Bayesian Analysis (2024) **19**, Number 1, pp. 293–318

Inference for Bayesian Nonparametric Models with Binary Response Data via Permutation Counting

Dennis Christensen[*,†]

Abstract. Since the beginning of Bayesian nonparametrics in the early 1970s, there has been a wide interest in constructing models for binary response data. Such data arise naturally in problems dealing with bioassay, current status data and sensitivity testing, and are equivalent to left and right censored observations if the inputs are one-dimensional. For models based on the Dirichlet process, inference is possible via Markov chain Monte Carlo (MCMC) simulations. However, there exist multiple processes based on different principles, for which such MCMC-based methods fail. Examples include logistic Gaussian processes and quantile pyramids. These require MCMC for posterior inference given exact observations, and thus become intractable when the data comprise both left and right censored observations. Here we present a new importance sampling algorithm for nonparametric models given exchangeable binary response data. It can be applied to any model from which samples can be generated, or even only approximately generated. The main idea behind the algorithm is to exploit the symmetries introduced by exchangeability. Calculating the importance weights turns out to be equivalent to evaluating the permanent of a certain class of $(0,1)$-matrix, which we prove can be done in polynomial time by deriving an explicit algorithm.

MSC2020 subject classifications: Primary 62N01, 62G05; secondary 15A15.

Keywords: Bayesian nonparametrics, binary response data, current status data, bioassay, permanents, importance sampling, binary classification.

1 Introduction

In many statistical applications, we only observe a Bernoulli random variable indicating whether a real-valued latent variable is below or above a certain threshold. Examples include problems in current status data and bioassay, where we aim to estimate a distribution P governing the probability that an individual has transitioned from state 0 to state 1 before time t (Albert and Chib, 1993; Keiding et al., 1996; Groeneboom and Jongbloed, 2014). Another example is sensitivity testing, in which we repeatedly choose an impact level E of energy and then observe whether a physical system is intact or broken after the impact. Such methods are used for studying the sensitivity of explosives or a material's resistance to stress (Dixon and Mood, 1948; Neyer, 1994; Christensen, 2022).

Since the beginning of Bayesian nonparametrics, there has been an interest in such

[*]Department of Mathematics, University of Oslo, Oslo, Norway.
[†]Norwegian Defence Research Establishment (FFI), Kjeller, Norway, dennis.christensen@ffi.no

 https://doi.org/10.1214/22-BA1353

binary response problems. Following the introduction of the Dirichlet process by Ferguson (1973), Antoniak (1974) showed that the posterior distribution of a Dirichlet process given censored data is a mixture of Dirichlet processes, and applied this to bioassay. Dirichlet processes with binary response data were further studied in Bhattacharya (1981); Kuo (1988); Gelfand and Kuo (1991); Doss (1994); Newton and Zhang (1999). These methods rely on one or more of the following particularly useful properties of the Dirichlet process: its conjugate posterior representation (Ferguson, 1973), its explicit marginal distribution (Antoniak, 1974) and its stick-breaking representation (Sethuraman, 1994). These three properties have also allowed for the development of Markov chain Monte Carlo (MCMC) sampling methods for Dirichlet process mixture models (DPMMs) (Ferguson, 1983; Lo, 1984). In the terminology of Papaspiliopoulos and Roberts (2008), this may be achieved either with marginal MCMC methods (Escobar and West, 1995; Neal, 2000) or conditional MCMC methods (Walker, 2007; Kalli et al., 2011). Due to the tractability of the mixture components, such MCMC techniques also apply when dealing with binary response data (see Paulon et al. (2020) for a recent application with current status data and dependent censoring). In principle, this is not only true for DPMMs, but for any mixture model with tractable marginal mixture distributions, such as normalised random measures with independent increments (NRMIs) (Regazzini et al., 2003; Lijoi et al., 2005, 2007), or with stick-breaking representations (Hjort, 1990; Paisley et al., 2010; Ishwaran and James, 2001). On the semiparametric side, Bayesian inference for the proportional hazards model (Cox, 1972) with current status data has been studied via Gibbs sampling (Cai et al., 2011) and expectation maximisation (Wang et al., 2015).

Although models based on Dirichlet processes are largely applicable to problems with binary response data, many nonparametric models are not. A notable example of this is the logistic Gaussian process (Leonard, 1978; Lenk, 1988, 1991) and the Gaussian process density sampler (Murray et al., 2008). In these models, the likelihood function will contain an integral of a Gaussian process due to the censoring. Thus, a direct implementation of MCMC-based inference is not feasible. Furthermore, since there is no conjugate posterior representation for such models, alternatives such as successive substitution sampling (Doss, 1994) are also out of reach. The same is true for other model choices in Bayesian nonparametrics, such as quantile pyramids (Hjort and Walker, 2009), normalised infinitely divisible multinomial (NIDM) processes (Lijoi et al., 2019) and Pitman-Yor multinomial processes (Lijoi et al., 2020). These examples are in line with Orbanz and Teh's (2011) prediction that over time, more Bayesian nonparametric models which are not based on the Dirichlet process will continue to arise.

In this paper, we introduce a new importance sampling algorithm which enables full Bayesian inference for models with exchangeable binary response data. The construction is highly general and applies to any model from which a data sample can be simulated. In particular, it does not rely on the tractable properties of Dirichlet processes. Unlike approximate methods such as the approximate Bayesian computation (ABC) rejection sampling algorithm (see Marin et al. (2012) for a review), our new simulation algorithm converges to the true posterior distribution, not just an approximation of it. As is illustrated in our simulation case study (see Section 4), this exact convergence result also holds when it is only possible to sample from a finite-dimensional truncation of the

model, as studied by Muliere and Secchi (1995); Campbell et al. (2019); Arbel et al. (2019); Lijoi et al. (2019, 2020).

The key to the new algorithm is to exploit the symmetry introduced by exchangeability of the data, and then essentially to correct for this exploitation by multiplying by an appropriate importance weight. Calculating the weight turns out to be equivalent to evaluating the permanent of a $(0,1)$-matrix, that is, a matrix whose entries are all either 0 or 1. For a general such matrix, this is known to be a #P-complete problem (Valiant, 1979). However, for the matrices arising in our setting, we are able to derive an explicit algorithm which computes their permanents in polynomial time. Code for implementing this new algorithm can be found in the publicly available GitHub repository.

The remainder of the paper is structured as follows. In Section 2, we set up the problem and introduce the importance sampling algorithm. We show how to calculate the marginal likelihood and how to carry out posterior inference. In Section 3, we derive an algorithm for calculating the importance weights in polynomial time. Next, in Section 4, we apply the new importance sampling algorithm to experiments with both simulated and real data. The theory is then extended in Section 5 to problems with multidimensional inputs. Finally, we briefly discuss extensions, limitations and consistency in Section 6.

2 Construction

2.1 Model

Let $([0,\infty),\mathscr{F})$ be the measurable space of non-negative real numbers equipped with the Borel σ-algebra. We use $[0,\infty)$ as our sample space to more conveniently illustrate the theory, although everything also applies to $\mathbb{R}$ or a real bounded interval. Let $P \sim \pi(\cdot)$ be a random probability distribution on $([0,\infty),\mathscr{F})$. Then P induces a random cumulative distribution function (cdf) F on $[0,\infty)$. Our binary data $y = (y_1,\ldots,y_n) \in \{0,1\}^n$ is assumed to be generated by

$$y_i \mid F \sim \text{Bernoulli}(F(t_i)),$$

independently for $i = 1,\ldots,n$, for some known thresholds $t_1,\ldots,t_n$. That is, $\pi(y_i \mid F) = F(t_i)^{y_i}\{1-F(t_i)\}^{1-y_i}$. We also write $\pi(y)$ for the marginal distribution of y, having marginalised over F.

It is useful to introduce the latent variables $x = (x_1,\ldots,x_n) \in [0,\infty)^n$ with $x_i \mid P \sim P$ independently for $i = 1,\ldots,n$. That is, $\pi(x_i \mid P) = P$. Then our binary variables y_i can be seen as indicator variables, $y_i = \mathbb{1}_{x_i \le t_i}$. Let $\pi(x)$ be the marginal distribution of x, marginalising over P. Note that the x_i need not be marginally independent. However, they will always form an exchangeable sequence. The same is true for the y_i. Since $\pi(\cdot)$ may both refer to the distribution of P and the marginal distributions of x or y, we will make it clear from context which distribution is in use.

We shall need to introduce some notation. Given $y_1, \ldots, y_n$, let

$$\mathcal{B}_i = \begin{cases} [0, t_i] & \text{if } y_i = 1, \\ (t_i, \infty) & \text{if } y_i = 0, \end{cases} \tag{2.1}$$

for $i = 1, \ldots, n$. Thus, observing y is equivalent to observing that $x_i \in \mathcal{B}_i$ for all $i = 1, \ldots, n$. Now, let $n_0 = \#\{i \mid y_i = 1\}$. Then, by exchangeability, we may without loss of generality order the y_i so that $y_1 = \cdots = y_{n_0} = 1$ and $y_{n_0+1} = \cdots = y_n = 0$, and further so that $t_1 \leq \cdots \leq t_{n_0}$ and $t_{n_0+1} \leq \cdots \leq t_n$. Note that this also induces an ordering of the sets $\mathcal{B}_1, \ldots, \mathcal{B}_n$. We write $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n$, so that observing y is equivalent to observing that $x \in \mathcal{B}$. For now, we will assume that there are no repeated values amongst the thresholds $t_1, \ldots, t_n$. Later on, in Section 2.2, we show how to account for situations where we have repeated values amongst them.

2.2 Estimating the marginal likelihood

Our first objective is to estimate the marginal likelihood $\pi(y) = \mathbb{P}(x \in \mathcal{B})$ of the model. In addition to being valuable in its own right, this will also guide how to perform posterior inference for P in general, to be covered in Section 2.3. For brevity of notation, define the measure $\mathscr{P}$ on $([0, \infty)^n, \mathscr{F}^n)$ by $\mathscr{P}(\mathcal{A}) = \mathbb{P}(x \in \mathcal{A})$. The marginal likelihood will be estimated via an importance sampling algorithm, exploiting the symmetries present as a result of the x_i being exchangeable.

Let $\mathbb{1}$ denote the indicator function, so $\mathbb{1}_{\mathcal{B}}(x)$ returns 1 if $x \in \mathcal{B}$ and 0 otherwise. Consider first the following naive estimator.

$$\hat{\mathscr{P}}_T(\mathcal{B}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\mathcal{B}}(x^{(t)}), \tag{2.2}$$

where $x^{(t)} \sim \pi(\cdot)$ independently for $t = 1, \ldots, T$.

By the law of large numbers, $\hat{\mathscr{P}}_T(\mathcal{B})$ is indeed a consistent estimator for the marginal likelihood $\mathscr{P}(\mathcal{B})$. However, in practice, we will never experience that $x \in \mathcal{B}$ if n is even moderately large, so $\hat{\mathscr{P}}_T(\mathcal{B})$ will always just be zero. This is also true even if parallel computing is employed, as the acceptance probability decreases exponentially with n. In order to adjust it to yield something practically feasible, we will have to loosen the condition that $x \in \mathcal{B}$ by replacing $\mathcal{B}$ with a larger space. We do this by exploiting the symmetries of the measure $\mathscr{P}$ due to the exchangeability of the x_i.

The group S_n of n-permutations acts on $[0, \infty)^n$ via permutations of indices. Specifically, for $x = (x_1, \ldots, x_n) \in [0, \infty)^n$, we write $\sigma(x) = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for the result of hitting x with the permutation $\sigma \in S_n$. Similarly, S_n acts on $\mathscr{F}^n$ via permutations, and we write $\sigma(\mathcal{B}) = \mathcal{B}_{\sigma(1)} \times \cdots \times \mathcal{B}_{\sigma(n)}$ for $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n \in \mathscr{F}^n$. We define the *orbit*[1] $\mathrm{Orb}(\mathcal{B})$ of $\mathcal{B}$ to be the set

$$\mathrm{Orb}(\mathcal{B}) = \bigcup_{\sigma \in S_n} \sigma(\mathcal{B}).$$

[1] Strictly speaking, this is the union of the orbit, where the orbit is usually defined as $\{\sigma(\mathcal{B}) \mid \sigma \in S_n\}$.

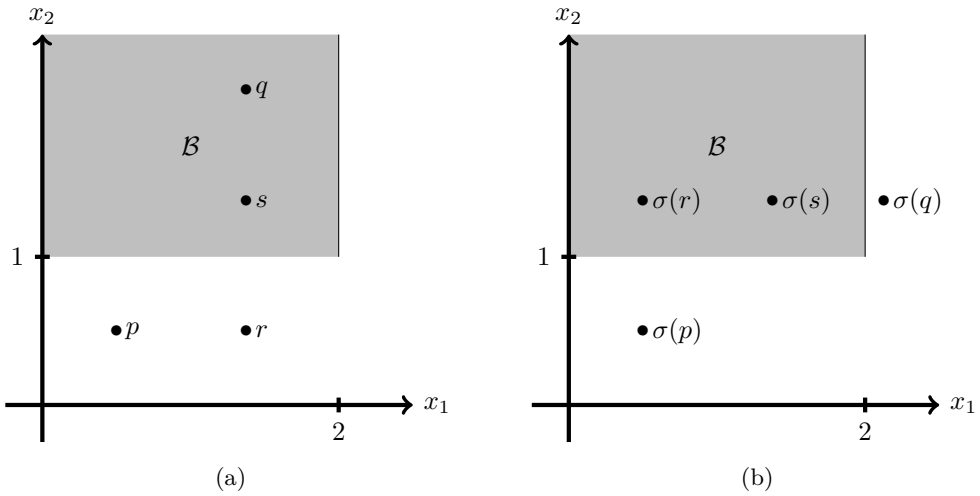

Figure 1: A two-dimensional example of calculating permutation numbers.

Next, we define the *permutation number* $w(x;\mathcal{B})$ *of* x *with respect to* $\mathcal{B}$ as

$$w(x;\mathcal{B}) = \#\{\sigma \in S_n \mid \sigma(x) \in \mathcal{B}\}.$$

Note that $0 \leq w(x;\mathcal{B}) \leq n!$ for all x, and furthermore that $w(x;\mathcal{B}) = 0$ if and only if $x \notin \mathrm{Orb}(\mathcal{B})$, that is, if and only if $\sigma(x) \notin \mathcal{B}$ for all permutations $\sigma \in S_n$.

Example. *We show how to calculate permutation numbers in a simple two–dimensional example. Consider the set* $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2 = [0,2] \times (1,\infty)$*, drawn in Figure 1a. In order to calculate the permutation numbers of the points* p, q, r, s*, we hit these points with the nontrivial permutation* $\sigma \in S_2$*, as shown in Figure 1b. That is, we reflect them across the diagonal* $x_1 = x_2$*. Now,* $p \notin \mathcal{B}$ *and* $\sigma(p) \notin \mathcal{B}$*, so* $w(p;\mathcal{B}) = 0$*. Next,* $q \in \mathcal{B}$ *and* $\sigma(q) \notin \mathcal{B}$*, so* $w(q;\mathcal{B}) = 1$*. Similarly,* $r \notin \mathcal{B}$ *and* $\sigma(r) \in \mathcal{B}$*, so* $w(r;\mathcal{B}) = 1$*. Finally,* $s \in \mathcal{B}$ *and* $\sigma(s) \in \mathcal{B}$*, so* $w(s;\mathcal{B}) = 2$*. Note in particular that* $w(r;\mathcal{B}) > 0$ *even though* $r \notin \mathcal{B}$*.*

In Section 3, we will derive an algorithm for efficiently computing the permutation numbers $w(x;\mathcal{B})$. For now, we shall show how they can be used to construct an importance sampling algorithm as an alternative to (2.2). Consider the modified estimator

$$\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B}) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{n!} w(x^{(t)};\mathcal{B}), \tag{2.3}$$

where $x^{(t)} \sim \pi(\cdot)$ independently for $t = 1,\ldots,T$. This is essentially an importance sampling estimator with proposal distribution $x \sim \pi(\cdot)$ and weights $W(x) = w(x;\mathcal{B})/n!$.

Proposition 2.1. $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$ *is an unbiased and consistent estimator for the marginal likelihood* $\mathscr{P}(\mathcal{B})$*.*

Proof. We have that $w(x;\mathcal{B}) = \#\{\sigma \in S_n \mid \sigma(x) \in \mathcal{B}\} = \#\{\sigma \in S_n \mid x \in \sigma(\mathcal{B})\}$, so that taking expectations, we get

$$\mathbb{E}\left[\frac{1}{n!}w(x;\mathcal{B})\right] = \frac{1}{n!}\int_{\mathrm{Orb}(\mathcal{B})} w(x;\mathcal{B})\,\mathrm{d}\pi(x) = \frac{1}{n!}\sum_{\sigma\in S_n}\int_{\sigma(\mathcal{B})}\mathrm{d}\pi(x)$$
$$= \frac{1}{n!}\sum_{\sigma\in S_n}\mathscr{P}(\sigma(\mathcal{B})) = \mathscr{P}(\mathcal{B}),$$

where we have used exchangeability for the final equality. This proves that the estimator is unbiased. Applying the law of large numbers to (2.3) establishes consistency. □

The benefit of calculating $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$ rather than the naive estimate $\hat{\mathscr{P}}_T(\mathcal{B})$ is that we only require x to land in $\mathrm{Orb}(\mathcal{B})$, which is a much larger set than $\mathcal{B}$. In practice, this means that we get way more contributing samples when calculating (2.3) rather than (2.2).

Given x and $\mathcal{B}$, it is not, a priori, easy to determine whether $x \in \mathrm{Orb}(\mathcal{B})$. However, by considering the order statistics of x, we can establish an easily verifiable criterion.

Definition 2.1. *Let $x \in [0,\infty)^n$ be fixed and let $\sigma \in S_n$ be any n-permutation such that $x_{\sigma(1)} \leq \cdots \leq x_{\sigma(n)}$. That is, $\sigma(x)$ are the order statistics of x. We say that x is* $\mathcal{B}$-admissible *if $\sigma(x) \in \mathcal{B}$.*

Proposition 2.2. *Let $x \in [0,\infty)^n$. Then $w(x;\mathcal{B}) > 0$ if and only if x is $\mathcal{B}$-admissible.*

Proof. See the Supplementary Material (Christensen, 2023). □

Proposition 2.2 simplifies computations significantly. For instance, when calculating $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$, we now have a simple criterion for checking whether $x \in \mathrm{Orb}(\mathcal{B})$. Namely, we check whether $\sigma(x) \in \mathcal{B}$, where σ is as in Definition 2.1.

Example. *We illustrate the volume of* $\mathrm{Orb}(\mathcal{B})$ *via a simple simulation study. For $n = 1,\ldots,300$, we let $0 < r_1 < \ldots r_n < 1$ be uniformly spaced and simulated $u_1,\ldots,u_n \sim \mathrm{Uniform}[0,1]$ independently. As in Section 2.1, we write $n_0 = \#\{i \mid u_i \leq r_i\}$ and let $t_1 < \cdots < t_{n_0}$ be those r_i satisfying $u_i \leq r_i$. Similarly, we let $t_{n_0+1} < \cdots < t_n$ be those r_i satisfying $u_i > r_i$. This defines the set $\mathcal{B}$.*

Two experiments were conducted, one where $x_1,\ldots,x_n \sim \mathrm{Uniform}[0,1]$ and another where $x_1,\ldots,x_n \sim \mathrm{Beta}(2,2)$, independently. The probability that $x \in \mathcal{B}$ is given by

$$\mathbb{P}(x \in \mathcal{B}) = \prod_{i=1}^{n_0} F(t_i) \times \prod_{i=n_0+1}^{n} \{1 - F(t_i)\},$$

where F denotes the cdf of the Uniform and Beta distribution in the first and second experiment, respectively. In either case, this probability decays exponentially and gets vanishingly small as n gets large. In order to compare $\mathbb{P}(x \in \mathcal{B})$ with the probability $\mathbb{P}(x \in \mathrm{Orb}(\mathcal{B}))$, we repeatedly simulated copies of x a total of 1000 times and counted

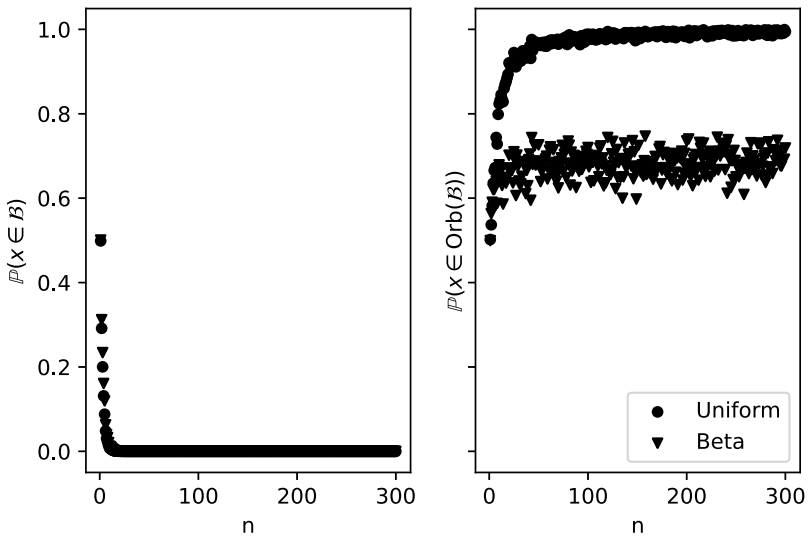

Figure 2: Empirical estimates of $\mathbb{P}(x \in \mathcal{B})$ and $\mathbb{P}(x \in \mathrm{Orb}(\mathcal{B}))$ for $x_1, \ldots, x_n \sim \mathrm{Uniform}[0,1]$ independently and $x_1, \ldots, x_n \sim \mathrm{Beta}(2,2)$ independently.

how many times we observed $x \in \mathcal{B}$ and how many times we observed $x \in \mathrm{Orb}(\mathcal{B})$ (using Proposition 2.2). We did this 100 *times for each n and averaged the results, which are plotted in Figure 2. We note that in both experiments, we get a much higher acceptance proportion when working with* $\mathrm{Orb}(\mathcal{B})$*. Also, this proportion does not seem to decrease with n. This makes sense intuitively, since the number of permutations increases exponentially with n.*

Remark. *The quality of the estimate $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$ is diagnosed by calculating the* effective sample size *(*ESS*), given by*

$$\mathrm{ESS} = \frac{\left(\sum_{t=1}^T w(x^{(t)};\mathcal{B})/n!\right)^2}{\sum_{t=1}^T \left(w(x^{(t)};\mathcal{B})/n!\right)^2} = \frac{\left(\sum_{t=1}^T w(x^{(t)};\mathcal{B})\right)^2}{\sum_{t=1}^T w(x^{(t)};\mathcal{B})^2}.$$

In (2.3), it may be useful not to treat T as a fixed sample size, but rather to keep adding terms until the ESS *reaches a fixed, predetermined value.*

In principle, $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$ can also be computed in cases where all observations are only right or left censored. That is, in cases where $n_0 = 0$ or $n_0 = n$. However, for such problems, we experience that the ESS *increases too slowly. By studying the geometry of the situation, we can gain some insight into why this is the case. The* ESS *will be small if only a few weights dominate. Now, if all the $\mathcal{B}_i$ extend to the right, say, then samples $x^{(t)}$ with all $x_i^{(t)}$ far to the right will yield large permutation numbers $w(x^{(t)};\mathcal{B})$. Indeed, the maximum value $w(x^{(t)};\mathcal{B}) = n!$ is attainable if the $x_i^{(t)}$ are sufficiently far to the right. However, when both left and right censored observations are present, the $\mathcal{B}_i$ are*

no longer nested, and so the weights tend to be more uniformly spread out, making a larger effective sample size obtainable.

Remark. *Some of the thresholds t_i may be equal. Let $r_1 \leq \cdots \leq r_l$ be the unique values of the set $\{t_1, \ldots, t_n\}$ and for each $j = 1, \ldots, l$, let $a_j = \#\{i \mid t_i = r_j\}$ and $b_j = \#\{i \mid t_i = r_j, i \leq n_0\}$. That is, a_j is the number of trials conducted at input r_j and b_j is the number of successes. The observations are now a sequence of binomial variables, and so the marginal likelihood takes the form $\mathbb{P}(x \in \mathcal{B}) \times \prod_{j=1}^{l} \binom{a_j}{b_j}$. Hence, if we have repeated trials, we may simply redefine $w(x; \mathcal{B})$ to be $w(x; \mathcal{B}) \times \prod_{j=1}^{l} \binom{a_j}{b_j}$ and carry out our analysis as normal. We continue without explicitly multiplying by this factor in our notation, but it should be kept in mind that the permutation numbers are multiplied by this factor if the data include repeated trials.*

Remark. *In (2.3), the samples come from the prior. Although the simulation study above indicates that this naive approach works sufficiently well, other choices of proposal distribution may be more efficient and increase performance. Examples of methods that might do so include sequential Monte Carlo (see Cappé et al. (2007) for a review), defensive mixture proposal distributions (Hesterberg, 1995) and population Monte Carlo (Cappé et al., 2004). For the sake of simplicity, in the present paper, we shall only consider the case where the samples are drawn from the prior.*

We conclude this section by showing that our estimate $\hat{\mathscr{P}}_T^{\mathrm{IS}}(\mathcal{B})$ yields a smaller variance than the naive estimate $\hat{\mathscr{P}}_T(\mathcal{B})$.

Proposition 2.3. *We have that*

$$\operatorname{Var}\left(\frac{1}{n!} w(x; \mathcal{B})\right) = \operatorname{Var}\left(\mathbb{1}_{\mathcal{B}}(x)\right) + \mathscr{P}(\mathcal{B}) - \frac{1}{n!} \sum_{\sigma \in S_n} \mathscr{P}(\sigma(\mathcal{B}) \cup \mathcal{B}).$$

In particular,

$$\operatorname{Var}\left(\frac{1}{n!} w(x; \mathcal{B})\right) \leq \operatorname{Var}\left(\mathbb{1}_{\mathcal{B}}(x)\right).$$

Proof. Using the same reasoning as in Proposition 2.1, we have that

$$w(x; \mathcal{B})^2 = \#\{(\sigma, \tau) \in S_n^2 \mid \sigma(x), \tau(x) \in \mathcal{B}\} = \#\{(\sigma, \tau) \in S_n^2 \mid x \in \sigma(\mathcal{B}) \cap \tau(\mathcal{B})\},$$

and so

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n!} w(x; \mathcal{B})\right)^2\right] &= \frac{1}{(n!)^2} \int_{\mathrm{Orb}(\mathcal{B})} w(x; \mathcal{B})^2 \, \mathrm{d}\pi(x) = \frac{1}{(n!)^2} \sum_{\sigma, \tau \in S_n} \int_{\sigma(\mathcal{B}) \cap \tau(\mathcal{B})} \mathrm{d}\pi(x) \\
&= \frac{1}{n!} \sum_{\sigma \in S_n} \mathscr{P}(\sigma(\mathcal{B}) \cap \mathcal{B}) = \frac{1}{n!} \sum_{\sigma \in S_n} \{\mathscr{P}(\sigma(\mathcal{B})) + \mathscr{P}(\mathcal{B}) - \mathscr{P}(\sigma(\mathcal{B}) \cup \mathcal{B})\} \\
&= 2\mathscr{P}(\mathcal{B}) - \frac{1}{n!} \sum_{\sigma \in S_n} \mathscr{P}(\sigma(\mathcal{B}) \cup \mathcal{B}).
\end{aligned}$$

Subtracting $\mathbb{E}[(1/n!) w(x; \mathcal{B})]^2$ from both sides yields the first result. For the second, note that $\mathscr{P}(\sigma(\mathcal{B}) \cup \mathcal{B}) \geq \mathscr{P}(\mathcal{B})$ for all $\sigma \in S_n$. □

2.3 Posterior inference

We now extend the results from the previous section to a general importance sampling algorithm targeting the posterior distribution of P given that $x \in \mathcal{B}$. Let $\theta = \theta(P) \in \mathbb{R}$ be some quantity related to P, such as a specific cdf value $F(t) = P([0,t])$ or a quantile $F^{-1}(q)$. Then consider the estimator

$$\hat{\theta}_T^{\mathrm{IS}} = \frac{\sum_{t=1}^T \theta^{(t)} w(x^{(t)}; \mathcal{B})}{\sum_{t=1}^T w(x^{(t)}; \mathcal{B})}, \tag{2.4}$$

where $\theta^{(t)} = \theta(P^{(t)})$, $x^{(t)} \sim P^{(t)}$ and $P^{(t)} \sim \pi(\cdot)$, independently for $t = 1, \ldots, T$.

Proposition 2.4. *The statistic $\hat{\theta}_T^{\mathrm{IS}}$ is a consistent estimator for the posterior mean $\mathbb{E}[\theta \mid x \in \mathcal{B}]$.*

Proof. The expression (2.4) is precisely the self-normalised importance sampling estimator targeting

$$\frac{\int_{[0,\infty)^n \times \mathbb{R}} \theta w(x;\mathcal{B}) \,\mathrm{d}\pi(x,\theta)}{\int_{[0,\infty)^n \times \mathbb{R}} w(x;\mathcal{B}) \,\mathrm{d}\pi(x,\theta)} = \frac{\int_{\mathrm{Orb}(\mathcal{B}) \times \mathbb{R}} \theta w(x;\mathcal{B}) \,\mathrm{d}\pi(x,\theta)}{\int_{\mathrm{Orb}(\mathcal{B}) \times \mathbb{R}} w(x;\mathcal{B}) \,\mathrm{d}\pi(x,\theta)} = \frac{\int_{\mathcal{B} \times \mathbb{R}} \theta \,\mathrm{d}\pi(x,\theta)}{\int_{\mathcal{B} \times \mathbb{R}} \mathrm{d}\pi(x,\theta)}$$
$$= \int_{\mathbb{R}} \theta \,\mathrm{d}\pi(\theta \mid x \in \mathcal{B}) = \mathbb{E}\left[\theta \mid x \in \mathcal{B}\right],$$

as required. □

3 Permutation numbers

We now outline how to calculate the permutation numbers $w(x;\mathcal{B})$. For the remainder of this section, assume that x is $\mathcal{B}$-admissible, so we know that $w(x;\mathcal{B}) > 0$. The first step of the derivation is to express the permutation number $w(x;\mathcal{B})$ as the *permanent* of a $(0,1)$-matrix.

Definition 3.1. *Let $A = (a_{ij})$ be an $m \times n$ matrix where $m \leq n$. Let $S_{n,m}$ denote the set of all m-permutations of the set $\{1, \ldots, n\}$. The* permanent $\mathrm{perm}(A)$ *of A is defined by*

$$\mathrm{perm}(A) = \sum_{\tau \in S_{n,m}} \prod_{i=1}^m a_{i,\tau(i)} \tag{3.1}$$

Note that the permanent is defined for rectangular matrices, not just square ones. In order to express the permutation number $w(x;\mathcal{B})$ as the permanent of a matrix, we make the following definition. Given $x \in [0,\infty)^n$, the *matching matrix* $A = (a_{ij})$ of x is the $n \times n$ $(0,1)$-matrix defined by

$$a_{ij} = \begin{cases} 1 & \text{if } x_i \in \mathcal{B}_j \\ 0 & \text{if } x_i \notin \mathcal{B}_j. \end{cases}$$

Lemma 3.1. *Let A be the matching matrix of x. Then*

$$w(x;\mathcal{B}) = \operatorname{perm}(A).$$

The key ingredient in the proof is to count the number of matchings in a bipartite graph. If the reader is unfamiliar with these notions, we recommend reading Bollobás (1979, Chapter 3).

Proof of Lemma 3.1. Let $\mathcal{G} = (V, E)$ be the bipartite graph with vertex set $V = \{x_1, \ldots, x_n\} \cup \{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$ and edges $E = \{(x_i, \mathcal{B}_j) \in V^2 \mid x_i \in \mathcal{B}_j\}$. Then the permutation number $w(x;\mathcal{B})$ is equal to the number of bijections $f : \{x_1, \ldots, x_n\} \to \{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$ such that $x_i \in f(x_i)$. That is, the number of perfect matchings in $\mathcal{G}$. But this is precisely the permanent of the biadjacency matrix of $\mathcal{G}$. That is, the permanent of the matrix A. □

Permanents are notoriously difficult to compute. Unlike the closely related determinant function (which is obtained by multiplying each term in (3.1) by $\operatorname{sign}(\tau)$), the permanent function is not multiplicative, and thus we cannot employ Gaussian elimination to compute permanents in polynomial time. In general, computing permanents of $(0,1)$-matrices is known to be a #P-complete problem (Valiant, 1979). The fastest general formula known for $(0,1)$-matrices is that by Ryser (1963), which requires $O(2^{n-1}n)$ operations for an $n \times n$ matrix. More recently, Huh (2022) has presented an efficient quantum algorithm for estimating permanents.

A $(0,1)$-matrix A is said to be *convertible* if there exists a matrix A' obtained by changing the signs of some of the entries in A such that $\operatorname{perm}(A) = \det(A')$. This means that $\operatorname{perm}(A)$ can be computed in polynomial time. Little (1975) provided a classification of all convertible matrices. Namely, a matrix A is convertible if and only if it can be realised as the biadjacency matrix of a bipartite graph $\mathcal{G}$ which does not contain an even subdivision J of the complete bipartite graph $K_{3,3}$ such that $\mathcal{G} - V(J)$ has a perfect matching. It is easy to construct examples of matching matrices which violate this criterion. For example, let $x = (1,1,1,3)$ and $\mathcal{B} = [0,2]^3 \times (2,\infty)$. If $\mathcal{G}$ denotes the graph constructed in the proof of Lemma 3.1, then $\mathcal{G} \cong K_{3,3} + K_{1,1}$, where the plus denotes disjoint union. Thus, $\mathcal{G}$ contains $K_{3,3}$ as a subgraph and still contains a perfect matching once this subgraph has been removed. Consequently, the corresponding matching matrix is not convertible.

The above example demonstrates that we cannot use convertibility to compute the permanents of matching matrices. However, we will show that matching matrices belong to a larger class of matrices, which we shall call *block rectangular*. We will then prove that the permanent of a block rectangular matrix can be calculated in polynomial time.

3.1 Block rectangular matrices

We begin with the definition of block rectangular matrices.

Definition 3.2. *Let k, m, n be natural numbers, where $k, m \leq n$. Let $\alpha \in \mathbb{Z}_{>0}^{k}$ and $\beta, \gamma \in \mathbb{Z}_{\geq 0}^{k-1}$ be three integer-valued vectors such that*

$$\sum_{r=1}^{k} \alpha_r = n, \quad \sum_{r=1}^{k-1} \beta_r \leq m, \quad \sum_{r=1}^{k-1} \gamma_r \leq m \tag{3.2}$$

and

$$\sum_{r=1}^{t} \beta_r \leq \sum_{r=1}^{t} \alpha_r, \quad \sum_{r=t}^{k-1} \gamma_r \leq \sum_{r=t+1}^{k} \alpha_r \tag{3.3}$$

for all $t = 1, \ldots, k$. The block rectangular matrix $M = (m_{ij})$ associated to α, β, γ, m *is the $m \times n$ $(0,1)$-matrix such that $m_{ij} = 1$ if and only if there exists $t \in \{1, \ldots, k\}$ such that*

$$\sum_{r=1}^{t-1} \beta_r < i \leq m - \sum_{r=t}^{k-1} \gamma_r, \quad \sum_{s=1}^{t-1} \alpha_s < j \leq \sum_{s=1}^{t} \alpha_s. \tag{3.4}$$

We say that a matrix M is block rectangular *if there exist α, β, γ, m such that M is the block rectangular matrix associated to α, β, γ, m. Note that we suppress k and n in the definition as these are implicitly defined through α.*

Example. *We consider three examples of constructing a block rectangular matrix from its associated parameters, as well as an example of a matrix which is not block rectangular.*

(a) Let $\alpha = (1,3,1,1,1), \beta = (0,1,2,1), \gamma = (1,1,1,0)$ and $m = 7$. This choice of α, β, γ satisfies conditions (3.2) and (3.3). Constructing the matrix from these parameters is done as follows. Firstly, $n = \sum_{i=1}^{5} \alpha_i = 7$, which, together with $m = 7$, determines the dimensions of the matrix. We consider each rectangular block separately. Letting $t = 1$ in (3.4), we obtain that $0 < i \leq 4$ and $0 < j \leq 1$. Repeating this step for $t = 2, \ldots, T$ establishes the dimensions of all the rectangular blocks, resulting in the matrix shown in Figure 3a.

(b) A block rectangular matrix need not be square. Let $\alpha = (3,2,2,1), \beta = (1,0,2), \gamma = (1,1,1)$, $m = 6$. Then α, β, γ satisfy conditions (3.2) and (3.3), but $n = \sum_{r=1}^{4} \alpha_r = 8 > 6$, so the resulting matrix, shown in Figure 3b, is not square.

(c) In the two examples above, we have $m = \sum_{r=1}^{k-1} \beta_r + \sum_{r=1}^{k-1} \gamma_r$. This need not be the case. Indeed, let $\alpha = (2,3,1,2), \beta = (1,0,2), \gamma = (0,1,1), m = 7$. Then α, β, γ satisfy condition (3.2) and (3.3), but $\sum_{r=1}^{3} \beta_r + \sum_{r=1}^{3} \gamma_r = 5 < 7$. This means that the matrix contains $7 - 5 = 2$ rows of ones, as can be seen in Figure 3c. This example illustrates why we need to include m as a separate parameter in order to describe the matrix uniquely.

(d) The matrix in Figure 3d is not block rectangular. Indeed, if it were, then it would be associated with the parameters $\alpha = (2,2,2,2), \beta = (2,3,1), \gamma = (3,1,0), m = 8$. But then we have that $\alpha_1 + \alpha_2 = 4 < 5 = \beta_1 + \beta_2$, which violates condition (3.3).

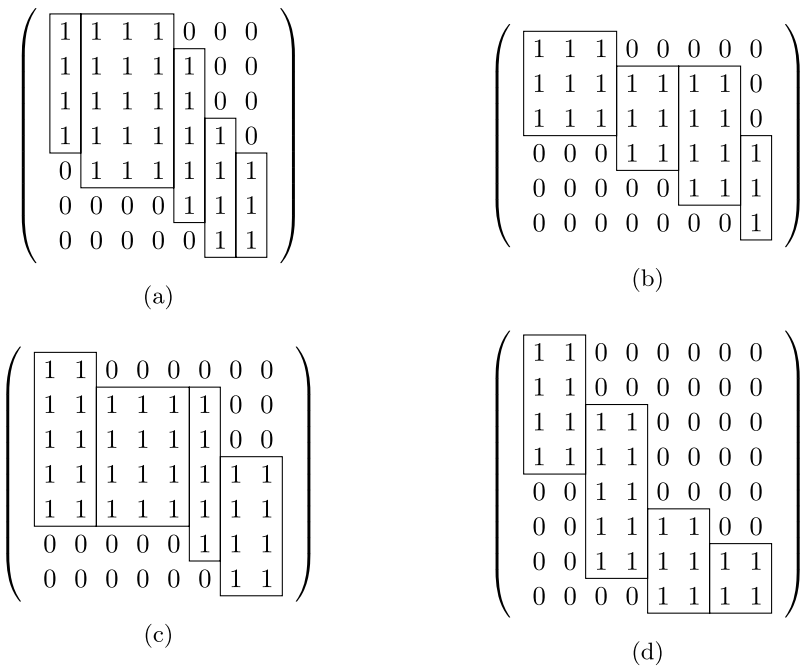

Figure 3: The matrices from the example above with the contours of the rectangular blocks highlighted.

As the definition stands, it is possible for a block rectangular matrix to contain a row of zeros. Indeed, consider for instance the matrix A parametrised by $\alpha = (1, 1, 1)$, $\beta = (1, 0)$, $\gamma = (0, 1)$, $m = 1$. Then A is the 1×3 zero matrix since no i, j, t will satisfy condition (3.4). We say that a block rectangular matrix A is *complete* if it does not contain a row of zeros.

Note that multiple parametrisations will give rise to the same block rectangular matrix. Indeed, we can always subdivide a rectangular block into more rectangular blocks of equal heights. In our notation, this would mean that for some $r \in \{1, \ldots, k-1\}$, we have $\beta_r = \gamma_r = 0$. However, by insisting that $k = \dim(\alpha)$ should always be minimal, we obtain a unique parametrisation for every block rectangular matrix. We refer to this as the *minimal parametrisation*.

There are examples of block rectangular matrices which cannot be realised as matching matrices. For example, let $\alpha = (1, 1, 1), \beta = (1, 1), \gamma = (1, 1), m = 3$. Then A is the 3×3 identity matrix, which is not a matching matrix. However, we have the following converse result.

Proposition 3.1. *Let x be $\mathcal{B}$-admissible and let A be the matching matrix of x. Then, after permuting its columns if necessary, A is a complete block rectangular matrix.*

Proof. See the Supplementary Material (Christensen, 2023). □

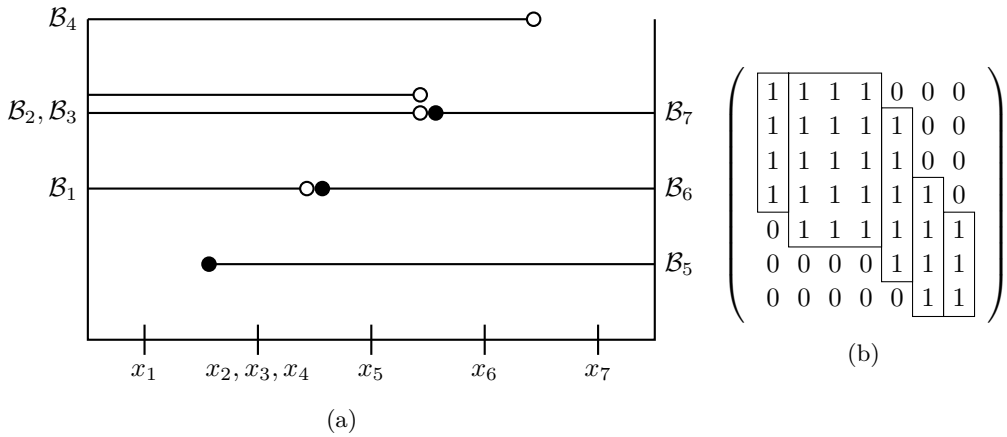

Figure 4: Example of a matching of x and $\mathcal{B}$. (a) Pictorial representation of x and $\mathcal{B}$. The x_i are marked along the horizontal axis. The $\mathcal{B}_i$ are drawn along the same axis, and stacked vertically for visual clarity. (b) The resulting matching matrix, with the contours of the rectangular blocks highlighted.

Example. *Suppose we have four left-censored observations, with* $s_1 = 2, s_2 = s_3 = 3, s_4 = 4$, *and three right-censored observations, with* $t_5 = 1, t_6 = 2, t_7 = 3$. *Next, let* $x = (x_1, \ldots, x_7) = (0.5, 1.5, 1.5, 1.5, 2.5, 3.5, 4.5)$. *See Figure 4a. To construct the matching matrix* A, *we first consider* x_1. *We observe that* $x_1 \in \mathcal{B}_1, \ldots, \mathcal{B}_4$, *but not* $\mathcal{B}_5, \mathcal{B}_6$ *or* $\mathcal{B}_7$. *Hence the first column of* A *consists of four ones followed by three zeros. We continue this way for all the* x_i, *which results in the matrix in Figure 4b. Note that* A *is block rectangular. In fact, we recognise it as the matrix from Figure 3a.*

We are now ready to state the main result of this section.

Theorem 3.1. *Let* A *be an* $m \times n$ *complete block rectangular matrix. Then there exists an implementable algorithm for computing* $\operatorname{perm}(A)$, *whose computational complexity grows polynomially with* n.

Proof. See the Supplementary Material (Christensen, 2023). □

Thus, with Theorem 3.1, we are able to compute the permutation numbers needed for the estimators (2.3) and (2.3). Code for computing permanents can be found in the publicly available GitHub repository. For a reasonably large value of n, say $n = 200$, this new approach is able to compute tens of thousands of permanents of $n \times n$ block rectangular matrices within a few hours. This is in contrast with more general approaches, such as the aforementioned Ryser's formula, which would not be able to handle even a single matrix of this dimension. In the following section, we illustrate the efficiency of the new approach with experiments.

4 Experiments

We now look at two simulation studies and a real data example in order to illustrate the performance of the new estimator. The first simulation study is a tractable bioassay problem involving a Dirichlet process model. Such models were first studied by Antoniak (1974). This problem is included to verify that the new algorithm agrees with existing methods. More precisely, we will compare it with the successive substitution sampling (SSS) algorithm introduced by Doss (1994). In the second simulation study, we employ the new algorithm to fit a quantile pyramids model, introduced by Hjort and Walker (2009), to binary response data. This is an example of a process from which data samples may only be simulated approximately, but where our new algorithm nevertheless works exactly. Finally, we employ a Pitman-Yor multinomial process model (Lijoi et al., 2020) to real seroprevalence data, originally studied by Keiding et al. (1996). All code was run on a computer running Windows 11 Pro with an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz and 16GB DDR4 RAM.

4.1 Simulations

A tractable Dirichlet process problem

For the first simulation study, we used the data reported in Table 1. This data was generated by simulating $n = 100$ points $u_1, \dots, u_n \sim \frac{1}{3}\mathcal{N}(-2, 0.7^2) + \frac{2}{3}\mathcal{N}(1, 0.7^2)$, a mixture of two normal distributions, and observing whether these points were below or above the respective thresholds in Table 1. Thus, for example, since the number of trials at the threshold -3 was 10, the number of successes there refers to the quantity $\#\{i \in \{1, \dots, 10\} \mid u_i \leq -3\}$. In the prior we let P be distributed according to a Dirichlet process with concentration parameter $\alpha = 1$ and base measure $\mathcal{N}(0, 1)$. The prior mean and individual realisations of the prior process are plotted in Figure 5. Using (2.3), the log marginal likelihood was calculated to be -12.861. Note that we have repeated thresholds in the data set. Repeating this calculation ten times yielded a standard deviation of 0.0137, showing that the estimate is stable. On average, it took $T = 438,606$ iterations to yield an ESS of 2000. Out of these, an average number of $411,837$ yielded a vanishing permanent which could immediately be discarded. The average computation time for calculating the permanents was 6 minutes and 33 seconds. The slowest run took 6 minutes and 57 seconds.

Due to the posterior tractability of the Dirichlet process, the posterior process can be simulated directly, for instance via the SSS method, introduced by Doss (1994). Table 2 show how the new importance sampling algorithm compares with the SSS method by comparing the values of the posterior mean at various quantiles. As we can see, the two methods are in agreement. Figure 5 shows plots of the posterior mean, calculated using the two different methods, along with individual realisations of the posterior process. This plot further verifies the agreement of the two approaches, illustrating that the new algorithm indeed converges to the posterior process.

Threshold	Number of successes	Number of trials
−3	0	10
−2.33	0	10
−1.67	2	10
−1	1	10
−0.33	4	10
0.33	6	10
1	9	10
1.67	10	10
2.33	10	10
3	10	10

Table 1: The thresholds, number of successes and number of trials for the Dirichlet process simulations.

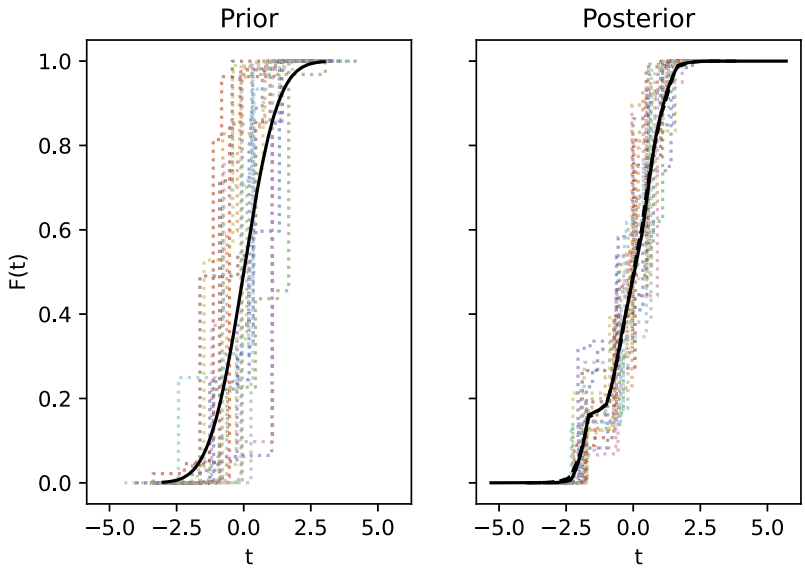

Figure 5: Prior and posterior estimates of the Dirichlet process model given $n = 100$ binary response data points. On the left, the solid curve is the prior mean and the dotted curves are realisations of the prior process. On the right, the solid and dashed curves are the posterior mean as obtained via permutation counting or successive substitution sampling, respectively, and the dotted curves are realisations of the posterior process.

Quantile pyramids simulations

Our next simulation study is a problem in which we wish to fit a quantile pyramids model, given binary response data. Such models were first studied by Hjort and Walker (2009), and provide an appealing alternative to Pólya trees, since they avoid the specification of a partition of the sample space. More specifically, we model P as a Beta

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
SSS	−1.851	−0.949	−0.572	−0.283	0.015	0.305	0.525	0.784	1.176
PC	−1.842	−0.950	−0.576	−0.290	0.040	0.349	0.558	0.789	1.130

Table 2: Posterior estimates of $\mathbb{E}[F^{-1}(q) \mid x \in \mathcal{B}]$ for different values of q, calculated via successive substitution sampling (SSS) and the new importance sampling algorithm based on permutation counting (PC).

quantile pyramid with parameters $(\frac{1}{2}a_m, \frac{1}{2}a_m)$, where $a_m = cm^3$ and $c = 2.5$. This is the same model as that considered in the simulation study by Hjort and Walker (2009). Given exact observations, the posterior process is intractable and so MCMC-based inference is required. As a result, there is, a priori, no straightforward way of simulating the posterior process given censored data. However, our new importance sampling algorithm will circumvent this issue.

As can be seen from (2.3), we require that it is possible to generate samples $x_i \sim P$. For the quantile pyramids model, this can only be done approximately. Indeed, the process is realised by simulating the quantiles $F^{-1}(j/2^K)$, $j = 1, \ldots, 2^K - 1$, for some finite number K. Increasing the value of K increases the precision of the realisation. If K were allowed to be infinite, then F would, by absolute continuity, be uniquely determined. Thus, for $i = 1, \ldots, n$, we could sample $x_i \sim P$ by first sampling uniform variables $u_i \sim \text{Unif}[0, 1]$ and then letting $x_i = F^{-1}(u_i)$. In practice, K is a finite number, and so this approach cannot determine the exact value of the x_i. However, by finding the numbers j_i such that $u_i \in (j_i/2^K, (j_i + 1)/2^K]$, we know that $x_i \in (F^{-1}(j_i/2^K), F^{-1}((j_i + 1)/2^K)]$. Thus, by increasing the value of K if necessary, we can make these intervals arbitrarily fine and thus know for certain whether $x_i \in \mathcal{B}_j$ for $j = 1, \ldots, n$. That is, we can sample the value of $w(x; \mathcal{B})$ *exactly*, even though x was only simulated approximately. As a result, the convergence results for the importance sampling algorithm still hold exactly.

The synthetic data were simulated as follows. For $n = 100$, we let $0 < r_1 < \cdots < r_n < 1$ be equally spaced points on the unit interval $[0, 1]$, and simulated $u_1, \ldots, u_n \sim \text{Beta}(1/2, 1)$ independently. Thus, the true underlying distribution is also the same as in the original simulation study undertaken by Hjort and Walker (2009). Writing $n_0 = \#\{i \mid u_i \leq r_i\}$, we let $t_1 < \cdots < t_{n_0}$ be those r_i such that $u_i \leq r_i$. Similarly, we let $t_{n_0+1} < \cdots < t_n$ be those r_i such that $u_i > r_i$. As in Section 2.1, we then let $\mathcal{B} = \mathcal{B}_1, \ldots, \mathcal{B}_n$, where $\mathcal{B}_i = [0, t_i]$ for $i = 1, \ldots, n_0$ and $\mathcal{B}_i = (t_i, 1]$ for $i = n_0 + 1, \ldots, n$.

Using (2.3), we calculated the log marginal likelihood to be −53.698. Performing this calculation 10 times yielded a standard deviation of 0.013, showing that the estimate is stable. On average, it took $T = 29,965$ iterations to obtain an effective sample size of 2000. Out of these, an average number of 8227 yielded vanishing permanents which could be discarded immediately. The average computation time for calculating the permanents was 6 minutes and 25 seconds. The slowest run took 8 minutes and 8 seconds. In Figure 6, we plot prior and posterior cdfs given the simulated data. We see that the posterior estimate has moved closer to the true cdf, and that the posterior variance is smaller than that of the prior. Indeed, Kolmogorov-Smirnov distances from the prior and the posterior means to the ground truth are 0.25 and 0.094, respectively.

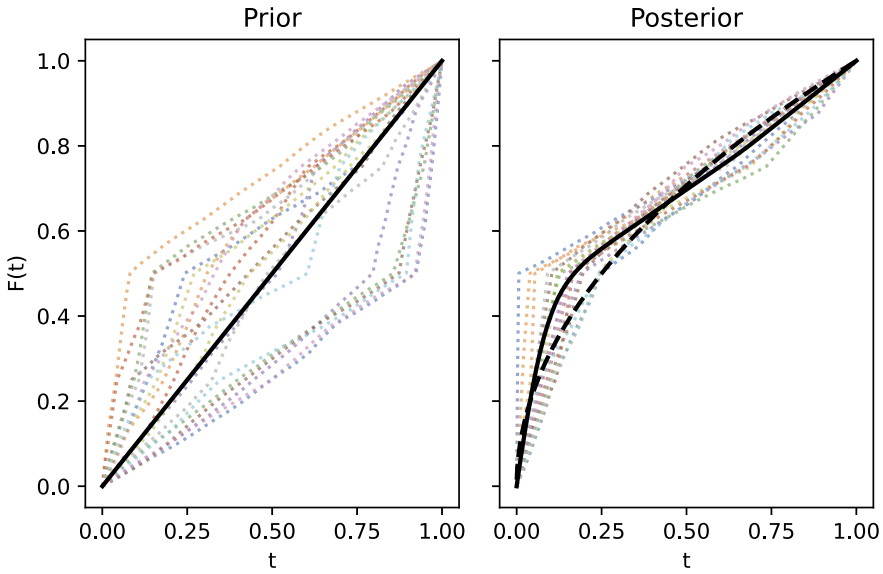

Figure 6: Prior and posterior estimates of the Beta quantile pyramid model given $n = 100$ binary response data points. On the left, the solid curve is the prior mean and the dotted curves are realisations of the prior process. On the right, the solid curve is the posterior mean, the dotted curves are realisations of the posterior process and the dashed curve is the true Beta$(1/2, 1)$ cdf.

4.2 Real current status data

We also applied the new importance sampling algorithm to a real rubella seroprevalence data set, originally studied by Keiding et al. (1996), provided by the Institute of Virology, Vienna. In this data set, the immunisation status of $n = 230$ Austrian males older than three months was tested during the period 1–25 March 1988. On a log scale, the data were scaled linearly so that the standard probit model $\mathbb{P}(y = 1) = \Phi(x)$ gave the best probit model fit. To model the time to infection, we used a Pitman-Yor multinomial (PYM) process (Lijoi et al., 2020) as prior for P. For the sake of simplicity, we used the realised probability distributions of the process to model the survival distribution directly, rather than imposing a PYM mixture model. Lijoi et al. (2020) showed that given exact observations, posterior simulation of PYM processes are possible without the use of MCMC methods, via the empirical marginalisation of a latent variable. Unfortunately, this algorithm does not apply to censored data. Although it would theoretically be possible to apply the SSS method (Doss, 1994) or similar algorithms to the PYM process model given censored data, each iteration of the sampling algorithm would require the aforementioned marginalisation. As a result, this approach would be computationally expensive and of questionable accuracy. On the other hand, since it is straightforward to generate samples from the PYM process model (Ridout, 2009; Lijoi et al., 2020), the new importance sampling algorithm can be directly applied

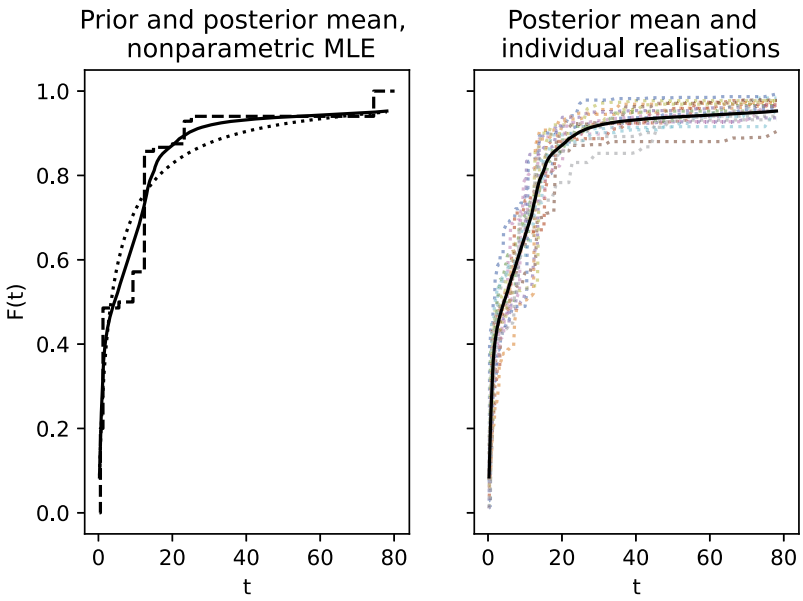

Figure 7: Posterior estimates of the Pitman-Yor multinomial process model given the Rubella data set ($n = 230$). On the left, the dotted curve is the prior mean, the solid curve is the posterior mean and the dashed curve is the nonparametric MLE. On the right, the solid curve is the posterior mean and the dotted curves are realisations of the posterior process.

for posterior inference. In the notation of Lijoi et al. (2020), we set the hyperparameters $\alpha = 2.0, \sigma = 0.7, H = 100$, along with the standard normal distribution $\mathcal{N}(0, 1)$ as base measure. These choices serve only as an illustration, and with more work, one could optimise this choice further, or impose hyperparameter priors in order to achieve a fully Bayesian approach. The log marginal likelihood was calculated to be -92.864. Reaching an ESS of 2000 required $T = 393,011$ iterations. Out of these, $219,492$ yielded a vanishing permanent and could be discarded. Calculating the permanents took 3 hours, 55 minutes and 27 seconds of computation time. In the left plot of Figure 7, we see the prior mean, the posterior mean and the nonparametric maximum likelihood estimator (MLE) (Ayer et al., 1955) of the cdf F. The nonparametric MLE is a frequentist estimator, analogous to the that by Kaplan and Meier (1958) for right censored data, as shown by Turnbull (1974). In the right plot, we see individual realisations from the posterior process, along with the posterior mean. The Kolmogorov-Smirnov distance from the prior mean and the posterior mean to the nonparametric MLE are 0.204 and 0.158, respectively, indicating an improved fit. Indeed, we see the same qualitative behaviour in the posterior mean as the penalty term models studied by Keiding et al. (1996). For the PYM process model, possible improvements may be achieved by optimising parameter choices, or indeed by introducing a similar penalty mechanism into the model.

5 Binary classification with multidimensional data

So far, we have assumed that the x_i are all one-dimensional. In this section, we show how the theory developed in the previous sections can be extended to problems where[2] $x_i \in \mathbb{R}^p$ for $p \geq 1$. This will both enable the addition of covariates, as well as provide an inference framework for binary classification models with multidimensional data.

Let $g \sim \pi(\cdot)$ be a (possibly random) function from $\mathbb{R}^p$ to $\mathbb{R}$ and let $F \sim \pi(\cdot)$ be a (possibly random) cdf on $\mathbb{R}$. We model the binary responses y_i as $y_i \mid F, g \sim \text{Bernoulli}(F(g(x_i)))$ independently for $i = 1, \ldots, n$.

Example. *We now consider three key examples which are covered by the above setup.*

- *Let g almost surely be a neural network and let F almost surely be the sigmoid activation function $F(a) = 1/(1+\exp(-a))$. Then the above model is the standard neural network model for binary classification (Bishop, 2006).*
- *Let g be a Gaussian process and let F almost surely be the sigmoid activation function. Then the above model is the standard Gaussian process model for binary classification (Rasmussen and Williams, 2006).*
- *Let g almost surely be a linear function with coefficients β and let F be the cdf of a random probability distribution P. Then the above model is semiparametric, and corresponds to the addition of covariates in the basic model introduced in Section 2.1.*

Again, we first provide an estimator for the marginal likelihood $\pi(y)$ of our n observations $y = (y_1, \ldots, y_n) \in \{0,1\}^n$, which now takes the form

$$\pi(y) = \mathbb{E}\left[\prod_{i=1}^{n} F(g(x_i))^{y_i} \{1 - F(g(x_i))\}^{1-y_i}\right].$$

We need to introduce some notation before we can write down our estimator. Given F, let $z = (z_1, \ldots, z_n) \sim \pi(\cdot \mid F)$ be distributed such that $\mathbb{P}(z_i \leq t \mid F) = F(t)$ independently for all $i = 1, \ldots, n$. Also, given g, write $\mathcal{B}_g = \mathcal{B}_{g,1} \times \cdots \times \mathcal{B}_{g,n}$, where

$$\mathcal{B}_{g,i} = \begin{cases} (-\infty, g(x_i)] & \text{if } y_i = 1, \\ (g(x_i), \infty) & \text{if } y_i = 0. \end{cases}$$

Then,

$$\mathbb{P}(z \in \mathcal{B}_g \mid g, F) = \int_{\mathcal{B}_g} \mathrm{d}\pi(z \mid F) = \prod_{i=1}^{n_0} F(g(x_i))^{y_i} \{1 - F(g(x_i))\}^{1-y_i}. \tag{5.1}$$

[2] In generative models for binary classification, there is also a distribution for the inputs x_i. However, for evaluation of the marginal likelihood and posterior inference of the hyperparameters, we condition on these input values, effectively treating them as constant. Hence, everything in this section also applied to generative models, but we omit conditioning on the value of x_i for the sake of clarity.

Now consider the following estimator.

$$\hat{\pi}_T^{\mathrm{IS}}(y) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n!}w(z^{(t)};\mathcal{B}_{g^{(t)}}), \tag{5.2}$$

where $z^{(t)} \sim \pi(\cdot \mid F^{(t)}), F^{(t)} \sim \pi(\cdot)$ and $g^{(t)} \sim \pi(\cdot)$.

Proposition 5.1. *The statistic $\hat{\pi}_T^{\mathrm{IS}}(y)$ is an unbiased and consistent estimator for the marginal likelihood $\pi(y)$.*

Proof. Using double expectation, we have that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n!}w(z;\mathcal{B}_g)\right] = \frac{1}{n!}\mathbb{E}\left[\mathbb{E}[w(z;\mathcal{B}_g)\mid g,F]\right] &= \frac{1}{n!}\mathbb{E}\left[\int_{\mathrm{Orb}(\mathcal{B}_g)} w(z;\mathcal{B}_g)\pi(\mathrm{d}z\mid F)\right]\\ &= \frac{1}{n!}\mathbb{E}\left[\sum_{\sigma\in S_n}\int_{\sigma(\mathcal{B}_g)}\pi(\mathrm{d}z\mid F)\right] = \mathbb{E}\left[\int_{\mathcal{B}_g}\pi(\mathrm{d}z\mid F)\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{n}F(g(x_i))^{y_i}\{1-F(g(x_i))\}^{1-y_i}\right] = \pi(y),\end{aligned}$$

where the penultimate equality follows from (5.1). This proves that the estimator is unbiased. Applying the law of large numbers to (5.2) establishes consistency. □

As in Section 2.3, it is possible to extend this result to a normalised importance sampling estimator for general posterior inference for g and P.

6 Discussion

We conclude the paper with a few points of discussion which shed light on directions for future work and further improvements.

In this paper, we have only considered binary responses, which for one-dimensional inputs corresponds to left and right censored observations. This is because such data yield block rectangular matching matrices, whose permanents are computable in polynomial time. However, for more complicated observations, such as interval censored data or polychotous responses (as opposed to binary), it is easy to construct examples of matching matrices which are not block rectangular. Hence, in order to apply the methods developed in this paper to such problems, it is necessary to develop an efficient and accurate estimation procedure for the permanents of the corresponding matching matrices. Further work in this direction is encouraged.

We have proved the consistency of the new estimator (2.4), in the sense that as $T \to \infty$, this converges to the posterior mean $\mathbb{E}[\theta \mid x \in \mathcal{B}]$. However, a separate question is whether this posterior mean itself is consistent. For parametric models, this is guaranteed via the Bernstein–von Mises theorem, which asserts consistency of the posterior mean and links Bayesian credibility sets with frequentist confidence intervals.

In contrast, Doss (1985a,b) and Diaconis and Freedman (1986) showed that for nonparametric models, there exist examples of reasonable choices of priors which lead to inconsistent posteriors. Hence, consistency does not automatically apply in Bayesian nonparametrics. Multiple positive consistency results have since been established for specific choices of nonparametric priors (Brunner and Lo, 1996; Ghosal et al., 1999), some of which also allow for censored data (Kim and Lee, 2004; De Blasi et al., 2009; Camerlenghi et al., 2021; Jongbloed et al., 2022). In general, the issue of consistency in Bayesian nonparametrics should be considered only a partially resolved question, especially for problems involving censored data. Further research in this area is needed to answer to which extent the asymptotic theory of the frequentist nonparametric MLE (Ayer et al., 1955; Groeneboom and Jongbloed, 2014) transfers to the Bayesian nonparametric setting.

Supplementary Material

Supplementary Material for "Inference for Bayesian nonparametric models with binary response data via permutation counting" (DOI: 10.1214/22-BA1353SUPP; .pdf). Proofs of Proposition 2.2, Proposition 3.1 and Theorem 3.1.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. MR1224394. 293

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2(6): 1152–1174. MR0365969. 294, 306

Arbel, J., De Blasi, P., and Prünster, I. (2019). "Stochastic approximations to the Pitman–Yor process." *Bayesian Analysis*, 14(4): 1201–1219. MR4136558. doi: https://doi.org/10.1214/18-BA1127. 295

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). "An empirical distribution function for sampling with incomplete information." *The Annals of Mathematical Statistics*, 26(4): 641–647. MR0073895. doi: https://doi.org/10.1214/aoms/1177728423. 310, 313

Bhattacharya, P. K. (1981). "Posterior distribution of a Dirichlet process from quantal response data." *The Annals of Statistics*, 9(4): 803–811. MR0619283. 294

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Cambridge: Springer. MR2247587. doi: https://doi.org/10.1007/978-0-387-45528-0. 311

Bollobás, B. (1979). *Graph Theory: An Introductory Course*, volume 63 of *Graduate Texts in Mathematics*. New York: Springer. MR0536131. 302

Brunner, L. J. and Lo, A. Y. (1996). "Limiting posterior distributions under mixture of conjugate priors." *Statistica Sinica*, 6(1): 187–197. MR1379056. 313

Cai, B., Lin, X., and Wang, L. (2011). "Bayesian proportional hazards model for current status data with monotone splines." *Computational Statistics & Data Analysis*, 55(9): 2644–2651. MR2802342. doi: https://doi.org/10.1016/j.csda.2011.03.013. 294

Camerlenghi, F., Lijoi, A., and Prünster, I. (2021). "Survival analysis via hierarchically dependent mixture hazards." *The Annals of Statistics*, 49(2): 863–884. MR4255111. doi: https://doi.org/10.1214/20-aos1982. 313

Campbell, T., Huggins, J. H., How, J. P., and Broderick, T. (2019). "Truncated random measures." *Bernoulli*, 25(2): 1256–1288. MR3920372. doi: https://doi.org/10.3150/18-bej1020. 295

Cappé, O., Godsill, S. J., and Moulines, E. (2007). "An overview of existing methods and recent advances in sequential Monte Carlo." In *Proceedings of the IEEE*, volume 95, 899–924. 300

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). "Population Monte Carlo." *Journal of Computational and Graphical Statistics*, 13(4): 907–929. MR2109057. doi: https://doi.org/10.1198/106186004X12803. 300

Christensen, D. (2022). "Nonparametric Bayesian sensitivity testing with optimal design." In *Proceedings of the 51st International Annual Conference of the Fraunhofer ICT*. Fraunhofer ICT. 293

Christensen, D. (2023). Supplementary Material for "Inference for Bayesian Nonparametric Models with Binary Response Data via Permutation Counting". *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1353SUPP. 298, 304, 305

Cox, D. R. (1972). "Regression models and life-tables." *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2): 187–202. MR0341758. 294

De Blasi, P., Peccati, G., and Prünster, I. (2009). "Asymptotics for posterior hazards." *The Annals of Statistics*, 37(4): 1906–1945. MR2533475. doi: https://doi.org/10.1214/08-AOS631. 313

Diaconis, P. and Freedman, D. (1986). "On the consistency of Bayes estimates." *The Annals of Statistics*, 14(1): 1–26. MR0829555. doi: https://doi.org/10.1214/aos/1176349830. 313

Dixon, W. J. and Mood, A. M. (1948). "A Method for obtaining and analyzing sensitivity data." *Journal of the American Statistical Association*, 43(241): 109–126. 293

Doss, H. (1985a). "Bayesian nonparametric estimation of the median; part I: Computation of the estimates." *The Annals of Statistics*, 13(4): 1432–1444. MR0811501. doi: https://doi.org/10.1214/aos/1176349746. 313

Doss, H. (1985b). "Bayesian nonparametric estimation of the median; part II: Asymptotic properties of the estimates." *The Annals of Statistics*, 13(4): 1445–1464. MR0811502. doi: https://doi.org/10.1214/aos/1176349747. 313

Doss, H. (1994). "Bayesian nonparametric estimation for incomplete data via succes-

sive substitution sampling." *The Annals of Statistics*, 22(4): 1763–1786. MR1329167. doi: https://doi.org/10.1214/aos/1176325756. 294, 306, 309

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 294

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 294

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In *Recent advances in statistics. Papers in honor of Herman Chernoff on his sixtieth birthday*, 287–302. Bibliohound. MR0736538. 294

Gelfand, A. E. and Kuo, L. (1991). "Nonparametric Bayesian bioassay including ordered polytomous response." *Biometrika*, 78(3): 657–666. MR1130934. doi: https://doi.org/10.1093/biomet/78.3.657. 294

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27(1): 143–158. MR1701105. doi: https://doi.org/10.1214/aos/1018031105. 313

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints*. New York: Cambridge University Press. MR3445293. doi: https://doi.org/10.1017/CBO9781139020893. 293, 313

Hesterberg, T. (1995). "Weighted average importance sampling and defensive mixture distributions." *Technometrics*, 37(2): 185–194. 300

Hjort, N. L. (1990). "Nonparametric Bayes estimators based on Beta processes in models for life history data." *The Annals of Statistics*, 18(3): 1259–1294. MR1062708. doi: https://doi.org/10.1214/aos/1176347749. 294

Hjort, N. L. and Walker, S. G. (2009). "Quantile pyramids for Bayesian nonparametrics." *The Annals of Statistics*, 37(1): 105–131. MR2488346. doi: https://doi.org/10.1214/07-AOS553. 294, 306, 307, 308

Huh, J. (2022). "A fast quantum algorithm for computing matrix permanent." *ArXiv preprint. Available at arXiv:*2205.01328. 302

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 294

Jongbloed, G., van der Meulen, F. H., and Pang, L. (2022). "Bayesian nonparametric estimation in the current status continuous mark model." *Scandinavian Journal of Statistics*, 49(3): 1329–1352. MR4471288. doi: https://doi.org/10.1111/sjos.12562. 313

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 294

Kaplan, E. L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations." *Journal of the American Statistical Association*, 53(282): 457–481. MR0093867. 310

Keiding, N., Begtrup, K., Scheike, T. H., and Hasibeder, G. (1996). "Estimation from current-status data in continuous time." *Lifetime Data Analysis*, 2: 119–129. 293, 306, 309, 310

Kim, Y. and Lee, J. (2004). "A Bernstein–von Mises theorem in the nonparametric right-censoring model." *The Annals of Statistics*, 32(4): 1492–1512. MR2089131. doi: https://doi.org/10.1214/009053604000000526. 313

Kuo, L. (1988). "Linear Bayes estimators of the potency curve in bioassay." *Biometrika*, 75(1): 91–96. MR0932821. doi: https://doi.org/10.1093/biomet/75.1.91. 294

Lenk, P. J. (1988). "The logistic normal distribution for Bayesian, nonparametric, predictive densities." *Journal of the American Statistical Association*, 83(402): 509–516. MR0971380. 294

Lenk, P. J. (1991). "Towards a practicable Bayesian nonparametric density estimator." *Biometrika*, 78(3): 531–543. MR1130921. doi: https://doi.org/10.1093/biomet/78.3.531. 294

Leonard, T. (1978). "Density estimation, stochastic processes and prior information." *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2): 113–132. MR0517434. 294

Lijoi, A., Mena, R. H., and Prünster, I. (2005). "Hierarchical mixture modeling with normalized inverse–Gaussian priors." *Journal of the American Statistical Association*, 100(472): 1278–1291. MR2236441. doi: https://doi.org/10.1198/016214505000000132. 294

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 294

Lijoi, A., Prünster, I., and Rigon, T. (2019). "Finite-dimensional discrete random structures and Bayesian clustering." *Collegio Carlo Alberto Working Paper*, No. 600. 294, 295

Lijoi, A., Prünster, I., and Rigon, T. (2020). "The Pitman–Yor multinomial process for mixture modelling." *Biometrika*, 107(4): 891–906. MR4186494. doi: https://doi.org/10.1093/biomet/asaa030. 294, 295, 306, 309, 310

Little, C. H. C. (1975). "A characterization of convertible (0, 1)-matrices." *Journal of Combinatorial Theory, Series B*, 18(3): 187–208. MR0424583. doi: https://doi.org/10.1016/0095-8956(75)90048-9. 302

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12(1): 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 294

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). "Approximate Bayesian computational methods." *Statistics and Computing*, 22: 1167–1180. MR2992292. doi: https://doi.org/10.1007/s11222-011-9288-2. 294

Muliere, P. and Secchi, P. (1995). "A note on a proper Bayesian bootstrap." Technical report, Dipartimento di economia politica e metodi quantitativi, Universita degli studi di Pavia. 295

Murray, I., MacKay, D., and Adams, R. P. (2008). "The Gaussian process density sampler." In *Advances in Neural Information Processing Systems*, volume 21. 294

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 294

Newton, M. A. and Zhang, Y. (1999). "A recursive algorithm for nonparametric analysis with missing data." *Biometrika*, 86(1): 15–26. MR1688068. doi: https://doi.org/10.1093/biomet/86.1.15. 294

Neyer, B. T. (1994). "A D-optimality-based sensitivity test." *Technometrics*, 36(1): 61–70. 293

Orbanz, P. and Teh, Y. W. (2011). "Bayesian nonparametric models." In Sammut, C. and Webb, G. I. (eds.), *Encyclopedia of Machine Learning*. Boston, Massachusetts: Springer. 294

Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). "A stick-breaking construction of the Beta process." In *International Conference on Machine Learning*, 847–854. PMLR. 294

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 294

Paulon, G., Müller, P., and Rosas, V. G. S. Y. (2020). "Bayesian nonparametric bivariate survival regression for current status data." *ArXiv preprint. Available at arXiv:*2009.06460. 294

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Massachusetts: MIT Press. MR2514435. 311

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *The Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 294

Ridout, M. S. (2009). "Generating random numbers from a distribution specified by its Laplace transform." *Statistics and Computing*, 19: 439–450. MR2565316. doi: https://doi.org/10.1007/s11222-008-9103-x. 309

Ryser, H. J. (1963). *Combinatorial Mathematics*, volume 14 of *Carus Mathematical Monographs*. American Mathematical Society. MR0150048. 302

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639–650. MR1309433. 294

Turnbull, B. W. (1974). "Nonparametric estimation of a survivorship function with doubly censored data." *Journal of the American Statistical Association*, 69(345): 169–173. MR0381120. 310

Valiant, L. G. (1979). "The complexity of computing the permanent." *Theoretical Computer Science*, 8(2): 189–201. MR0526203. doi: https://doi.org/10.1016/0304-3975(79)90044-6. 295, 302

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics - Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 294

Wang, N., Wang, L., and McMahan, C. S. (2015). "Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm." *Computational Statistics & Data Analysis*, 83: 140–150. MR3281802. doi: https://doi.org/10.1016/j.csda.2014.10.013. 294

Acknowledgments

I extend my sincere gratitude to my supervisors, Professor Nils Lid Hjort and Dr Erik Unneberg, for their assistance with finalising the present paper. Their comments and suggestions helped to extend its scope and improve its clarity. I also thank the anonymous reviewers, the Associate Editor and the Editor-in-Chief for helpful and constructive suggestions.