  $r$  ( $< n$ ) observed values of  $X_1, \dots, X_r$  was considered from a Bayesian point of view by Campbell and Hollander (1978). WLOG assume that  $x_1, \dots, x_r$  are the first  $r$  values in the sample. Let  $K$ ,  $L$  and  $M$  denote the number of observations among  $X_1, \dots, X_n$  that are less than, equal to and greater than  $X_1$ , respectively. Then the rank order  $G$  of  $X_1$  is taken as the average value of the ranks that would be assigned to the  $L$  values tied at  $X_1$ , in the ascending order, i.e.  $G = \frac{1}{L} \sum_{i=1}^L (K + i) = K + (L + 1)/2$ .

Let  $K'$ ,  $L'$  and  $M'$  be defined respectively, as the corresponding numbers of observations among  $x_1, \dots, x_r$ . Then the rank order  $G'$  of  $x_1$  among  $x_1, \dots, x_r$  is given by  $G' = K' + (L' + 1)/2$ . Given  $x_1, \dots, x_r$ , the problem is to estimate  $G$  which is clearly a function of  $K$ ,  $L$  and  $M$ . Assuming  $F \in D(\alpha)$ , Campbell and Hollander obtained the posterior mean,

$$\hat{G} = \mathcal{E}(G|x_1, \dots, x_r) = G' + (n - r) \left\{ \alpha'((-\infty, x_1)) + \frac{1}{2} \alpha'(\{x_1\}) / \alpha'(R) \right\}, \quad (2.66)$$

where  $\alpha' = \alpha + \sum_{i=1}^r \delta_{x_i}$ .  $\hat{G}$  depends on  $x_1, \dots, x_r$  only through  $G'$  and  $x_1$ . In comparison, the non-Bayesian estimators are given by  $G_F = G' + (n - r)F(x_1)$  in the case of a known continuous function  $F$  and  $G_U = G' + (n - 1)G'/(n - r)$ , when  $F$  is unknown.

## 2.6 Bivariate Distribution Function

Ferguson's (1973) definition of the Dirichlet process on an arbitrary space of probability measures makes it amenable for its extension to higher dimensions in a straight forward manner. In presenting the applications of Dirichlet process in bivariate situation, we will be concerned with the distribution and survival functions defined on  $R^2 = R \times R$  and a finite non-null measure  $\alpha$  on  $(R^2, \mathcal{B}^2)$  where  $\mathcal{B}^2$  represents the  $\sigma$ -field of Borel subsets of  $R^2$ .

Let  $P$  be a random probability measure on  $(R^2, \mathcal{B}^2)$  and  $F(x, y)$  be the corresponding bivariate distribution function. Assume that we have a random sample  $(\mathbf{X}, \mathbf{Y}) = (X_1, Y_1), \dots, (X_n, Y_n)$  from  $F(x, y)$ . Then the Bayesian estimators are presented first for the distribution function  $F$  and then for its functionals.

### 2.6.1 Estimation of $F$ w.r.t. the Dirichlet Process Prior

For the Bayesian estimation of  $F(x, y)$ , we assume that  $F$  has a Dirichlet process prior with parameter  $\alpha$ . As in the univariate case, we take the weighted loss function  $L(F, \hat{F}) = \int_{R^2} (F - \hat{F})^2 dW$ , where  $W$  now is a nonnegative weight function on  $R^2$ . The Bayesian estimator of  $F(x, y)$  with respect to the Dirichlet process prior and the loss function  $L$ , is a direct extension of Ferguson's Bayesian estimator in one-dimension, and is given by

$$\begin{aligned} \hat{F}_\alpha(x, y) &= \frac{\alpha((-\infty, x] \times (-\infty, y]) + \sum_{i=1}^n \delta_{(X_i, Y_i)}((-\infty, x] \times (-\infty, y])}{\alpha(R^2) + n} \\ &= p_n \frac{\alpha((-\infty, x] \times (-\infty, y])}{\alpha(R^2)} \\ &\quad + (1 - p_n) \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}((-\infty, x] \times (-\infty, y]). \end{aligned} \quad (2.67)$$

Empirical Bayes estimation of  $F(x, y)$  when  $\alpha(\cdot)$  is unknown but  $\alpha(R^2)$  is known can be carried out as in the univariate case. Also, following Zehnwirth's (1981) lead, an estimator for unknown  $\alpha(R^2)$  was developed in Dalal and Phadia (1983) and was used when  $\alpha(R^2)$  is assumed to be unknown.

### 2.6.2 Estimation of $F$ w.r.t. a Tailfree Process Prior

In Chap. 1, the tailfree processes were introduced and their properties as well as the bivariate extension (Phadia 2007) were discussed. Here the Bayes estimator of  $F$  with respect to the bivariate tailfree process prior is derived under the weighted loss function. If  $x$  and  $y$  are binary rationals, then the estimate can be written as a finite

sum; if either  $x$  or  $y$  is not a binary rational, then the estimate involves an infinite sum.

In view of the conjugacy property of tailfree processes, it is sufficient to derive the estimate for the no-sample problem. Then for a sample of size  $n$ , all we have to do is to update the parameters. Consider for example,  $(x, y) = (\frac{1}{2}, \frac{3}{4})$ . Following the notation of Sect. 1.16,

$$\begin{aligned}\widehat{F}\left(\frac{1}{2}, \frac{3}{4}\right) &= \mathcal{E}\left[F\left(\frac{1}{2}, \frac{3}{4}\right)\right] \\ &= \mathcal{E}[P(B_{11}) + P(B_{21}) + P(B_{23})] \\ &= \mathcal{E}[Z_1 + Z_2 Z_{21} + Z_2 Z_{23}] \\ &= \mathcal{E}[Z_1 + Z_{21} + Z_{23}] \\ &= \frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2} \frac{\alpha_{21}}{\gamma_{21}} + \frac{\alpha_2}{\gamma_2} \frac{\alpha_{23}}{\gamma_{23}}.\end{aligned}$$

On the other hand if  $(x, y) = (\frac{1}{3}, \frac{1}{2})$ , say, then

$$\begin{aligned}\widehat{F}\left(\frac{1}{3}, \frac{1}{2}\right) &= \mathcal{E}\left[F\left(\frac{1}{3}, \frac{1}{2}\right)\right] \\ &= \mathcal{E}\left[P(B_{11} \cup B_{12}) + P\left(\bigcup_{i=1}^2 \bigcup_{j=1}^2 (B_{13ij} \cup B_{14ij}) + \dots\right)\right] \\ &= \mathcal{E}\left[Z_{11} + Z_{12} + \sum_{i=1}^2 \sum_{j=1}^2 (Z_{13ij} + Z_{14ij}) + \dots\right] \\ &= \frac{\alpha_{11}}{\gamma_{11}} + \frac{\alpha_{12}}{\gamma_{12}} + \sum_{i=1}^2 \sum_{j=1}^2 \left(\frac{\alpha_{13ij}}{\gamma_{13ij}} + \frac{\alpha_{14ij}}{\gamma_{14ij}}\right) + \dots\end{aligned}$$

Now given a sample  $\mathbf{X}$ , all we have to do is to update the  $\alpha$ 's.

### 2.6.3 Estimation of a Covariance

The covariance of  $P$  is defined for  $(x, y) \in R^2$  by the formula

$$\text{Cov } P = \int xy dP - \int x dP \int y dP. \quad (2.68)$$

Assuming the squared error loss  $L_2$  and  $P \in \mathcal{D}(\alpha)$ , Ferguson (1973) derived its Bayesian estimator. For the no-sample problem we have,

$$\mathcal{E}(\text{Cov } P) = \frac{\alpha(R^2)}{\alpha(R^2) + 1} \sigma_{12}, \quad (2.69)$$

where  $\sigma_{12}$  is the covariance of  $\mathcal{E}(P)$  given by  $\sigma_{12} = [\int xy d\alpha(x, y) - \mu_1 \mu_2] / \alpha(R^2)$ ,  $\mu_1 = \int x d\alpha(x, y) / \alpha(R^2)$ , and  $\mu_2 = \int y d\alpha(x, y) / \alpha(R^2)$ . Now for the sample of size  $n$ , we update  $\alpha$  and obtain the Bayes estimate as

$$\widehat{\text{Cov } P_\alpha} = \frac{\alpha(R^2) + n}{\alpha(R^2) + n + 1} \times (p_n \sigma_{12} + (1 - p_n) s_{12} + p_n (1 - p_n) (\mu_1 - \bar{X}_n)(\mu_2 - \bar{Y}_n)), \quad (2.70)$$

where  $s_{12} = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$  is the sample covariance. This is again a mixture of three relevant quantities.

### 2.6.4 Estimation of the Concordance Coefficient

The problem of estimation of concordance coefficient in a bivariate distribution was treated in Dalal and Phadia (1983). Let  $(X, Y)$  and  $(X', Y')$  be two independent observations from a joint distribution function  $F(x, y)$ . A quantity of interest is  $\Delta = P\{(X - X')(Y - Y') > 0\}$ , which is related to Kendall's  $\tau = \mathcal{E}(\text{sign})(X - X')(Y - Y')$ , by the equation  $\tau = 2\Delta - 1$ . It is used as a measure of the dependence between  $X$  and  $Y$  as well as a measure of the degree of concordance among observations from  $F(x, y)$ . Let

$$\begin{aligned} T_1 &= \{(x, y, x', y') : (x - x')(y - y') > 0\} \quad \text{and} \\ T_2 &= \{(x, y, x', y') : (x - x')(y - y') = 0\}. \end{aligned} \quad (2.71)$$

Since  $F$  is allowed to be discrete, a slight modification of  $\Delta$ , namely,

$$\Delta = P_F\{(X - X')(Y - Y') > 0\} + \frac{1}{2} \cdot P_F\{(X - X')(Y - Y') = 0\} \quad (2.72)$$

is preferred. The rationale is that the tied pairs are evenly distributed among concordants  $(X - X')(Y - Y') > 0$  and discordant  $(X - X')(Y - Y') < 0$ . When  $X$  and  $Y$  are independent,  $\Delta = 0$ , and its estimator serves as a statistic to test the hypothesis of independence of  $X$  and  $Y$ . Now,

$$\Delta = \Delta_F = \int \left( I_{T_1} + \frac{1}{2} \cdot I_{T_2} \right) d(F(x, y)F(x', y')). \quad (2.73)$$

Assuming  $F \in \mathcal{D}(\alpha)$ , and  $\alpha$  defined on  $(R^2, \mathcal{B}^2)$ , the Bayes estimator of  $\Delta$  for the no sample problem is given by

$$\widehat{\Delta}_{\alpha 0} = \mathcal{E}_{\mathcal{D}(\alpha)}(\Delta_F) = \int \left( I_{T_1} + \frac{1}{2} \cdot I_{T_2} \right) d\mathcal{E}_{\mathcal{D}(\alpha)}(F(x, y)F(x', y')). \quad (2.74)$$

Let  $\alpha = MQ$  and let  $G$  be a CDF corresponding to the measure  $Q$ . Applying Theorem 4 of Ferguson (1973) in evaluating  $\mathcal{E}_{\mathcal{D}(\alpha)}(F(x, y)F(x', y'))$  and simplifying we get,

$$\widehat{\Delta}_{\alpha 0} = \frac{M}{M+1} \Delta_G + \frac{1}{2(M+1)}, \quad (2.75)$$

where  $\Delta_G = P_G[(X - X')(Y - Y') > 0] + \frac{1}{2} P_G[(X - X')(Y - Y') = 0]$ .

When  $X$  and  $Y$  are independent,  $\Delta_G = \frac{1}{2}$ , and therefore,  $\widehat{\Delta}_{\alpha 0} = \frac{1}{2}$  also.

Now for the case of  $n$  observations,  $(X_1, Y_1), \dots, (X_n, Y_n) \sim F(x, y)$ , the posterior distribution of  $F$  given the data is again a Dirichlet process with the parameter  $\alpha$  updated as  $\alpha + \sum_{i=1}^n \delta_{(X_i, Y_i)}$ , which can be rewritten as

$$\begin{aligned} \alpha + \sum_{i=1}^n \delta_{(X_i, Y_i)} &= (M+n) \left[ \frac{M}{M+n} Q + \frac{1}{M+n} \sum_{i=1}^n \delta_{(X_i, Y_i)} \right] \\ &= (M+n) Q^*, \quad \text{say.} \end{aligned} \quad (2.76)$$

If  $G^*$  is a CDF corresponding to  $Q^*$ , then  $G^* = p_n G + (1 - p_n) \widehat{G}_n$ , where  $\widehat{G}_n$  is the empirical CDF based on the  $n$  observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , and  $p_n = M/(M+n)$ . Hence the Bayes estimator is given by,

$$\begin{aligned} \widehat{\Delta}_{\alpha n} &= \frac{M+n}{M+n+1} \int \left( I_{T_1} + \frac{1}{2} \cdot I_{T_2} \right) d(p_n G + (1 - p_n) \widehat{G}_n)(p_n G + (1 - p_n) \widehat{G}_n) \\ &\quad + \frac{1}{2(M+n+1)} \\ &= \frac{M+n}{M+n+1} [p_n^2 \Delta_G + 2p_n(1 - p_n) \Delta(G, \widehat{G}_n) + (1 - p_n)^2 \Delta_{\widehat{G}_n}] \\ &\quad + \frac{1}{2(M+n+1)}, \end{aligned} \quad (2.77)$$

where

$$\Delta_{\widehat{G}_n} = \frac{1}{n^2} \sum_{i,j=1}^n \left( I_{[(x-x')(y-y')>0]} + \frac{1}{2} I_{[(x-x')(y-y')=0]} \right), \quad (2.78)$$

and  $\Delta(G, \widehat{G}_n) = \frac{1}{n} \sum_{i=1}^n \Delta_G(x_i, y_i)$  with

$$\Delta_G(x_i, y_i) = \left\{ P_G[(X - x_i)(Y - y_i) > 0] + \frac{1}{2} P_G[(X - x_i)(Y - y_i) = 0] \right\}. \quad (2.79)$$

Here  $\Delta_G$  and  $\Delta_{\widehat{G}_n}$  can be interpreted as the natural estimates of the coefficient of concordance for the idealized model, and the sample and a single observation, respectively; whereas  $\Delta_G(x_i, y_i)$  is the theoretical concordance probability of the pair  $(x_i, y_i)$ .

The authors evaluated explicitly the Bayesian estimator for two interesting models, namely the bivariate normal and Gumbel's bivariate exponential distribution.

They extended the above result to the empirical Bayes estimate of  $\Delta$  with  $M$  known, and used Zehnwirth's (1981) technique to estimate  $M$ , when  $M$  is unknown. In both cases, they showed that the estimates are asymptotically optimal with rate of convergence  $O(n^{-1})$ . The details can be found in their paper.

## 2.7 Estimation of a Function of $P$

The examples of Sects. 2.4 and 2.6 can be generalized to any measurable function  $\phi(P)$  of  $P$ . Let  $\mathfrak{X}^k$  denote the product space. A real valued function  $\phi : \Pi \rightarrow R$  is said to be *estimable* with kernel  $h$  if there exists a statistics  $h(X_1, \dots, X_k)$  such that  $\phi(P) = \int_{\mathfrak{X}^k} h(x_1, \dots, x_k) \prod_{i=1}^k dP(x_i)$ . The degree of an estimable parameter  $\phi(P)$  is the least sample size for which there is such an  $h$  (p. 151 in Zacks 1971). The Bayes and empirical Bayes estimation of estimable parameters of degree 1 and 2 under loss function  $L_2$  and with respect to the Dirichlet and Dirichlet Invariant processes as priors were investigated by Yamato (1977a, 1977b), Tiwari (1981) and Tiwari and Zalkikar (1985). Their results are as follows.

**Dirichlet Process Prior** Based on a random sample  $\mathbf{X}$  from  $P$ , the Bayesian estimator  $\hat{\phi}$  of  $\phi$  under  $L_2$  loss is given by the posterior mean  $\mathcal{E}(\phi(P) \mid X_1, \dots, X_n)$ . In particular, suppose  $\phi(P) = \phi_h(P)$  and  $P \in \mathcal{D}(\alpha)$ , where

$$\phi_h(P) = \int_{\mathfrak{X}^k} h(x_1, \dots, x_k) dP(x_1) \dots dP(x_k), \quad (2.80)$$

and  $h$  is a symmetric measurable function from  $\chi^k$  into  $R$  satisfying

$$\int_{\mathfrak{X}^k} |h(x_1, \dots, x_k)| d\bar{\alpha}(x_1) \dots d\bar{\alpha}(x_m) < \infty, \quad (2.81)$$

where as before,  $\bar{\alpha}(\cdot) = \alpha(\cdot)/\alpha(R)$ . Under a further assumption concerning the second moment of  $h$  with respect to  $\bar{\alpha}^m$ ,  $m \leq k$ , namely

$$\int_{\mathfrak{X}^m} |h(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m)|^2 d\bar{\alpha}(x_1) \dots d\bar{\alpha}(x_m) < \infty, \quad (2.82)$$

for all possible combinations of arguments  $(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_m, \dots, x_m)$ ,  $m \leq k$ , from all distinct ( $m = k$ ) to all identical ( $m = 1$ ), the Bayes estimator of  $\phi_h(P)$  with respect to  $\mathcal{D}(\alpha)$  for the no sample problem is  $\hat{\phi}_{h,\alpha}^0 = \mathcal{E}_{\mathcal{D}(\alpha)}(\phi_h(P))$ , and for the sample  $X_1, \dots, X_n$  it is

$$\hat{\phi}_{h,\alpha}^n = \mathcal{E}_{\mathcal{D}(\alpha + \sum_{i=1}^n \delta_{X_i})}(\phi_h(P)) = \hat{\phi}_{h,\alpha + \sum_{i=1}^n \delta_{X_i}}^0. \quad (2.83)$$

Thus using this expression and Property 9 of Sect. 1.2, Yamato (1977a, 1977b) and Tiwari (1981) derived the following result. Based on a sample  $X_1, \dots, X_n$ , the

Bayes estimator of  $\phi_h(P)$  with respect to the prior  $\mathcal{D}(\alpha)$  and loss  $L_2$  is given by

$$\begin{aligned} \hat{\phi}_{h,\alpha}^n &= \sum_{C(\sum i m_i = k)} \frac{k! [\alpha(\mathfrak{X}) + n]^{\sum m_i}}{\prod_{i=1}^k [i^{m_i} (m_i)!] [\alpha(\mathfrak{X}) + n]^{(k)}} \\ &\times \int_{\mathfrak{X}^{\sum m_i}} h(x_{11}, \dots, x_{1m_1}, x_{21}, \dots, x_{2m_2}, \dots, x_{k1}, \dots, x_{km_k}) \\ &\times \prod_{i=1}^k \prod_{j=1}^{m_i} d\hat{F}_\alpha(x_{ij}), \end{aligned} \quad (2.84)$$

where  $\hat{F}_\alpha(\cdot) = p_n \bar{\alpha}(\cdot) + (1 - p_n) \hat{F}_n(\cdot)$  is the Bayes estimator of  $F$  corresponding to  $P$ . Sethuraman and Tiwari (1982) showed that  $\hat{\phi}_{h,\alpha}^n \rightarrow \hat{\phi}_{h,\sum_{i=1}^n \delta_{x_i}}$  as  $\alpha(\mathfrak{X}) \rightarrow 0$ .

Also, if  $h(x_1, \dots, x_k)$  is such that it vanishes whenever two coordinates are equal, then

$$\hat{\phi}_{h,\sum_{i=1}^n \delta_{x_i}} = \frac{n(n-1) \dots (n-k+1)}{n^{(k)}} U_{h,n}, \quad (2.85)$$

where  $U_{h,n}$  is the usual  $U$ -statistic based on the sample  $X_1, \dots, X_n$ . Yamato (1977b) has proved that the asymptotic distribution of  $\hat{\phi}_{h,\sum_{i=1}^n \delta_{x_i}}^0$  is the same as that of  $U_{h,n}$ .

Using the above result and based on a sample  $\mathbf{X}$ , the Bayes estimators with respect to  $\mathcal{D}(\alpha)$  of estimable functions of degree 1 and 2, namely  $\phi_1(P) = \int h(x) dP(x)$  and  $\phi_2(P) = \int \int h(x, y) dP(x) dP(y)$  are obtained in Tiwari and Zalkikar (1985) as

$$\hat{\phi}_1(P) = p_n \int h(x) d\bar{\alpha}(x) + \frac{(1 - p_n)}{n} \sum_{i=1}^n h(X_i) \quad (2.86)$$

and

$$\begin{aligned} \hat{\phi}_2(P) &= \frac{M + n}{M + n + 1} \left\{ p_n^2 \int \int h(x, y) d\bar{\alpha}(x) d\bar{\alpha}(y) \right. \\ &\quad + \frac{2p_n(1 - p_n)}{n} \sum_{i=1}^n \int h(x, x_i) d\bar{\alpha}(x) \\ &\quad \left. + \frac{(1 - p_n)^2}{n^2} \sum_{i \neq j} h(X_i, X_j) \right\}. \end{aligned} \quad (2.87)$$

From these two expressions, the Bayes estimators of parameters such as the mean, variance, covariance and the probability that  $X$  is stochastically smaller than  $Y$  can be derived. Explicit expressions were given earlier.

Tiwari and Zalkikar also extended Dalal and Phadia's (1983) result for the Bayes and empirical Bayes estimators of the concordance coefficient to a general param-

ter of degree 2, namely,

$$\varsigma = \int h(x, y; x', y') dP(x, y) dP(x', y'), \quad (2.88)$$

where  $h(x, y; x', y')$  is a real valued function defined on  $(R^4, \mathcal{B}^4)$ , where  $\mathcal{B}^4$  stands for the corresponding Borel sets of  $R^4$ . The Bayes estimator of  $\varsigma$  with respect to the Dirichlet process prior defined on  $(R^2, \mathcal{B}^2)$  is given by

$$\begin{aligned} \widehat{\varsigma}_\alpha &= \frac{M+m}{M+m+1} [p_m^2 \varsigma_{\bar{\alpha}} + 2p_m(1-p_m)\varsigma(\bar{\alpha}, F_m) + (1-p_m)^2 \varsigma_{F_m}] \\ &+ \frac{1}{M+m+1} \left\{ p_m \int h(x, y; x, y) d\bar{\alpha}(x, y) \right. \\ &\left. + \frac{(1-p_m)}{m} \sum_{i=1}^m h(X_i, Y_i; X_i, Y_i) \right\}, \end{aligned} \quad (2.89)$$

where  $\varsigma_{\bar{\alpha}} = \int h(x, y; x', y') d\bar{\alpha}(x, y) d\bar{\alpha}(x', y')$ ,  $\varsigma_{F_m} = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m h(X_i, Y_i; X_j, Y_j)$ ,  $\varsigma(\bar{\alpha}, F_m) = \frac{1}{m} \sum_{i=1}^m \varsigma_{\bar{\alpha}}(X_i, Y_i)$ , and  $\varsigma_{\bar{\alpha}}(x, y) = \int h(x, y; x', y') d\bar{\alpha}(x', y')$ .

Note that by taking  $h(x, y; x', y') = I_{T_1} + \frac{1}{2}I_{T_2}$  where

$$T_1 = \{(x, y; x', y') : (x - x')(y - y') > 0\} \quad (2.90)$$

and

$$T_2 = \{(x, y; x', y') : (x - x')(y - y') = 0\}, \quad (2.91)$$

the results of Dalal and Phadia (1983) can be obtained.

**Dirichlet Invariant Process Prior** Yamato (1986, 1987) carried out similar estimation procedures using the Dirichlet Invariant process with parameter  $\alpha$  and under the same loss  $L_2$ . Let  $\alpha = M Q$  and  $M = \alpha(\mathfrak{X})$  and assume the same finite group  $\mathcal{G} = \{g_1, \dots, g_k\}$  of transformations as used in Dalal (1979a).

In particular, if we take  $\mathcal{G} = \{e, g\}$  with  $e(x) = x$ ,  $g(x) = 2\eta - x$ , for  $x \in R$  and  $\eta$  a constant, and  $h(x) = I[x \leq t]$ , then  $F^* = P((-\infty, t])$  and its Bayes estimate yields

$$\widehat{F}_\alpha^*(t) = p_n F_0(t) + (1 - p_n) \widehat{F}_n^*(t), \quad (2.92)$$

where,  $F_0(t) = Q((-\infty, t])$  and  $\widehat{F}_n^*(t)$  is  $\eta$ -symmetrized version of the empirical distribution,

$$\widehat{F}_n^*(t) = \frac{1}{2n} \sum_{i=1}^n \delta_{X_i}((-\infty, t]) + \delta_{2\eta - X_i}((-\infty, t]). \quad (2.93)$$

$\widehat{F}_\alpha^*$  is identical to the one obtained by Dalal (1979a).

In the second paper, Yamato (1987) using the alternative definition of the Dirichlet Invariant process generalizes the above treatment to an arbitrary degree  $s$  of estimable parameters in one sample case. As an example of this result, the Bayes estimate of  $\phi_1$  under  $L_2$  loss is obtained as

$$\widehat{\phi}_{1\alpha}^* = p_n \int h(x) dQ(x) + \frac{(1-p_n)}{nk} \sum_{i=1}^n \sum_{j=1}^k h(g_j X_i), \quad (2.94)$$

wherein  $\frac{1}{nk} \sum_{i=1}^n \sum_{j=1}^k h(g_j X_i)$  is the  $\mathcal{G}$ -invariant U-statistic based on kernel  $h$ .

Similarly, the Bayesian estimator for an estimable parameter of degree 2,  $\varphi_2$  is obtained. Assume that

$$\begin{aligned} \int_{\mathfrak{X}^2} h(x, y) dQ(x) dQ(y) &< \infty, \\ \int_{\mathfrak{X}} h(x, gx) dQ(x) &< \infty \quad \text{for any } g \in \mathcal{G}, \end{aligned} \quad (2.95)$$

and let  $X_1, \dots, X_n$  be a sample from  $P$ ,  $P \in \mathcal{DGI}(\alpha)$ . Then the Bayes estimate of  $\varphi_2$  under  $L_2$  loss is (Yamato 1986, 1987)

$$\begin{aligned} \widehat{\phi}_{2\alpha}^* &= \frac{M+n}{M+n+1} \left[ p_n^2 \int_{\mathfrak{X}^2} h(x, y) dQ(x) dQ(y) \right. \\ &\quad + \frac{2p_n(1-p_n)}{nk} \sum_{i=1}^n \sum_{j=1}^k \int_{\mathfrak{X}} h(x, g_j X_i) dQ(x) \\ &\quad \left. + \frac{(1-p_n)^2}{n^2 k^2} \sum_{i_1, i_2} \sum_{j_1, j_2} h(g_{j_1} X_{i_1}, g_{j_2} X_{i_2}) \right] \\ &\quad + \frac{1}{k(M+n+1)} \left[ p_n \sum_{j=1}^k \int_{\mathfrak{X}} h(x, g_j x) dQ(x) \right. \\ &\quad \left. + \frac{(1-p_n)}{nk} \sum_i \sum_{j_1, j_2} h(g_{j_1} X_i, g_{j_2} X_i) \right]. \end{aligned} \quad (2.96)$$

If we let  $M$  go to zero, the above estimator reduces to

$$\widehat{\phi}_2^{**} = \frac{1}{n(n+1)k^2} \sum_{j_1, j_2} \left[ \sum_{i_1, i_2} h(g_{j_1} X_{i_1}, g_{j_2} X_{i_2}) + \sum_i h(g_{j_1} X_i, g_{j_2} X_i) \right], \quad (2.97)$$

and if we replace Dirichlet Invariant with Dirichlet process, clearly the estimator reduces to

$$\hat{\phi}_{2D}^{**} = \frac{1}{n(n+1)} \left[ \sum_{i_1, i_2} h(X_{i_1}, X_{i_2}) + \sum_i h(X_i, X_i) \right]. \quad (2.98)$$

For illustrative purposes, Yamato takes several different forms of  $h(x, y)$  and derives the Bayes estimates of the resulting parameters. For example if we take  $h(x, y) = |x - y|$  and  $\mathfrak{X} = R$ , then  $\theta = \int_{R^2} |x - y| dP(x) dP(y)$  is the coefficient of mean difference of the distribution  $P$ . On the other hand if  $h(x, y) = (x_1 - y_1)(x_2 - y_2)/2$  with  $x = (x_1, x_2)$  and  $y = (y_1, y_2)$ , then

$$\theta = \int_{R^2} x_1 x_2 dP(x_1, x_2) - \int_R x_1 dP(x_1, x_2) \int_R x_2 dP(x_1, x_2) \quad (2.99)$$

is the covariance of the distribution  $P$ .

In another example, he takes  $h(x, y) = \psi((x_1 - y_1)(x_2 - y_2))$  with  $x = (x_1, x_2)$  and  $y = (y_1, y_2)$ , and  $\psi = I[t > 0]$ . Then  $\theta = 2P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1$  is a measure of the correlation between  $(X_1, Y_1)$  and  $(X_2, Y_2)$  or concordance. Taking  $\mathcal{G} = \{e, g\}$  with  $e(x_1, x_2) = (x_1, x_2)$ ,  $g(x_1, x_2) = (x_2, x_1)$ , for  $(x_1, x_2) \in R^2$ , he derives the Bayes estimate. When  $M \rightarrow 0$ , this estimator reduces to the non-Bayesian estimator (Randles and Wolf 1979), namely

$$\begin{aligned} \hat{\theta} = & \frac{1}{n(n+1)} [\# \text{ of } \{ \text{pairs } (i, j) : (X_i - X_j)(Y_i - Y_j) > 0, 1 \leq i < j \leq n \} \\ & + \# \text{ of } \{ \text{pairs } (i, j) : (X_i - Y_j)(Y_i - X_j) > 0, 1 \leq i < j \leq n \} \\ & + \# \{ i : X_i = Y_i, 1 \leq i \leq n \} - n]. \end{aligned} \quad (2.100)$$

**Empirical Bayes Estimation of  $\phi(P)$**  Earlier empirical Bayes estimation results derived for  $F(t)$  by Korwar and Hollander (1976), Hollander and Korwar (1976), and for  $P(X \leq Y)$  by Phadia and Susarla (1979) under  $\mathcal{D}(\alpha)$  prior were reported. Tiwari and Zalkikar (1985) generalize these results by replacing the indicator function of the sets  $(-\infty, x]$  and  $[X \leq Y]$  by arbitrary measurable functions  $h(x)$  and  $h(x, y)$ . Specifically, the empirical Bayes estimation of estimable parameters of degree one and two of an unknown probability measure on  $(R, \mathcal{B})$  is treated, and asymptotically optimal results with rate of convergence  $O(n^{-1})$  of these estimators were established. In proving these results they used the Sethuraman (1994) representation for the Dirichlet process.

The Bayesian estimator of  $\phi_1$  based on a sample  $\mathbf{X}_{n+1}$  of size  $m$  at the  $(n+1)$ -th stage was obtained earlier as

$$\hat{\phi}_{1\alpha} = p_m \phi_0 + (1 - p_m) U_{n+1}, \quad (2.101)$$

where  $\phi_0 = \int h d\bar{\alpha}$ ,  $p_m = M/(M + nm)$  and  $U_{n+1} = \frac{1}{m} \sum_{j=1}^m h(X_{n+1, j})$ .

To estimate  $\phi_1$  at the  $(n+1)$ -th stage on the basis of  $(\mathbf{X}_1, \dots, \mathbf{X}_{n+1})$ , we may use the techniques of Sect. 2.2.4 to estimate first  $\phi_0$  from the previous  $n$  copies

and  $M$  by Zehnwirth's approach. Substituting these estimates, the empirical Bayes estimator of  $\phi_1$  at the  $(n+1)$ -th stage is given by

$$\hat{\phi}_{\alpha 1, n+1} = \hat{p}_m \sum_{i=1}^n \frac{U_i}{n} + (1 - \hat{p}_m) U_{n+1}, \quad (2.102)$$

where, for the samples  $\mathbf{X}_i$ ,  $i = 1, 2, \dots, n$ ,  $U_i = \frac{1}{m} \sum_{j=1}^m h(X_{ij})$ ,  $\hat{p}_m = \hat{M}_n / (\hat{M}_n + m)$ ,  $\hat{M}_n = \max(0, m(F_n - 1)^{-1})$ , and  $F_n$  is the F-ratio statistics in one-way ANOVA table based on the observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . For this estimator, they also established the asymptotic optimality relative to  $\alpha$  with rate of convergence  $O(n^{-1})$ .

Similarly they consider the empirical Bayes estimation of  $\phi_2$  with  $h(x, y) = 0$  whenever  $x = y$ . If  $M$  is known, the empirical Bayes estimator of  $\phi_2$  at the  $(n+1)$ -th stage is

$$\begin{aligned} \hat{\phi}_{\alpha 2, n+1}(P) = & \frac{M+m}{M+m+1} \left[ p_m^2 \frac{M+1}{M} \sum_{i=1}^n \frac{U_{2i}}{n} + 2p_m(1-p_m) \sum_{k=1}^n \sum_{i=1}^n \frac{U_{n+1, i, k}}{mn} \right. \\ & \left. + (1-p_m)^2 \sum_{1 \leq j \neq k \leq m} \frac{h(X_{n+1, j}, X_{n+1, k})}{m^2} \right], \end{aligned} \quad (2.103)$$

where for the  $i$ -th sample,  $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,m})$ ,  $i = 1, 2, \dots, n$  and,

$$U_{2i} = \frac{1}{m(m-1)} \sum_{1 \leq j \neq k \leq m} h(X_{i,j}, X_{i,k}), \quad (2.104)$$

$$U_{n+1, i, k} = \frac{1}{m} \sum_{j=1}^m h(X_{n+1, k}, X_{i, j}), \quad i = 1, 2, \dots, n. \quad (2.105)$$

Under the assumption that  $\int h^2(x, y) d\bar{\alpha}(x) d\bar{\alpha}(y)$  exists and is finite, they show that the sequence  $\{\hat{\phi}_{2, n+1}\}$  is asymptotically optimal relative to  $\alpha$  with the rate of convergence  $O(n^{-1})$ .

Ghosh (1985) also considered the empirical Bayes estimation of  $\phi_2(P)$  and computed exact Bayes risk for Bayes and empirical Bayes estimators. He showed that Dalal and Phadia (1983) result for the estimation of concordance coefficient can be obtained as a special case of his result.

Again for the case when  $M$  is unknown, Tiwari and Zalkikar (1985) use Zehnwirth's estimate for  $M$  and established similar result (see their paper for details). They also obtained the empirical Bayes estimator for  $\zeta$  and proved its asymptotic optimality with rate of convergence  $O(n^{-1})$ .

*Remark 2.4* Asymptotic optimality of the empirical Bayes estimators of variance and the mean deviation about the mean of  $P$  can be derived from the above result by taking  $h(x, y) = \frac{1}{2}(x - y)^2$  and  $|x - y|$ , respectively.

Similar empirical Bayes treatment is also given in Ghosh et al. (1989), where the main idea was to use past as well as the current data in estimating the parameters of the Dirichlet process prior. They show that by doing this, we get improved estimators in terms of smaller risks.

## 2.8 Two-Sample Problems

Suppose we have two independent samples,  $X_1, \dots, X_{n_1}$  from  $F$  and  $Y_1, \dots, Y_{n_2}$  from  $G$ . In this section we consider the Bayesian estimation of certain functionals of  $F$  and  $G$  with respect to the Dirichlet priors  $\mathcal{D}(\alpha_1)$  and  $\mathcal{D}(\alpha_2)$ , respectively.

### 2.8.1 Estimation of $P(X \leq Y)$

Ferguson (1973) derived the Bayesian estimator of  $\Delta = P(X \leq Y) = \int F dG$  under the squared error loss  $L_2$ . Let  $F \in \mathcal{D}(\alpha_1)$  and independently,  $G \in \mathcal{D}(\alpha_2)$ . Then for the no-sample problem the estimate of  $\Delta$  is given by  $\Delta_0 = \mathcal{E}(\Delta) = \int F_0 dG_0$  where  $F_0 = \mathcal{E}(F)$  and  $G_0 = \mathcal{E}(G)$ , and the expectation is taken with respect to the Dirichlet priors. Given the samples  $X_1, \dots, X_{n_1} \sim F$  and  $Y_1, \dots, Y_{n_2} \sim G$ , we update the estimate  $\Delta_0$  and obtain the Bayesian estimate as  $\hat{\Delta} = \int \hat{F}_{\alpha_1} d\hat{G}_{\alpha_2}$ , where  $\hat{F}_{\alpha_1}$  and  $\hat{G}_{\alpha_2}$  are Bayes estimators of  $F$  and  $G$ , respectively as obtained in Sect. 2.2.1. Simplifying further we get

$$\begin{aligned} \hat{\Delta}_{\alpha_1\alpha_2}(\mathbf{X}, \mathbf{Y}) &= p_{1n_1} p_{2n_2} \Delta_0 + p_{1n_1} (1 - p_{2n_2}) \frac{1}{n_2} \sum_{i=1}^{n_2} F_0(Y_i) \\ &\quad + (1 - p_{1n_1}) p_{2n_2} \frac{1}{n_1} \sum_{i=1}^{n_1} G_0(X_i -) + (1 - p_{1n_1})(1 - p_{2n_2}) \frac{1}{n_1 n_2} U, \end{aligned} \quad (2.106)$$

where  $p_{1n_1} = \alpha_1(R)/(\alpha_1(R) + n_1)$ ;  $p_{2n_2} = \alpha_2(R)/(\alpha_2(R) + n_2)$ , and

$$U = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} I_{(-\infty, Y_j]}(X_i)$$

is the Mann-Whitney statistic. When both  $\alpha_1(R)$  and  $\alpha_2(R)$  tend to zero,  $\hat{\Delta}_{\alpha_1\alpha_2}$  reduces to the usual nonparametric estimate  $(1/(n_1 n_2))U$ .

Hollander and Korwar (1976) extend this estimator to the empirical Bayes estimator. Assume that  $\alpha_1$  and  $\alpha_2$  are unknown except for  $\alpha_1(R)$  and  $\alpha_2(R)$  which are specified, and that we have  $n$  copies of data available from the first  $n$  stages and are required to estimate  $\Delta$  at the  $(n+1)$ -th stage. As in one sample case, they estimate  $\alpha_1$  and  $\alpha_2$  from the first  $n$ -stage data  $\mathbf{X}_i = (X_{i1}, \dots, X_{in_1})$  and  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_2})$

for  $i = 1, 2, \dots, n$  and propose the following estimator

$$\begin{aligned}
 \hat{\Delta}_{\alpha_1\alpha_2n}(\mathbf{X}, \mathbf{Y}) = & p_{1n_1}p_{2n_2}\frac{1}{n^2n_1n_2}\sum_{j=1}^n\sum_{k=1}^{n_2}\sum_{i=1}^n\sum_{l=1}^{n_1}I_{(-\infty, Y_{jk}]}(X_{il}) \\
 & + p_{1n_1}(1 - p_{2n_2})\frac{1}{nn_1n_2}\sum_{k=1}^{n_2}\sum_{i=1}^n\sum_{l=1}^{n_1}I_{(-\infty, Y_{n+1k}]}(X_{il}) \\
 & + (1 - p_{1n_1})p_{2n_2}\frac{1}{n_1}\sum_{l=1}^{n_1}\left\{1 - \frac{1}{nn_2}\sum_{j=1}^n\sum_{k=1}^{n_2}I_{(-\infty, X_{n+1l}]}(Y_{jk})\right\} \\
 & + (1 - p_{1n_1})(1 - p_{2n_2})\frac{1}{n_1n_2}\sum_{l=1}^{n_1}\sum_{k=1}^{n_2}I_{(-\infty, Y_{n+1k}]}(X_{n+1l}).
 \end{aligned} \tag{2.107}$$

Finally, they show that  $\hat{\Delta}_{\alpha_1\alpha_2n}$  is asymptotically optimal with respect to  $\alpha_1$  and  $\alpha_2$ . Clearly, when  $\alpha_1(R)$  and  $\alpha_2(R)$  are also unknown, they could be estimated as indicated earlier, and the above estimator may be adjusted accordingly.

### 2.8.2 Estimation of the Difference Between Two CDFs

A measure of the difference between two distributions functions  $F$  and  $G$ , is defined by

$$d(F, G) = \int (F(t) - G(t))^2 d\left(\frac{F(t) + G(t)}{2}\right), \tag{2.108}$$

which is somewhat difficult to handle. However, if the distributions are continuous on  $R$ , then it can be written as

$$d(F, G) = \frac{4}{3} - \left[ \int G(t)dF^2(t) + \int F(t)dG^2(t) \right]. \tag{2.109}$$

Based on two independent samples,  $X_1, \dots, X_{n_1}$  from  $F$  and  $Y_1, \dots, Y_{n_2}$  from  $G$ , Yamato (1975) considered the problem of Bayesian estimation of  $d(F, G)$  under the squared error loss  $L(d(F, G), \hat{d}(F, G)) = (d(F, G) - \hat{d}(F, G))^2$ . In order to use the latter version of the definition, he defines linearized Dirichlet process as priors for  $F$  and  $G$  which are assumed to be continuous. Following Doksum (1972), he defines a linearized Dirichlet process as follows. For reals  $a < b$ , consider the partition  $\pi$  of  $(a, b)$ ,  $a = t_1 < t_2 < \dots < t_k = b$  and denote the norm of the partition as  $\|\Delta\pi\| = \max_{1 \leq i \leq k-1} |t_{i+1} - t_i|$ . Let  $\alpha$  be a finite measure on  $(R, \mathcal{B})$  with support  $(a, b)$ . Let  $H_0$  be a realization of the Dirichlet process with parameter  $\alpha$  such that  $H_0(a) = 0$  and  $H_0(b) = 1$  with probability one. Given the partition  $\pi$ , the joint distribution of the corresponding increments of the distribution function

has a Dirichlet distribution. With this formulation, he defines a linearized Dirichlet process as follows:

**Definition 2.5** (Yamato)  $H$  is said to be a linearized Dirichlet process with parameter  $\alpha$  and partition  $\pi$ , when  $H$  is linear between the points  $(t_1, H_0(t_1)), \dots, (t_k, H_0(t_k))$  and  $H_0(t_i)$ ,  $i = 1, 2, \dots, k$  are the realization of the Dirichlet process with parameter  $\alpha$  having support  $(a, b)$  and partition  $\pi$ , with  $a = t_1$  and  $b = t_k$ .

Assume  $F$  and  $G$  as independent linearized Dirichlet processes with parameters  $\alpha_1$  and  $\alpha_2$ , respectively, and partition  $\pi$ . Then under the squared error loss, the Bayes estimate is given by the posterior mean,

$$\mathcal{E}[d(F, G) | X_1, \dots, X_{n_1} \text{ and } Y_1, \dots, Y_{n_2}]. \quad (2.110)$$

To evaluate this expectation, he defines (pseudo Bayesian estimators)

$$\begin{aligned} \widehat{F}_{\alpha_1 n_1}(t) &= p_{n_1} F_0(t) + (1 - p_{n_1}) \widehat{F}_{n_1}(t), \\ \widehat{G}_{\alpha_2 n_2}(t) &= p_{n_2} G_0(t) + (1 - p_{n_2}) \widehat{G}_{n_2}(t), \end{aligned} \quad (2.111)$$

on the interval  $(a, b)$ , with  $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 0$  for  $t \leq a$ , and  $\widehat{F}_{\alpha_1 n_1}(t) = \widehat{G}_{\alpha_2 n_2}(t) = 1$  for  $t \geq b$ , with probability one, where  $p_{n_1} = \alpha_1(R)/(\alpha_1(R) + n_1)$ ,  $p_{n_2} = \alpha_2(R)/(\alpha_2(R) + n_2)$ ,  $\widehat{F}_{n_1}$  and  $\widehat{G}_{n_2}$  are the empirical distribution functions of the samples  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively,  $F_0(t) = \alpha_1(t)/\alpha_1(R)$  and  $G_0(t) = \alpha_2(t)/\alpha_2(R)$ . Denoting by  $\widehat{F}_{\alpha_1 n_1, \Delta}$  and  $\widehat{G}_{\alpha_2 n_2, \Delta}$ , the linearized versions of  $\widehat{F}_{\alpha_1 n_1}$  and  $\widehat{G}_{\alpha_2 n_2}$ , respectively on the partition  $\pi$ , he evaluates the above expectation obtaining the Bayesian estimator of  $d(F, G)$  on the interval  $(a, b)$  as

$$\begin{aligned} \widehat{d}_{\alpha_1 \alpha_2}(F, G) &= \frac{4}{3} - \frac{\alpha_1(R) + n_1}{\alpha_1(R) + n_1 + 1} \int_a^b \widehat{G}_{\alpha_2 n_2, \Delta}(t) d\widehat{F}_{\alpha_1 n_1, \Delta}^2(t) \\ &\quad - \frac{1}{\alpha_1(R) + n_1 + 1} \left\{ \frac{2}{3} \int_a^b \widehat{G}_{\alpha_2 n_2, \Delta}(t) d\widehat{F}_{\alpha_1 n_1, \Delta}(t) \right. \\ &\quad \left. + \frac{1}{3} \sum_{i=1}^{k-1} \widehat{G}_{\alpha_2 n_2}(t_{i+1}) [\widehat{F}_{\alpha_1 n_1}(t_{i+1}) - \widehat{F}_{\alpha_1 n_1}(t_i)] \right\} \\ &\quad - \frac{\alpha_2(R) + n_2}{\alpha_2(R) + n_2 + 1} \int_a^b \widehat{F}_{\alpha_1 n_1, \Delta}(t) d\widehat{G}_{\alpha_2 n_2, \Delta}^2(t) \\ &\quad - \frac{1}{\alpha_2(R) + n_2 + 1} \left\{ \frac{2}{3} \int_a^b \widehat{F}_{\alpha_1 n_1, \Delta}(t) d\widehat{G}_{\alpha_2 n_2, \Delta}(t) \right. \\ &\quad \left. + \frac{1}{3} \sum_{i=1}^{k-1} \widehat{F}_{\alpha_1 n_1}(t_{i+1}) [\widehat{G}_{\alpha_2 n_2}(t_{i+1}) - \widehat{G}_{\alpha_2 n_2}(t_i)] \right\}. \end{aligned} \quad (2.112)$$

Finally, taking the limit  $\|\Delta\pi\| \rightarrow 0$ , the above estimator reduces to

$$\begin{aligned} \widehat{d}_{\alpha_1\alpha_2}(F, G) = & \frac{4}{3} - \frac{1}{\alpha_1^* + 1} \left\{ \int_a^b \widehat{G}_{\alpha_2 n_2}(t) d\widehat{F}_{\alpha_1 n_1}(t) + \alpha_1^* \int_a^b \widehat{G}_{\alpha_2 n_2}(t) d\widehat{F}_{\alpha_1 n_1}^2(t) \right\} \\ & - \frac{1}{\alpha_2^* + 1} \left\{ \int_a^b \widehat{F}_{\alpha_1 n_1}(t) d\widehat{G}_{\alpha_2 n_2}(t) + \alpha_2^* \int_a^b \widehat{F}_{\alpha_1 n_1}(t) d\widehat{G}_{\alpha_2 n_2}^2(t) \right\}, \end{aligned} \quad (2.113)$$

where  $\alpha_1^* = \alpha_1(R) + n_1$  and  $\alpha_2^* = \alpha_2(R) + n_2$ .

This estimator is derived on the basis of a particular prior distribution with the interval  $(a, b)$  as its support. In general, when  $F$  and  $G$  are continuous, the author proposes an estimator of  $d(F, G)$  as  $\widehat{d}(F, G)$  with the range of integrals replaced in the above formula by  $-\infty$  to  $\infty$ . By letting  $\alpha_1(R)$  and  $\alpha_2(R)$  tend to zero, we get a non-Bayesian estimator of  $d(F, G)$ .

It should be noted that the above formula (2.113) for the difference between two distributions is valid if and only if  $F$  and  $G$  are continuous. For this reason, the author used the linearized Dirichlet processes as priors in deriving the Bayes estimate, and passing through the limit of the Bayes estimate yielded the above estimator  $\widehat{d}_{\alpha_1\alpha_2}(F, G)$  (with the integrals  $\int_{-\infty}^{\infty}$ ). However, the author argues that if we define the difference as the above quantity regardless of the distribution functions being continuous or not, and assign Dirichlet priors to them, the direct computation will show that the resulting estimate is equal to the above estimate.

### 2.8.3 Estimation of the Distance Between Two CDFs

When  $F$  and  $G$  are continuous distribution functions, the horizontal distance between  $F$  and  $G$  is defined as  $\Delta(x) = G^{-1}(F(x)) - x$ , for a real number  $x$ . Hollander and Korwar (1982) consider a one-sample problem where  $G$  is assumed to be known and only a random sample of size  $n$  from  $F$  is available to estimate  $\Delta$ . Although  $F$  is continuous, they assume  $F \in \mathcal{D}(\alpha)$ . Under the loss function  $L_1$ , the Bayes estimator for the no-sample problem is found by minimizing the integrand of

$$\mathcal{E}(L(\Delta, \widehat{\Delta})) = \int \mathcal{E}(\Delta(x) - \widehat{\Delta}(x))^2 dW(x) \quad (2.114)$$

yielding  $\widehat{\Delta}_0(x) = \mathcal{E}(\Delta(x)) = \mathcal{E}\{G^{-1}F(x)\} - x$ . For a sample of size  $n$  from  $F$ , the Bayesian estimator is obtained simply by updating  $\alpha$ .

If  $G$  is assumed to be an exponential distribution,  $G(x) = 1 - e^{-\lambda x}$ ,  $x > 0$ ,  $\lambda > 0$ , then  $\widehat{\Delta}_0$  is

$$\begin{aligned} \widehat{\Delta}_0(x) &= \frac{1}{\lambda \cdot B(\alpha', \beta')} \cdot \int_0^1 \sum_{j=1}^{\infty} \frac{y^{\alpha'+j-1} (1-y)^{\beta'-1}}{j} dy - x \\ &= \frac{1}{\lambda} \cdot \sum_{j=1}^{\infty} \frac{B(\alpha' + j, \beta')}{j \cdot B(\alpha', \beta')} - x, \end{aligned} \quad (2.115)$$

where  $\alpha' = \alpha((-\infty, x])$ ,  $\beta' = \alpha((x, \infty)) = \alpha(R) - \alpha'$ . Now for a sample of size  $n$  from  $F$ , the Bayes estimator is the above expression  $\hat{\Delta}_0(x)$  with  $\alpha'$  and  $\beta'$  replaced by  $\alpha^* = \alpha' + \sum_{i=1}^n \delta_{X_i}$  and  $\beta^* = \alpha(R) + n - \alpha^*$ , respectively, their updated versions.

## 2.9 Hypothesis Testing

In applications of the Dirichlet process prior so far, we have discussed mainly the estimation of an unknown distribution function  $F$  or a parameter  $\varphi$  which is a function of the unknown probability measure  $P$ . Ferguson (1973) pointed out the difficulty of using the Dirichlet Process prior in hypothesis testing problems. However, Susarla and Phadia (1976) were able to show how such problems can be handled. The idea was to replace the usual 0–1 loss with a smoother loss function based on a known weight function  $W$ . Thus their approach to the problem of the hypothesis testing was from a decision theoretic point of view—a first as far as we know. Their method can be extended to treat multiple decision theoretic problems as well. This is described now.

### 2.9.1 Testing $H_0 : F \leq F_0$

Let  $\mathbf{X} = (X_1, \dots, X_m)$  be a random sample from the distribution function  $F$ . Let  $F_0$  be a known distribution function. Consider the problem of testing hypothesis  $H_0 : F \leq F_0$  against the alternative  $H_1 : F \not\leq F_0$  when the loss function  $L$  is given by

$$L(F, a_0) = \int (F - F_0)^+ dW \quad \text{and} \quad L(F, a_1) = \int (F - F_0)^- dW, \quad (2.116)$$

where  $L(F, a_i)$  indicates the loss incurred when action  $a_i$  (deciding in favor of  $H_i$ ) is taken for  $i = 0, 1$ ,  $W$  is a weight function,  $a^+ = \max\{a, 0\}$  and  $a^- = -\min\{a, 0\}$  for any real number  $a$ . Assume  $F \in \mathcal{D}(\alpha)$ . Let  $\delta(\mathbf{X}) = \mathcal{P}\{\text{taking action } a_0 \mid \mathbf{X}\}$ . Then the Bayes risk of  $\delta$  against  $\mathcal{D}(\alpha)$  is

$$r_m(\delta, \alpha) = \int \mathcal{E}[L(F, a_0) - L(F, a_1) \mid \mathbf{X}] \delta(\mathbf{X}) dQ_m(\mathbf{X}) + \mathcal{E}[L(F, a_1)], \quad (2.117)$$

where  $Q_m$  is the unconditional distribution of  $\mathbf{X}$  and the expectation is taken with respect to  $\mathcal{D}(\alpha)$ . Hence a Bayes rule against  $\mathcal{D}(\alpha)$  which minimizes the above risk is given by  $\delta_m(\mathbf{X}) = I[\Delta_m(\mathbf{X}) \leq 0]$  where  $\Delta_m(\mathbf{X}) = \int \mathcal{E}[F(u) - F_0(u) \mid \mathbf{X}] dW(u)$  and the minimum Bayes risk is

$$r_m^*(\alpha) = \int_{[\Delta_m(\mathbf{X}) \leq 0]} \Delta_m(\mathbf{X}) dQ_m(\mathbf{X}) + \mathcal{E}[L(F, a_1)]. \quad (2.118)$$

If  $\alpha$  is known,  $\Delta_m(\mathbf{X})$  can be easily evaluated since for each  $u$ ,  $F(u)|\mathbf{X} = \mathbf{x} \sim Be(\alpha(-\infty, u] + \sum_{i=1}^m I[X_i \leq u], \alpha(u, \infty) + \sum_{i=1}^m I[X_i > u])$ .

When  $\alpha$  is unknown, we can use the empirical Bayes method. Let  $\alpha(R) = 1$  and assume the usual set up for the empirical Bayes estimation with sample size  $m_i$  at the  $i$ -th stage. Then an empirical Bayes rule at the  $(n+1)$ -th stage is given by

$$\xi_{n+1}(\mathbf{X}_{n+1}) = \mathcal{P}\{\text{accepting } a_0 \mid \mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}_{n+1}\}, \quad (2.119)$$

Let  $\widehat{\Delta}_n(\mathbf{X}_{n+1})$  be an estimate based on  $(\mathbf{X}_1, \dots, \mathbf{X}_n)$  of

$$\Delta_{m_{n+1}}(\mathbf{X}_{n+1}) = \int \mathcal{E}[F_{n+1}(u) - F_0(u) \mid \mathbf{X}_{n+1}] dW(u)$$

given by

$$\begin{aligned} & \widehat{\Delta}_n(\mathbf{x}_{n+1}) + \int F_0 dW \\ &= \int \frac{\{\widehat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} I[X_{n+1,i} \leq u]\} dW(u)}{(1 + m_{n+1})}, \end{aligned} \quad (2.120)$$

where  $\widehat{\alpha}(-\infty, u] = n^{-1} \sum_{j=1}^n m_j^{-1} \sum_{i=1}^{m_j} I[X_{j,i} \leq u]$ . Let

$$\xi_n(\mathbf{x}_{n+1}) = I[\widehat{\Delta}_n(\mathbf{X}_{n+1}) \leq 0],$$

and let  $r_{n+1}(\xi_n)$  denote the risk of using  $\xi_n$  to decide about  $F_{n+1}$ . Then Susarla and Phadia (1976) have proved that  $r_{n+1}(\xi_n) - r_{m_{n+1}}^*(\alpha) \leq n^{-1/2}$ .

When  $\alpha(R)$  is unknown, they estimate it by a consistent estimator  $\widehat{\alpha}(R) = (\log m_n)^{-1} \{\text{\# of distinct observations in } \mathbf{X}_n\}$  (see property 16 of Sect. 1.2). Let  $\xi_n^*$  be the rule obtained by substituting this estimator in  $\xi_n$  with

$$\begin{aligned} & \widehat{\Delta}_n(\mathbf{X}_{n+1}) + \int F_0 dW \\ &= \int \frac{\{\widehat{\alpha}(R)\widehat{\alpha}(-\infty, u] + \sum_{i=1}^{m_{n+1}} I[X_{n+1,i} \leq u]\} dW(u)}{(\widehat{\alpha}(R) + m_{n+1})}. \end{aligned} \quad (2.121)$$

Then they prove the following asymptotic result. Let  $\alpha$  be nonatomic and  $m_{n+1} \rightarrow \infty$  as  $n \rightarrow \infty$ . Then  $r_{n+1}(\xi_n^*) - r_{m_{n+1}}^*(\alpha) = O((m_{n+1})^{-1} (\min\{\log m_n, n\})^{-1/2})$ .

In addition, they have shown that some of these procedures are component-wise admissible and have also discussed the extension of their results to the multiple action problem.

## 2.9.2 Testing Positive Versus Nonpositive Dependence

In the bivariate distribution case, we come across the problem of testing positive dependence versus nonpositive dependence. Let  $F(x, y)$  be a bivariate distribution

function defined on  $(R^2, \mathcal{B}^2)$  with marginal CDFs  $F_X(x)$  and  $F_Y(y)$ , respectively. The objective is to test the following hypotheses.

$$\begin{aligned} H_0 : F(x, y) &\geq F_X(x)F_Y(y) \quad \text{for all } (x, y) \text{ in } R^2 \\ H_1 : F(x, y) &< F_X(x)F_Y(y) \quad \text{for all } (x, y) \text{ in } R^2, \end{aligned} \quad (2.122)$$

under the loss function

$$\begin{aligned} L(F, a_0) &= \int (F(x, y) - F_X(x)F_Y(y))^- dW(x, y) \\ L(F, a_1) &= \int (F(x, y) - F_X(x)F_Y(y))^+ dW(x, y), \end{aligned} \quad (2.123)$$

where the actions  $a_0$  and  $a_1$  are to accept  $H_0$  and  $H_1$ , respectively,  $W$  is a known weight function on  $R^2$ . For given observations  $(\mathbf{x}, \mathbf{y})$ , denote by  $\theta(\mathbf{x}, \mathbf{y})$  the probability of taking action  $a_0$ . Then Dalal and Phadia (1983) have shown that the Bayes rule against  $\mathcal{D}(\alpha)$  is given by

$$\theta(\mathbf{x}, \mathbf{y}) = I_{[\Delta_n(\mathbf{x}, \mathbf{y})]}, \quad (2.124)$$

where

$$\begin{aligned} \Delta_n(\mathbf{x}, \mathbf{y}) &= \mathcal{E}[L(F, a_0) - L(F, a_1) \mid (\mathbf{X}, \mathbf{Y})] \\ &= \int [\mathcal{E}(F(x', y') - F_X(x')F_Y(y')) \mid (\mathbf{X}, \mathbf{Y})] dW(x', y'). \end{aligned} \quad (2.125)$$

Here the expectation is taken with respect to the posterior Dirichlet process with parameter  $\alpha + \sum_{i=1}^n \delta_{(x_i, y_i)}$ . Let  $\alpha = M\mathcal{Q}$ ,  $G_0$  be a CDF corresponding to  $\mathcal{Q}$ ,  $G^* = p_n G_0 + (1 - p_n)\widehat{G}_n$ , where  $\widehat{G}_n$  is the empirical CDF based on the  $n$  observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , and  $p_n = M/(M + n)$ . Then the integrand can be evaluated as

$$\begin{aligned} &\frac{MG_0(x', y') + \sum_{i=1}^n \delta_{(x_i, y_i)}((-\infty, x'] \times (-\infty, y'])}{M + n} \\ &- \frac{G^*(x', y') + MG_X^*(x')G_Y^*(y')}{M + n + 1}, \end{aligned} \quad (2.126)$$

and hence  $\Delta_n(\mathbf{x}, \mathbf{y})$  can be evaluated. As in the case of estimating the concordance coefficient above, the empirical Bayes solution can be carried out here as well when  $\alpha$  is not known, with  $M$  known or unknown.

**Testing the Hypothesis  $H_0 : F \leq G$  Against the Alternative  $H_1 : F \not\leq G$**  An analog of the test discussed in Sect. 2.9.1 in a two-sample situation is to test the hypothesis  $H_0 : F \leq G$  against the alternative  $H_1 : F \not\leq G$ . This topic is covered more generally in Sect. 3.6 based on samples with right-censored observations. Its application to the uncensored data as a special case is obvious and therefore it will not be presented here.

### 2.9.3 A Selection Problem

Consider the following selection problem. We are given  $k$  samples,  $\mathbf{X}_i = (X_{i1}, \dots, X_{ik_i})$  distributed according to  $F_i$ ,  $i = 1, 2, \dots, k$ , and a sample  $\mathbf{Y} = (Y_1, \dots, Y_n)$  known to have come from one of the  $k$  distributions. The problem is to find from which one. Antoniuk (1974) considered this problem and provided a Bayes solution. Let  $\mathfrak{X}$  be a set of nonnegative integers and  $\sigma(\mathfrak{X})$  be the corresponding  $\sigma$ -algebra generated by the singleton sets. Assume that for  $i = 1, 2, \dots, k$ ,  $F_i \sim \mathcal{D}(\alpha_i)$ . For technical reasons, each  $\alpha_i$  is taken to be a discrete measure with the same support and defined on  $\sigma(\mathfrak{X})$  with  $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \dots)$  and  $\alpha_i(\{j\}) = \alpha_{ij}$ ,  $|\alpha_i| = \sum_{j=0}^{\infty} \alpha_{ij}$ . Let  $\pi_j$  be the prior probability that the sample  $Y_1, \dots, Y_n$  came from  $F_j$ ,  $j = 1, 2, \dots, k$ . Let  $L(i, j)$  be the associated loss function in deciding  $\mathbf{Y}$  as coming from  $F_i$  when in fact it is from  $F_j$ . The goal is to seek a non-randomized decision rule which minimizes the expected loss. First note that  $F_i|\mathbf{X}_i \sim \mathcal{D}(\alpha_i^*)$ , where  $\alpha_i^* = (\alpha_{i0}^*, \alpha_{i1}^*, \dots)$ , with  $\alpha_{ij}^* = \alpha_{ij} + m_{ij}$ , and  $m_{ij}$  is the number of  $X_i$ 's equal to  $j$ ,  $j = 0, 1, \dots$ . The Bayes risk  $r_i$  is given by

$$r_i(\boldsymbol{\pi}, \boldsymbol{\alpha}) = \sum_{j=1}^k L(i, j) P(j|Y_1, \dots, Y_n) = \sum_{j=1}^k L(i, j) \frac{\pi_j P(\mathbf{Y}|j)}{\sum_{j=1}^k \pi_j P(\mathbf{Y}|j)}, \quad (2.127)$$

where

$$P(\mathbf{Y}|j) = \prod_{l=0}^{\infty} \frac{\alpha_{jl}^{*(k_l)}}{\alpha_l^{*(n)}}, \quad a^{(n)} = a(a+1) \dots (a+n-1), n > 0, \quad (2.128)$$

and  $k_l$  is the number of  $Y$ 's equal to  $l$ . The Bayes decision rule selects  $s$ , where  $r_s = \min r_i$ . For the 0–1 loss and uniform prior  $\pi_j = 1/k$ , the Bayes decision rule is to choose  $s$  for which  $P(\mathbf{Y}|s) = \max_j P(\mathbf{Y}|j)$ .



<http://www.springer.com/978-3-642-39279-5>

Prior Processes and Their Applications

Nonparametric Bayesian Estimation

Phadia, E.G.

2013, XIV, 207 p., Hardcover

ISBN: 978-3-642-39279-5