Biometrika (2015), **102**, 4, *pp.* 829–842
Printed in Great Britain
doi: 10.1093/biomet/asv032
Advance Access publication 28 July 2015

Shared kernel Bayesian screening

By ERIC F. LOCK

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A
elock@umn.edu

and DAVID B. DUNSON

Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A
dunson@duke.edu

Summary

This article concerns testing for equality of distribution between groups. We focus on screening variables with shared distributional features such as common support, modes and patterns of skewness. We propose a Bayesian testing method using kernel mixtures, which improves performance by borrowing information across the different variables and groups through shared kernels and a common probability of group differences. The inclusion of shared kernels in a finite mixture, with Dirichlet priors on the weights, leads to a simple framework for testing that scales well for high-dimensional data. We provide closed asymptotic forms for the posterior probability of equivalence in two groups and prove consistency under model misspecification. The method is applied to DNA methylation array data from a breast cancer study, and compares favourably to competitors when Type I error is estimated via permutation.

Some key words: Epigenetics; Independent screening; Methylation array; Misspecification; Multiple comparisons; Multiple testing; Nonparametric Bayes inference.

1. Introduction

1·1. *Motivation*

In modern biomedical research, it is common to screen for differences between groups in many variables. These variables are often measured using the same technology and are not well characterized using a simple parametric distribution. As an example, we consider DNA methylation arrays. Methylation is an epigenetic phenomenon that can affect transcription and occurs at genomic locations where a cytosine nucleotide is followed by a guanine nucleotide, called CpG sites. High-throughput microarrays are commonly used to measure methylation levels for thousands of CpG sites genome-wide. Measurements are typically collected from a tissue that contains several distinct cell types, and at a given CpG site each cell type is typically either methylated or unmethylated (Reinius et al., 2012). Arrays therefore give continuous measurements for discrete methylation states, and the resulting values are between 0, no methylation, and 1, fully methylated. Figure 1 shows the distribution of methylation measurements over individuals for three CpG sites based on data from the Cancer Genome Atlas Network (2012). Multimodality and skewness are common; kernel mixtures are useful for modelling such complexity.

Methylation variables share several distributional features, such as common support, common modes and common patterns of skewness. The use of kernels that are shared across variables thus

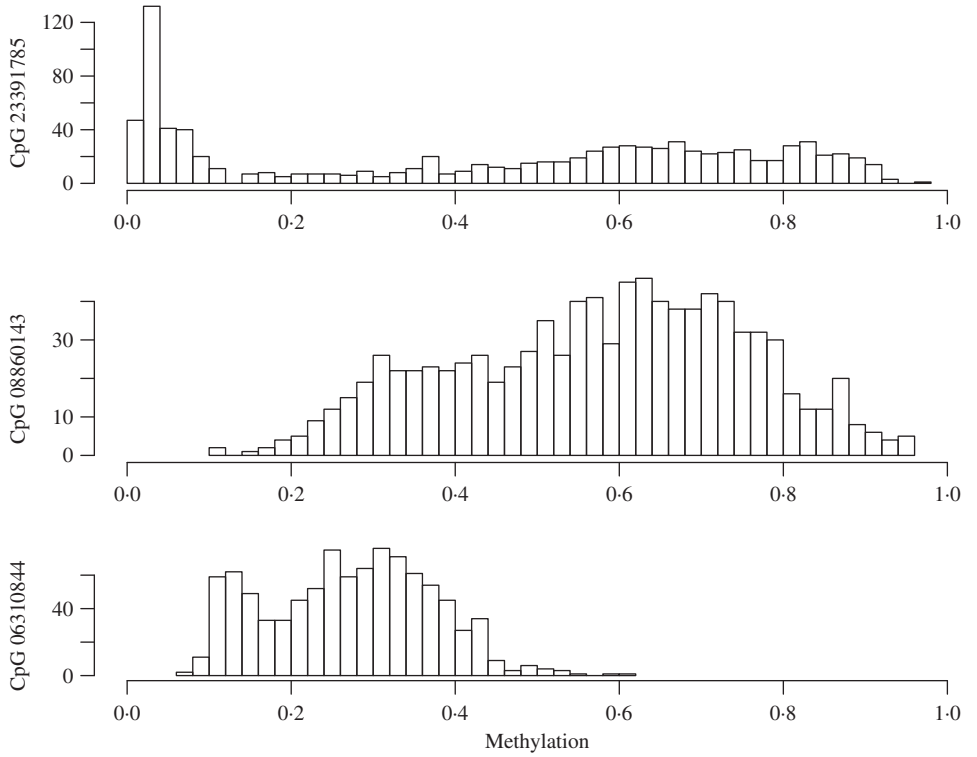

Fig. 1. Distribution of methylation measurements at three CpG sites; for each histogram the vertical scale gives the frequency of occurrence among individuals.

not only reduces computational burden but can also improve performance. It is also natural to share kernels across groups, with the interpretation that two groups arise from the same discrete process but in potentially different proportions.

We introduce a simple, computationally efficient, and theoretically supported Bayesian method for screening using shared kernels across groups and, if appropriate, across variables. The population distribution for each variable is approximated using a mixture of kernels $\{F_k\}_{k=1}^K$. For two groups 0 and 1, we test whether the groups have different kernel weights. Specifically, for group distributions $F_m^{(0)}$ and $F_m^{(1)}$ at variable m, $F_m^{(0)} = \sum_{k=1}^K \pi_{mk}^{(0)} F_k$ and $F_m^{(1)} = \sum_{k=1}^K \pi_{mk}^{(1)} F_k$, the competing hypotheses are

$$H_{0m} : \pi_{mk}^{(0)} = \pi_{mk}^{(1)} \text{ for all } k \quad \text{versus} \quad H_{1m} : \pi_{mk}^{(0)} \neq \pi_{mk}^{(1)} \text{ for some } k. \tag{1}$$

In practice, $F_1, \ldots, F_K$ and a shared Dirichlet prior distribution for the weights $\Pi_m^{(0)}$ and $\Pi_m^{(1)}$ are estimated empirically. A simple and tractable Gibbs sampling procedure is then used to estimate the posterior probability of H_{0m} for each variable.

While methylation array data provide excellent motivation, our framework addresses the general statistical problem of testing for equality between two groups that are drawn from the same strata but in potentially different proportions. We argue that the method may also be useful for applications that do not have such a clear interpretation, and this is supported with theoretical results in § 4.

1·2. *Related work*

The multimodality of methylation measurements is widely recognized (Laird, 2010) but often not accounted for in practice. The two-sample t-test is most commonly used to identify sites of differential expression in case-control studies (Bock, 2012). Alternative testing approaches are rank-based or discretize the data based on arbitrary thresholds (Chen et al., 2011; Qiu & Zhang, 2012). Other statistical models have been proposed to identify CpG sites that are hypomethylated, hypermethylated or undifferentiated with respect to normal cells (Khalili et al., 2007; Akalin et al., 2012). The focus on differential methylation levels between groups may miss other important differences between group distributions; for example, certain genomic regions have been shown to exhibit more variability in methylation, and hence greater epigenetic instability, among cancer cells than among normal cells (Hansen et al., 2011).

Although our model involves finite mixtures, it is intended to be robust with respect to parametric assumptions and so is comparable to nonparametric methods. There is a literature on nonparametric Bayes testing of equivalence in distribution between groups. Dunson & Peddada (2008) used a dependent Dirichlet process to test for equality against stochastically ordered alternatives; they employed an interval test based on total variation distance, and the framework is easily extended to unordered alternatives. Pennell & Dunson (2008) also used a Dirichlet process model for multiple groups and an interval test. Ma & Wong (2011) and Holmes et al. (2015) used Polya tree priors to test for exact equality. Existing nonparametric Bayes tests do not exploit shared features among variables, in the form of shared kernels or otherwise.

If kernel memberships are known, our testing framework (1) is equivalent to a test for association with a $2 \times K$ contingency table. For this there are standard frequentist methods such as Fisher's exact test and Pearson's chi-squared test, as well as established Bayesian methods (Good & Crook, 1987; Albert, 1997). In our context the component memberships are unknown and are inferred probabilistically. Xu et al. (2010) addressed this as part of a series of comparisons for Bayesian mixture distributions between groups. They compared marginal likelihoods for models with and without assuming constant weights between groups. Our focus is instead on screening settings in which there are many variables, and it is important to borrow information while adjusting for multiple testing. Shared kernels facilitate borrowing of information and computational scaling, and in our implementation a shared prior for the probability of equality at each variable induces a multiplicity adjustment with favourable properties (Scott & Berger, 2006, 2010; Muller et al., 2007).

2. Model

2·1. *Shared kernel mixtures*

Below we describe the general model for shared kernel Bayesian screening. Details that are specific to our implementation for methylation array data, including estimation techniques that facilitate posterior computation in high dimensions, are given in § 5.

First we describe a shared kernel mixture model, to lay the groundwork for the two-group screening model in § 2·2. Given data x_{mn} for M variables ($m = 1, \ldots, M$) and N subjects ($n = 1, \ldots, N$), the shared kernel model assumes that observations x_{mn} are realized from one of K component distributions $F_1, \ldots, F_K$. Typically x_{mn} is a continuous and unidimensional observation, but we present the model in sufficient generality to allow for more complex data structures. We assume that $F_1, \ldots, F_K$ have corresponding likelihoods from the same parametric family $f(\cdot, \theta_k)$.

Let $c_{mn} \in \{1, \ldots, K\}$ represent the component generating x_{mn}, and let $\pi_{mk} = \mathrm{pr}(c_{mn} = k)$ be the probability that an arbitrary subject belongs to component k in variable m. The generative

model is $x_{mn} \sim F_k$ with probability π_{mk}. Under a Bayesian framework, one puts a prior distribution on $\{\Pi_m = (\pi_{m1}, \ldots, \pi_{mK})\}_{m=1}^M$ and, if they are unspecified, the kernels $F_1, \ldots, F_K$. It is natural to use a Dirichlet conjugate prior for Π_m, characterized by a K-dimensional parameter α of positive real numbers. Small values of α, with $\alpha_k \leqslant 1$, will favour small values for a subset of the π_{mk} values. Thus, some kernels may have negligible impact for a given variable.

2·2. *Two-group screening*

We extend the shared kernel model above to allow for two sample groups: $X^{(0)}$ with data $x_{mn}^{(0)}$ for N_0 subjects ($n = 1, \ldots, N_0$; $m = 1, \ldots, M$), and $X^{(1)}$ with data $x_{mn}^{(1)}$ for N_1 subjects ($n = 1, \ldots, N_1$; $m = 1, \ldots, M$). Observations for all M variables are realized from a common set of kernels $F_1, \ldots, F_K$, but the two groups have potentially different weights $\{\Pi_m^{(0)}\}_{m=1}^M$ and $\{\Pi_m^{(1)}\}_{m=1}^M$.

The weights $\Pi_m^{(0)}$ and $\Pi_m^{(1)}$ each have prior distribution $\mathrm{Dir}(\alpha)$, whether they are identical or not. Let H_{0m} be the event that the mixing weights are the same for both groups: $\Pi_m^{(0)} = \Pi_m^{(1)}$. Under H_{1m}, $\Pi_m^{(0)}$ and $\Pi_m^{(1)}$ are considered independent realizations from $\mathrm{Dir}(\alpha)$. Let $F_m^{(0)}$ be the distribution for group 0 and $F_m^{(1)}$ the distribution for group 1. We consider a dummy variable $\mathbb{1}(H_{0m}) \sim \mathrm{Ber}\{\mathrm{pr}(H_{0m})\}$ and independent realizations $\tilde{\Pi}_{mk}$, $\tilde{\Pi}_{mk}^{(0)}$, $\tilde{\Pi}_{mk}^{(1)} \sim \mathrm{Dir}(\alpha)$ to give the joint distribution for groups $i = 0, 1$:

$$F_m^{(i)} = \sum_{k=1}^K \left[\mathbb{1}(H_{0m})\tilde{\pi}_{mk} + \{1 - \mathbb{1}(H_{0m})\}\tilde{\pi}_{mk}^{(i)}\right] F_k.$$

As $\mathrm{pr}(H_{0m}) \to 1$, $F_m^{(0)}$ and $F_m^{(1)}$ share the same mixing weights, and as $\mathrm{pr}(H_{0m}) \to 0$ the weights are independent.

Let $\vec{n}_m^{(0)} = (n_{m1}^{(0)}, \ldots, n_{mK}^{(0)})$ represent the number of subjects in group 0 that belong to each kernel k in variable m, and define $\vec{n}_m^{(1)}$ similarly for group 1. Then $\vec{n}_m = \vec{n}_m^{(0)} + \vec{n}_m^{(1)}$ gives the total number of subjects allocated to each component. Under H_{0m}, the distribution for the component memberships $C_m^{(0)}$ and $C_m^{(1)}$ is

$$\begin{aligned}
\mathrm{pr}\left(C_m^{(0)}, C_m^{(1)} \mid H_{0m}\right) &= \int_\Pi \mathrm{pr}\left(C_m^{(0)}, C_m^{(1)} \mid \Pi\right) f(\Pi \mid \alpha)\, \mathrm{d}\Pi \\
&= \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{mk} + \alpha_k\right)} \prod_{k=1}^K \frac{\Gamma(n_{mk} + \alpha_k)}{\Gamma(\alpha_k)} \\
&= \beta(\vec{n}_m + \alpha)/\beta(\alpha),
\end{aligned}$$

where Γ is the gamma function and $\beta(\alpha) = \prod_{k=1}^K \Gamma(\alpha_k)/\Gamma(\sum_{k=1}^K \alpha_k)$ is the multivariate beta function. Similarly, under H_{1m},

$$\begin{aligned}
\mathrm{pr}\left(C_m^{(0)}, C_m^{(1)} \mid H_{1m}\right) &= \int_\Pi \mathrm{pr}(C_m^{(0)} \mid \Pi) f(\Pi_m \mid \alpha)\, \mathrm{d}\Pi \int_\Pi \mathrm{pr}\left(C_m^{(1)} \mid \Pi\right) f(\Pi \mid \alpha)\, \mathrm{d}\Pi \\
&= \frac{\beta\left(\vec{n}_m^{(0)} + \alpha\right)\beta\left(\vec{n}_m^{(1)} + \alpha\right)}{\beta(\alpha)^2}.
\end{aligned}$$

Let the shared prior probability of no difference be $P_0 = \text{pr}(H_{0m})$ for all m. The posterior probability of H_{0m} given $C_m^{(0)}$ and $C_m^{(1)}$ is

$$\begin{aligned}\text{pr}\left(H_{0m} \mid C_m^{(0)}, C_m^{(1)}\right) &= \frac{P_0 \,\text{pr}\left(C_m^{(0)}, C_m^{(1)} \mid H_{0m}\right)}{P_0 \,\text{pr}\left(C_m^{(0)}, C_m^{(1)} \mid H_{0m}\right) + (1-P_0)\,\text{pr}\left(C_m^{(0)}, C_m^{(1)} \mid H_{1m}\right)} \\ &= \frac{P_0 \beta(\alpha)\beta\left(\vec{n}_m + \alpha\right)}{P_0 \beta(\alpha)\beta\left(\vec{n}_m + \alpha\right) + (1-P_0)\beta\left(\vec{n}_m^{(0)} + \alpha\right)\beta\left(\vec{n}_m^{(1)} + \alpha\right)},\end{aligned} \tag{2}$$

but in practice the kernel memberships are unknown, and the kernels may be unknown as well. There is no analogous closed form that accounts for uncertainty in $(C_m^{(0)}, C_m^{(1)})$, and direct computation is usually infeasible. We instead employ a Gibbs sampling procedure that uses (2) to approximate the full posterior distribution. Under multiple related tests, $M > 1$, we infer P_0 using a $\text{Be}(a, b)$ prior, where by default $a = b = 1$. The mean of the realized values of $\text{pr}(H_{0m} \mid C_m^{(0)}, C_m^{(1)})$ over the sampling iterations is used to estimate the posterior probability of H_{0m} for each variable.

While this article focuses on the two-group case, extensions to multiple groups are straightforward. A natural approach is to define a prior to cluster the groups. For example, we could use a Dirichlet process as in Gopalan & Berry (1998), but instead of clustering group means we would be clustering group distributions. Each cluster would then have a separate weight vector drawn from a Dirichlet distribution.

The above approach is presented in the context of shared kernels for high-dimensional screening, i.e., large M. The framework is also useful in the simple case where $M = 1$, and is particularly well motivated when two groups have the same strata but are in potentially different proportions. The theoretical results presented in §§ 3 and 4 are not specific to high-dimensional screening, and we will drop the variable subscript m for simplicity.

3. Asymptotic forms

We investigate the asymptotic forms that result from (2) as the number of observations tends to infinity. The proofs are given in the Supplementary Material.

Let $N = N_0 + N_1$ and fix $\lambda_0 = N_0/(N_0 + N_1)$. In Theorem 1 we derive the asymptotic form of the conditional Bayes factor $\text{pr}(H_0 \mid C^{(0)}, C^{(1)})/\,\text{pr}(H_1 \mid C^{(0)}, C^{(1)})$.

Theorem 1. *Let $\vec{p}_0 = \vec{n}^{(0)}/N_0$, $\vec{p}_1 = \vec{n}^{(1)}/N_1$, $\vec{p} = (\vec{n}^{(0)} + \vec{n}^{(1)})/N$, $r_{0k} = p_{0k}/p_k$ and $r_{1k} = p_{1k}/p_k$. Then, as $N_0, N_1 \to \infty$,*

$$\frac{\text{pr}\left(H_0 \mid C^{(0)}, C^{(1)}\right)}{\text{pr}\left(H_1 \mid C^{(0)}, C^{(1)}\right)} \sim c N^{(K-1)/2} \prod_{k=1}^{K} r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}},$$

where

$$c = \frac{P_0}{1-P_0}\left\{\frac{\lambda_0(1-\lambda_0)}{2\pi}\right\}^{(K-1)/2} \prod_{k=1}^{K} p_k^{\alpha_k+1/2}(r_{0k}r_{1k})^{1/2-\alpha_k}.$$

The asymptotic form given in Theorem 1 does not depend on the generative distribution. In the following we consider corollaries under H_0 and H_1.

COROLLARY 1. *Under* $H_0 : \Pi^{(0)} = \Pi^{(1)} = \Pi$,

$$\frac{\text{pr}\left(H_0 \mid C^{(0)}, C^{(1)}\right)}{\text{pr}\left(H_1 \mid C^{(0)}, C^{(1)}\right)} \sim cN^{(K-1)/2} \prod_{k=1}^{K} \exp\left[-\frac{\{\lambda_0(1-\lambda_0)\}^{1/2}}{2\pi_k} N(p_{0k} - p_{1k})^2\right],$$

where $\{\lambda_0(1-\lambda_0)\}^{1/2} N(p_{0k} - p_{1k})^2 \sim \chi^2_1$.

It follows that under H_0 the log Bayes factor has order $(K-1)\log(N)/2 + O_{\text{p}}(1)$, and therefore $\text{pr}(H_0 \mid C^{(0)}, C^{(1)})$ converges to 1 at a sublinear rate.

COROLLARY 2. *Under* $H_1 : \Pi^{(0)} \neq \Pi^{(1)}$, *let* $\Pi^* = \lambda_0 \Pi^{(0)} + (1-\lambda_0)\Pi^{(1)}$. *Then*

$$\frac{\text{pr}\left(H_0 \mid C^{(0)}, C^{(1)}\right)}{\text{pr}\left(H_1 \mid C^{(0)}, C^{(1)}\right)} \sim cN^{(K-1)/2} \prod_{k=1}^{K} \left(\frac{\pi_k^{(0)}}{\pi_k^*}\right)^{-N\lambda_0\pi_k^{(0)}} \left(\frac{\pi_k^{(1)}}{\pi_k^*}\right)^{-N(1-\lambda_0)\pi_k^{(1)}} \exp\left\{O_{\text{p}}\left(N^{1/2}\right)\right\}.$$

It follows that under H_1 the log Bayes factor has order

$$-N \sum \left\{ \lambda_0 \pi_k^{(0)} \log\left(\frac{\pi_k^{(0)}}{\pi_k^*}\right) + (1-\lambda_0)\pi_k^{(1)} \log\left(\frac{\pi_k^{(1)}}{\pi_k^*}\right) \right\} + O_{\text{p}}\left(N^{1/2}\right),$$

and therefore $\text{pr}(H_0 \mid C^{(0)}, C^{(1)})$ converges to zero at an exponential rate.

Exponential convergence under H_1 and sublinear convergence under H_0 have been observed for many Bayesian testing models (Kass & Raftery, 1995; Walker, 2004). Johnson & Rossell (2010) discuss this property for local prior densities, in which regions of the parameter space consistent with H_0 also have nonnegligible density under H_1; they give general conditions for the Bayes factor to have order $N/2$ under H_0 and to converge exponentially under H_1 when testing a point null hypothesis for a scalar parameter. In our view, the asymmetry in asymptotic rates under H_0 and H_1 is reasonable in our case and in most other models, as H_0 is much more precise. In practice, we still obtain strong evidence in favour of H_0 for moderate samples.

The exact asymptotic distributions given in Corollaries 1 and 2 are derived under the assumption that the component memberships $C^{(0)}$ and $C^{(1)}$ are known, but in practice they are unknown. Additionally, the component distributions $F_1, \ldots, F_K$ may be unknown. A simulation study presented in the Supplementary Material suggests that the asymptotic rates derived above also hold with a prior on $C^{(0)}$, $C^{(1)}$ and $F_1, \ldots, F_K$.

4. CONSISTENCY

We establish consistency of our method as a test for equality of distribution under very general conditions. The following results allow for misspecification in that the true data-generating distribution may not fall within the support of the prior. For example, $F^{(0)}$ and $F^{(1)}$ may not be finite mixtures of simpler component distributions. Such misspecified models clearly do not provide a consistent estimator for the full data-generating distribution, but, as we will show, they can still be consistent as a test for equality of distribution. The proofs of all theorems, corollaries and remarks in this section are given in the Appendix.

First, we derive asymptotic results for a one-group finite mixture model under misspecification. The proof of our first result uses the general asymptotic theory for Bayesian posteriors under misspecification given in Kleijn & van der Vaart (2006), and we borrow their notation where appropriate. Theorem 2 below implies that the posterior for a mixture distribution will

converge to the convex combination of component distributions f^* that is closest in terms of Kullback–Leibler divergence to the true density f_0. First, we define $B(\epsilon, f^*; f_0)$ to be a neighbourhood of the density f^* under the measure induced by the density f_0,

$$B(\epsilon, f^*; f_0) = \left\{ f \in \mathbb{F} : -\int f_0 \log \frac{f}{f^*} \leqslant \epsilon^2, \ \int f_0 \left(\log \frac{f}{f^*} \right)^2 \leqslant \epsilon^2 \right\},$$

and define $d(f_1, f_2)$ to be the weighted Hellinger distance,

$$d^2(f_1, f_2) = \frac{1}{2} \int \left(f_1^{1/2} - f_2^{1/2} \right)^2 \frac{f_0}{f^*}.$$

THEOREM 2. *Let $x_1, \ldots, x_N$ be independent with density f_0. Let $\mathbb{F}$ be the set of all convex combinations of dictionary densities $\{f_k\}_{k=1}^K$, and let P define a prior on $\mathbb{F}$. Assume that $f^* = \arg\min_{f \in \mathbb{F}} \mathrm{KL}(f_0 || f^*)$ exists and that $\mathrm{pr}\{B(\epsilon, f^*; f_0)\} > 0$ for all $\epsilon > 0$. Then, for any fixed $\epsilon > 0$,*

$$\mathrm{pr}\{f \in \mathbb{F} : d(f, f^*) \geqslant \epsilon \mid x_1, \ldots, x_N\}, \quad N \to 0.$$

The prior support condition $\mathrm{pr}\{B(\epsilon, f^*; f_0)\} > 0$ for all $\epsilon > 0$ is satisfied for all priors that have positive support over $\mathbb{F}$. This includes priors for Π with positive support over the unit simplex $\mathbb{S}^{K-1}$, such as Dirichlet priors. Although the weighted Hellinger distance d is nonstandard, convergence in d implies convergence of the component weights, as shown in Corollary 3.

COROLLARY 3. *In the setting of Theorem 2, let $\Pi^* = (\pi_1^*, \ldots, \pi_K^*)$ be the component weights corresponding to f^*. Assume Π^* is unique in that $\sum \pi_k f_k = \sum \pi_k^* f_k = f^*$ only if $\Pi = \Pi^*$. Then, for any fixed $\epsilon > 0$,*

$$\mathrm{pr}\big(\Pi \in \mathbb{S}^{K-1} : \|\Pi - \Pi^*\| \geqslant \epsilon \mid x_1, \ldots, x_N\big), \quad N \to 0.$$

Uniqueness of the component weights at f^* is trivially satisfied if distinct mixture weights yield distinct distributions in $\mathbb{F}$. Such identifiability has been established in general for Gaussian mixtures with variable means and variances, as well as for several other common cases (Teicher, 1963; Yakowitz & Spragins, 1968).

Kullback–Leibler divergence over $\mathbb{F}$ is convex, and its minimizer f^* satisfies interesting conditions.

Remark 1. In the setting of Theorem 2, assume that $\pi_k^* > 0$ for $k = 1, \ldots, K$ and $\sum \pi_k^* = 1$. Then $f^* = \sum \pi_k^* f_k$ achieves the minimum Kullback–Leibler divergence in $\mathbb{F}$ with respect to f_0 if and only if

$$\int \frac{f_1}{f^*} f_0 = \cdots = \int \frac{f_K}{f^*} f_0.$$

If some $\pi_k^* = 0$, the minimum divergence is achieved where $\int (f_k / f^*) f_0$ are equivalent for all $\pi_k^* > 0$.

We now give the result on consistency as a test for equality of distribution.

THEOREM 3. *Assume that* $x_1^{(0)}, \ldots, x_{N_0}^{(0)}$ *are independent with density* $f^{(0)}$ *and* $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ *are independent with density* $f^{(1)}$*, and let*

$$f^{*(0)} = \arg\min_{f \in \mathbb{F}} \mathrm{KL}\big(f^{(0)} || f\big), \quad f^{*(1)} = \arg\min_{f \in \mathbb{F}} \mathrm{KL}\big(f^{(1)} || f\big).$$

Assume that the uniqueness condition in Corollary 3 holds for $f^{*(0)}$ *and* $f^{*(1)}$*. If* $f^{(0)} = f^{(1)}$*, then* $\mathrm{pr}(H_0 \mid X) \to 1$ *as* $N \to \infty$*. If* $f^{*(0)} \neq f^{*(1)}$*, then* $\mathrm{pr}(H_0 \mid X) \to 0$ *as* $N \to \infty$*.*

Theorem 3 implies that the posterior probability of equality is consistent under H_0, even under misspecification. Consistency under H_1 holds generally under misspecification, but fails if $f^{(0)}$ and $f^{(1)}$ are both closest in Kullback–Leibler divergence to the same $f^* \in \mathbb{F}$. This can occur if $f^{(0)}$ and $f^{(1)}$ are both closer to the same component distribution f_k than they are to any other distribution in the convex hull.

5. APPLICATION TO METHYLATION DATA

5·1. *Data and estimation*

We illustrate our approach on a methylation array dataset for $N = 597$ breast cancer samples and $M = 21\,986$ CpG sites. These data are publicly available from TGCA Data Portal (Cancer Genome Atlas Network, 2012). We focus on testing for a difference between tumours that are identified as basal-like ($N_0 = 112$) and those that are not ($N_1 = 485$) at each site. Basal-like samples have a relatively poor clinical prognosis and a distinct gene expression profile, but the role of DNA methylation in this distinction has not been well characterized.

For scalability and to borrow information across sites, we apply a two-stage procedure. First, a set of dictionary kernels is estimated. Specifically, for $k = 1, \ldots, K$, f_k is the density of a normal distribution with mean μ_k and precision τ_k truncated to fall within the interval $[0, 1]$. We use a normal-gamma prior for μ_k and τ_k. For computational reasons we estimate the posterior for $f_1, \ldots, f_K$ from a subsample of 500 sites, for an effective sample size of $597 \times 500 = 298\,500$ observations. We employ a Gibbs sampler and update the common Dirichlet prior parameter α at each iteration using maximum likelihood estimation (Ronning, 1989). Alternatively, one could use a hyperprior for α, but this complicates posterior estimation and probably has little influence on posterior estimates as the effective sample size is very large. Similarly, we find that there is little uncertainty in the posterior mean and variance for each kernel; we can ignore the error in estimating these densities and fix them in the second stage.

The number of kernels $K = 9$ is chosen by crossvalidation based on the mean loglikelihood for held-out observations. Estimates for the dictionary densities $f_1, \ldots, f_9$ are shown in Fig. 2; to address the label-switching problem, we order the kernels by their means and then average over Gibbs samples. For fixed $f_1, \ldots, f_9$, we compute the posterior for the two-group model at each CpG site using a simple and efficient Gibbs sampler and a uniform hyperprior for P_0. We calculate the component likelihoods $f_k(x_{mn})$ for all sites m, samples n, and components k in advance, which greatly reduces the computational burden.

5·2. *Results*

We run Gibbs sampling for the two-group model for all 21 986 CpG sites, with 5000 iterations, after a 1000-iteration burn-in. The draws mix well and converge quickly; mixing is considerably improved by fixing the dictionary densities.

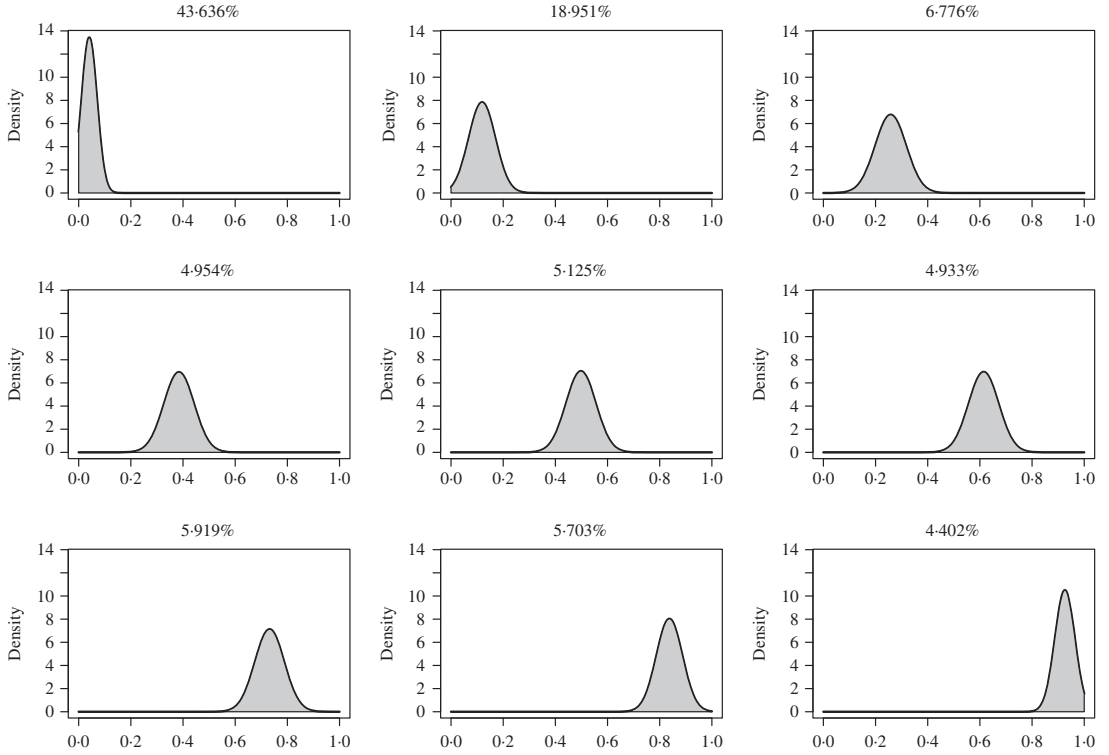

Fig. 2. Truncated normal dictionary densities for $K = 9$; above each panel the percentage of samples allocated to each density (over all sites) is shown.

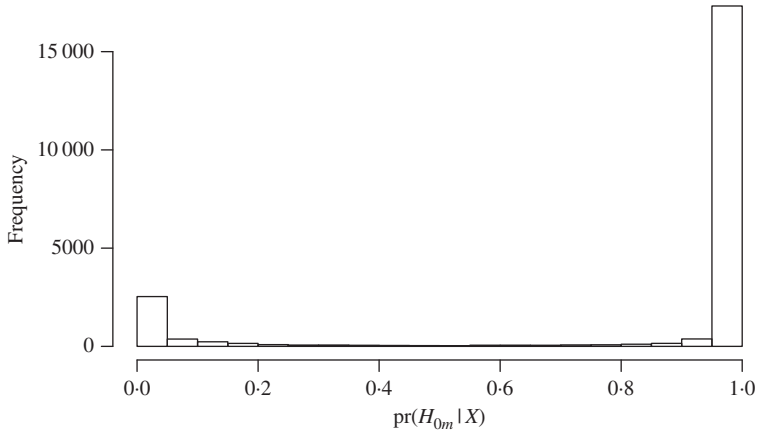

Fig. 3. Histogram of posterior probabilities of H_0 at 21 986 CpG sites with $N_0 = 112$ basal and $N_1 = 485$ nonbasal tumours.

The global prior probability of no difference, inferred using a uniform hyperprior, was $\hat{P}_0 = 0{\cdot}821$. The estimated posterior probabilities pr($H_{0m} \mid X$) are shown in Fig. 3. These have a U-shaped distribution, with 91% of values falling below 0·05 or above 0·95. Many values are close to 1, suggesting that these methylation sites play no role in the distinction between basal and nonbasal tumours.

Figure 4 shows the sample distributions and mixture density fits for basal and nonbasal tumours at four CpG sites. These four sites were selected to show a range of estimated differences

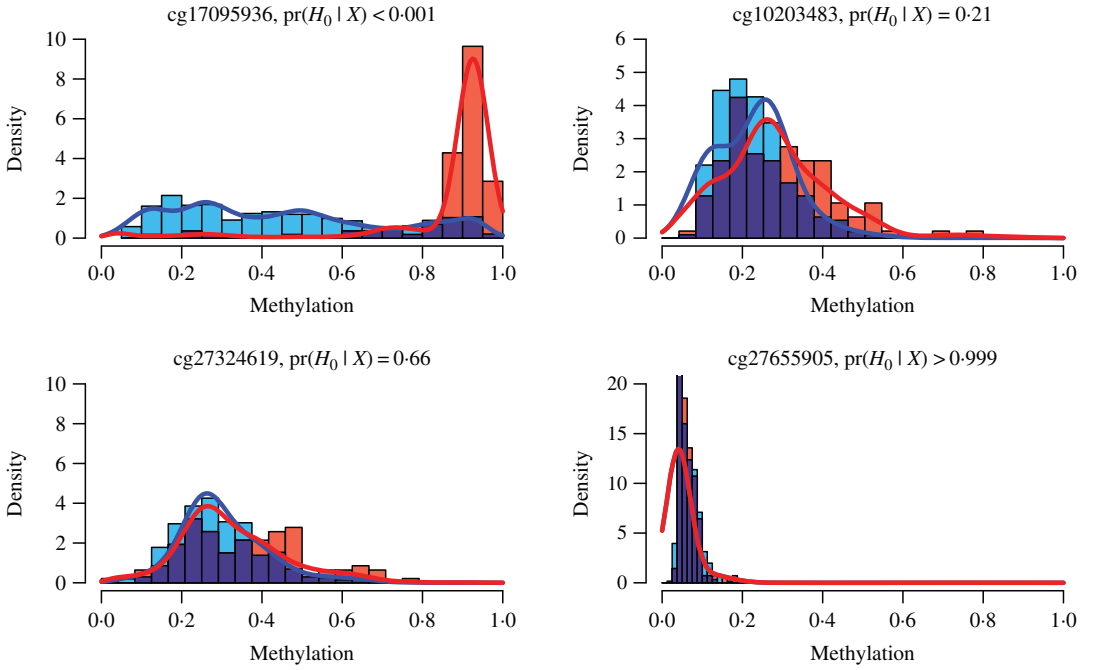

Fig. 4. Estimated densities for basal (red) and nonbasal (blue) samples for four CpG sites with different posterior probabilities of H_0. Histograms are shown for both groups, and their overlap is coloured violet.

between the distributions for basal and nonbasal tumours. In general, the estimated mixture densities appear to fit the data well. Some CpG sites with posterior probabilities $\text{pr}(H_{0m} \mid X)$ that are very small have dramatically different distributions between the two groups. For the majority of CpG sites, the estimated distributions for the two groups are nearly identical. The method naturally borrows strength across groups to estimate a common density when $\text{pr}(H_{0m} \mid X) \to 1$, and estimates the two densities separately when $\text{pr}(H_{0m} \mid X) \to 0$.

We investigated the potential relevance of differentially distributed CpG sites by considering the expression of the gene at their genomic location. DNA methylation is thought to primarily inhibit transcription and therefore suppress gene expression. Of the 2117 CpG sites with $\text{pr}(H_{0m} \mid X) < 0{\cdot}01$, 1256 have a significant negative association with gene expression according to Spearman's rank correlation, with p-value less than 0·01. For these cases, methylation provides a potential mechanistic explanation for well-known differences in gene transcription levels between basal and nonbasal tumours. In particular, these include five genes from the well-known PAM50 gene signature for breast cancer subtyping (Parker et al., 2009): *MYBL2*, *EGFR*, *MIA*, *SFRP1* and *MLPH*. A spreadsheet included with the Supplementary Material gives the posterior probability $\text{pr}(H_{0m} \mid X)$ and the corresponding gene expression statistics for all CpG sites.

6. METHODS COMPARISON ON METHYLATION DATA

We use data from § 5 to compare the power of testing methods on methylation array data. We consider the following methods: (a) the shared kernel test, as implemented in § 5 but with P_0 fixed at 0·5 so that the Bayes factors are independent; (b) the two-sample Anderson–Darling test (Scholz & Stephens, 1987); (c) a dependent optional Polya tree test (Ma & Wong, 2011), using code provided by the authors under default specifications; (d) a Polya tree test (Holmes et al., 2015), using code provided by the authors under default specifications; (e) the Wilcoxon rank

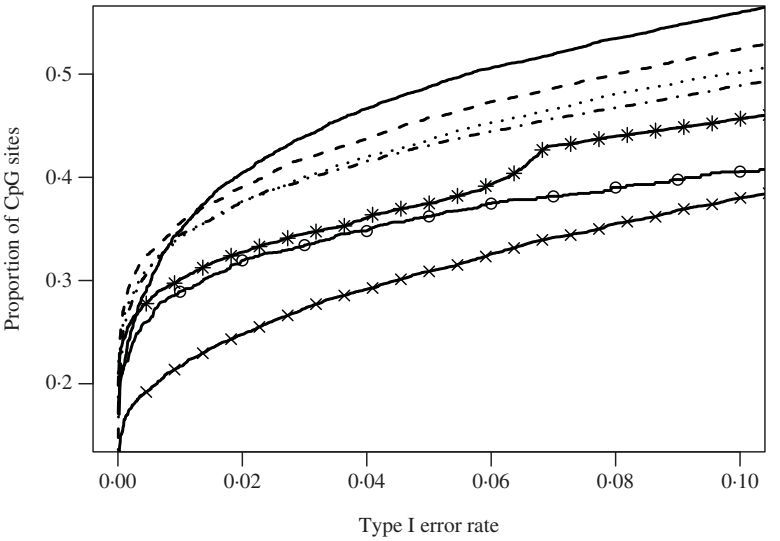

Fig. 5. Proportion of CpG sites identified as different between groups plotted against the proportion of sites identified as different under permutation, i.e., the Type I error rate, for seven different testing methods: the shared kernel test (solid); the Anderson–Darling test (dashed); the t-test (dotted); the Wilcoxon test (dot-dash); the dependent optional Polya tree test (asterisks); the restricted dependent Dirichlet process test (circles); and the Polya tree test (crosses).

sum test; (f) the two-sample t-test with unequal variance; (g) a restricted dependent Dirichlet process test (Dunson & Peddada, 2008) using the interval null hypothesis $d_{\mathrm{TV}} \in [0, 0{\cdot}05]$, where d_{TV} is the total variation distance. Methods (a)–(d) are general tests for equality of distribution, while methods (e)–(g) test for different levels of methylation.

We use each method to test for a difference between basal and nonbasal tumours at all 21 986 CpG sites. For comparison, we also apply each method under random permutation of the group labels separately at each site to generate a null distribution. The curves shown in Fig. 5 are obtained by varying the threshold on the Bayes factor or p-value, depending on the method. We compare the proportion of the 21 986 CpG sites that are identified as different with the proportion of sites that are identified as different under permutation. The proportion under permutation gives a robust estimate of the Type I error rate, so this is a frequentist approach to assessing discriminatory power. The shared kernel test exceeds other Bayesian nonparametric tests by a wide margin. It also generally performs as well as or better than frequentist approaches, although the Anderson–Darling test is competitive. Unlike nonparametric frequentist competitors, the shared kernel approach admits a full probability model to assess strength of evidence for both the null and the alternative hypotheses, which can be used in larger Bayesian models. Moreover, the shared kernel approach facilitates interpretation by modelling the full distribution, with uncertainty, for each group.

Acknowledgement

This work was supported in part by the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes R code and results of the analysis in § 5, as well as additional computational details and simulation studies.

APPENDIX

Proof of Theorem 2

The result follows from Corollary 2.1 of Kleijn & van der Vaart (2006). The space $\mathbb{F}$ is compact relative to total variation distance, and hence is bounded with respect to d. Therefore the covering numbers $N(\epsilon, \mathbb{F}, d)$ are finite for all $\epsilon > 0$. The space $\mathbb{F}$ is also convex, and so it follows from Lemmas 2.2 and 2.3 of Kleijn & van der Vaart (2006) that the entropy condition of Corollary 2.1 is satisfied for the metric d.

Proof of Corollary 3

Fix $\epsilon > 0$. Because $\mathrm{KL}(f^*||f)$ is defined, f^* and f_0 have common support. Therefore $d(f, f^*) = 0$ implies that $H(f, f^*) = 0$, where H is the Hellinger distance

$$H^2(f, f^*) = \frac{1}{2}\int (f^{1/2} - f^{*1/2})^2.$$

Hence, $d(\sum \pi_k f_k, f^*) = 0$ implies that $f = f^*$, and therefore that $\Pi = \Pi^*$ by the uniqueness assumption. Because $d(\sum \pi_k f_k, f^*)$ is continuous with respect to Π, there exists $\delta > 0$ such that $d(\sum \pi_k f_k, f^*) \leqslant \delta$ implies that $\|\Pi - \Pi^*\| \leqslant \epsilon$. Therefore

$$\mathrm{pr}\big(\Pi \in \mathbb{S}^{K-1} : \|\Pi - \Pi^*\| < \epsilon \mid X\big) \leqslant \mathrm{pr}\{f \in \mathbb{F} : d(f, f^*) > \delta \mid X\} \to 0$$

by Theorem 2.

Proof of Remark 1

As $\mathrm{KL}(f_0||\sum \pi_k f_k)$ is globally convex with respect to Π, the minimum divergence is achieved when all first-order derivatives are zero. Fix $\pi_3, \ldots, \pi_K$, and let $\pi_1 = a$ and $\pi_2 = 1 - a - \sum_{k=3}^K \pi_k$ for $0 \leqslant a \leqslant 1 - \sum_{k=3}^K \pi_k$. Let

$$f^{(a)} = af_1 + \left(1 - a - \sum_{k=3}^K \pi_k\right) f_2 + \sum_{k=3}^K \pi_k f_k.$$

Then

$$\frac{\partial}{\partial a}\mathrm{KL}(f_0||f^{(a)}) = -\int \frac{\partial}{\partial a}\log(f^{(a)}) f_0 = -\int \frac{f_1}{f^{(a)}} f_0 + \int \frac{f_2}{f^{(a)}} f_0.$$

Hence, $\partial\, \mathrm{KL}(f_0||f^{(a)})/\partial a = 0$ implies that

$$\int \frac{f_1}{f^{(a)}} f_0 = \int \frac{f_2}{f^{(a)}} f_0.$$

An analogous result holds for any π_i and π_j with $i \neq j$. Therefore, if $f^* = \arg\min_{f \in \mathbb{F}} \mathrm{KL}(f_0||f)$ with $\pi_k^* > 0$ for all k, then

$$\int \frac{f_1}{f^*} f_0 = \cdots = \int \frac{f_K}{f^*} f_0.$$

If $\pi_k^* = 0$ for some k, a similar argument shows that $\int (f_k/f^*) f_0$ must be equivalent for all $\pi_k^* > 0$.

Proof of Theorem 3

Let C indicate group membership, so that the generative distribution for $x_n \in \{X^{(0)}, X^{(1)}\}$ is

$$g(f^{(0)}, f^{(1)}, C) \sim \begin{cases} f^{(0)}, & C = 0, \\ f^{(1)}, & C = 1. \end{cases}$$

Note that

$$\begin{aligned}\mathrm{KL}\{g(\hat{f}^{(0)}, \hat{f}^{(1)}, C) \,||\, g(f^{(0)}, f^{(1)}, C)\} &= \int (1-C) f^{(0)} \log \frac{f^{(0)}}{\hat{f}^{(0)}} + \int C f^{(1)} \log \frac{f^{(1)}}{\hat{f}^{(1)}} \\ &= \lambda_0 \mathrm{KL}(\hat{f}^{(0)} || f^{(0)}) + (1-\lambda_0)\mathrm{KL}(\hat{f}^{(1)} || f^{(1)}).\end{aligned}$$

So, for $(\hat{f}^{(0)}, \hat{f}^{(1)}) \in \mathbb{F}^2$, the divergence with the generative distribution is minimized at $\hat{f}^{(0)} = f^{*(0)}$ and $\hat{f}^{(1)} = f^{*(1)}$. As $P_0 < 1$, the prior has positive support over $\mathbb{F}^2$ and so the concentration conditions of Theorem 2 are satisfied. It follows from Corollary 3 that

$$\mathrm{pr}(\|\hat{\Pi}^{(0)} - \Pi^{*(0)}\| \geqslant \epsilon \mid X) \to 0, \quad \mathrm{pr}(\|\hat{\Pi}^{(1)} - \Pi^{*(1)}\| \geqslant \epsilon \mid X) \to 0, \quad \epsilon > 0. \tag{A1}$$

Assume that $f^{*(0)} \neq f^{*(1)}$ and fix $\epsilon < \|\Pi^{*(0)} - \Pi^{*(1)}\|$. From (A1), $\mathrm{pr}(\|\hat{\Pi}^{(0)} - \hat{\Pi}^{*(1)}\| < \epsilon \mid X) \to 0$. This implies that $\mathrm{pr}(H_0 \mid X) \to 0$, as $\mathrm{pr}(H_0 \mid X) < \mathrm{pr}(\|\hat{\Pi}^{(0)} - \hat{\Pi}^{*(1)}\| < \epsilon \mid X)$.

Assume that $f^{*(0)} = f^{*(1)} = f^*$, where f^* has weights Π^*. Let

$$A_\delta = \{\Pi^{(0)}, \Pi^{(1)} : \|\Pi^{(0)} - \Pi^*\| < \delta, \ \|\Pi^{(1)} - \Pi^*\| < \delta\}.$$

Let f_α be the density for a $\mathrm{Dir}(\alpha)$ distribution and let $f(x \mid \Pi) = \sum_{k=1}^K \pi_k f_k$. For large N,

$$\begin{aligned}\mathrm{pr}(A_\delta, X \mid H_1) &= \iint_{\Pi^{(0)}, \Pi^{(1)} \in A_\delta} \prod_{i=1}^{N_0} f(x_i \mid \Pi^{(0)}) \prod_{j=1}^{N_1} f(x_j \mid \Pi^{(1)}) f_\alpha(\Pi^{(0)}) f_\alpha(\Pi^{(1)}) \\ &\leqslant \iint_{\Pi^{(0)}, \Pi^{(1)} \in A_\delta} \prod_{i=1}^{N_0} f(x_i \mid \Pi^{(0)}) \prod_{j=1}^{N_1} f(x_j \mid \Pi^{(0)}) f_\alpha(\Pi^{(0)}) f_\alpha(\Pi^{(1)}) \\ &= \mathrm{pr}(A_\delta, X \mid H_0)\, \mathrm{pr}(A_\delta \mid H_0),\end{aligned}$$

and so

$$\begin{aligned}\mathrm{pr}(H_1 \mid A_\delta, X) &= \frac{\mathrm{pr}(H_1)\, \mathrm{pr}(A_\delta, X \mid H_1)}{\mathrm{pr}(H_1)\, \mathrm{pr}(A_\delta, X \mid H_1) + P_0\, \mathrm{pr}(A_\delta, X \mid H_0)} \\ &\leqslant \frac{\mathrm{pr}(H_1)\, \mathrm{pr}(A_\delta \mid H_0)}{\mathrm{pr}(H_1)\, \mathrm{pr}(A_\delta \mid H_0) + P_0}.\end{aligned}$$

Clearly $\mathrm{pr}(A_\delta \mid H_0) \to 0$ as $\delta \to 0$, and therefore

$$\mathrm{pr}(H_1 \mid A_\delta, X) \to 0, \quad \delta \to 0. \tag{A2}$$

Result (A1) implies that for all $\delta > 0$,

$$\mathrm{pr}(\bar{A}_\delta \mid X) \to 0, \quad N \to 0. \tag{A3}$$

Fix $\epsilon > 0$. It follows from (A2) and (A3) that we may take δ sufficiently small to ensure that

$$\mathrm{pr}(H_1 \mid X) = \mathrm{pr}(\bar{A}_\delta \mid X)\, \mathrm{pr}(H_1 \mid \bar{A}_\delta X) + \mathrm{pr}(A_\delta \mid X)\, \mathrm{pr}(H_1 \mid A_\delta X) < \epsilon$$

for large N. Therefore, $\mathrm{pr}(H_0 \mid X) \to 1$ as $N \to \infty$.

References

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A. & Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87.

Albert, J. H. (1997). Bayesian testing and estimation of association in a two-way contingency table. *J. Am. Statist. Assoc.* **92**, 685–93.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Rev. Genet.* **13**, 705–19.

CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70.

CHEN, M., ZAAS, A., WOOD, C., GINSBERG, G., LUCAS, J., DUNSON, D. & CARIN, L. (2011). Predicting viral infection from high-dimensional biomarker trajectories. *J. Am. Statist. Assoc.* **106**, 1259–79.

DUNSON, D. B. & PEDDADA, S. D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859–74.

GOOD, I. J. & CROOK, J. F. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *Ann. Statist.* **15**, 670–93.

GOPALAN, R. & BERRY, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Am. Statist. Assoc.* **93**, 1130–9.

HANSEN, K. D., TIMP, W., BRAVO, H. C., SABUNCIYAN, S., LANGMEAD, B., MCDONALD, O. G., WEN, B., WU, H., LIU, Y., DIEP, D. et al. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genet.* **43**, 768–75.

HOLMES, C. C., CARON, F., GRIFFIN, J. E. & STEPHENS, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal.* **10**, 297–320.

JOHNSON, V. E. & ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Statist. Soc.* B **72**, 143–70.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

KHALILI, A., POTTER, D., YAN, P., LI, L., GRAY, J., HUANG, T. & LIN, S. (2007). Gamma-normal-gamma mixture model for detecting differentially methylated loci in three breast cancer cell lines. *Cancer Info.* **3**, 43–54.

KLEIJN, B. J. K. & VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34**, 837–77.

LAIRD, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Rev. Genet.* **11**, 191–203.

MA, L. & WONG, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Am. Statist. Assoc.* **106**, 1553–65.

MULLER, P., PARMIGIANI, G. & RICE, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics* 8, J. M. Bernardo, R. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West, eds. Oxford: Oxford University Press, pp. 349–70.

PARKER, J. S., MULLINS, M., CHEANG, M. C., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X., HU, Z. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–7.

PENNELL, M. L. & DUNSON, D. B. (2008). Nonparametric Bayes testing of changes in a response distribution with an ordinal predictor. *Biometrics* **64**, 413–23.

QIU, P. & ZHANG, L. (2012). Identification of markers associated with global changes in DNA methylation regulation in cancers. *BMC Bioinformatics* **13**, S7.

REINIUS, L. E., ACEVEDO, N., JOERINK, M., PERSHAGEN, G., DAHLÉN, S.-E., GRECO, D., SÖDERHÄLL, C., SCHEYNIUS, A. & KERE, J. (2012). Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361.

RONNING, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *J. Statist. Comp. Simul.* **32**, 215–21.

SCHOLZ, F. W. & STEPHENS, M. A. (1987). K-sample Anderson–Darling tests. *J. Am. Statist. Assoc.* **82**, 918–24.

SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plan. Infer.* **136**, 2144–62.

SCOTT, J. G. & BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–619.

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34**, 1265–9.

WALKER, S. G. (2004). Modern Bayesian asymptotics. *Statist. Sci.* **19**, 111–7.

XU, L., HANSON, T., BEDRICK, E. J. & RESTREPO, C. (2010). Hypothesis tests on mixture model components with applications in ecology and agriculture. *J. Agric. Biol. Envir. Statist.* **15**, 308–26.

YAKOWITZ, S. J. & SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39**, 209–14.

[*Received October* 2014. *Revised April* 2015]