Bayesian Analysis (2017) **12**, Number 1, pp. 89–112

# Bayesian Nonparametric Tests via Sliced Inverse Modeling

Bo Jiang[§*], Chao Ye[¶†], and Jun S. Liu[‖‡]

**Abstract.** We study the problem of independence and conditional independence tests between categorical covariates and a continuous response variable, which has an immediate application in genetics. Instead of estimating the conditional distribution of the response given values of covariates, we model the conditional distribution of covariates given the discretized response (aka "slices"). By assigning a prior probability to each possible discretization scheme, we can compute efficiently a Bayes factor (BF)-statistic for the independence (or conditional independence) test using a dynamic programming algorithm. Asymptotic and finite-sample properties such as power and null distribution of the BF statistic are studied, and a stepwise variable selection method based on the BF statistic is further developed. We compare the BF statistic with some existing classical methods and demonstrate its statistical power through extensive simulation studies. We apply the proposed method to a mouse genetics data set aiming to detect quantitative trait loci (QTLs) and obtain promising results.

**AMS 2000 subject classifications:** primary 62G10; secondary 62C10, 62P10.

**Keywords:** Bayes factor, dynamic programming, non-parametric tests, sliced inverse model, variable selection.

## 1 Introduction

Statistical tools for analyzing data sets with categorical covariates and continuous response have been extensively used in many areas such as genetics, clinical trials, social science, and Internet commerce. By grouping individual observations according to combinatoric configurations of covariates, classical regression-based methods are derived from conditional models of response given configurations of covariates. Recent demands for analyzing large-scale, high-dimensional data sets pose new challenges to these traditional methods. For example, in quantitative trait loci (QTL) mapping (Lander and

*Two Sigma Investments, LLC, 100 Avenue of the Americas, Floor 16, New York, NY 10013, USA, bojiang83@gmail.com

†MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China, yechao1009@gmail.com

‡Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA, jliu@stat.harvard.edu

§The views expressed herein are the authors alone and are not necessarily the views of Two Sigma Investments, LLC or any of its affliates.

¶This research was supported in part by the National Basic Research Program of China (2012CB316504).

‖This research was supported in part by NSF grants DMS-1007762 and DMS-1120368, and NIH R01 GM113242-01.

 DOI: 10.1214/16-BA993

Schork, 1994; Brem et al., 2002; Morley et al., 2004), scientists wish to discover genomic loci associated with a continuous quantitative trait such as human height, crop yield, or gene expression level, by sequencing hundreds or thousands of genetic markers (encoded as categorical variables) on a genome-wide scale. Regression-based approaches such as analysis of variance (ANOVA) are sensitive to distributional assumption of quantitative traits, and ineffective in detecting individual markers with heteroscedastic or other higher-order effects. Recently, Aschard et al. (2013) proposed a non-parametric method to test whether the distribution of quantitative traits differs by genotypes of a genetic marker. However, for complex traits such as human height, some important genetic markers may have different effects in combination than individually (i.e., the *epistasis* effect in genetic terminology) and thus need to be considered jointly. The number of possible genotype configurations grows exponentially with the number of genetic markers under consideration and, furthermore, markers located on the same chromosome can be highly correlated. Even with a sample size of several hundreds and a moderate number of genetic markers, it is very likely that some genotype configurations contain very few or even no observations. Traditional non-parametric methods have limited power in such situations.

In the past two decades, there has been a considerable interest in studying non-parametric testing problems from a Bayesian perspective. Several methods have been proposed for testing a parametric model versus a non-parametric alternative (the goodness-of-fit problem). For some notable examples see Carota and Parmigiani (1996); Florens et al. (1996); Berger and Guglielmi (2001); Basu and Chib (2003); Hanson (2006); McVinish et al. (2009). As for Bayesian non-parametric two-sample test, Holmes et al. (2015) introduced a method to compute the Bayes factor for testing the null through the marginal likelihood of the data with Pólya tree priors. Recently, Ma and Wong (2011) developed a testing method for two-sample problems based on coupling optional Pólya tree prior, which can simultaneously learn the partition of the data. In this paper, instead of modeling probability distribution on infinite dimensional objects (that is, modeling the conditional distribution of a continuous variable given population indicator), we propose to model the frequencies of population indicator conditional on a discretization of the continuous response variable, and develop a dynamic programming algorithm to compute the test statistic with computational complexity quadratic or even linear in sample size.

The *inverse* modeling perspective is motivated by the naïve Bayes method and the "Bayesian epistatic association mapping" (BEAM) model of Zhang and Liu (2007). BEAM was developed to detect epistatic interactions in genome-wide case-control association studies. Both methods prefer the response variable to be discrete, and their extensions to cases with a continuous response often relies on an *ad hoc* discretization strategy. Recently, Jiang et al. (2015) proposed a non-parametric $K$-sample test from the inverse modeling perspective and developed a dynamic slicing (DS) algorithm to determine the optimal discretization (aka "slicing") that maximizes a regularized likelihood. In this paper, we employ a full-Bayesian view on the inverse modeling approach and further generalize the framework to testing conditional independence between a continuous response and a categorical variable given a set of (previously selected) categorical variables. Instead of constructing test statistics based on regularized likelihood ratios as

in DS of Jiang et al. (2015), we calculate the Bayes Factor (BF) by marginalizing over all possible slicing schemes under the inverse model. From numerical studies, we observed that the BF approach has a superior power in detecting both unconditional and conditional dependences compared with DS and other non-parametric testing methods. The proposed conditional dependence test is further used to construct a stepwise searching strategy for categorical variable selection and applied to QTL mapping analysis.

The rest of this paper is organized as follows. In Section 2.1, we construct a non-parametric test based on the Bayes Factor of a sliced inverse model, which we refer to as the BF statistic, to detect the conditional dependence between a continuous response and a categorical covariate. An efficient dynamic programming algorithm is developed in Section 2.2 to compute the BF statistic and its asymptotic and finite-sample properties are studied in Section 2.3. We investigate the sensitivity of the BF statistic to choices of hyper-parameters and fit an empirical formula that links the value of the BF with type-I error in Sections 2.4 and 2.5, respectively. A forward stepwise variable selection procedure based on the proposed BF statistic is described in Section 2.6. In Section 3, we use simulations to evaluate the powers of different methods for both unconditional and conditional dependence tests, and compare the BF statistic with classic stepwise regression in detecting interaction on synthetic QTL data sets. In Section 4, we further illustrate the proposed methodology on a mouse QTL data set and demonstrate its advantage over traditional QTL mapping methods. Additional remarks in Section 5 conclude the paper. Proofs of the theorems and other technical derivations are provided in the online supplement (Jiang et al., 2016).

## 2 An inverse model for non-parametric dependence test

Suppose $Y$ is a continuous response variable, and both $X$ and $Z$ are categorical variables with $|X|$ and $|Z|$ levels, respectively. Assume that we have known that $Y$ is dependent of $Z$. Note that $Z$ can be a "super" variable if there are more than one actual variables having been previously selected, in which case we encode each possible configuration of the selected variables as a level of $Z$. For example, if we have selected $Z_1$ and $Z_2$, both of which have support in $\{0, 1\}$, then, we can define $Z = Z_1 + 2Z_2$. Define $Z \equiv 0$ (and thus $|Z| = 1$) if we are interested in testing the marginal independence between $Y$ and $X$. We consider the following hypothesis testing problem:

$$\begin{aligned} & H_0 \quad : X \text{ and } Y \text{ are conditionally independent given } Z \\ \text{v.s.} \quad & H_1 \quad : X \text{ and } Y \text{ are not conditionally independent given } Z \end{aligned}$$

Note again that all the results in this section is directly applicable to testing the unconditional dependence between $X$ and $Y$ (i.e., by letting $Z \equiv 0$).

Suppose $\{(x_i, y_i, z_i)\}_{i=1}^n$ are independent observations of $(X, Y, Z)$. Without loss of generality, henceforth we assume that observations have been sorted according to the $Y$ values so that $y_i = y_{(i)}$. We divide the sorted list of observations into slices and define a function $S(y_i)$ taking values in $\{1, 2, \ldots, |S|\}$ as the slice membership of $y_i$, where $|S|$ denotes the total number of slices. Under the null hypothesis, the conditional

distribution of $X$ given $Z$ does not depend on $Y$ and

$$X \mid \ Y = y, Z = j \sim \text{Multinomial}\left(1, p_j\right), \tag{1}$$

where $p_j = (p_{j,1}, \ldots, p_{j,|X|})$ and $\sum_{k=1}^{|X|} p_{j,k} = 1$ for $j = 1, \ldots, |Z|$. Under the alternative hypothesis, the distribution of $X$ conditional on $Z = j$ and $S(Y) = h$ $(1 \le h \le |S|)$ is given by

$$X \mid \ Z = j, S(Y) = h \sim \text{Multinomial}\left(1, p_j^{(h)}\right), \tag{2}$$

where $p_j^{(h)} = (p_{j,1}^{(h)}, \ldots, p_{j,|X|}^{(h)})$ and $\sum_{k=1}^{|X|} p_{j,k}^{(h)} = 1$ for $j = 1, \ldots, |Z|$ and $h = 1, \ldots, |S|$. Jiang et al. (2015) proposed a dynamic slicing (DS) statistic to test the above hypotheses with $Z \equiv 0$ based on a regularized likelihood ratio. The DS statistic can be generalized to test conditional dependence with $|Z| > 1$, but the number of parameters in the model increases dramatically as $|Z|$ increases, which may impair the power of the method. The details of the generalized DS statistic, algorithms and theoretical results are provided in the online supplement (Jiang et al., 2016). Here, we explore a different testing approach based on the Bayes factor (BF). In Section 3, we will show that the proposed BF statistic consistently outperforms the DS statistic under a variety of scenarios in simulations.

## 2.1 Bayes factor under inverse model

Under the null model (1) and the alternative model (2), we further assume the following priors on $p_j$ and $p_j^{(h)}$ (whose dimensionalities are $|X|$), respectively:

$$p_j \sim \text{Dirichlet}\left(\frac{\alpha_0}{|X|}, \ldots, \frac{\alpha_0}{|X|}\right), \tag{3}$$

and

$$p_j^{(h)} \sim \text{Dirichlet}\left(\frac{\alpha_0}{|X|}, \ldots, \frac{\alpha_0}{|X|}\right),$$

where $\alpha_0 > 0$ is a hyper-parameter. First, we randomly draw a discretization of $Y$, $\{S(y_i)\}_{i=1}^n$, and then conditional on this discretization, the distribution of $X$ then depends jointly on $Z$ and on the slice containing $Y$. With a slight abuse of notation, we let $\Pr_{H_1}(X \mid S(Y), Z)$ denote the shorthand of the probability of observing $\{X_i = x_i\}_{i=1}^n$ under $H_1$ given $\{z_i\}_{i=1}^n$ and the slicing scheme $\{S(y_i)\}_{i=1}^n$. After integrating out $p_j^{(h)}$, we can write down the probability

$$\Pr_{H_1}\left(X \mid S(Y), Z\right) = \prod_{j=1}^{|Z|} \prod_{h=1}^{|S|} \left[ \frac{\Gamma(\alpha_0)}{\Gamma\left(\alpha_0 + n_j^{(h)}\right)} \prod_{k=1}^{|X|} \frac{\Gamma\left(n_{j,k}^{(h)} + \frac{\alpha_0}{|X|}\right)}{\Gamma\left(\frac{\alpha_0}{|X|}\right)} \right],$$

where $n_{j,k}^{(h)}$ is the number of observations with $z_i = j$, $x_i = k$ and $S(y_i) = h$, and $n_j^{(h)} = \sum_{k=1}^{|X|} n_{j,k}^{(h)}$ is the number of observations with $z_i = j$ and $S(y_i) = h$. Similarly,

since $X$ is independent of any slicing of $Y$ conditional on $Z$ under $H_0$, by integrating out $p_j$ we have

$$\Pr_{H_0}(X \mid Y, Z) = \Pr_{H_0}(X \mid Z) = \prod_{j=1}^{|Z|} \left[ \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \prod_{k=1}^{|X|} \frac{\Gamma\left(n_{j,k} + \frac{\alpha_0}{|X|}\right)}{\Gamma\left(\frac{\alpha_0}{|X|}\right)} \right],$$

where $n_{j,k}$ is the number of observations with $z_i = j$ and $x_i = k$, and $n_j = \sum_{k=1}^{|X|} n_{j,k}$ is the number of observations with $z_i = j$.

Given $n$ observations ranked by their response values, we denote the collection of all possible slicing schemes as $\Omega_n(S)$ and the probability for choosing a slicing scheme $S(\cdot)$ from $\Omega_n(S)$ *a priori* as $\Pr(S(Y))$. For a slicing scheme $S(\cdot)$ with $|S|$ slices, we assume here that

$$\Pr(S(Y)) = \pi_0^{|S|-1}(1-\pi_0)^{n-|S|}. \tag{4}$$

That is, with probability $\pi_0$, a "slice" is "inserted" independently between the $i$th and $(i+1)$th ranked observations, for $i = 1, \ldots, n-1$. Given $n$ observations, we re-parameterize the prior as $\pi_0 \equiv 1/(1+n^{\lambda_0})$ so that on the log-odds scale we have:

$$\log(\pi_0/(1-\pi_0)) \equiv -\lambda_0 \log(n). \tag{5}$$

Under $H_1$ and the slicing prior in (4), we have

$$\Pr_{H_1}(X|Z,Y) = \sum_{S(Y)\in\Omega_n(S)} \Pr_{H_1}(X|Z, S(Y)) \Pr(S(Y)). \tag{6}$$

Finally, the BF statistic for comparing the model under the alternative hypothesis and the null is defined as the Bayes factor for testing $H_1$ against $H_0$:

$$\mathrm{BF}(X|Z,Y) = \frac{\Pr_{H_1}(X|Z,Y)}{\Pr_{H_0}(X|Z,Y)} = \sum_{S(Y)\in\Omega_n(S)} \mathrm{BF}(X|Z,S(Y)) \Pr(S(Y)), \tag{7}$$

where

$$\mathrm{BF}(X|Z,S(Y)) = \frac{\Pr_{H_1}(X|Z,S(Y))}{\Pr_{H_0}(X|Z,Y)}.$$

We will describe an efficient algorithm to calculate the BF statistic (7) next.

## 2.2 A dynamic programming algorithm

To avoid a bruteforce summation over $2^{n-1}$ possible slicing schemes in $\Omega_n(S)$, we use a dynamic programming algorithm to calculate $\mathrm{BF}(X|Z,Y)$ in (7) as follows:

**Algorithm 1.**

- *Step* 1: Rank observations according to the observed values of $Y$, $\{y_i\}_{i=1}^n$. Slicing is only allowed along the ranked list of observations.

- *Step* 2*:* For $1 \le s \le t \le n$, calculate

$$\psi_{s,t} = \prod_{j=1}^{|Z|} \left[ \frac{\Gamma(\alpha_0)}{\Gamma\left(\alpha_0 + n_j^{(s:t)}\right)} \prod_{k=1}^{|X|} \frac{\Gamma\left(n_{j,k}^{(s:t)} + \frac{\alpha_0}{|X|}\right)}{\Gamma\left(\frac{\alpha_0}{|X|}\right)} \right],$$

  where $n_{j,k}^{(s:t)}$ is the number of observations with $z_i = j$ and $x_i = k$ for $s \le i \le t$, and $n_j^{(s:t)} = \sum_{k=1}^{|X|} n_{j,k}^{(s:t)}$ is the number of observations with $z_i = j$ for $s \le i \le t$.

- *Step* 3: Fill in entries of the table $\{w_t\}_{t=1}^n$ (define $w_1 \equiv (1-\pi_0)/\pi_0$) recursively for $t = 2, \ldots, n$,

$$w_t = \sum_{s=2}^{t} w_{s-1}(1-\pi_0)^{t-s+1} \left[ \frac{\psi_{1,s-1}\psi_{s,t}}{\psi_{1,t}} \right],$$

  where $w_s$ stores a partial sum of (7) until the $s$th ranked observation. Then,

$$\mathrm{BF}\left(X|Z,Y\right) = w_n.$$

The computational complexity of the dynamic slicing algorithm is $O(n^2)$.

## 2.3 Asymptotic and finite-sample properties of the BF statistic

In this section, we evaluate the frequency properties of the BF statistic under different priors on data generating mechanisms. Specifically, we derive theoretical bounds on type-I errors of the BF statistic under three different generative schemes of replicate data: conditional permutation (where replicate data are generated by permuting group indicators conditioning on observed values), unconditional sampling under the sharp null (where the frequency parameters of replicate data are generated from a degenerate point prior), and unconditional sampling under the hierarchical null (where the frequency parameters of replicate data are generated from arbitrary distributions with bounded densities). We show that, as sample size $n \to \infty$ and with appropriate choice of hyper-parameters, the BF statistic is almost surely smaller than or equal to 1 for all three data generating schemes under the null. Moreover, when the alternative hypothesis $H_1$ is true, we prove that the BF statistic goes to infinity at an exponential rate with the increase of sample size.

Given $n$ observations $\{x_i, y_i, z_i\}_{i=1}^n$, we can calculate the conditional null distribution of the observed BF statistic by shuffling the observed values of $X$ within each group of observations indexed by $\{i : z_i = j\}$, independently for $j \in \{1, \ldots, |Z|\}$. We call this shuffling scheme the conditional permutation null. Let $\Pr_{\text{shuffle}}(\mathrm{BF}(X|Y,Z) > b)$ denote the probability of observing a BF value of $b$ or larger using the conditional permutation scheme. We prove the following theorem in the online supplement (Jiang et al., 2016).

**Theorem 1.** *Assume that the hyper-parameter* $0 < \alpha_0 \le |X|$ *and observed sample size* $n \ge |X|$*. There exists a constant* $C_1 > 0$*, which only depends on* $|X|$ *and* $|Z|$*, such that*

$$\Pr_{\text{shuffle}}\left(\mathrm{BF}\left(X|Y,Z\right) > b\right) \le C_1 n^{|Z|(|X|-1)} \min\left\{ \frac{1}{(\log(b)+1)n^{\lambda_0-3}}, \frac{1}{b} \right\},$$

*for any* $b \geq 1$ *and* $\lambda_0$ *as defined in (5). Thus,*

$$\Pr_{\text{shuffle}}\left(\text{BF}\left(X|Y,Z\right)>1\right) \leq \frac{C_1}{n^{\lambda_0-|Z|(|X|-1)-3}}$$

*and* $\text{BF}(X|Y,Z) \leq 1$ *a.s. as* $n \to \infty$ *for* $\lambda_0 > |Z|(|X|-1)+4$.

The above definition of type-I error is conditioning on $\{n_{j,k} : j = 1,\ldots,|Z|, k = 1,\ldots,|X|\}$, *i.e.* the number of observations with $z_j = j$ and $x_i = k$. When the total number of observations $n$ is large, the conditional permutation null can be approximated by the following *sharp* null hypothesis:

$$H_0^{\text{sharp}} : X|Y = y, Z = j \sim \text{Multinomial}\left(1, p_j\right), j = 1,\ldots,|Z|,$$

where $p_j$'s are fixed but unknown. The following corollary, whose proof is given in the online supplement (Jiang et al., 2016), provides a similar finite-sample bound on unconditional type-I error under the sharp null hypothesis.

**Corollary 1.** *Assume that the hyper-parameter* $0 < \alpha_0 \leq |X|$. *When the sharp null hypothesis* $H_0^{\text{sharp}}$ *is true, there exists a constant* $C_2 > 0$, *which only depends on* $|X|$ *and* $|Z|$, *such that*

$$\Pr_{\text{sharp}}\left(\text{BF}\left(X|Y,Z\right)>b\right) \leq C_2 n^{|Z|(|X|-1.5+\gamma_0)} \min\left\{\frac{1}{(\log(b)+1)n^{\lambda_0-3}}, \frac{1}{b}\right\}.$$

*for any* $b \geq 1$, *where* $\gamma_0 = 0.57722\ldots$ *is the Euler-Mascheroni constant and* $\lambda_0$ *is defined in (5). Thus,*

$$\Pr_{\text{sharp}}\left(\text{BF}\left(X|Y,Z\right)>1\right) \leq \frac{C_2}{n^{\lambda_0-|Z|(|X|-1.5+\gamma_0)-3}}$$

*and* $\text{BF}(X|Y,Z) \leq 1$ *a.s. as* $n \to \infty$ *for* $\lambda_0 > |Z|(|X|-1.5+\gamma_0)+4$.

The sharp null hypothesis assumes that the nuisance frequency parameters $\{p_j\}_{j=1}^{|Z|}$ are fixed but unknown. We may further consider a hierarchical data generating scheme where the frequency parameters are sampled from some unknown distributions with bounded densities. This is especially relevant when we repeat the dependence test on a collection of covariates (*e.g.* genetic markers) with the same number of categories but varying marginal frequencies. Specifically, we consider the hierarchical null hypothesis as follows:

$$\begin{aligned} H_0^{\text{hierar}} : \quad & X|Y = y, Z = j \sim \text{Multinomial}\left(1, p_j\right), \\ & \text{and } p_j \sim f_j\left(p_j\right), j = 1,\ldots,|Z|, \end{aligned}$$

where there exists $M_0 > 0$ such that the unknown prior density $f_j(p_j) \leq M_0$ for $j = 1,\ldots,|Z|$. When $H_0^{\text{hierar}}$ is true, the Dirichlet prior (3) with $\alpha_0 \leq |X|$ has a positive probability of overlapping with the *true* distribution of $p_j$. Thus, we can obtain a tighter type-I error bound under the hierarchical null hypothesis according to the following corollary proved in the online supplement (Jiang et al., 2016).

**Corollary 2.** *Assume that the hyper-parameter* $0 < \alpha_0 \leq |X|$*. When the hierarchical null hypothesis* $H_0^{\text{hierar}}$ *is true, there exists a constant* $C_3 > 0$*, which only depends on* $|X|$ *and* $|Z|$*, such that*

$$\Pr_{\text{hierar}}\left(\text{BF}\left(X|Y,Z\right) > b\right) \leq C_3 \min\left\{\frac{1}{(\log(b)+1)n^{\lambda_0-3}}, \frac{1}{b}\right\},$$

*for any* $b \geq 1$ *and* $\lambda_0$ *as defined in (5). Thus,*

$$\Pr_{\text{hierar}}\left(\text{BF}\left(X|Y,Z\right) > 1\right) \leq \frac{C_3}{n^{\lambda_0-3}},$$

*and* $\text{BF}(X|Y,Z) \leq 1$ *a.s. as* $n \to \infty$ *for* $\lambda_0 > 4$ *and* $\alpha_0 \leq |X|$*.*

Notably, compared with Theorem 1 and Corollary 1, the finite-sample bound in Corollary 2 depends on the number of categories $|X|$ and $|Z|$ only through $C_3$, which is a constant with respect to sample size $n$ and cutoff $b$.

Next, we show that under $H_1$, $\text{BF}(X|Y,Z)$ goes to infinity with an exponential rate proportional to the sample size and the conditional mutual information between $X$ and $Y$ given $Z$, $\text{MI}(X,Y|Z)$.

**Theorem 2.** *Assume that hyper-parameters* $\alpha_0$ *and* $\lambda_0$ *as defined in (3) and (5) satisfying* $0 < \alpha_0 \leq |X|$*,* $\lambda_0 \geq 1$ *and* $\lambda_0 = o(n^{\frac{1}{3}}/\log(n))$*. Under the regularity condition in the online supplement (Jiang et al., 2016),*

$$\Pr\left(\text{BF}\left(X|Y,Z\right) \geq e^{n[\text{MI}(X,Y|Z)-\delta(n)]}\right) \geq 1 - 4n^{-\frac{1}{32}\log(n)},$$

*where*

$$\delta(n) = O\left(\frac{(\lambda_0 + |Z|(|X| - 1.5 + \gamma_0)/3)|Z|\log(n)}{n^{1/3}}\right) \to 0$$

*as* $n \to \infty$*. Thus,* $\text{BF}(X|Y,Z) \geq e^{n[\text{MI}(X,Y|Z)-\epsilon]}$ *a.s. for any* $\epsilon > 0$ *as* $n \to \infty$*.*

Note that the conditional mutual information $\text{MI}(X,Y|Z) > 0$ if and only if $X$ and $Y$ are not conditionally independent given $Z$. Theorem 1 and 2 guarantee the consistency of the BF statistic in testing dependence given any finite threshold.

The requirement of $\lambda_0 \geq 1$ in Theorem 2 is sufficient but not necessary. In Section 2.4, through simulation studies, we show that the BF statistic can approach infinity as sample size increases under some $\lambda_0 < 1$. However, when the value of $\lambda_0$ is small enough, the BF statistic will converge to zero as shown in Figure 1. Intuitively, this phenomenon can be explained by the fact that too much weight is given to configurations with bad slicings (swamped by the "entropy" effect). On the other hand, when $\lambda_0$ as defined in (5) is large relative to the sample size (implying a very small $\pi_0$), the $\delta(n)$ term in Theorem 2 will no longer converge to zero and the BF statistic will not be able to differentiate $H_1$ from $H_0$. For example, one can show that when $\lambda_0 = O(n)$, the BF statistic $\text{BF}(X|Y,Z) \to 1$ almost surely as $n \to \infty$. Unlike the DS statistic, which is monotonically increasing as $\lambda_0$ becomes smaller, the relationship between the BF statistic and the hyper-parameter $\lambda_0$ is not monotonic. In the following section, we study the sensitivity of the BF statistic and its type-I error to the choice of $\lambda_0$ based on numerical simulations.

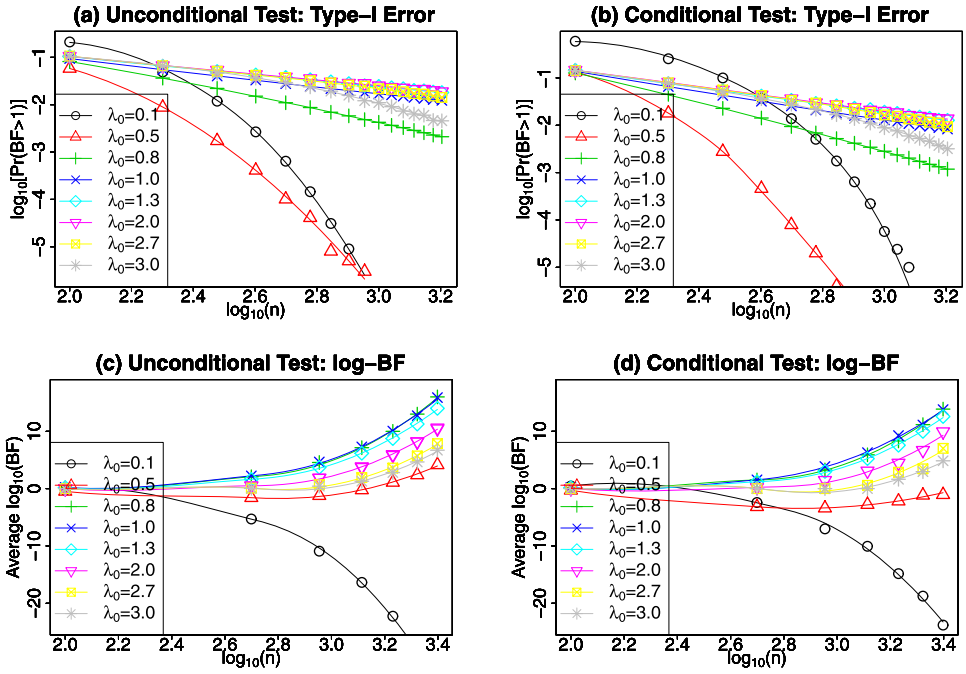


Figure 1: Sensitivity of type-I error (corresponding to a cutoff of 1) to different values of $\lambda_0$ for unconditional (a) and conditional (b) BF statistics, and average logarithm of unconditional (c) and conditional (d) BF statistics given different values of $\lambda_0$ when an alternative hypothesis is true. The lines connecting the points are from the LOESS fit. We fixed $\alpha_0 = 1$ in this analysis.

## 2.4 Choice of hyper-parameter $\lambda_0$ and $\sqrt{n}$-partition prior

Theorems 1 and 2 suggests that we should generally choose hyper-parameters $\alpha_0$ and $\lambda_0$ as defined in (3) and (5) such that $\alpha_0 \leq |X|$ and $\lambda_0 \geq 1$. Using numerical simulations, we further study the sensitivity of the BF statistic and its type-I error to the choice of hyper-parameter $\lambda_0$ in both unconditional and conditional dependence tests. For unconditional test, we generate equal number observations with binary indicator $X = 0$ and $X = 1$, and simulate the continuous response $Y \sim N(\mu X, 1)$ with $\mu = 0.4$. For conditional test, we generate binary covariates $X$ and $Z$ independently, and simulate the response $Y \sim N(\mu X + \mu Z, 1)$ with $\mu = 0.4$. We calculate the average logarithmic value of BF statistics under the alternative hypothesis, as well as type-I error of BF statistic given a cutoff of 1 under the null hypothesis, which is obtained by shuffling the observed values of $X$ (while retaining the association between $Z$ and $Y$ for conditional test). We use $\alpha_0 = 1$ in all the simulations in this section and will conduct a sensitivity analysis on $\alpha_0$ in Section 3.1.

Figure 1 shows the type-I error (given a cutoff of 1) and average logarithmic value of BF statistic under varying sample size $n$. As we can see, type-I errors under different

sample sizes are insensitive to a wide range of $\lambda_0$ from 1 to 3. Furthermore, under the alternative hypothesis, values of BF statistics are comparable for choice of $\lambda_0$ between 0.8 and 1.3. We also observed that the type-I error of the critical region $\{BF > b\}$ is not monotonic to the value of $\lambda_0$. For example, given the same sample size $n$ and a critical value $b = 1$, the BF statistic with $\lambda_0 = 2$ has a larger type-I error than that with $\lambda_0 = 1$ and $\lambda_0 = 3$. On the other hand, the BF statistic with $\lambda_0 = 2$ is on average smaller than the BF statistic with $\lambda_0 = 1$ when $X$ and $Y$ are (unconditionally or conditionally) dependent. Finally, we note that when $\lambda_0$ is too small (*e.g.*, 0.1), which results in a relatively large $\pi_0$ ($\approx n^{-\lambda_0}$) and a large number of expected slices (*i.e.*, $n^{1-\lambda_0}$), the logarithm of BF tends to negative infinity even when $H_1$ is true. Furthermore, it appears that as we vary $\lambda_0$ from 1 to 0, the "phase transition" phenomena (*i.e.*, the logarithm of BF diverges to positive infinity versus negative infinity) occurs at around 0.5. Given these observations, unless noted otherwise, we will choose $\lambda_0 = 1$ and $\alpha_0 = 1$ (see Section 3.1 for simulation results using different $\alpha_0$'s) for the following studies. The prior proposed in (4) has an independent prior to slice between each pair of observations and allows "thin" slices containing very few observations. We have seen that if we choose $\lambda_0$ too small, allowing too many "thin" slices may reduce the power of the method. When $n$ is large enough, we can first divide ranked observations into $\lfloor\sqrt{n}\rfloor$ bins such that each bin contains approximately $\lfloor\sqrt{n}\rfloor$ observations. Then, we can define a test statistic to have the same form as (7) except that the summation is taken over slicing schemes restricted on the fixed $\lfloor\sqrt{n}\rfloor$ bins (slicing is not allowed within a bin), and we call this variant of the method BF with $\sqrt{n}$-partition prior. Note that by using the $\sqrt{n}$-partition prior, we can further reduce the computational complexity of the dynamic programming algorithm from $O(n^2)$ to $O(n)$.

## 2.5 Empirical formulas for type-I errors

Theorem 1 provides finite-sample bounds for type-I errors under the conditional shuffling scheme, that is,

$$\Pr_{\text{shuffle}}\left(\text{BF}\left(X|Y,Z\right) > b\right) \leq C_1 n^{|Z|(|X|-1)} \min\left\{\frac{1}{(\log(b)+1)n^{\lambda_0-3}}, \frac{1}{b}\right\}.$$

Based on numerical simulations, we found that the relationship between the value of BF statistic and its significance level can be further refined. Specifically, we simulate observations for both conditional and unconditional tests using the same procedure as described in Section 2.4 with binary covariate $X$ (or $Z$) and varying sample size $n$. Then, we calculate the BF statistic with hyper-parameters $\lambda_0 = 1$ and $\alpha_0 = 1$ on shuffled samples. For unconditional test (*i.e.* $Z \equiv 0$), we obtain the following empirical formula of type-I error given the cutoff of BF statistic $b$ and sample size $n$:

$$\Pr_{\text{shuffle}}\left(\text{BF}\left(X|Y,Z\equiv 0\right) > b\right) \approx \frac{\gamma_p}{b^{\alpha_p} n^{\beta_p}}, \tag{8}$$

where $\alpha_p$, $\beta_p$ and $\gamma_p$ only depend on $p$, the proportion of observations with $X = 1$. For example, when $p = 0.5$, $\alpha_{0.5} \approx 1.12$, $\beta_{0.5} \approx 0.6$ and $\gamma_{0.5} \approx 0.76$ for $|X| = 2$ and $|Z| = 1$. Figure 2(a)–(b) illustrates the fitting between empirical formula (8) and observed values

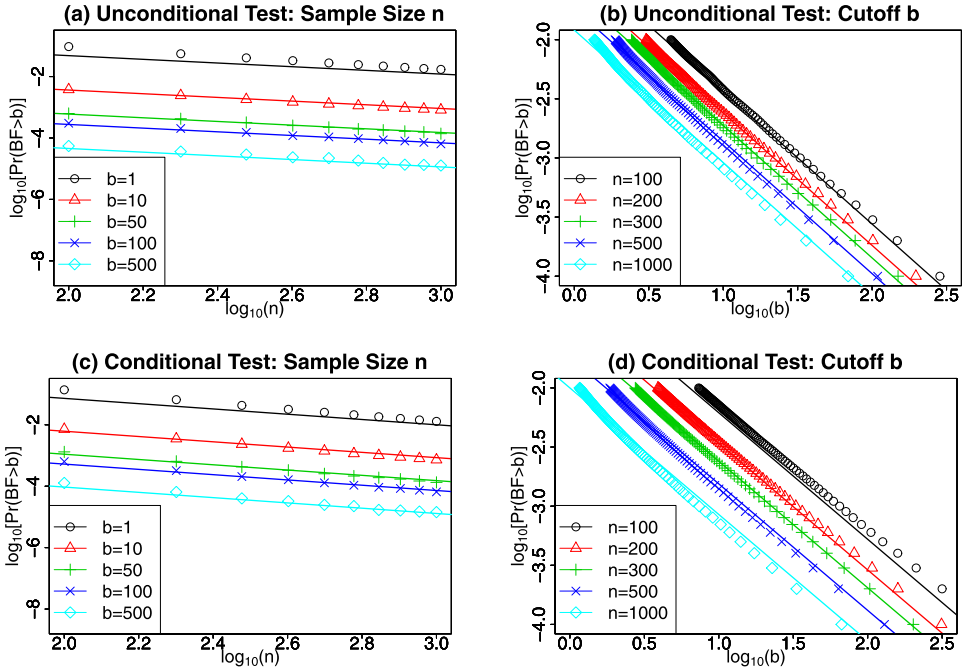


Figure 2: Empirical fitting of type-I error $\Pr_{\text{shuffle}}(\text{BF}(X|Y,Z) > b)$ given sample size $n$ and cutoff $b$ for unconditional and conditional test. Straight lines in (a)–(b) and (c)–(d) are calculated from empirical formula (8) and (9), respectively.

of type-I error when $p = 0.5$. Fitted $\alpha_p$, $\beta_p$ and $\gamma_p$ for other values of $p$ are given in the online supplement (Jiang et al., 2016).

Moreover, for conditional test, we obtain the following empirical formula of type-I error given the cutoff of BF statistic $b$ and sample size $n$:

$$\Pr_{\text{shuffle}}\left(\text{BF}\left(X|Y,Z\right) > b\right) \approx \frac{\gamma_{\mathbf{f}}}{b^{\alpha_{\mathbf{f}}} n^{\beta_{\mathbf{f}}}}, \tag{9}$$

where $\alpha_{\mathbf{f}}$, $\beta_{\mathbf{f}}$ and $\gamma_{\mathbf{f}}$ only depend on $\mathbf{f}$, the vector of observed frequencies for configurations of $(X, Z)$. For example, given $|X| = |Z| = 2$ and $\mathbf{f} = (0.25, 0.25, 0.25, 0.25)$, $\alpha_{\mathbf{f}} \approx 1.07$, $\beta_{\mathbf{f}} \approx 0.86$, and $\gamma_{\mathbf{f}} \approx 3.8$. Figure 2(c)–(d) illustrates the fitting between empirical formula (9) and observed values of type-I error. These fitting formulas are useful when one has to deal with many similar hypotheses simultaneously or is interested in very small p-values.

## 2.6 Forward stepwise selection based on the BF statistic

Given a continuous response $Y$ and a set of categorical covariates $\{X_j\}_{j=1}^m$, variable selection procedures aim to select a subset of covariates indexed by $\mathcal{A}$ such that $\{X_j :$

$j \in \mathcal{A}\}$ are associated with the response $Y$ while the other covariates $\{X_j : j \notin \mathcal{A}\}$ are independent of $Y$ given $\{X_j : j \in \mathcal{A}\}$. Here, we propose to use a forward stepwise procedure based on conditional BF statistic, preceded by an independent screening stage based on unconditional BF statistic. Throughout this paper, we assume that the number of categorical covariates, $m$, is fixed and does not increase with sample size $n$.

**Algorithm 2.**

- *Independent Marginal Screening*: calculate unconditional BF statistic denoted as $\mathrm{BF}(X_j \mid Y, Z \equiv 0)$ for $j = 1, \ldots, m$. Let $\mathcal{B}$ denote the index set of covariates with the corresponding BF statistics larger than a pre-specified threshold $b_0$, and $j_0 = \underset{j \in \{1,\ldots,m\}}{\operatorname{argmax}} \{\mathrm{BF}(X_j|Y, Z \equiv 0)\}$. Proceed if $\mathrm{BF}(X_{j_0}|Y, Z \equiv 0) > b_0$ (*i.e.* $\mathcal{B} \neq \emptyset$).

- *Forward Stepwise Selection*: let $\mathcal{C}_t$ denote the index set of covariates that have been selected at iteration $t$. Initialize $\mathcal{C}_1 = j_0$ and $Z_1 = X_{j_0}$, and iterate the following steps for $t \geq 2$:

  - At iteration $t$ ($t \geq 2$), encode the configurations of selected variables in $\mathcal{C}_{t-1}$ into a "super" variable $Z_{t-1}$.
  - Calculate conditional BF statistic $\mathrm{BF}(X_j|Y, Z_{t-1})$ for $j \in \mathcal{B} - \mathcal{C}_{t-1}$, and let $j_t = \underset{j \in \mathcal{B}-\mathcal{C}_{t-1}}{\operatorname{argmax}} \{\mathrm{BF}(X_j|Y, Z_{t-1})\}$.
  - Let $\mathcal{C}_t = \mathcal{C}_{t-1} \cup \{j_t\}$ if $\mathrm{BF}(X_{j_t}|Y, Z_{t-1}) > b_t$. Otherwise, stop and output $\mathcal{C}_{t-1}$.

We may decide the threshold $b_t$ at iteration $t$ according to our prior belief in the null hypothesis or a pre-specified interpretation on the relationship between Bayes factor and strength of evidence. For example, Kass and Raftery (1995) viewed a Bayes factor of $> 150$ as very strong evidence against the null hypothesis. Alternatively, we can choose the threshold $b_t$ to control for type I errors. Specifically, at each iteration, we estimate the null distribution of the maximum BF statistics under $H_0$ by using a conditional permutation scheme as follows:

- To generate a permuted data set at iteration $t$, shuffle the observed values of $Y$ within each group of observations indexed by $\{i : z_{t-1,i} = k\}$, independently for $k \in \{1, \ldots, |Z_{t-1}|\}$.

- Estimate a null distribution of $\mathrm{BF}(X_{j_t}|Y, Z_{t-1})$ by calculating the maximum of BF statistics for $j \in \mathcal{B} - \mathcal{C}_{t-1}$ on each permuted data set.

Then, we can use the empirical null distribution to calculate a $p$-value for the observed value of $\mathrm{BF}(X_{j_t}|Y, Z_{t-1})$, and terminate the iterative variable selection procedure if the $p$-value is larger than a threshold (*e.g.*, 0.05).

# 3 Simulation studies

## 3.1 Unconditional dependence testing

We first compare different methods in testing unconditional dependence between a binary indicator $X$ and a continuous response $Y$. Note that this testing problem is equivalent to the classic two-sample testing problem. Methods under comparison considerations include: the BF statistic with hyper-parameters $\lambda_0 = 1$ and $\alpha_0 = 1$ or 2 (which we call "BF ($\alpha_0 = 1$ or 2)"), BF with $\sqrt{n}$-partition prior with $\alpha_0 = 1$ and $\lambda_0 = 1$ ("BF ($\sqrt{n}$-p)"), dynamic slicing ("DS") test statistic (see online supplement (Jiang et al., 2016) for details), the Wilcoxon rank-sum test ("rank-sum"; also known as the Mann-Whitney $U$ test; Wilcoxon, 1945; Mann and Whitney, 1947), Welch's $t$-test ("$t$-test"; Welch, 1947), Kolmogorov-Smirnov ("KS") test and Anderson-Darling ("AD") test (Anderson and Darling, 1952). The null hypothesis of Welch's $t$-test is that the means of two normally distributed populations are equal (but with possibly unequal variance), and the null hypothesis of rank-sum test is that the probability of an observation from one population exceeding an observation from the second population equals to 0.5. All other methods test the null hypothesis that the distributions of two populations are the same against a completely general alternative hypothesis that the binary indicator $X$ and the quantity of interest $Y$ are not independent.

We generated binary variable $X \sim \text{Bern}(0.5)$, and simulated the continuous variable $Y$ under the alternative hypothesis according to following scenarios with sample size $n = 400$:

Scenario 1 (Gaussian with mean shift): $Y \sim N(-\mu, 1)$ when $X = 0$; and $Y \sim N(\mu, 1)$ when $X = 1$; $\mu = 0.1$.

Scenario 2 (Cauchy with mean shift): $Y \sim \text{Cauchy}(-\mu, 1)$ when $X = 0$; and $Y \sim \text{Cauchy}(\mu, 1)$ when $X = 1$; $\mu = 0.2$.

Scenario 3 (Gaussian with scale change): $Y \sim N(0, 1)$ when $X = 0$; and $Y \sim N(0, \sigma^2)$ when $X = 1$; $\sigma = 1.2$.

Scenario 4 (Gaussian with mean shift and scale change): $Y \sim N(\mu, 1)$ when $X = 0$; and $Y \sim N(-\mu, \sigma^2)$ when $X = 1$; $\mu = 0.1$ and $\sigma = 1.2$.

Scenario 5 (Symmetric Gaussian mixture): $Y \sim$ a mixture of $N(-\mu, 1)$ and $N(\mu, 1)$ with probabilities $1 - \theta$ and $\theta$ when $X = 0$; and $Y \sim N((2\theta-1)\mu, 1+4\theta(1-\theta)\mu^2)$ when $X = 1$; $\theta = 0.5$ and $\mu = 1.2$.

Scenario 6 (Asymmetric Gaussian mixture): same as Scenario 3 except that $\theta = 0.9$.

The receiver operating characteristic (ROC) curves in Figure 3 illustrates how true positive rates, the fraction of true positives out of the total actual positives, trade against false positive rates, the fraction of false positives out of the total actual negatives, of different methods at varying thresholds.

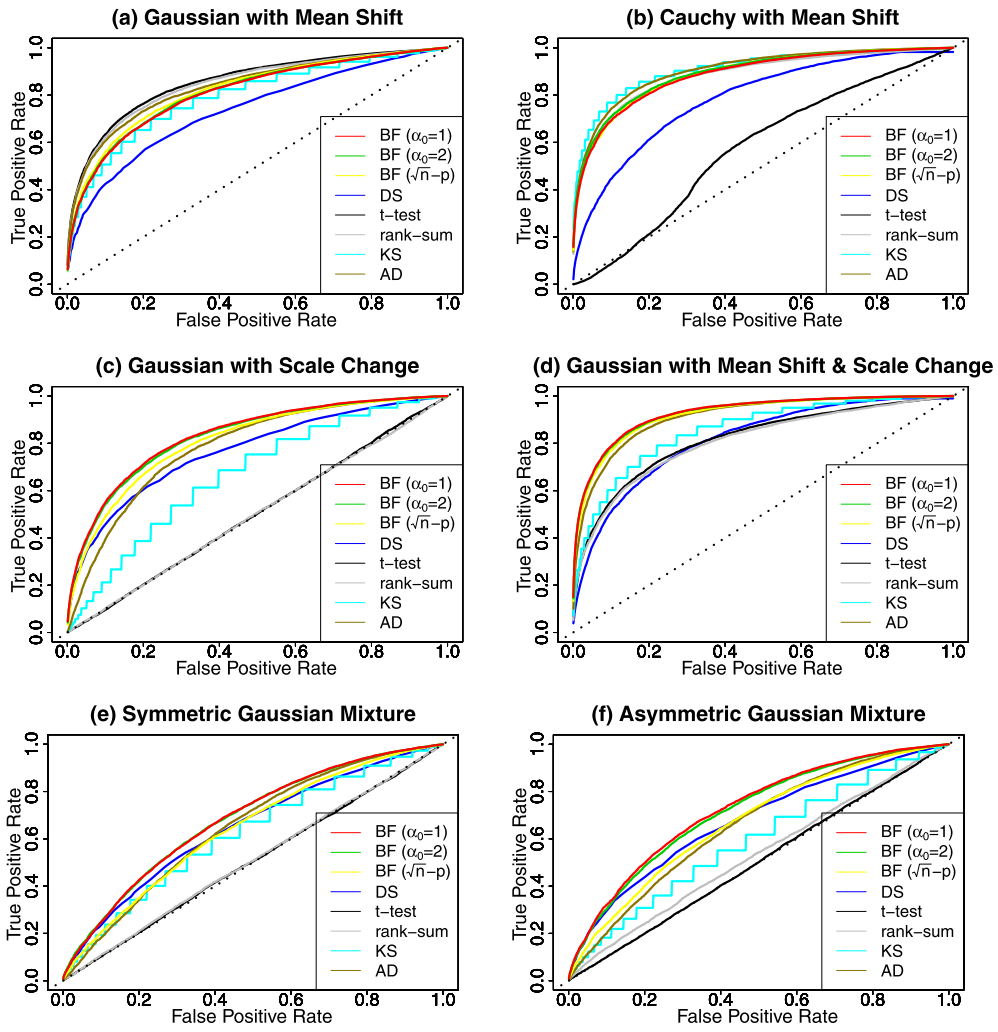


Figure 3: The ROC curves that compare true positive rate, the fraction of true positives out of the total actual positives, and false positive rate, the fraction of false positives out of the total actual negatives, of different methods in Scenarios 1–6 of Section 3.1.

In Scenario 1, two populations corresponding to $X = 0$ and $X = 1$ follow Gaussian distributions with different means but the same variance, which satisfies all the parametric assumptions of the two-sample $t$-test. As expected, in Figure 3(a), Welch's $t$-test achieved the highest power in this scenario, which was followed closely by the rank-sum test and the Anderson-Darling test. The BF statistics had slightly lower power under this scenarios but still outperformed the Kolmogorov-Smirnov test. The dependence test based on dynamic slicing had the lowest power in this case. In Scenario 2, the two samples were generated from Cauchy distribution with different location parameters

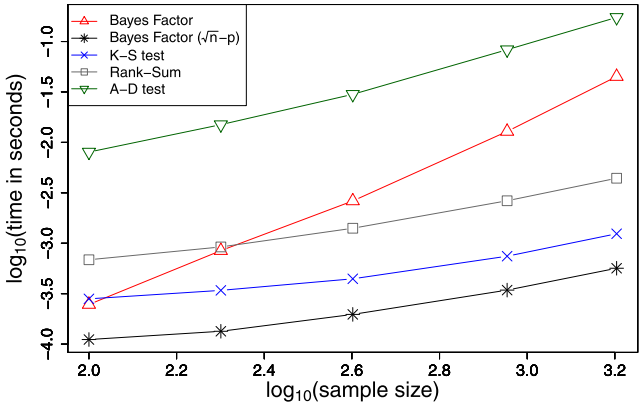


Figure 4: Timing comparisons for two-sample test.

(population medians) and the same scale parameter. The $t$-test was completely powerless. As shown in Figure 3(b), given the same false positive rate, Kolmogorov-Smirnov test achieved the true positive rate, followed by the Anderson-Darling, the BF methods and rank-sum test. Tests based on dynamic slicing had significantly lower true positive rate.

When two Gaussian populations have the same mean but different variances (Scenario 2), the BF statistics had a superior power compared with others in Figure 3(c). Among the other methods, the Anderson-Darling test and dynamic slicing ("DS") test outperformed the Kolmogorov-Smirnov tests, while the rank-sum test and $t$-test had almost no power under this scenario. Figure 3(d) illustrates the scenario when both the means and the variances of two Gaussian populations are different. BF methods dominated all other methods in this scenario.

Scenarios 5 and 6 demonstrate the performances of different methods when two populations have both the same mean and the same variance, but different skewness and kurtosis. In both scenarios, the BF statistic with $\alpha_0 = 1$ achieved the highest power as shown in Figure 3(e)–(f).

The ROC curves of BF statistics with $\alpha_0 = 1$ and $\alpha_0 = 2$ were similar in Scenarios 1–5, while the BF statistic with $\alpha_0 = 1$ had a slightly better performance under Scenario 6. The BF statistic with $\alpha_0 = 1$ dominated the dynamic slicing ("DS") test statistic in all the six scenarios. The performance of BF with $\sqrt{n}$-partition prior ("BF ($\sqrt{n}$-p)") were similar to that of the original BF with independent prior. It slightly outperformed the original BF in Scenario 1, but performed worse than the original version in Scenarios 3–6.

We further compare run-time performance of two-sample testing methods. Figure 4 shows the logarithmic running time (with base 10) of different methods with increasing sample sizes (also on the logarithmic scale in base 10). When sample size n is smaller than 1600, the BF with independent prior ("Bayes Factor") has better run-time per-

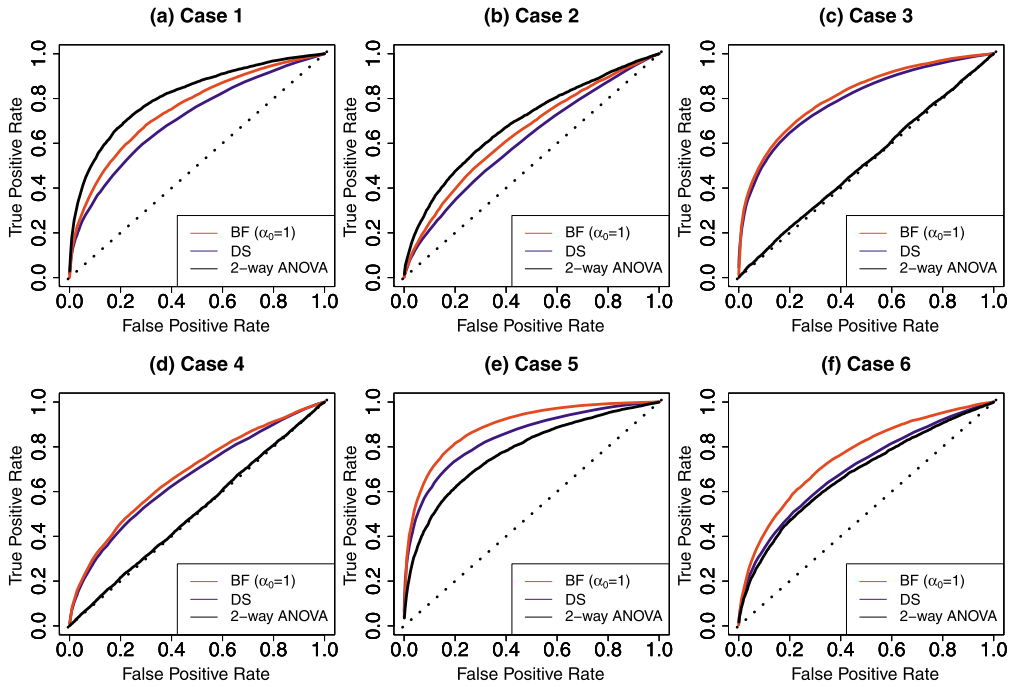


Figure 5: The ROC curves that compare true positive rate, the fraction of true positives out of the total actual positives, and false positive rate, the fraction of false positives out of the total actual negatives, of different methods in Cases 1–6 of Section 3.2 with uncorrelated covariates $X$ and $Z$.

formance compared with the Anderson-Darling test, and its computationally efficient variant, BF with $\sqrt{n}$-partition prior ("Bayes Factor ($\sqrt{n}$-p)"), has a smaller computational cost than the Kolmogorov-Smirnov test and the rank-sum test. From Figure 4, we can also see that as sample size increases, the running time of the original BF method increase quadratically with sample size, while the running time of the BF with $\sqrt{n}$-partition prior increase linearly with sample size.

## 3.2 Conditional dependence testing

Next, we compare different methods in testing the conditional dependence between a binary covariate $X$ and a continuous response $Y$ given another binary covariate $Z$. Methods under comparison consideration include: the BF statistic (with $\lambda_0 = 1$ and $\alpha_0 = 1$), dynamic slicing ("DS") statistic, and two-way ANOVA test, which tests for main and interaction effects of $X$ conditioning on $Z$.

In our study, we generate $n = 400$ samples with binary covariates $Z \sim \text{Bern}(0.5)$, and conditioning on $Z$, $X|Z = 0 \sim \text{Bern}(p_0)$ and $X|Z = 1 \sim \text{Bern}(1 - p_0)$. We choose $p_0 = 0.5$ for conditional tests with uncorrelated covariates, and $p_0 = 0.75$ for conditional

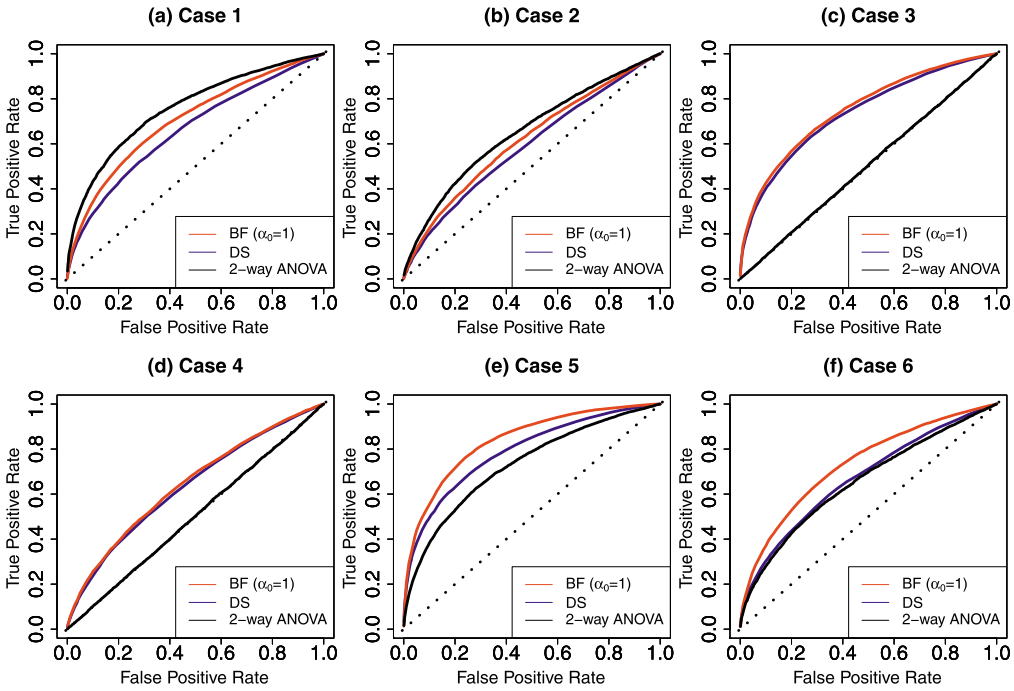


Figure 6: The ROC curves that compare true positive rate, the fraction of true positives out of the total actual positives, and false positive rate, the fraction of false positives out of the total actual negatives, of different methods in Cases 1–6 of Section 3.2 with correlated covariates $X$ and $Z$.

tests with correlated covariates. Then, we simulate the response $Y$ according to the following models under the alternative hypothesis:

Case 1: $Y = \mu Z + \mu X + \epsilon$; $\epsilon \sim N(0,1)$, $\mu = 0.2$.

Case 2: $Y = \mu ZX + \epsilon$; $\epsilon \sim N(0,1)$, $\mu = 0.2$.

Case 3: $Y = \mu Z + \mu X + \epsilon$; $\epsilon \sim \text{Cauchy}(0,1)$, $\mu = 0.4$.

Case 4: $Y = \mu ZX + \epsilon$; $\epsilon \sim \text{Cauchy}(0,1)$, $\mu = 0.4$.

Case 5: $Y = \mu Z + \mu X + \epsilon$; $\epsilon \sim N(0,(1+\gamma X)^2)$, $\mu = 0.2, \gamma = 0.2$.

Case 6: $Y = \mu ZX + \epsilon$; $\epsilon \sim N(0,(1+\gamma ZX)^2)$, $\mu = 0.2, \gamma = 0.2$.

Our goal here is to test whether $X$ is independent of $Y$ given $Z$. The ROC curves of different methods under Cases 1–6 with uncorrelated ($p_0 = 0.5$) and correlated ($p_0 = 0.75$) $X$ and $Z$ are given in Figure 5 and Figure 6, respectively.

In Cases 1 and 2, the two samples were generated from homoscedastic normal distribution with either linear combination or multiplicative interaction of covariates, which satisfies all the parametric assumptions of the two-way ANOVA test. As we have expected, the two-way ANOVA test achieved highest power in Figure 5(a)–(b) and Figure 6(a)–(b), followed by the BF and DS test statistics.

However, as shown in Figure 5(c)–(d) and Figure 6(c)–(d), when the two samples were generated from Cauchy distribution, the two-way ANOVA test was completely powerless in Cases 3 and 4, while the BF and DS statistics had considerable powers with Cauchy noises.

Cases 5 and 6 illustrate the scenarios when the response has heteroscedastic variances depending on covariates. In both cases, the BF statistic achieved better powers than the DS and two-way ANOVA test. Among all the conditional dependence testing scenarios we have considered, the BF test statistic always outperformed the DS test statistic. The relative performances of different methods were consistent whether covariates $X$ and $Z$ are correlated or not.

## 3.3 Interaction detection on synthetic QTL data

Traditional QTL studies are based on linear regression models (Lander and Botstein, 1989) in which each (continuous) trait variable is regressed against each (discrete) marker variable. The $p$-value of the regression slope is reported as a measure of significance for association. Storey et al. (2005) developed a stepwise regression method to search for pairs of markers that are associated with the gene expression quantitative trait. This procedure, however, tends to miss QTL pairs with small marginal effects but a strong interaction effect.

In this section, we compare the proposed variable selection method based on the BF statistic with the stepwise regression (SR) method in identifying genetic markers with interaction effects in synthetic QTL data sets. Using the R package *qtl*, we generated 100 binary markers with sample size $n = 400$ such that adjacent markers are correlated with each other, and then, we randomly select two markers and simulate quantitative traits according to Cases 1–6 in the previous section. We evaluate the performance of the BF and SR methods using the following procedure. First, in the screening step, we calculate the unconditional test statistic for each marker and obtain a list of candidate markers with test statistic above a given threshold $T_1$. Second, conditioning on the most significant candidate marker, we select other candidate markers with conditional test statistics above another threshold $T_2$. Finally, we vary the thresholds $T_1$ and $T_2$ simultaneously to generate the ROC curves in Figure 7.

As we can see from Figure 7, the SR had a better power when its underlying assumptions, *i.e.*, linearity, normality, and homoscedasticity, were satisfied as in Cases 1 and 2. However, with the presence of extreme values or heteroscedastic effects in Cases 3–6, the BF statistic was much more powerful than the SR.

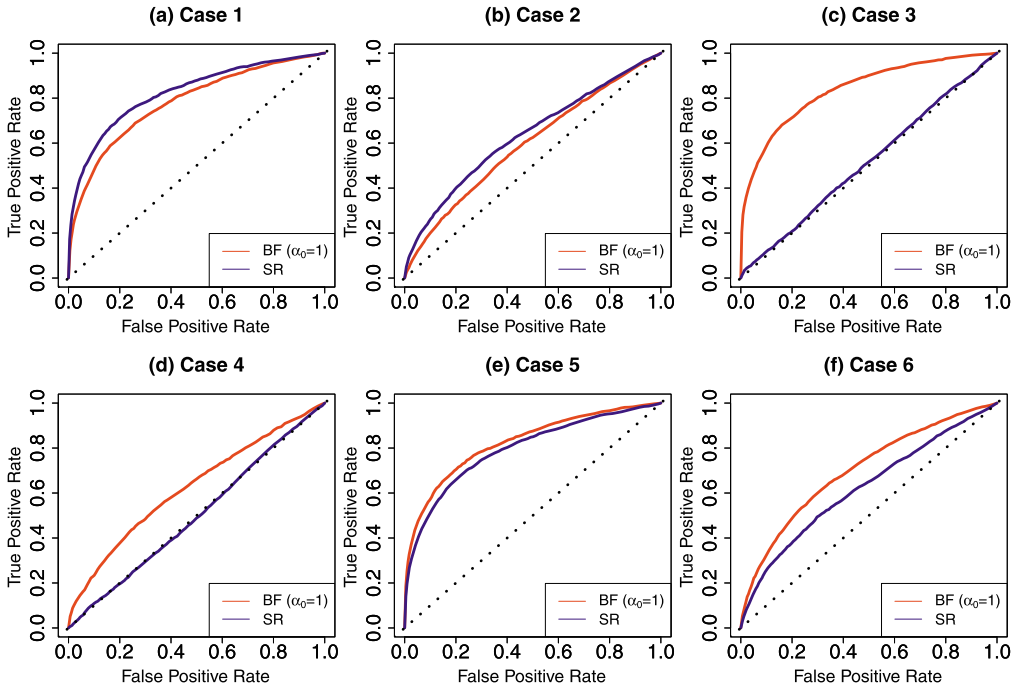


Figure 7: The ROC curves that compare true positive rate, the fraction of the true gene-marker pairs detected, and false positive rate, the fraction of unrelated gene-marker pairs falsely selected, of stepwise regression (SR) and BF statistic on synthetic QTL data sets in Section 3.3.

# 4 Application to QTL study in mouse

Burke et al. (2012) measured a mouse population for complex adult phenotypes, related to body size and bone structure, and conducted a genome-wide search for QTLs that are marginally associated with quantitative phenotypes. Each mouse in the population was genotyped at 558 biallelic loci, *i.e.*, binary genetic markers. For each trait, Burke et al. (2012) performed a single-locus genome-wide search using one-way ANOVA model and permutation-based test of significance.

We applied the screening and the BF-based stepwise selection procedure proposed in Section 2.6 to search for effective loci associated with two quantitative traits, femur length and vertebra length. At the screening step and each forward selection step, we permuted sample labels conditioning on the observed values of previously selected QTLs, and in each permuted data set, we recorded the maximum value of BF statistics among all the candidate QTLs. Then, a *genome-wide* $p$-value is evaluated by comparing the observed BF statistic with these maximum BF values from 1000 permuted data sets. In the screening step, we retained 35 and 49 loci, respectively for femur length and vertebra length, with unconditional BF value larger than 10 (corresponding to genome-

| QTL | Location | Bayes Factor | $p$-value | Reported in Burke et al. (2012) |
|---|---|---|---|---|
| rs3091203 | CH13·22 | $2.7 \times 10^7$ | $< 0.001$ | $p$-value $< 0.001$ |
| *D2Mit285* | CH2·152 | $1.8 \times 10^5$ | $< 0.001$ | CH2·157 (rs4223627), $p$-value $< 0.001$ |
| rs3657845 | CH17·17 | 63.7 | 0.004 | $p$-value $= 0.011$ |
| *D5Mit25* | CH5·114 | 2.3 | 0.192 | $p$-value $< 0.001$ |
| *D9Mit110* | CH9·91 | 46.9 | 0.062 | $p$-value $< 0.001$ |

Table 1: Identified QTLs (ranked according to their orders in forward stepwise selection) associated with mouse femur length, their Bayes factor values and corresponding genome-wide $p$-values, and relationships with significant loci reported in Burke et al. (2012). Genomic location is in *Chromosome·Mb* format.

wide $p$-value of 0.03 and 0.05) as candidate QTLs for forward stepwise selection. QTLs identified through stepwise selection on each trait, together with their BF values and genome-wide $p$-values in forward selection steps, as well as relationships with significant loci found in Burke et al. (2012), are given in Tables 1 and 2.

Table 1 shows femur length QTLs that are selected from the first 5 iterations of the proposed stepwise method, together with their BF values and genome-wide $p$-values at each forward selection step. Burke et al. (2012) reported the same 5 genomic regions as marginally associated with the trait and their genome-wide $p$-values from the paper are given in Table 1. Using a cutoff of 0.05 for $p$-values, the BF-based stepwise procedure was terminated after the third iteration, *i.e.*, we could not reject the null hypothesis that *D5Mit25* and mouse femur length are conditionally independent given the previously selected loci rs3091203, *D2Mit285* and rs3657845. This is confirmed on an independent replicate population with femur length measurement and genotypes on a subset of loci (356 of 558 loci) provided by Burke et al. (2012). Although *D5Mit25* (located at CH5·114) was not genotyped in the replicate population, genotypes of its neighboring locus rs13478469 (also located at CH5·114) were available. Note that in the original population, both *D5Mit25* and rs13478469 have genome-wide $p$-values smaller than 0.001 according to Burke et al. (2012), but rs13478469 was not reported in Burke et al. (2012) due to its adjacency and high correlation (about 0.95) with *D5Mit25*. In the replicate population, 3-way ANOVA test shows that the top 3 QTLs in Table 1 all have significant main effects with $p$-values $< 0.001$. On the other hand, rs13478469 does not have significant main effect or interaction effects with other 3 QTLs ($p$-values $> 0.1$) according to 4-way ANOVA test on the replicate population. These results from an independent replicate population are consistent with the conclusions of our testing procedure applied to the original population.

In Table 2, the proposed forward stepwise procedure based on the BF statistic detected 6 QTLs associated with vertebra length under a significance level of 0.05, which include all of the 5 genomic regions (either the locus itself or the neighboring locus located next to it) reported in Burke et al. (2012). Besides, our analysis identified an additional locus, *D16Mit36*. From the first plot in Figure 8, we can see that although the distributions of vertebra length given two alleles of *D16Mit36* have similar means (one-way ANOVA test of equal means has $p$-value $= 0.19$), the variances are quite different (an F-test of equal variances has $p$-value $= 7.16 \times 10^{-5}$). Because of its heteroscedastic

| QTL | Location | Bayes Factor | $p$-value | Reported in Burke et al. (2012) |
|---|---|---|---|---|
| rs4222738 | CH1·158 | $7.0 \times 10^9$ | $< 0.001$ | $p$-value $< 0.001$ |
| *D1Mit105* | CH1·162 | $4.6 \times 10^7$ | $< 0.001$ | CH1·166 (rs4222769), $p$-value $< 0.001$ |
| *D2Mit58* | CH2·108 | $1.6 \times 10^3$ | 0.001 | CH2·111 (rs3023543), $p$-value $= 0.006$ |
| *D16Mit36* | CH16·31 | 991.8 | 0.001 | **Not reported** |
| *D7Mit76* | CH7·18 | 247.9 | 0.019 | $p$-value 0.0026 |
| rs13481706 | CH13·16 | $2.8 \times 10^5$ | 0.039 | $p$-value 0.019 |

Table 2: Identified QTLs (ranked according to their orders in forward stepwise selection) associated with mouse vertebra length, their Bayes factor values and corresponding genome-wide $p$-values, and relationships with significant loci reported in Burke et al. (2012). Genomic location is in *Chromosome·Mb* format.

effect, the screening based on the BF statistic was able to retain *D16Mit36* as a candidate QTL, while one-way ANOVA test missed the locus completely. Further analysis shows that QTLs *D16Mit36* and *D2Mit58*, which was identified in the previous step, have positive *epistasis* effect. From Figure 8, we can see that *D16Mit36* and *D2Mit58* have non-additive effects, that is, the effect of *D2Mit58* is larger when *D16Mit36* has allele B6. Two-way ANOVA test of interaction effect between *D16Mit36* and *D2Mit58* has a $p$-value of $= 0.001$. This example demonstrates that the proposed selection procedure based on the BF statistic is particularly effective in detecting QTLs with interaction and heteroscedastic effects.

## 5 Discussion

We have developed a non-parametric dependence testing method for categorical covariates and continuous response, and implemented the proposed method in R package *bfslice*, which can be downloaded from `http://www.people.fas.harvard.edu/~junliu/BF/` or requested from the authors directly. As a dependence testing tool, the proposed Bayes factor-based statistic achieves a higher statistical power compared with traditional non-parametric methods such as the Kolmogorov-Smirnov test, and is more robust to outliers and various distributional assumptions compared with classical ANOVA based approaches. Furthermore, the stepwise variable selection method based on the BF statistic is particularly effective in detecting covariates with interaction or heteroscedastic effects, especially when the combined number of covariate categories is relatively large compared with sample size. Theoretically, we proved upper bounds on p-values (type-I errors) of the BF statistic under a variety of null hypothesis assumptions, and showed that the proposed BF statistic asymptotically grows to infinity at an exponential rate under the alternative hypothesis and with proper choices of hyperparameters. We also fitted a fairly accurate empirical formula for the type-I error of any given BF cutoff value. But a theoretical derivation of its exact form remains an open question.

The method described in this paper can be easily used to deal with categorical or discrete ordinal response variables. For categorical response, different response categories naturally define the "slicing" scheme, and for discrete ordinal response (or continuous

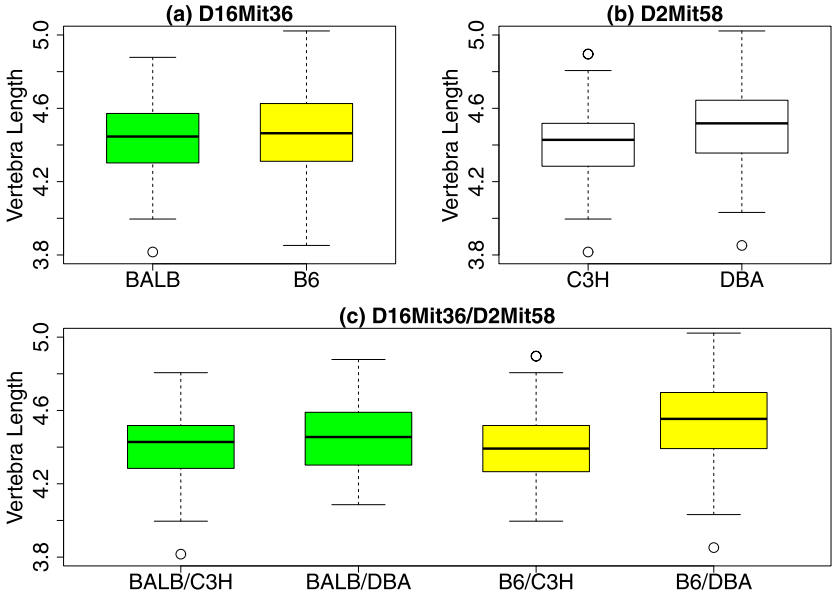


Figure 8: Box-plots showing the heteroscedastic effect on vertebra length given two alleles (BALB and B6) of *D16Mit36* (a), the mean shift effect on vertebra length given two alleles (C3H and DBA) of *D2Mit58* (b), and the epistasis effect between *D16Mit36* and *D2Mit58* (c).

response with ties), we can arbitrarily rank observations with the same value of response and only allow slicing between ranked observations that have different observed response values. A potential research direction is to extend the Bayes factor approach for variable selection with continuous covariates under the sliced inverse regression framework (Jiang and Liu, 2014).

For applications with multivariate response, we can further generalize the concept of "slices" to unobserved clusters (aka "partitions") of samples, and model the distribution of response and covariates independently given hidden cluster labels. Combined with a Markov Chain Monte Carlo strategy, we are currently developing a Bayesian partition procedure (Jiang and Liu, 2015) for detecting expression quantitative trait loci (eQTLs) with variable selection on both responses (gene expression) and covariates (genetic variations). For a large data set with several millions of SNPs, one can use the method proposed in this paper as a fast yet powerful screening tool to pre-select a subset of SNPs, and then apply the full Bayesian model on the selected SNPs.

## Supplementary Material

Supplement to "Bayesian Nonparametric Tests via Sliced Inverse Modeling" (DOI: 10.1214/16-BA993SUPP; .pdf). We provide additional supporting materials that include detailed proofs and additional empirical results.

## References

Anderson, T. W. and Darling, D. A. (1952). "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes." *The Annals of Mathematical Statistics*, 23: 193–212. MR0050238. 101

Aschard, H., Zaitlen, N., Tamimi, R. M., Lindström, S., and Kraft, P. (2013). "A non-parametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes." *Genetic Epidemiology*, 37(4): 323–333. 90

Basu, S. and Chib, S. (2003). "Marginal likelihood and Bayes factors for Dirichlet process mixture models." *Journal of the American Statistical Association*, 98(461): 224–235. MR1965688. doi: http://dx.doi.org/10.1198/01621450338861947. 90

Berger, J. O. and Guglielmi, A. (2001). "Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives." *Journal of the American Statistical Association*, 96(453): 174–184. MR1952730. doi: http://dx.doi.org/10.1198/016214501750333045. 90

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). "Genetic dissection of transcriptional regulation in budding yeast." *Science*, 296(5568): 752–755. 89

Burke, D. T., Kozloff, K. M., Chen, S., West, J. L., Wilkowski, J. M., Goldstein, S. A., Miller, R. A., and Galecki, A. T. (2012). "Dissection of complex adult traits in a mouse synthetic population." *Genome Research*, 22(8): 1549–1557. 107, 108, 109

Carota, C. and Parmigiani, G. (1996). "On Bayes Factors for Nonparametic Alternatives." In: Bernardo, J., Berger, J. O., Dawid, A., and Smith, A. (eds.), *Bayesian statistics 5*, 507–511. Oxford University Press. MR1425421. 90

Florens, J.-P., Richard, J.-F., and Rolin, J. (1996). "Bayesian encompassing specification tests of a parametric model against a nonparametric alternative." Technical Report 96.08, Université Catholique de Louvain, Institut de Statistique. 90

Hanson, T. E. (2006). "Inference for mixtures of finite Polya tree models." *Journal of the American Statistical Association*, 101(476): 1548–1565. MR2279479. doi: http://dx.doi.org/10.1198/016214506000000384. 90

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). "Two-sample Bayesian nonparametric hypothesis testing." *Bayesian Analysis*, 10(2): 297–320. 90

Jiang, B. and Liu, J. S. (2014). "Variable selection for general index models via sliced inverse regression." *The Annals of Statistics*, 42(5): 1751–1786. MR3262467. doi: http://dx.doi.org/10.1214/14-AOS1233. 110

Jiang, B. and Liu, J. S. (2015). "Bayesian partition models for identifying expression quantitative trait loci." *Journal of the American Statistical Association*, (just-accepted): 00–00. 110

Jiang, B., Ye, C., and Liu, J. S. (2016). "Supplement to "Bayesian Nonparametric Tests via Sliced Inverse Modeling"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA993SUPP. 91, 92, 94, 95, 96, 99, 101

Jiang, B., Ye, C., and Liu, J. S. (2015). "Nonparametric K-sample tests via dynamic slicing." *Journal of the American Statistical Association*, 110(510): 642–653. MR3367254. doi: http://dx.doi.org/10.1080/01621459.2014.920257. 90, 91, 92

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. 100

Lander, E. S. and Botstein, D. (1989). "Mapping mendelian factors underlying quantitative traits using RFLP linkage maps." *Genetics*, 121(1): 185–199. 106

Lander, E. S. and Schork, N. J. (1994). "Genetic dissection of complex traits." *Science*, 265(5181): 2037–2048. 89

Ma, L. and Wong, W. H. (2011). "Coupling optional Pólya trees and the two sample problem." *Journal of the American Statistical Association*, 106(496): 1553–1565. MR2896856. doi: http://dx.doi.org/10.1198/jasa.2011.tm10003. 90

Mann, H. B. and Whitney, D. R. (1947). "On a test of whether one of two random variables is stochastically larger than the other." *The Annals of Mathematical Statistics*, 18: 50–60. MR0022058. 101

McVinish, R., Rousseau, J., and Mengersen, K. (2009). "Bayesian goodness of fit testing with mixtures of triangular distributions." *Scandinavian Journal of Statistics*, 36(2): 337–354. MR2528988. doi: http://dx.doi.org/10.1111/j.1467-9469.2008.00620.x. 90

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004). "Genetic analysis of genome-wide variation in human gene expression." *Nature*, 430(7001): 743–747. 89

Storey, J. D., Akey, J. M., and Kruglyak, L. (2005). "Multiple locus linkage analysis of genomewide expression in yeast." *PLoS Biology*, 3(8): e267. 106

Welch, B. L. (1947). "The generalization of "student's" problem when several different population variances are involved." *Biometrika*, 34: 28–35. MR0019277. 101

Wilcoxon, F. (1945). "Individual comparisons by ranking methods." *Biometrics Bulletin*, 1: 80–83. 101

Zhang, Y. and Liu, J. S. (2007). "Bayesian inference of epistatic interactions in case-control studies." *Nature Genetics*, 39(9): 1167–1173. 90