Shared kernel Bayesian screening

BY E. F. LOCK AND D. B. DUNSON

Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A

Eric.F.Lock@gmail.com dunson@duke.edu

SUMMARY

This article concerns testing for differences between groups in many related variables. For example, the focus may be on identifying genomic sites with differential methylation between tumor subtypes. Standard practice in such applications is independent screening using adjustments for multiple testing to maintain false discovery rates. We propose a Bayesian nonparametric testing methodology, which improves performance by borrowing information adaptively across the different variables through the incorporation of shared kernels and a common probability of group differences. The inclusion of shared kernels in a finite mixture, with Dirichlet priors on the different weight vectors, leads to a simple and scalable methodology that can be routinely implemented in high dimensions. We provide some theoretical results, including closed asymptotic forms for the posterior probability of equivalence in two groups and consistency even under model misspecification. The method is shown to compare favorably to frequentist and Bayesian competitors, and is applied to methylation array data from a breast cancer study.

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

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 related and measured using the same technology, such as a microarray, but may not be well characterized using simple parametric distributions. In DNA methylation data, for example, measurements are usually skewed and multi-modal. Kernel mixtures, in which observations are realized from a collection of simpler distributions, are useful for modeling such complexity. When variables from the same technology share distributional features, the use of a common set of kernels has the dual advantage of improving performance and reducing computational burden.

We introduce a simple, computationally efficient, and theoretically supported Bayesian approach for screening using shared kernels. 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,

$$F_m^{(0)} = \sum_{k=1}^K \pi_{mk}^{(0)} F_k \text{ and } F_m^{(1)} = \sum_{k=1}^K \pi_{mk}^{(1)} F_k,$$

and the competing hypotheses are

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

In practice $F_1, ..., F_K$ and a shared Dirichlet prior distribution for the weights $\Pi_m^{(0)}, \Pi_m^{(1)}$ 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 the ideas in this article are broadly applicable, methylation array data provide excellent motivation. Methylation is an epigenetic phenomenon that occurs at CpG dinucleotide sites along the genome and can affect transcription. High-throughput bisulphate microarrays such as the Illumina® HumanMethylation450 Beadchip are now commonly used to measure DNA methylation levels for thousands of CpG dinucleotide sites genome-wide. Measurements are typically collected from a complex tissue containing several cell types, and at a given site each cell type is either methylated or unmethylated. Arrays therefore give continuous measurements for discrete methylation states, and our testing model has the mechanistic interpretation that two groups arise from the same discrete process but in potentially different proportions. All measurements are between 0 (no methylation) and 1 (complete methylation), and so tissue mixtures also motivate the use of shared kernels. Even without this mechanistic interpretation we argue that the method is useful for screening related variables with complex distributions, and this is supported with theoretical results under misspecification in Section 4.

1·2. *Related work*

The multi-modality 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). Mixture models for methylation data have been introduced with the goal of clustering (Houseman et al., 2008; Zhang et al., 2012; Kormaksson et al., 2012), but these models have not been used to test for equality between groups.

If the kernel that generated each observation is known, (1) is equivalent to testing for association with a $2 \times K$ contingency table, for which there is a well-developed frequentist (Fisher's exact test, Pearson's chi-square test) and Bayesian (Good & Crook, 1987; Albert, 1997) literature. 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 compare marginal likelihoods for models with and without assuming constant weights between groups. Our focus is instead on screening settings in which there are many related variables, and it is important to borrow information while adjusting for multiple testing. Shared kernels facilitate borrowing of information and computational scaling, while an unknown global $\text{pr}(H_{0m})$ induces a multiplicity adjustment with favorable properties (Scott & Berger, 2006; Muller et al., 2007; Scott et al., 2010).

Our approach is related to a small but growing literature on hypothesis testing using Bayesian nonparametrics. Such tests are attractive because they model the full distribution within each group. Hence, they have flexibility over tests with strong distributional assumptions, and power over more conventional nonparametric tests. Dunson & Peddada (2008) use a dependent Dirichlet process to test for equality of distributions against stochastically ordered alternatives. Their approach allows for the kernels of the Dirichlet process to differ by a location shift between populations. They use an interval test for the total variation distance between population distributions, and the framework is easily extended to unordered alternatives. Ma & Wong (2011) and Holmes

et al. (unpublished manuscript, available at `http://arxiv.org/abs/0910.5060`) use Polya tree priors to test for equality between two distributions. Holmes et al. evaluate evidence for the hypothesis that two population distributions are equal against the alternative that they are independent draws from the same Polya tree prior; Ma & Wong (2011) introduce a coupling optional Polya tree to model two population distributions as dependent draws from a Polya tree prior.

Existing nonparametric Bayes tests do not have a clear mechanistic motivation for methylation array data. More generally, they do not exploit shared features on many related variables, in the form of shared kernels or otherwise. To our knowledge, we are the first to address the problem of nonparametric Bayes screening of group differences in many related variables.

2. MODEL

2·1. *Shared kernel mixtures*

We describe a shared kernel mixture model to lay the groundwork for the two-group screening model in Section 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 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 $F_1, \ldots, F_K$ have corresponding likelihoods from the same parametric family $f(\cdot, \theta_k)$. For the application in Section 5 we use a location-scale mixture of truncated normal kernels.

Let $c_{mn} \in \{1, \ldots, K\}$ represent the component generating x_{mn}, and π_{mk} be the probability that an arbitrary subject belongs to component k in variable m: $\pi_{mk} = \mathrm{pr}(c_{mn} = k)$. 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. The Dirichlet is characterized by a K-dimensional concentration parameter α of positive reals. Small values of α, with $\alpha_k \leq 1$, will favor small values for a subset of the π_{mk} values, so that a given variable may not represent all K kernels. We use an empirical Bayes method of estimating α in our applications.

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 $\Pi_m^{(1)}$ have prior distribution $\mathrm{Dir}(\alpha)$. Let H_{0m} be the event that the mixing weights are the same for both groups: $\Pi_m^{(0)} = \Pi_m^{(1)} = \Pi$. Under H_{1m}, $\Pi_m^{(0)}$ and $\Pi_m^{(1)}$ are considered independent realizations from $\mathrm{Dir}(\alpha)$. The population distributions $F_m^{(i)}$, $i = 0, 1$ are therefore

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

where $\tilde{\Pi}_{mk}, \tilde{\Pi}_{mk}^{(0)}, \tilde{\Pi}_{mk}^{(1)} \sim \mathrm{Dir}(\alpha)$ and $\mathbb{1}(H_{0m}) \sim \mathrm{Bernoulli}\{\mathrm{pr}(H_{0m})\}$. 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 $\vec{n}_m^{(0)} = (n_{m1}^{(0)}, \ldots, n_{mK}^{(0)})$ give 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

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

where Γ is the gamma function and β is the multivariate beta function

$$\beta(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}.$$

Similarly, under H_{1m},

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

Let $P_0 = \text{pr}(H_{0m})$ be a global prior probability of no difference. The posterior probability of H_{0m} given $C_m^{(0)}$ and $C_m^{(1)}$ has the closed form

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

This posterior probability is computationally and analytically helpful. However, 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 not feasible. We instead employ a Gibbs sampling procedure that makes use of (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. Details of the Gibbs sampler are provided in the Appendix B·8.

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

3. ASYMPTOTIC FORMS

We investigate the asymptotic forms that result from Equation (2) as the number of observations tends to infinity. Proofs for all theorems and corollaries herein are given in Appendix A.

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 $\mathrm{pr}(H_0 \mid C^{(0)}, C^{(1)})/\,\mathrm{pr}(H_1 \mid C^{(0)}, C^{(1)})$.

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* $N_0, N_1 \to \infty$,

$$\frac{\mathrm{pr}(H_0 \mid C^{(0)}, C^{(1)})}{\mathrm{pr}(H_1 \mid C^{(0)}, C^{(1)})} \sim cN^{\frac{K-1}{2}} \prod_{k=1}^{K} r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}$$

where

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

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

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

where $N(p_{0k} - p_{1k})^2$ *is proportional to a chi-square distribution with 1 degree of freedom:*

$$\{\lambda_0(1-\lambda_0)\}^{1/2} N(p_{0k} - p_{1k})^2 \sim \chi_1^2.$$

It follows that under H_0 the log of the Bayes factor has order

$$\frac{K-1}{2} \log(N) + O_p(1),$$

and therefore $\mathrm{pr}(H_0 \mid C^{(0)}, C^{(1)})$ converges to 1 at a linear rate.

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

$$\Pi^* = \lambda_0 \Pi^{(0)} + (1 - \lambda_0)\Pi^{(1)}.$$

Then,

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

It follows that under H_1 the log of the Bayes factor has order

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

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

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. In practice we are interested

in cases where the component memberships are unknown. Additionally, the component distributions $F_1, \ldots, F_K$ may be unknown. A simulation study presented in Section 6·2 suggests that the general rates of convergence that result from Corollaries 1 and 2 still hold approximately with uncertainty in $C^{(0)}$, $C^{(1)}$, and $F_1, \ldots, F_K$, provided the generative model lies within the support of the prior.

4. CONSISTENCY

We establish consistency of the method as a test for equality of distribution under very general conditions. The following results pertain to robustness when the generative model is misspecified. For example, if the distributions $F^{(0)}$ and $F^{(1)}$ can not be represented as a finite mixture of simpler component distributions and therefore do not fall in the support of the prior. Proofs for all theorems, corollaries, and remarks herein are given in appendix A.

First, we derive asymptotic results for a one group finite mixture model under misspecification. These results follow from 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 (KL) divergence to the true density f_0. First, we define $B(\epsilon, f^*; f_0)$ to be a neighborhood of the density f^* under the measure induced by the density f_0:

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

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

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

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 $f^* = \underset{f \in \mathbb{F}}{\text{argmin}}\ \mathrm{KL}(f_0 || f^*)$ exists and $P\{B(\epsilon, f^*; f_0)\} > 0$ for all $\epsilon > 0$. Then, for any fixed $\epsilon > 0$,*

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

The prior support condition $P\{B(\epsilon, f^*; f_0)\} > 0$ 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 non-standard, convergence in d implies convergence of the component weights, as shown in Corollary 3.

COROLLARY 3. *Under 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 $\epsilon > 0$,*

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

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 (Yakowitz et al., 1968).

KL-divergence over $\mathbb{F}$ is convex, and its minimizer f^* satisfies interesting conditions given in Remark 2 below.

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

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

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

We now give the result for consistency as a test for equality of distribution in the two-group case.

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

$$f^{*(0)} = \underset{f \in \mathbb{F}}{\textit{argmin}}\ \mathrm{KL}(f^{(0)} || f)\ ,\ \ f^{*(1)} = \underset{f \in \mathbb{F}}{\textit{argmin}}\ \mathrm{KL}(f^{(1)} || f).$$

Assume the uniqueness condition for Corollary 3 *holds for $f^{*(0)}$ and $f^{*(1)}$. Then,*

- *if $f^{(0)} = f^{(1)}$, $\mathrm{pr}(H_0 \mid X) \to 1$ as $N \to \infty$ and*
- *if $f^{*(0)} \neq f^{*(1)}$, $\mathrm{pr}(H_0 \mid X) \to 0$ 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 KL-divergence to the same $f^* \in \mathbb{F}$. This can occur if $f^{(0)}$ and $f^{(1)}$ are both closer in KL-divergence 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*

We illustrate our approach on a methylation array dataset for $N = 597$ breast cancer samples and $M = 21,986$ CpG sites. These data were curated by The Cancer Genome Atlas and are publicly available from the TCGA Data Portal (Cancer Genome Atlas Network, 2012). Breast cancer is a heterogeneous disease, and a deeper understanding of the epigenetic contributions to this heterogeneity may lead to more targeted therapies. Several comprehensive studies of breast cancer genomics have incorporated methylation arrays (Curtis et al., 2012; Cancer Genome Atlas Network, 2012; Lock & Dunson, 2013). We focus on testing for a difference between tumors that are identified as Basal-like ($N_0 = 112$) against those that are not ($N_1 = 485$) at each CpG 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. DNA methylation is thought to play an important role in regulating gene expression, so it is reasonable to expect differences between the two groups at some CpG sites.

R code to peform the following analysis, as well as instructions on how to download these data, are available at `http://people.duke.edu/~el113/MethTestingTCGA.zip`.

5·2. *Estimation*

For scalability and to allow for borrowing of information across sites, we apply a two-stage procedure. First, a set of dictionary kernels are estimated as a location-scale mixture of truncated normals. 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 sub-sample of 500 sites ($N = 597 \times 500 \approx 300,000$). We employ a Gibbs sampler and update the common Dirichlet prior parameter α at each iteration using maximum likelihood (Ronning, 1989). The effective sample size is very large and so there is little uncertainty in the posterior mean and variance for each kernel; we can obtain a very accurate estimate of the dictionary, and hence can ignore the negligible error in estimating these densities and fix them in the second stage.

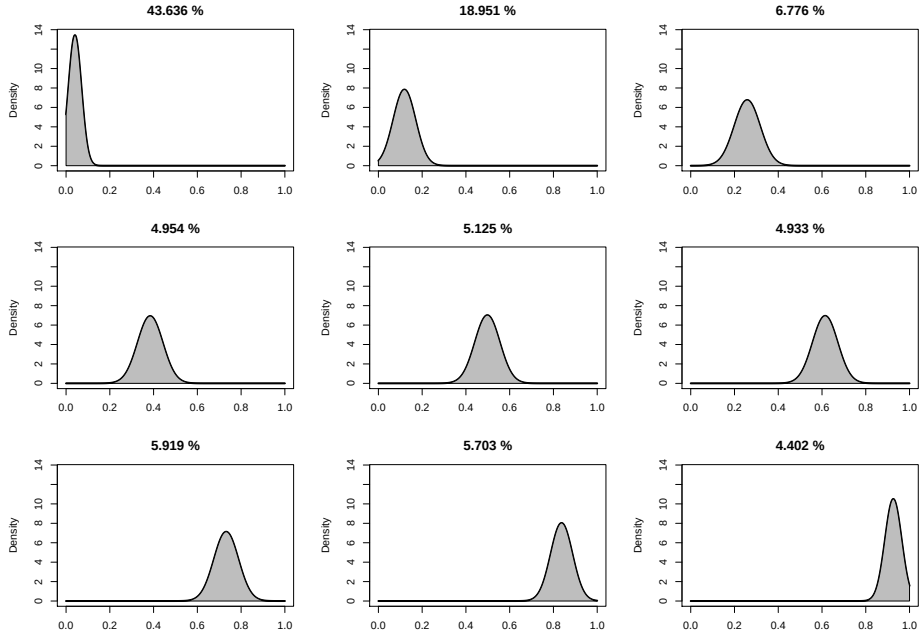

Fig. 1. Truncated normal dictionary densities for $K = 9$, with the percentage of samples allocated to each density (over all sites).

The number of kernels $K = 9$ is chosen by cross validation based on the mean log-likelihood for held out observations. Point estimates for the dictionary densities $f_1, \ldots, f_9$ are shown in Figure 1; 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 prior 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. For further estimation details see Appendix B·8.

5·3. *Results*

We run the Gibbs sampling scheme 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.

The estimated prior probability of no difference was $\hat{P}_0 = 0.821$. The estimated posterior probabilities $\mathrm{pr}(H_{0m} \mid X)$ are shown in Figure 2. These have a U-shaped distribution, with 91% of values falling below 0.05 or above 0.95. Most values are approximately 1, suggesting that most potential methylation sites play no role in the distinction between Basal and non-Basal tumors. This is as expected, as only a small proportion of genes are thought to play a significant role in breast tumor heterogeneity.

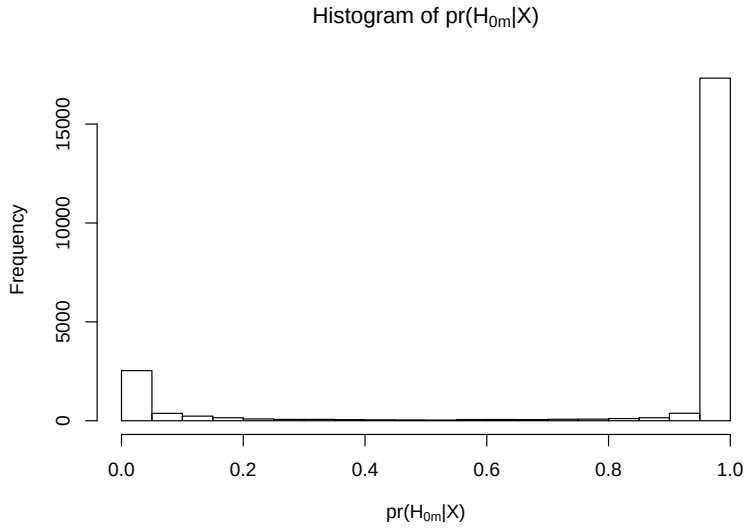

Fig. 2. Histogram of posterior probabilities of H_0 at 21,986 CpG sites with $N_0 = 112$ Basal and $N_1 = 485$ non-Basal tumors.

Figure 3 shows the sample distributions and mixture density fits for Basal and non-Basal tumors at four CpG sites. These four sites were selected to show a range of estimated differences between the distributions for Basal and non-Basal tumors. In general, the estimated mixture densities appear to fit the data well. Some CpG sites with posterior probabilities $\mathrm{pr}(H_{0m} \mid X)$ that are very close to 0 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 $\mathrm{pr}(H_{0m} \mid X) \to 1$, and estimates the two densities separately when $\mathrm{pr}(H_{0m} \mid X) \to 0$.

6. VALIDATION

6·1. *Methods comparison on methylation data*

We use data from Section 5 to compare standard frequentist and Bayesian non-parametric testing methods on methylation array data. We focus on a set of 574 CpG sites that were previously identified as relevant to breast cancer heterogeneity (Cancer Genome Atlas Network, 2012). We expect a high percentage of these sites to have different methylation distributions between Basal and non-Basal tumors. We use these sites to compare power of the following testing methods: (1) shared kernel test, as implemented in Section 5, (2) shared kernel test with P_0 fixed at 0.5, (3) two-sample Anderson-Darling test (Scholz & Stephens, 1987), (4) dependent optional Polya tree test (Ma & Wong, 2011), using code provided by the authors under default specifications, (5) Wilcoxon rank sum test, (6) two-sample t-test with unequal variance, (7) restricted

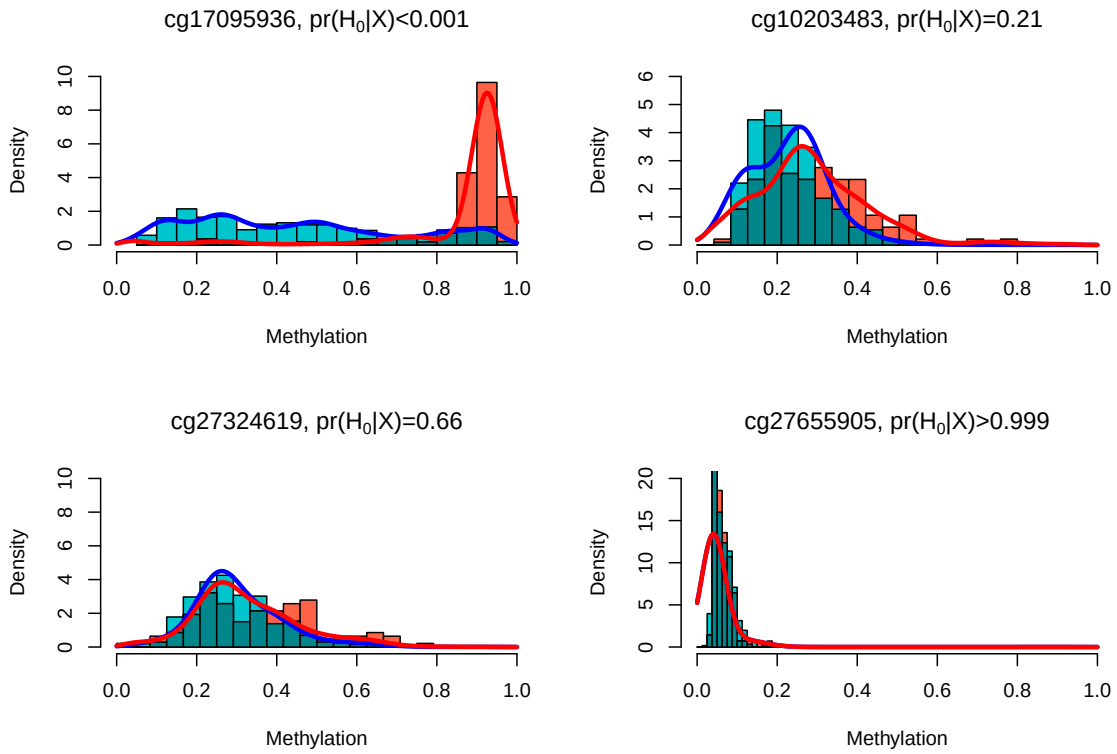

Fig. 3. The estimated densities for Basal (red) and non-Basal (blue) samples for four CpG sites with different posterior probabilities of H_0.

dependent Dirichlet process test (Dunson & Peddada, 2008), using the interval null hypothesis $d_{TV} \in [0, 0.05]$, where d_{TV} is total variation distance, (8) Polya tree test (Holmes et al.), using code provided by the authors under default specifications.

We estimate power by the proportion of the 574 CpG sites that are identified as different, and we estimate the false positive rate by the proportion of CpG sites that are identified as different under random permutation of the group labels. The receiver operating characteristic curve shown in Figure 4 is obtained by varying the threshold on the Bayes factor or p-value, depending on the method. The shared kernel test with learned P_0 is the only approach that borrows information across multiple sites and outperforms all independent screening methods by a wide margin. The independent shared kernel test with fixed $P_0 = 0.5$ also performs as well or better than existing methods, although the Anderson-Darling test is competitive.

6·2. *Simulation study: Bayes factor convergence*

We investigate the performance of the shared kernel test for simulated data. We simulate 200 data sets from the proposed model as described in the Appendix B·10. For each simulation, we estimate the posterior for both the component distributions and group weights simultaneously. We fix $P_0 = 0.5$ and use a truncated normal-gamma prior for the component densities, as described in Section 5·2.

To investigate the behavior of the posterior probability of H_0 as a function of N, we normalize the log of the posterior Bayes factor as suggested by the dominating term from the asymptotic

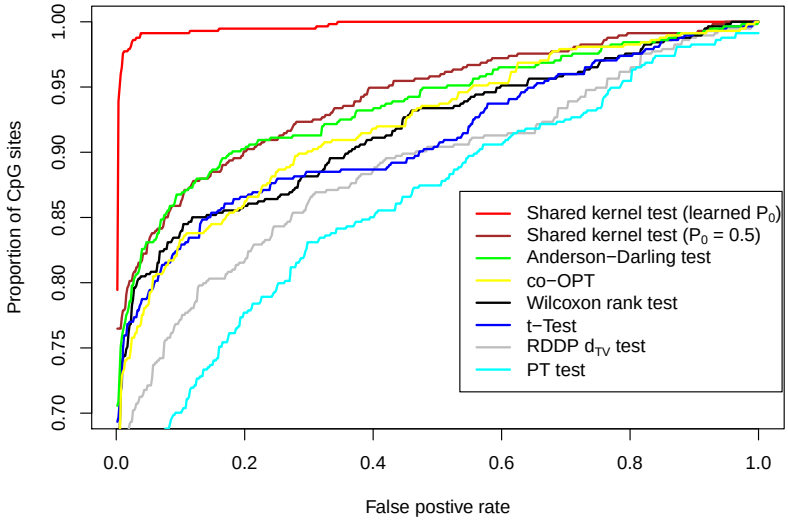

Fig. 4. Proportion of true CpG sites identified as different between groups, as a function of false positive rate, for eight different testing methods.

rates in Section 3. Specifically, under H_0 the normalized Bayes factor is

$$\frac{2}{K-1}\log\left\{\frac{\mathrm{pr}(H_0|X)}{\mathrm{pr}(H_1|X)}\right\} \tag{3}$$

and under H_1 the normalized Bayes factor is

$$\frac{\log\left\{\frac{\mathrm{pr}(H_0|X)}{\mathrm{pr}(H_1|X)}\right\}}{\sum\left\{\lambda_0\pi_k^{(0)}\log\left(\frac{\pi_k^{(0)}}{\pi_k^*}\right)+(1-\lambda_0)\pi_k^{(1)}\log\left(\frac{\pi_k^{(1)}}{\pi_k^*}\right)\right\}}. \tag{4}$$

Figure 5 shows the normalized Bayes factor for each of 200 simulations. As expected, $\mathrm{pr}(H_0|X)$ tends to 1 under H_0 and tends to 0 under H_1, as $N\to\infty$. Moreover, the log of the Bayes factor tends to $-\infty$ at an approximately linear rate with N under H_1, and tends to $+\infty$ at an approximately log-linear rate with N under H_0. These rates agree with the asymptotic forms derived in Section 3.

ACKNOWLEDGMENT

This work was supported in part by grants from the National Institutes of Health.

A. PROOFS

A·1. *Proof of Theorem* 1

Proof. Consider

$$\begin{aligned}\frac{\beta(\vec{n})}{\beta(\vec{n}^{(0)})\beta(\vec{n}^{(1)})} &= \frac{\Gamma(N_0)\Gamma(N_1)}{\Gamma(N)}\prod_{k=1}^{K}\frac{\Gamma(n_k)}{\Gamma(n_k^{(0)})\Gamma(n_k^{(1)})}\\ &= \frac{\beta(N_0,N_1)}{\prod_{k=1}^{K}\beta(n_k^{(0)},n_k^{(1)})}.\end{aligned}$$

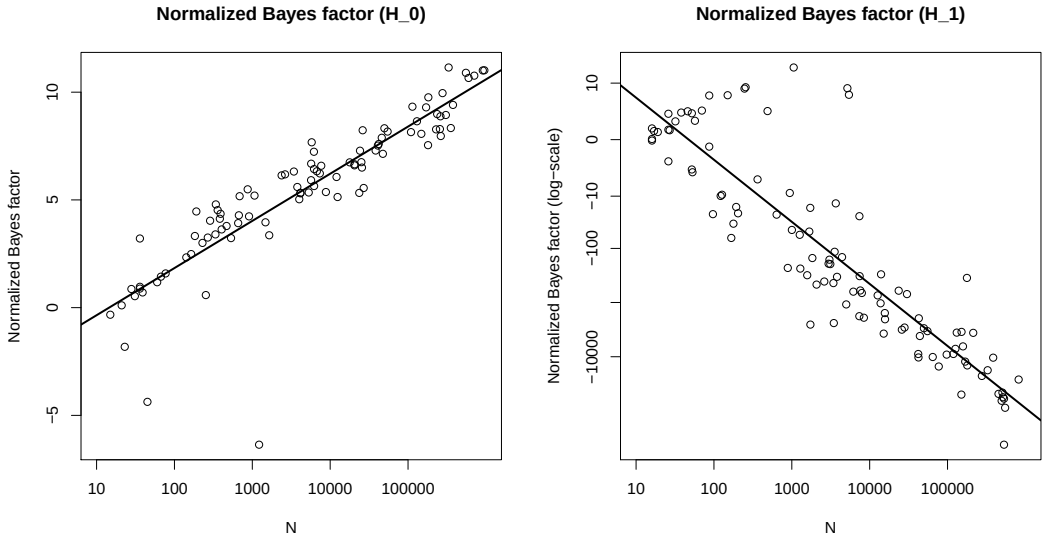

Fig. 5. Normalized Bayes factors (see (3) and (4)) versus N, under H_0 (left plot) and H_1 (right plot). Under H_0 the normalized Bayes factors are given on the original scale, under H_1 they are given on a log scale. Both plots show a linear trend.

Stirling's approximation gives

$$\beta(x,y) \sim (2\pi)^{1/2} x^{x-\frac{1}{2}} y^{y-\frac{1}{2}} (x+y)^{\frac{1}{2}-x-y},$$

and so

$$\begin{aligned}
\frac{\beta(\vec{n})}{\beta(\vec{n}^{(0)})\beta(\vec{n}^{(1)})} &\sim \frac{(2\pi)^{1/2} N_0^{N_0-\frac{1}{2}} N_1^{N_1-\frac{1}{2}} N^{\frac{1}{2}-N}}{\prod_{k=1}^K (2\pi)^{1/2} (p_{0k}N_0)^{p_{0k}N_0-\frac{1}{2}} (p_{1k}N_1)^{p_{1k}N_1-\frac{1}{2}} (p_k N)^{\frac{1}{2}-p_k N}} \\
&= \frac{(2\pi)^{1/2} N_0^{N_0-\frac{1}{2}} N_1^{N_1-\frac{1}{2}} (N)^{\frac{1}{2}-N}}{(2\pi)^{\frac{K}{2}} N_0^{N_0-\frac{K}{2}} N_1^{N_1-\frac{K}{2}} N^{\frac{K}{2}-N} \prod_{k=1}^K r_{0k}^{p_{0k}N_0-1/2} r_{1k}^{p_{1k}N_1-1/2} p_k^{-1/2}} \\
&= \left(\frac{N_0 N_1}{2\pi N}\right)^{\frac{K-1}{2}} \prod_{k=1}^K r_{0k}^{1/2-n_k^{(0)}} r_{1k}^{1/2-n_k^{(1)}} p_k^{1/2}. \qquad \text{(A1)}
\end{aligned}$$

Also

$$\begin{aligned}
\beta(\vec{n}+\alpha) &= \frac{\prod_{k=1}^K \Gamma(p_k N + \alpha_i)}{\Gamma(N + \sum_{k=1}^K \alpha_k)} \sim \frac{\prod_{k=1}^K (p_k N)^{\alpha_k} \Gamma(p_k N)}{N^{\sum \alpha_k} \Gamma(N)} \qquad \text{(A2)} \\
&= \frac{N^{\sum \alpha_k} \prod_{k=1}^K p_k^{\alpha_k}}{N^{\sum \alpha_k}} \beta(\vec{n}) = \beta(\vec{n}) \prod_{k=1}^K p_k^{\alpha_k} \qquad \text{(A3)}
\end{aligned}$$

where (A2) uses the approximation $\Gamma(x+c) \sim x^c \Gamma(x)$ as $x \to \infty$. Similarly,

$$\beta(\vec{n}^{(0)}+\alpha) \sim \beta(\vec{n}^{(0)}) \prod_{k=1}^K p_{0k}^{\alpha_k} \quad \text{and} \quad \beta(\vec{n}^{(1)}+\alpha) \sim \beta(\vec{n}^{(1)}) \prod_{k=1}^K p_{1k}^{\alpha_k}. \qquad \text{(A4)}$$

Therefore

$$\frac{\beta(\vec{n}+\alpha)}{\beta(\vec{n}^{(0)}+\alpha)\beta(\vec{n}^{(1)}+\alpha)} \sim \frac{\beta(\vec{n})}{\beta(\vec{n}^{(0)})\beta(\vec{n}^{(1)})} \prod_{k=1}^{K} \left(\frac{p_k}{p_{0k}p_{1k}}\right)^{\alpha_k} \tag{A5}$$

$$\sim \left(\frac{N_0 N_1}{2\pi N}\right)^{\frac{K-1}{2}} \prod_{k=1}^{K} p_k^{\alpha_k+1/2} r_{0k}^{1/2-n_k^{(0)}-\alpha_k} r_{1k}^{1/2-n_k^{(1)}-\alpha_k} \tag{A6}$$

Where (A5) follows from (A3) & (A4), and (A6) follows from (A1). It follows from (2) and (A6) that

$$\begin{aligned}\frac{\mathrm{pr}(H_0|C^{(0)},C^{(1)})}{\mathrm{pr}(H_1|C^{(0)},C^{(1)})} &= \frac{\beta(\alpha)P_0}{\mathrm{pr}(H_1)} \frac{\beta(\vec{n}+\alpha)}{\beta(\vec{n}^{(0)}+\alpha)\beta(\vec{n}^{(1)}+\alpha)} \\ &\sim cN^{\frac{K-1}{2}} \prod_{k=1}^{K} r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}\end{aligned}$$

where

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

A·2. *Proof of Corollary* 1

Proof. Let $\lambda_1 = 1-\lambda_0$ and consider

$$\log(r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}) = N\{p_k\log(p_k) - \lambda_0 p_{0k}\log(p_{0k}) - \lambda_1 p_{1k}\log(p_{1k})\}.$$

Under H_0, as $N\to\infty$,

$$p_{0k} = \pi_k + \frac{Z_0}{(N\lambda_0)^{1/2}} \quad \text{and} \quad p_{1k} = \pi_k + \frac{Z_1}{(N\lambda_1)^{1/2}},$$

where Z_0, Z_1 are independent standard normal variables. It follows that $\log(r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}) = A + B$, where

$$\begin{aligned}A &= N\pi_k\left\{\log\left(\pi_k + \frac{\lambda_0^{1/2}Z_0}{N^{1/2}} + \frac{\lambda_1^{1/2}Z_1}{N^{1/2}}\right) - \lambda_0\log\left(\pi_k + \frac{Z_0}{(N\lambda_0)^{1/2}}\right) - \lambda_1\log\left(\pi_k + \frac{Z_1}{(N\lambda_1)^{1/2}}\right)\right\} \\ &= N\pi_k\left\{\log\left(1 + \frac{\lambda_0^{1/2}Z_0}{\pi_k N^{1/2}} + \frac{\lambda_1^{1/2}Z_1}{\pi_k N^{1/2}}\right) - \lambda_0\log\left(1 + \frac{Z_0}{\pi_k(N\lambda_0)^{1/2}}\right) - \lambda_1\log\left(1 + \frac{Z_1}{\pi_k(N\lambda_1)^{1/2}}\right)\right\}\end{aligned}$$

and

$$\begin{aligned}B &= \left\{(N\lambda_0)^{1/2}Z_0 + (N\lambda_1)^{1/2}Z_1\right\}\log\left(\pi_k + \frac{\lambda_0^{1/2}Z_0}{N^{1/2}} + \frac{\lambda_1^{1/2}Z_1}{N^{1/2}}\right) \\ &\quad - (N\lambda_0)^{1/2}Z_0\log\left(\pi_k + \frac{Z_0}{(N\lambda_0)^{1/2}}\right) - (N\lambda_1)^{1/2}Z_1\log\left(\pi_k + \frac{Z_1}{(N\lambda_1)^{1/2}}\right) \\ &= \left\{(N\lambda_0)^{1/2}Z_0 + (N\lambda_1)^{1/2}Z_1\right\}\log\left(1 + \frac{\lambda_0^{1/2}Z_0}{\pi_k N^{1/2}} + \frac{\lambda_1^{1/2}Z_1}{\pi_k N^{1/2}}\right) \\ &\quad - (N\lambda_0)^{1/2}Z_0\log\left(1 + \frac{Z_0}{\pi_k(N\lambda_0)^{1/2}}\right) - (N\lambda_1)^{1/2}Z_1\log\left(1 + \frac{Z_1}{\pi_k(N\lambda_1)^{1/2}}\right).\end{aligned}$$

The Maclaurin expansion $\log(1+x) = \sum_{i=1}^{\infty}(-1)^{1+i}(x^i/i)$ gives

$$\begin{aligned} A &= 0 - \frac{1}{2\pi_k}\left\{\left(\lambda_0^{1/2}Z_0 + \lambda_1^{1/2}Z_1\right)^2 - Z_0^2 - Z_1^2\right\} + O_p\left(N^{-\frac{1}{2}}\right) \\ &= \frac{1}{2\pi_k}(\lambda_1^{1/2}Z_0 - \lambda_0^{1/2}Z_1)^2 + O_p\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

and

$$\begin{aligned} B &= \frac{1}{\pi_k}\left\{(\lambda_0^{1/2}Z_0 + \lambda_1^{1/2}Z_1)^2 - Z_0^2 - Z_1^2\right\} + O_p\left(N^{-\frac{1}{2}}\right) \\ &= -\frac{1}{\pi_k}(\lambda_1^{1/2}Z_0 - \lambda_0^{1/2}Z_1)^2 + O_p\left(N^{-\frac{1}{2}}\right). \end{aligned}$$

Thus,

$$\begin{aligned} \log(r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}) &= A + B \\ &= -\frac{1}{2\pi_k}(\lambda_1^{1/2}Z_0 - \lambda_0^{1/2}Z_1)^2 + O_p\left(N^{-\frac{1}{2}}\right) \\ &= -\frac{(\lambda_0\lambda_1)^{1/2}}{2\pi_k}N(p_0 - p_1)^2 + O_p\left(N^{-\frac{1}{2}}\right). \end{aligned}$$

It follows from Theorem 1 that

$$\frac{\mathrm{pr}(H_0|C^{(0)}, C^{(1)})}{\mathrm{pr}(H_1|C^{(0)}, C^{(1)})} \sim cN^{\frac{K-1}{2}} \prod_{k=1}^{K} \exp\left\{-\frac{(\lambda_0\lambda_1)^{1/2}}{2\pi_k}N(p_{0k} - p_{1k})^2\right\}$$

as $N \to \infty$, where $(\lambda_0\lambda_1)^{1/2}N(p_0 - p_1)^2 \sim \chi_1^2$. □

A·3. *Proof of Corollary* 2

Proof. Let $\lambda_1 = 1 - \lambda_0$, and $\pi_k^* = \lambda_0\pi_k^{(0)} + \lambda_1\pi_k^{(1)}$. As $N \to \infty$,

$$p_{0k} = \pi_k^{(0)} + O_p\left(N^{-\frac{1}{2}}\right), \qquad p_{1k} = \pi_k^{(1)} + O_p\left(N^{-\frac{1}{2}}\right) \quad \text{and} \quad p_k^* = \pi_k^* + O_p\left(N^{-\frac{1}{2}}\right).$$

Consider

$$\begin{aligned} \log(r_{0k}^{-n_k^{(0)}} r_{1k}^{-n_k^{(1)}}) &= N[p_k\log(p_k) - \lambda_0 p_{0k}\log(p_{0k}) - \lambda_1 p_{1k}\log(p_{1k})] \\ &= N\{\pi_k^*\log(p_k) - \lambda_0\pi_k^{(0)}\log(p_{0k}) - \lambda_1\pi_k^{(1)}\log(p_{1k})\} + O_p\left(N^{1/2}\right) \\ &= N\{\pi_k^*\log(\pi_k^*) - \lambda_0\pi_k^{(0)}\log(\pi_k^{(0)}) - \lambda_1\pi_k^{(1)}\log(\pi_k^{(1)})\} + O_p\left(N^{1/2}\right) \\ &= -N\left\{\lambda_0\pi_k^{(0)}\log\left(\frac{\pi_k^{(0)}}{\pi_k^*}\right) + \lambda_1\pi_k^{(1)}\log\left(\frac{\pi_k^{(1)}}{\pi_k^*}\right)\right\} + O_p\left(N^{1/2}\right). \end{aligned}$$

Thus, by Theorem 1,

$$\frac{\mathrm{pr}(H_0|C^{(0)}, C^{(1)})}{\mathrm{pr}(H_1|C^{(0)}, C^{(1)})} \sim cN^{\frac{K-1}{2}} \prod_{k=1}^{K} \left(\frac{\pi_k^{(0)}}{\pi_k^*}\right)^{-N\lambda_0\pi_k^{(0)}} \left(\frac{\pi_k^{(1)}}{\pi_k^*}\right)^{-N\lambda_1\pi_k^{(1)}} \exp\left\{O_p\left(N^{1/2}\right)\right\}.$$

A·4. *Proof of Theorem* 2

Proof. 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 totally bounded with respect to d. Hence 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. □

A·5. *Proof of Corollary* 3

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

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

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

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

by Theorem 2. □

A·6. *Proof of Remark* 2

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

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

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

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

An analogous result holds for any π_i, π_j with $i \neq j$. Therefore, if

$$f^* = \operatorname*{argmin}_{f \in \mathbb{F}} \mathrm{KL}(f_0||f)$$

with $\pi_k^* > 0$ for all k,

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

If $\pi_k^* = 0$ for some k, a similar argument shows that $\int \frac{f_k}{f^*} f_0$ must be equivalent for all $\pi_k^* > 0$. □

A·7. *Proof of Theorem* 3

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

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

Note that

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

So, for $(\hat{f}^{(0)}, \hat{f}^{(1)}) \in \mathbb{F}^2$ the KL-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 therefore the concentration conditions of Theorem 2 are satisfied. It follows from Corollary 3 that

$$\text{pr}(||\hat{\Pi}^{(0)} - \Pi^{*(0)}|| \geq \epsilon \mid X) \to 0\,,\ \text{pr}(||\hat{\Pi}^{(1)} - \Pi^{*(1)}|| \geq \epsilon \mid X) \to 0 \tag{A7}$$

for all $\epsilon > 0$.

Assume $f^{*(0)} \neq f^{*(1)}$. Fix $0 < \epsilon < ||\Pi^{*(0)} - \Pi^{*(1)}||$, and it follows from (A7) that

$$\text{pr}(||\hat{\Pi}^{(0)} - \hat{\Pi}^{*(1)}|| < \epsilon \mid X) \to 0.$$

This implies $\text{pr}(H_0 \mid X) \to 0$, as $\text{pr}(H_0 \mid X) < \text{pr}(||\hat{\Pi}^{(0)} - \hat{\Pi}^{*(1)}|| < \epsilon \mid X)$.

Assume $f^{*(0)} = f^{*(1)} = f^*$, 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 densiy for a Dir(α) distribution and $f(x|\Pi) = \sum_{k=1}^K \pi_k f_k$. For large N,

$$\begin{aligned}
\text{pr}(A_\delta, X|H_1) &= \int\int_{\Pi^{(0)},\Pi^{(1)} \in A_\delta} \prod_{i=1}^{N_0} f(x_i|\Pi^{(0)}) \prod_{j=1}^{N_1} f(x_j|\Pi^{(1)}) f_\alpha(\Pi^{(0)}) f_\alpha(\Pi^{(1)}) \\
&\leq \int\int_{\Pi^{(0)},\Pi^{(1)} \in A_\delta} \prod_{i=1}^{N_0} f(x_i|\Pi^{(0)}) \prod_{j=1}^{N_1} f(x_j|\Pi^{(0)}) f_\alpha(\Pi^{(0)}) f_\alpha(\Pi^{(1)}) \\
&= \text{pr}(A_\delta, X|H_0)\,\text{pr}(A_\delta|H_0),
\end{aligned}$$

and so

$$\begin{aligned}
\text{pr}(H_1|A_\delta, X) &= \frac{\text{pr}(H_1)\,\text{pr}(A_\delta, X|H_1)}{\text{pr}(H_1)\,\text{pr}(A_\delta, X|H_1) + P_0\,\text{pr}(A_\delta, X|H_0)} \\
&\leq \frac{\text{pr}(H_1)\,\text{pr}(A_\delta|H_0)}{\text{pr}(H_1)\,\text{pr}(A_\delta|H_0) + P_0}.
\end{aligned}$$

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

$$\text{pr}(H_1|A_\delta, X) \to 0 \text{ as } \delta \to 0. \tag{A8}$$

Result (A7) implies that for all $\delta > 0$,

$$\text{pr}(\bar{A}_\delta|X) \to 0 \text{ as } N \to 0. \tag{A9}$$

Fix $\epsilon > 0$. It follows from (A8) and (A9) that we may take δ sufficiently small to ensure

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

for large N. Therefore, $\text{pr}(H_0|X) \to 1$ as $N \to \infty$. □

B·8. *Pseudocode*

Here we give the details of the estimation procedure for the application to TCGA data, as introduced in Section 5. To estimate the kernel parameters $\theta_k = (\mu_k, 1/\sigma_k^2)$ we use the flexible normal-gamma prior

$$\text{NG}(\mu_0 = 0.5, \lambda_0 = 1, a_0 = 1, b_0 = 0.5).$$

After randomly selecting a subsample of 500 probes, the kernels are estimated via Gibbs sampling, where each iteration proceeds as follows:

- Draw $C^{(0)}, C^{(1)} \mid \Pi, X, \Theta$. The probability that for variable m subject n is allocated to kernel k is

$$\text{pr}(C_{mn} = k \mid X^{(i)}, \Pi^{(i)}, \Theta) \propto \pi_k f(X_{mn} \mid \mu_k, \sigma_k, [0,1]),$$

where $f(\cdot \mid \mu_k, \sigma_k, [0,1])$ is the density of a normal distribution with mean μ_k and standard deviation σ_k, truncated to fall within the interval $[0,1]$.

- Draw $\{\Pi_m\}_{m=1}^M \mid \mathbb{C}^{(0)}, \mathbb{C}^{(1)}$, where $\tilde{\pi}_{mk} \sim \text{Dir}(\alpha + \vec{n}_m)$.

$$\Pi^{(i)} = \text{pr}(H_0 \mid C^{(0)}, C^{(1)})\tilde{\Pi}_k + (1 - \text{pr}(H_0 \mid C^{(0)}, C^{(1)}))\tilde{\Pi}_k^{(i)}.$$

- Draw $\Theta \mid C, X$. The posterior distribution for θ_k, $k = 1, \ldots, K$ is

$$\theta_k \sim \text{NG}(\mu_{0k}, \lambda_{0k} = 1, a_{0k}, b_{0k}).$$

  Let X_k be the collection of values, over all probes, belonging to kernel k. To account for truncation, generate $\mathbb{Y}_k = \tilde{F}_k^{-1} F_k(X_k)$, where $\tilde{F}_k$ is the CDF for $\text{N}(\mu_k, \sigma_k)$ without truncation and F_k is the CDF with truncation. Let N_k be the total number of values belonging to kernel k, $\bar{Y}_k$ be the sample mean of $\mathbb{Y}_k$, and S_k the sample variance for $\mathbb{Y}_k$. The posterior normal-gamma parameters are
  - $\mu_{0k} = \frac{\lambda_0 \mu_0 + N_k \bar{Y}_k}{\lambda_0 + N_k}$
  - $\lambda_{0k} = \lambda_0 + N_k$
  - $a_{0k} = a_0 + \frac{N_k}{2}$
  - $b_{0k} = b_0 + \frac{N_k S_k}{2} + \frac{\lambda_0 N_k (\bar{Y}_k - \mu_0)^2}{2(\lambda_0 + N_k)}$.
- Estimate α from $\{\Pi_m\}_{m=1}^M$ as in Ronning (1989).

For each Gibbs iteration we relabel kernels if necessary to maintain the order $\mu_1 < \mu_2 < \ldots < \mu_K$. We average over the resulting Gibbs samples to obtain point estimates for $\{\theta_k\}_{k=1}^K$ and α.

For two-class testing, we put a uniform $\text{Be}(a_0 = 1, b_0 = 1)$ prior on P_0, and Gibbs sample as follows:

- Draw $C^{(0)}, C^{(1)} | \{\Pi_m^{(0)}, \Pi_m^{(1)}\}_{m=1}^M, X^{(0)}, X^{(1)}$ The probability that for variable m subject n in class i is realized from component k is

$$\text{pr}(C_{mn}^{(i)} = k \mid X_m^{(i)}, \Pi_m^{(i)}) \propto \pi_{mk}^{(i)} f(X_{mn}^{(i)} | \mu_k, \sigma_k, [0,1]).$$

- Compute $p_m = \text{pr}(H_{0m} \mid \mathbb{C}_m^{(0)}, \mathbb{C}_m^{(1)})$ as in Section 2.2, Equation (2) for $m = 1, \ldots, M$.
- Draw $\{\Pi_m^{(0)}, \Pi_m^{(1)}\}_{m=1}^M \| \mathbb{C}^{(0)}, \mathbb{C}^{(1)}, P_0$. For $\tilde{\Pi}_k \sim \text{Dir}(\alpha + \vec{n})$, $\tilde{\Pi}_k^{(0)} \sim \text{Dir}(\alpha + \vec{n}^{(0)})$, $\tilde{\Pi}_k^{(1)} \sim \text{Dir}(\alpha + \vec{n}^{(1)})$, for class $i = 0, 1$

$$\Pi_m^{(i)} = P_m \tilde{\Pi}_k + (1 - P_m)\tilde{\Pi}_k^{(i)}.$$

- Draw P_0 from $\text{Be}(1 + \sum_{m=1}^M P_m, 1 + M - \sum_{m=1}^M P_m)$.

B·9. *Methylation function*

We investigated the potential relevance of CpG identified in Section 5·3 by considering the expression of the gene at their genomic location. DNA methylation is thought to inhibit transcription and therefore repress gene expression. Of 2117 CpG sites with $P(H_{0m}|X) < 0.01$, 842 have both a significant negative association with gene expression (p-value < 0.01; Spearman's rank correlation) and significant differential gene expression between the Basal and non-Basal classes (p-value < 0.01; two-sample t-test). For these cases methylation gives a potential mechanistic explanation for well-known differences in gene transcription levels between Basal and non-Basal tumors. 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***. The spreadsheet `http://people.duke.edu/~el113/TCGA_ME-GE_results.csv` gives the posterior probability $P(H_{0m}|X)$ and corresponding gene expression statistics for all CpG sites.

B·10. *Data simulation*

For Section 6·2, we simulate 200 separate univariate datasets from the assumed model as follows:

1. Draw N uniformly on a log-scale from 10 to $1,000,000$.

2. Draw K uniformly from $\{2, \ldots, 9\}$.
3. Draw $\mu_1, \ldots, \mu_K$ independently from Uniform$[0, 1]$.
4. Draw $\sigma_1, \ldots, \sigma_K$ independently from Uniform$[0, \frac{1}{K}]$
5. Draw H_0 from Bernoulli(0.5)
6. If $H_0 = 1$
   - Draw Π from a uniform, K-dimensional Dirichlet distribution
   - For $n = 1, \ldots, N$ assign x_n to either class 0 or class 1 with equal probability.
   - Draw $x_1, \ldots, x_N \in \mathbb{X}$ from density

$$\sum_{k=1}^{K} \pi_k \text{Tnorm}(\mu_k, \sigma_k, [0, 1]),$$

     where Tnorm defines the density of a truncated normal distribution.
7. If $H_0 = 0$
   - Draw $\Pi^{(0)}$ and $\Pi^{(1)}$ independently from a uniform, K-dimensional Dirichlet distribution
   - For $n = 1, \ldots, N$ assign x_n to either class 0 or class 1 with equal probability.
   - Draw $x_1, \ldots, x_{N_0} \in \mathbb{X}^{(0)}$ from

$$\sum_{k=1}^{K} \pi_k^{(0)} \text{Tnorm}(\mu_k, \sigma_k, [0, 1])$$

   - Draw $x_1, \ldots, x_{N_1} \in \mathbb{X}^{(1)}$ from

$$\sum_{k=1}^{K} \pi_k^{(1)} \text{Tnorm}(\mu_k, \sigma_k, [0, 1]).$$

B·11. *Simulation study: Distribution recovery*

Here we investigate the ability of the two-class method to recover the generative distribution. We compare distribution recovery under the two-class model versus

1. Fitting a Dirichlet mixture model separately for each class, and
2. Fitting a Dirichlet mixture model that ignores class distinctions.

Ideally, we would like the two-class model to have similar performance to approach 1 under H_1 and similar performance to approach 2 under H_0.

We simulate 200 univariate datasets and estimate the posterior for the two-class model as in Section B·10. Separate and common models (approaches 1 and 2 above) are estimated analogously except for the dependence between classes. Figure B·11 shows the total variation distance between the mean posterior distribution and the generative distribution for each simulation and using each of the three estimation approaches. Not surprisingly, separate estimation performs much better under H_1 and common estimation performs marginally better under H_0. The flexible two-class approach performs similarly to common estimation under H_0 and separate estimation under H_1, even for smaller sample sets.

REFERENCES

ALBERT, J. H. (1997). Bayesian testing and estimation of association in a two-way contingency table. *J. Am. Statist. Assoc.* **92**, 685–693.

BOCK, C. (2012). Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719.

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**, 1259–1279.

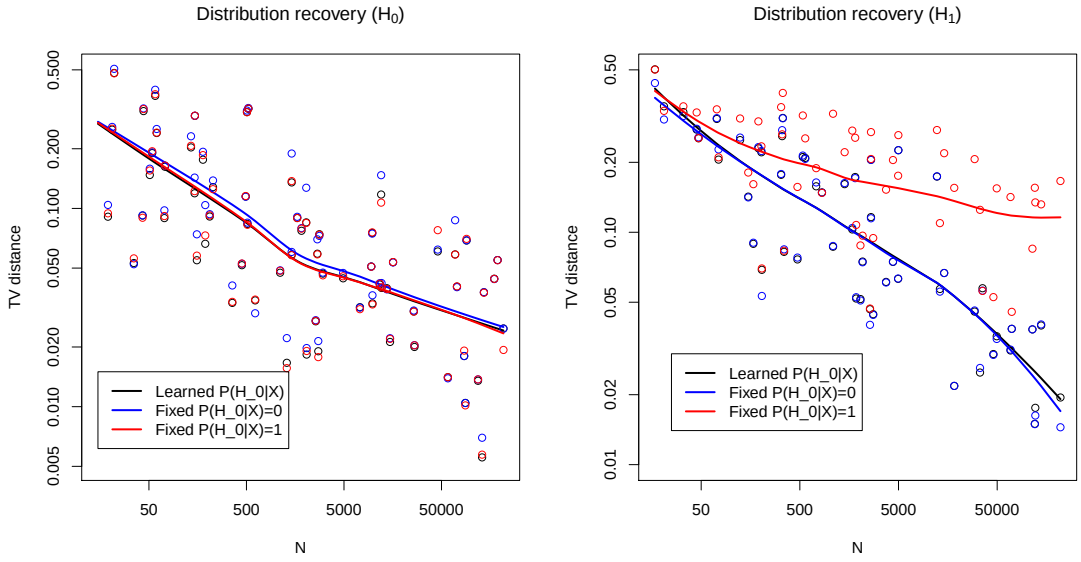

Fig. 6. Total variation distance between the estimated posterior and the generative model for 200 randomly generated simulations under H_0 (left plot) and H_1 (right plot). Results are shown for the two-class model (learned $P(H_0|X)$), separate estimation of each class (fixed $P(H_0|X) = 0$), and common estimation ignoring class labels (fixed $P(H_0|X) = 1$).

CURTIS, C., SHAH, S. P., CHIN, S.-F., TURASHVILI, G., RUEDA, O. M., DUNNING, M. J., SPEED, D., LYNCH, A. G., SAMARAJIWA, S., YUAN, Y. et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

DUNSON, D. B. & PEDDADA, S. D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859–874.

GOOD, I. & CROOK, J. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *Ann. Statist.* **15**, 670–693.

HOUSEMAN, E. A., CHRISTENSEN, B. C., YEH, R.-F., MARSIT, C. J., KARAGAS, M. R., WRENSCH, M., NELSON, H. H., WIEMELS, J., ZHENG, S., WIENCKE, J. K. & KESLEY, K. T. (2008). Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**, 365.

KLEIJN, B. J. & VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34**, 837–877.

KORMAKSSON, M., BOOTH, J. G., FIGUEROA, M. E. & MELNICK, A. (2012). Integrative model-based clustering of microarray methylation and expression data. *Ann. Appl. Statist.* **6**, 1327–1347.

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

LOCK, E. F. & DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29**, 2610–2616.

MA, L. & WONG, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Am. Statist. Assoc.* **106**, 1553–1565.

MULLER, P., PARMIGIANI, G. & RICE, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*, J. Bernardo, R. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, eds. Oxford: Oxford University Press, pp. 349–370.

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. *Journal of Clinical Oncology* **27**, 1160–1167.

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

RONNING, G. (1989). Maximum likelihood estimation of Dirichlet distributions. *J. Stat. Comput. Simul.* **32**, 215–221.

SCHOLZ, F. & STEPHENS, M. (1987). K-sample Anderson–Darling tests. *J. Am. Statist. Assoc.* **82**, 918–924.

SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136**, 2144–2162.

SCOTT, J. G., BERGER, J. O. et al. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–2619.

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. Environ. Stat.* **15**, 308–326.

YAKOWITZ, S. J., SPRAGINS, J. D. et al. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39**, 209–214.

ZHANG, L., MENG, J., LIU, H. & HUANG, Y. (2012). A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles. *BMC Genomics* **13**, S20.