Published in final edited form as:
J Am Stat Assoc. 2018 ; 113(523): 1362–1371. doi:10.1080/01621459.2017.1328361.

On the Null Distribution of Bayes Factors in Linear Regression

Quan Zhou and **Yongtao Guan***
Baylor College of Medicine

Abstract

We show that under the null, the 2 log(Bayes factor) is asymptotically distributed as a weighted sum of chi-squared random variables with a shifted mean. This claim holds for Bayesian multilinear regression with a family of conjugate priors, namely, the normal-inverse-gamma prior, the g-prior, and the normal prior. Our results have three immediate impacts. First, we can compute analytically a p-value associated with a Bayes factor without the need of permutation. We provide a software package that can evaluate the p-value associated with Bayes factor efficiently and accurately. Second, the null distribution is illuminating to some intrinsic properties of Bayes factor, namely, how Bayes factor quantitatively depends on prior and the genesis of Bartlett's paradox. Third, enlightened by the null distribution of Bayes factor, we formulate a novel scaled Bayes factor that depends less on the prior and is immune to Bartlett's paradox. When two tests have an identical p-value, the test with a larger power tends to have a larger scaled Bayes factor, a desirable property that is missing for the (unscaled) Bayes factor.

Keywords

p-value; weighted sum of chi-squared random variables; scaled Bayes factor

1 Introduction

Bayesian methods have been sidelined by most practitioners in genetic association studies. The main reason is that p-value, although often misinterpreted, is entrenched among practitioners [Sellke et al., 2001, Nuzzo, 2014]. A Bayesian method that performs genetic association tests, such as that of Guan and Stephens [2008], often faces an inconvenient demand to produce a p-value associated with an extraordinarily large Bayes factor. Because the null distribution of Bayes factor is unknown, a previous solution has been to obtain the p-value through permutation [Servin and Stephens, 2007]. In genome-wide association studies, however, the significance threshold for p-values is exceedingly small, owing to the burden of multiple testing; thus, the number of permutations required is often prohibitively

*Quan Zhou is a PhD student in the program of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine (quan.zhou@alumni.bcm.edu). Yongtao Guan is an Assistant Professor of USDA/ARS Children’s Nutrition Research Center, Department of Pediatrics, and Department of Molecular and Human Genetics of Baylor College of Medicine, 1100 Bates, Room 2070, Houston TX 77030 (yongtaog@bcm.edu).

Supplementary
The Supplementary online includes proofs of theorems, R code for computing Bayes factors and scaled Bayes factors, and a table to summarize top single SNP association results using $\sigma_b = 0.5$ for the IOP dataset presented in Section 5. Our software package to compute p-values for a weighted sum of chi-squared random variables is freely available at https://github.com/haplotype/BACH and http://www.haplotype.org.

large and hence impractical. This motivates us to quantify the null distribution of Bayes factors, from which we can compute a p-value associated with Bayes factor analytically, without the need of permutation.

Our second motivation is to understand Bayes factor itself, first and foremost to understand, quantitatively, the prior-dependence nature of Bayes factors. Such prior-dependency often turns away naive practitioners. The second is to investigate the root of inconsistency of Bayes factor, namely, Bartlett's paradox [Bartlett, 1957]. Bartlett discovered that a diffusive prior tends to favor, unintentionally, the null model. In other words, if one were uncertain about the prior effects, one would automatically favor the null. (On the other hand, if one were too certain about the prior effects, one would risk prior-misspecification, which also unintentionally favors the null.) We identified the dominant term in Bayes factor that is affected by the prior, which motivates us to systematically scale Bayes factor. The scaled Bayes factor depends weakly on the prior and no longer suffers from the Barlett's paradox.

Our third motivation is to emphasize the necessity of taking into account power to interpret p-values. The probability that a positive report is false depends on both the p-value and the power of a test [Wacholder et al., 2004]. A plainer reiteration of this insightful observation is that a small p-value alone cannot provide a strong evidence for a true association, and it has to be interpreted in light of the statistical power [Burton et al., 2007]. A large Bayes factor, however, by itself provides a strong evidence for a true association [Stephens and Balding, 2009]. And Sawcer [2010] related Bayes factor to the ratio between the power and the p-value. It is therefore beneficial for a study to report both Bayes factors and their associated p-values. The idea of computing a p-value associated with Bayes factor dates back to Good [1957], as a symbol of "Bayes/non-Bayes compromise" [Good, 1992]. The p-values will satisfy the practical mandate imposed by the research community, and the companion Bayes factors will remind one to account for power when interpreting p-values. For example, tests may be ranked differently by their p-values than by their Bayes factors, and two identical p-values may associate with different Bayes factors. Both examples highlight the necessity of taking into account the statistical power to interpret p-values. To this end, our scaled Bayes factor becomes more appealing. When two tests produce identical p-values, the scaled Bayes factor tends to assign a larger value to the test with a larger power, while the (unscaled) Bayes factor does not.

Our main result states that, under the null, the logarithm of Bayes factor has the same distribution as a weighted sum of chi-squared random variables with a shifted mean. The results hold asymptotically for Bayesian multi-linear regression. For simple linear regression, we have $2\ \log(\text{Bayes factor}) = \lambda\chi_1^2 + \log(1-\lambda)$, where λ is a quantity related to the prior and the data, and χ_1^2 denotes a chi-squared random variable of one degree of freedom (denoted by *d.f.*). The undesirable properties of Bayes factor, namely, its over-dependence on the prior and Bartlett's paradox, find their roots in the shift term log(1 − λ). Our scaled Bayes factor effectively substitutes this term with −λ to achieve $2\ \log(\text{scaled Bayes factor}) = \lambda(\chi_1^2 - 1)$. For simple linear regression, the p-value associated with a Bayes factor is the same as the p-value of the likelihood ratio test. For multi-linear regression, computing the p-value associated with a Bayes factor requires evaluation of the distribution function of a weighted sum of chi-squared random variables. Based on a recently published polynomial algorithm [Bausch, 2013], we developed a software package to evaluate the p-values analytically, which can efficiently achieve an arbitrary precision.

The paper is structured as follows. In Section 2 we formulate the model and the priors, and provide our main result on the null distribution of Bayes factors. In Section 3 we demonstrate how to compute a p-value associated with a Bayes factor. In Section 4 we introduce the scaled Bayes factor and demonstrate its benefits. In Section 5 we analyze a real dataset to compute and compare Bayes factor, the scaled Bayes factor, and the p-values associated with Bayes factors. In the last section we summarize our findings and discuss relevant (future) topics. All proofs are in the Supplementary online.

2 The Null Distribution of Bayes factor

We consider the standard hypothesis testing problem in linear regression with independent normal errors.

$$H_0 : \boldsymbol{y} \mid \boldsymbol{a}, \tau \sim \mathrm{MVN}(\boldsymbol{Wa}, \tau^{-1}\boldsymbol{I}_n), H_1 : \boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{b}, \tau \sim \mathrm{MVN}(\boldsymbol{Wa} + \boldsymbol{Lb}, \tau^{-1}\boldsymbol{I}_n), \quad (1)$$

where MVN stands for the multivariate normal distribution, $\boldsymbol{I}_n$ is an $n \times n$ identity matrix, $\boldsymbol{W}$ is a full-rank $n \times q$ matrix representing the nuisance covariates, including a column of **1, *a*** is a q-vector, $\boldsymbol{L}$ is an $n \times p$ matrix representing the covariates of interest, $\boldsymbol{b}$ is a p-vector, and $\boldsymbol{\tau}^{-1}$ is the error variance. The two models H_0 and H_1 are nested and the null model H_0 represents no effect of $\boldsymbol{L}$.

The Bayesian linear regression comes with three forms of conjugate priors in the literature. The first is the normal-inverse-gamma (NIG) prior [O'Hagan and Forster, 2004, Chap. 9], detailed below:

$$\boldsymbol{a} \mid \tau \sim \mathrm{MVN}(\boldsymbol{0}, \tau^{-1}\boldsymbol{V}_a), \boldsymbol{b} \mid \tau \sim \mathrm{MVN}(\boldsymbol{0}, \tau^{-1}\boldsymbol{V}_b), \tau \sim \mathrm{Gamma}(\kappa_1/2, \kappa_2/2), \quad (2)$$

where $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$ are some positive definite matrices, and the gamma distribution is in the shape-rate parameterization. Following the standard treatment [c.f., Servin and Stephens, 2007] to let $\boldsymbol{V}_a^{-1} \to \boldsymbol{0}$ and κ_1, $\kappa_2 \to 0$, we can compute Bayes factor in the closed form

$$\mathrm{BF} = |\boldsymbol{V}_b|^{-1/2} |\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1}|^{-1/2} \left\{ 1 - \frac{\boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{y}}{\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{W}(\boldsymbol{W}^t\boldsymbol{W})^{-1}\boldsymbol{W}^t\boldsymbol{y}} \right\}^{-n/2} \quad (3)$$

where

$$\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{W}(\boldsymbol{W}^t\boldsymbol{W})^{-1}\boldsymbol{W}^t)\boldsymbol{L} \quad (4)$$

is the residuals of $\boldsymbol{L}$ after regressing out $\boldsymbol{W}$, and $|\cdot|$ denotes the determinant. Since $\boldsymbol{W}$ is assumed to contain a column of $\mathbf{1}$, each column in $\boldsymbol{X}$ is therefore centered.

Bayes factor in (3) uses the null as the base model and is thus called the null-based Bayes factor [c.f., Liang et al., 2008], which has been widely used in genetic association studies [Balding, 2006, Marchini et al., 2007, Guan and Stephens, 2008, Xu and Guan, 2014].

The use of the improper prior $\boldsymbol{V}_a^{-1} \to 0$, κ_1, $\kappa_2 \to 0$ has two merits. First, the limiting prior distributions for $\boldsymbol{a}$ and τ is equivalent to Jeffreys' prior [Ibrahim and Laud, 1991, O'Hagan and Forster, 2004], $p(\boldsymbol{a}, \tau) \propto \tau^{(q-2)/2}$, which is the standard choice of the non-informative prior for the nuisance parameters in the literature. Moreover, the posterior distributions for $\boldsymbol{a}$ and τ are proper. Second, Bayes factor in (3), which can be written as the limit of a sequence of Bayes factors with proper priors (see proof in Supplementary), is invariant to the shifting and scaling of $\boldsymbol{y}$ (or independent of $\boldsymbol{a}$ and τ). To see this, replace $\boldsymbol{y}$ with the standardized random variable

$$z \overset{\text{def}}{=} \tau^{1/2}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{a}). \quad (5)$$

and one can check (3) still holds.

We assume *a priori* that the expectation of $\boldsymbol{b}$ is zero so that the direction of the effect has no influence on Bayes factor. This prior for $\boldsymbol{b}$ is commonly adopted in practice [c.f., Jeffreys, 1961, Chap. 5]. For the NIG prior, we further assume independence between the effects and the covariates. Henceforth when we refer to the NIG prior we mean

$$\boldsymbol{V}_b = \sigma_b^2\boldsymbol{I}_p, \quad (6)$$

unless otherwise stated. The NIG prior and the corresponding Bayes factor will be the primary focus of this paper.

The second conjugate prior is Zellner's g-prior [Zellner, 1986, Liang et al., 2008]:

$$p(\boldsymbol{a}, \tau) \propto 1/\tau, \boldsymbol{b} \mid \tau \sim \text{MVN}(\mathbf{0}, \frac{g}{\tau}(\boldsymbol{X}^t\boldsymbol{X}) - 1). \quad (7)$$

This can be seen as a special case of the NIG prior with $\boldsymbol{V}_b = g(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ and thus Bayes factor under the g-prior can be derived straightforwardly from (3).

The third conjugate prior, the normal prior, can also be viewed as a special case of the NIG prior, when the error variance $\boldsymbol{\tau}^{-1}$ is assumed known:

$$\boldsymbol{a} \mid \tau \sim \mathrm{MVN}(\mathbf{0}, \tau^{-1}\boldsymbol{V}_a), \boldsymbol{b} \mid \tau \sim \mathrm{MVN}(\mathbf{0}, \tau^{-1}\boldsymbol{V}_b)\,. \quad (8)$$

Under this prior, and letting $\boldsymbol{V}_a^{-1} \to \boldsymbol{0}$, Bayes factor can also be computed analytically

$$\mathrm{BF} = |\boldsymbol{V}_b|^{-1/2} |\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1}|^{-1/2} \exp\left\{\frac{1}{2} z^t \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1})^{-1}\boldsymbol{X}^t z\right\}, \quad (9)$$

where $\boldsymbol{X}$ is defined in (4) and $\boldsymbol{z}$ is defined in (5). Because the error variance in most applications is unknown, the normal prior is more of theoretical interest. But, as we will see shortly, Bayes factor with the normal prior is approximately equal to that with the NIG prior. This is not too surprising because the data are very informative on the error variance.

Theorem 1. *For multi-linear regression* (1) *with the NIG prior* (2), *the g-prior* (7), *and the normal prior* (8), *under the null Bayes factors* (BF) *given in* (3) *and* (9) *follow*

$$2\log(\mathrm{BF}) = \sum_{i=1}^{p} [\lambda_i Q_i + \log(1 - \lambda_i)] + \varepsilon, \quad (10)$$

where $Q_i = (\boldsymbol{u}_i^t z)^2 \underset{\sim}{i.i.d} \chi_1^2$ *with* $\boldsymbol{z}$ *defined in* (5), *and* $(\lambda_i, \boldsymbol{u}_i)$ *is the ith eigenvalue-eigenvector pair of* $\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1})^{-1}\boldsymbol{X}^t$ *with* $\boldsymbol{X}$ *defined in* (4). *For the NIG prior and the g-prior,* $\boldsymbol{\varepsilon} = o_P(1)$ *vanishes in probability when the sample size* $n \to \infty$. *For the normal prior* $\boldsymbol{\varepsilon} = 0$.

Theorem 1 states that under the null 2 log BF is distributed as a weighted sum of chi-squared random variables with a shifted mean, and the weights and the mean-shift can be computed. By evaluating the distribution function, we can obtain a p-value associated with Bayes factor. For simple linear regression, Q_1 is asymptotically equal to the test statistic of the likelihood ratio test. Thus, the p-value associated with Bayes factor is the same as the p-value of the likelihood ratio test.

It is easy to see that $\lambda_i \in [0, 1)$. When the leading eigenvalue approaches 1, $\sum_{i=1}^{p} \log(1 - \lambda_i)$ goes to negative infinity, and so does the 2 log BF. Under two scenarios the leading eigenvalue does approach 1: the sample size goes to infinity or the prior diffuses indefinitely. Thus the prior-dependence nature of Bayesfactor and the Barlett's paradox both find their roots in the term $\sum_{i=1}^{p} \log(1 - \lambda_i)$. Moreover, when the sample size gets extraordinarily large, every λ_i approaches 1 and $\sum_{i=1}^{p} \lambda_i Q_i$ behaves like the likelihood ratio test statistic, which is distributed as a chi-squared random variable with p degrees of freedom, a special case of Wilks's [1938] theorem.

3 The P-value Associated with Bayes factor

Using Theorem 1 we can compute a p-value associated with Bayes factor given in (3), which henceforth is denoted by p_B. Since Bayes factor is a test statistic, p_B is naturally a Frequentist p-value. We point out p_B is also a Bayesian p-value. The p-values, or tail probabilities, are frequently computed in Bayesian context to check whether the model provides a good fit to the data. Bayesians p-values can be computed by comparing the observed test statistic to a predictive distribution obtained by integrating out the nuisance parameters over a reference distribution. Different Bayesian p-values can be computed using different reference distributions [Robins et al., 2000, Table 1]. Two well-known examples are the prior predictive p-values [Box, 1980] and the posterior predictive p-values [Rubin, 1984, Meng, 1994], which use the prior and the posterior as the reference distributions respectively. In our case, Bayes factors in (3) and (9) are independent of the nuisance parameters; thus, p_B can be viewed as a posterior predictive p-value. This convenience can be viewed as a bonus from the improper prior we used.

Corollary 1. *Denote by p_F the p-value from the likelihood ratio test, then for simple linear regression, we have asymptotically $p_B = p_F$.*

When $p = 1$, the right-hand side of (10) contains a single chi-squared random variable Q_1, which is asymptotically equal to the likelihood ratio test statistic, and therefore the two p-values are equal. In addition, for simple linear regression (10) explains the linear relationship between log BF and the likelihood ratio test statistic observed in Wakefield [2008] and Guan and Stephens [2008].

3.1 Weighted Sum of χ_1^2

For multi-linear regression, the right hand side of (10) contains a weighted sum of chi-squared variables. The weights $\lambda_1, ..., \lambda_p$ are functions of the prior effect size σ_b and the eigenvalues of the matrix $\boldsymbol{X}$ defined in (4). In contrast, the likelihood ratio test statistic is asymptotically equal to $\sum_{i=1}^{p} Q_i$ and distributed as χ_p^2, which does not take into account the eigenvalues of $\boldsymbol{X}$. Consequently, p_B no longer equals to p_F in general. One exception, however, is Bayes factor under the g-prior, where we have $\lambda_i = g/(g+1)$ for every i.

To compute p_B for multi-linear regression requires evaluating the distribution function of a weighted sum of chi-squared random variables, a challenging problem. Fortunately, a recent polynomial method by Bausch [2013] provides an efficient solution. Our contribution here is its implementation. We have implemented Bausch's method in C++, which allows one to compute p-values (tail probabilities) to an arbitrary precision efficiently.

First we briefly summarize Bausch's method and then provide more details of our implementation. Bausch pairs the chi-squared variables to take advantage of the identity

$$f_{\lambda_1 Q_1 + \lambda_2 Q_2}(x) = (4\lambda_1\lambda_2)^{-1/2} \exp\left(-\frac{\lambda_1 + \lambda_2}{4\lambda_1\lambda_2} x\right) I_0\left(\frac{\lambda_2 - \lambda_1}{4\lambda_1\lambda_2} x\right),$$

where Q_1 and Q_2 are independent χ_1^2 random variables and I_0 is the modified Bessel function of the first kind. I_0 can be approximated, to an arbitrary precision, by its Taylor expansion, and the series obtained can be integrated *algebraically* in the subsequent convolutions. The error of this algorithm only depends on the remainder terms of the Taylor expansions and thus can be quantified. Bausch showed that the complexity of this algorithm is polynomial in p.

In our implementation, the weights (λ_1, ..., λ_p) are sorted in a descending order and the chi-squared variables are then paired consecutively. If p is odd, the term with the smallest weight is retained for a numerical convolution in the last step. This pairing strategy aims to minimize the number of terms needed in Taylor expansions for a pre-specified precision. After Taylor expansion, we are faced with convolving gamma density functions (up to a normalizing constant). The order of the convolutions is determined by a single-linkage hierarchical clustering [Murtagh and Contreras, 2012] on the rate parameters of the gamma densities. Convolving two gamma densities of similar rates is computationally more efficient.

Our program has four features outstanding. First, we adopted GNU Multi-Precision Library so that our program can produce an arbitrarily small p-value without suffering underflow or overflow. Second, for an even number of chi-squared variables, p_B can be computed at an arbitrary precision; for an odd number of chi-squared variables, the error introduced at the last step of numerical convolution can be made arbitrarily small. Third, the terms of Taylor expansion are determined by a pre-specified precision and a strict error bound is provided. Last, since the program is written in C++, it is fast and suitable for studies that evaluate millions of tests, such as genetic association studies. Figure 1 demonstrates the efficiency of our program, for example, when $p = 10$ our program can evaluate ≈ 150 p-values per second. The speed appears to be quadratic in p. The weighted sum of chi-squared variables occurs frequently in statistical applications, such as the ridge regression, the variance component model, and recently the association testing for rare variants [Wu et al., 2011, Epstein et al., 2015]. We believe our program has a wide range of applications. The source code and executables of our program BACH (Bausch's Algorithm for CHi-square weighted sum) are freely available at http://haplotype.org.

3.2 Accuracy and Calibration of p_B for Finite Sample Sizes

Using Theorem 1, we can evaluate extremely small p-values, an important advantage in applications such as genome-wide association studies (GWAS) compared to the permutation method described in [Servin and Stephens, 2007]. Since p_B is quantified asymptotically, we are compelled to evaluate the accuracy and calibration of p_B for small sample sizes. We also computed the likelihood ratio test p-value p_F as a comparison because of its intrinsic connection to Bayes factor (and hence p_B) shown in Theorem 1.

We used a GWAS dataset to perform our simulation studies. The details of the dataset are provided in Section 5. For given n and p, we randomly sampled a subset of genotypes of n individuals and p SNPs. Then we simulated $\boldsymbol{y}$ under the null, that is, $\boldsymbol{y} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I_n})$. For each pair of sampled genotypes and simulated phenotypes, we computed p_B, using $\sigma_b = 0.2$, and p_F. We chose $n = 100$, 300 and $p = 1, 5, 10, 20$. For every combination of n and p we repeated the simulations 10^7 times. For $p = 1$, we can compute the exact p-value associated with Bayes factor using the F test (see proof in Supplementary). The comparison between exact values of p_B and p_B obtained from asymptotic approximation is shown in the top row of Figure 2. For $p > 1$, true values of p_B cannot be obtained analytically, we thus compared our asymptotic results against the theoretical uniform distribution. The two bottom rows show that for $n = 100$, the asymptotic results are conservative for small p-values; but for $n = 300$, the asymptotic results appear to be well-aligned with the theoretical prediction. Overall, Fig. 2 demonstrates that p_B is well calibrated, and the calibration is better than the p_F at the tail. We thus conclude that our asymptotic method for obtaining p_B is accurate and well-calibrated when the sample size is more than a few hundred.

4 Scaled Bayes Factors

Bayes factors are sensitive to the specification of priors. Let's consider the NIG prior with $\boldsymbol{V}_b = \sigma_b^2 \boldsymbol{I}_p$ and denote the singular values of $\boldsymbol{X}$ by δ_i for $i = 1, \ldots, p$. Then λ_i in (10) becomes $\lambda_i = \delta_i^2 / (\delta_i^2 + 1/\sigma_b^2)$, and thus

$$2 \log \mathrm{BF} \; = \sum_{i=1}^{p} \left\{ \frac{\delta_i^2 Q_i}{\delta_i^2 + 1/\sigma_b^2} - \log\left(1 + \delta_i^2 \sigma_b^2\right\} . \quad (11)\right.$$

Here we assume the sample size is sufficiently large such that the $o_P(1)$ error term can be safely omitted. The term $\lambda_i Q_i$ is bounded by Q_i (because $\lambda_i < 1$), but the term $\log(1/\delta_i^2\sigma_b^2)$ is monotonically increasing with respect to both δ_i and σ_b. When the sample size goes to infinity, δ_i goes to infinity; when the prior becomes more diffusive, σ_b goes to infinity. These properties give rise to the prior-dependence nature of Bayes factor and Bartlett's paradox.

By (10), $\mathbb{E}_0[2 \log \mathrm{BF}] = \sum_{i=1}^{p} (\lambda_i + \log(1 - \lambda_i))$, where $\mathbb{E}_0$ is the expectation evaluated under the null. Centering 2 log BF to obtain

$$2 \log \mathrm{sBF} \overset{\mathrm{def}}{=} 2 \log \mathrm{BF} - \mathbb{E}_0[2 \log \mathrm{BF}] = \sum_{i=1}^{p} \lambda_i (Q_i - 1) . \quad (12)$$

We call the quantity log BF − $\mathbb{E}_0$[log BF] the logarithm of the scaled Bayes factor (sBF). By definition, evaluating sBF requires computing $\mathbb{E}_0$[2 log BF]. In addition to direct computation, $\mathbb{E}_0$[2 log BF] can also be evaluated by simulating $\boldsymbol{y}$ under the null. A valid and convenient approach to simulating under the null was proposed by Kennedy [1995]. The approach is to permute $\boldsymbol{y}_{\mathrm{W}}$, the residuals of $\boldsymbol{y}$ regressing out covariates $\boldsymbol{W}$. Since 2 log BF is a weighted sum of chi-squared random variables, a modest number of permutations of $\boldsymbol{y}_W$ provide an accurate estimation of its mean under the null. The permutation might be advantageous over the analytical computation when the model is mis-specified.

Proposition 2. *The scaled Bayes factor has the following properties.*

- **a.** $\mathbb{E}_0[2\log \text{sBF}] = 0; \text{sBF}/\text{BF} = \prod_{i=1}^{p}\{\exp(-\lambda_i/2)/\sqrt{(1-\lambda_i)}\} > 1.$
- **b.** sBF *and* BF *have the same (Bayesian) p-value p_B.*
- **c.** *Let $\tilde{\boldsymbol{y}}$ be a permutation of* $\boldsymbol{y}$. *Then* $\text{BF}(\boldsymbol{y})/\text{BF}(\tilde{\boldsymbol{y}}) = \text{sBF}(\boldsymbol{y})/\text{sBF}(\tilde{\boldsymbol{y}}) \overset{\text{def}}{=} D(\tilde{\boldsymbol{y}})$, *and* $\mathbb{E}_P[\log D(\tilde{\boldsymbol{y}})] = \log \text{sBF}(\boldsymbol{y})$, *and* $\text{sBF}(\boldsymbol{y}) < \mathbb{E}_P[D(\tilde{\boldsymbol{y}})]$.

Comparing (11) and (12), the scaling removes from sBF the over-dependence on prior and Bartlett's paradox observed in BF (Fig. 3). The scaling is a function of $\boldsymbol{\lambda}$ which takes value in $[0, 1)^p$. If there is a gap between λ_i and 1, then the i-th component contributes modestly to the scaling. For example, when $p = 1$ the scaling is 1:5 when $\lambda_1 = 0.8$ and 2.0 when $\lambda_1 = 0.9$. When all $\lambda_i \rightarrow 0$, the scaling approaches 1 and meanwhile sBF → 1, as expected; when all $\lambda_i \rightarrow 1$, although the scaling factor blows up (sBF=BF → ∞), 2 log sBF is stable and $2\log \text{sBF} \rightarrow \chi_p^2 - p$.

Consider a multiple testing problem that tests $H_1, H_2, \ldots$ against H_0. Each alternative model is concerned with testing a single covariate in association with the response variable, and each covariate has the same λ_1. Then sBF and BF produce the same ranking for alltests, because the scaling coefficient is determined solely by λ_1. In light of the connection between BF and p_F [Wakefield, 2008, Guan and Stephens, 2008], we have that, asymptotically, sBF and p_F produce the same ranking for all tests that have the same λ_1. However, when λ_1 differs, the three statistics BF, sBF, and p_F (or p_B) produce different rankings.

4.1 sBF disregards informativeness of covariates under the null

Let us focus on simple linear regression. The treatment of multi-linear regression can be found in Supplementary. Let $\boldsymbol{V}_b = \sigma_b^2$ and $\boldsymbol{X}^t\boldsymbol{X} = \delta_1^2$. Then we have $\lambda_1 = \delta_1^2/(\delta_1^2 + 1/\delta_b^2)$. So λ_1 can be taken as a measurement of the *informativeness* of a covariate. In genetic association studies, a SNP's λ_1 is determined by the minor allele frequency and the prior effect size, and for a fixed prior effect size, the larger the minor allele frequency, the larger the λ_1.

Proposition 3. *Suppose two models H_1 and H_2 are each concerned with a single but different covariate, and H_1 associates with a larger λ_1. Denote* BF_j *and* sBF_j *of* BF *and* sBF *for model H_j* $(j = 1, 2)$. *We have*

$$\mathbb{E}_0[\log \text{BF}_1 - \log \text{BF}_2] < 0, \quad \mathbb{E}_o[\log \text{sBF}_1 - \log \text{sBF}_2] = 0\,. \qquad (13)$$

Although it is rudimentary to prove Proposition 3 (see Supplementary), the result is illuminating with respect to the difference between BF and sBF. Under the null, BF has the propensity to assign a larger value to a less informative covariate. In other words, BF penalizes more heavily to a more informative covariate. On the other hand, sBF disregards the informativeness of a covariate under the null. This indifference to the informativeness of sBF is advantageous under the alternative model (next section), because, loosely speaking, the over-penalty of BF on more informative covariates applies not just under the null, but also under the alternative.

4.2 BF and sBF under the local alternatives

The local alternatives are a sequence of alternatives that scale down the effect size when sample size increases so that the test statistic converges for large samples [c.f. Ferguson, 1996, Chap. 22]. The following theorem quantifies BF (and hence sBF) under the local alternatives.

Theorem 4. *For multi-linear regression* (1) *with the NIG prior* (2), *the g-prior* (7), *and the normal prior* (8), *under the local alternatives* $\boldsymbol{b} = \boldsymbol{\beta}/\sqrt{n\tau}$ *and assuming either* $\mathbf{L}^t\mathbf{L}$*/nconverges or* $\mathbf{L}$ *is entry-wise bounded, Bayes factors given in* (3) *and* (9) *follow*

$$2\log \mathrm{BF} = \sum_{i=1}^{p} \lambda_i Q_i + \log(1-\lambda_i) + \varepsilon, Q_i \sim X_1^2(\rho_i), \quad (14)$$

where $\varepsilon \sim o_P(1)$, *Qi is a noncentral chi-squared random variable with d.f.* = 1 *and thenoncentrality parameter* $\rho_i = (\boldsymbol{u}_i^t \boldsymbol{L}\boldsymbol{\beta})^2/n$, *and* $(\lambda_i, \boldsymbol{u}_i)$ is the ith eigenvalue-eigenvector pair of $\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{V}_b^{-1})^{-1}\boldsymbol{X}^t$. For the normal prior $\varepsilon = 0$.

Note in the above theorem 2 log BF has the same form as in (10). The only difference is that under the local alternatives $Q_1, \ldots, Q_p$ become noncentral chi-squared random variables instead of central chi-squared under the null. The assumptions on $\boldsymbol{L}$ guarantees the stochastic boundedness of ρ_i, permitting a Taylor expansion that leads to the linear approximation. These two assumptions are usually satisfied in practice, particularly in genetic association studies, where each entry of $\boldsymbol{L}$ is the allele counts of an individual at a marker and thus bounded by two.

Let's assume that the sample size is large enough so that the error term ε in Theorem 4 can be safely ignored. For simple linear regression, we have 2 log BF = $\lambda_1 Q_1 + \log(1-\lambda_1)$ and 2 log sBF = $\lambda_1(Q_1-1)$, where $Q_1 \sim \chi_1^2(\rho_1)$ is a noncentral chi-squared random variable. Because $\mathbb{E}[Q_1] = \rho_1 + 1$, we have $\mathbb{E}[2 \log \mathrm{sBF}] = \lambda_1\rho_1$, which is proportional to λ_1 for a fixed ρ_1. In other words, under the local alternatives, sBF tends to assign larger values to more informative covariates. On the other hand, $\mathbb{E}[2 \log \mathrm{BF}] = \lambda_1(\rho_1 + 1) + \log(1-\lambda_1)$ is not a monotonic function of λ_1 for a fixed ρ_1. Thus, the (unscaled) Bayes factor does not respect the informativeness of covariates under the alternative model.

Proposition 5. *Consider simple linear regression in the context of Theorem 4,*

- **a.** *Given* $b = \sigma_b/\sqrt{\tau}, Q_1 \sim \chi_1^2(\rho_1)$.
- **b.** *Let* $b \sim N(0, \sigma_b^2/\tau)$, *the marginal distribution (over b) of* Q_1 *is* $(1-\lambda_1)^{-1}\chi_1^2$, *a scaled central chi-squared distribution.*

The above proposition says that under the local alternatives the distribution of Q_1 (and hence BF and sBF) is determined by λ_1. In (a), the alternative is evaluated at a fixed point, while in (b) it is averaged over the prior distribution of b. Because the power of a test is determined by the alternative distribution of Q_1 (for a fixed null), Proposition 5 suggests that the statistical power is positively correlated with λ_1. This result is simple yet profound. Suppose we are faced with two tests with equal p-values that suggest the null should be rejected. Without knowing the powers of the tests, we cannot decide which rejection is more reliable or carries more evidence [Stephens and Balding, 2009]. Suppose two tests have different λ_1's but the same Q_1's, then the two tests have the same p-value. From 2 log sBF = $\lambda_1(Q_1-1)$, the scaled Bayes factor has a propensity to assign a largervalue to the test that has a larger power (or a larger λ_1), a desirable property. On the other hand, this desirable property is missing for the unscaled Bayes factor.

5 Applying to Ocular Hypertension GWAS Datasets

To illustrate how the scaled Bayes factor performs in real data analysis, we applied for access and downloaded two GWAS datasets from the database of Genotypes and Phentoypes (dbGaP). Both studies were funded by the National Eye Institute: one is the Ocular Hypertension Treatment Study [Kass et al., 2002] (henceforth OHTS, dbGaP accession number: phs000240.v1.p1), the other is National Eye Institute Human Genetics Collaboration Consortium Glaucoma Genome-Wide Association Study [Ulmer et al., 2012] (henceforth NEIGHBOR, dbGaP accession number: phs000238.v1.p1).

The phenotype of interest is the intraocular pressure (IOP). The OHTS dataset only contains individuals of high IOP (> 21). The NEIGHBOR dataset is a case-control design for glaucoma [Ulmer et al., 2012, Weinreb et al., 2014], in which many samples have IOP measurements because a high IOP is considered a major risk factor and a precursor phenotype for glaucoma. The NEIGHBOR dataset, however, contains case samples with small IOP and control samples with large IOP. Since IOP and glaucoma evidently have different genetic basis, though many are overlapping, we removed those samples. We also removed samples whose IOP measurements differ by more than 10 between the two eyes since such a large difference is likely to be caused by a different mechanism, e.g., physical accidents. The average IOP of the two eyes was used as the raw phentoype. We then performed the routine quality control for the genotypes using the same procedure described in Xu and Guan [2014]. OHTS and NEIGHBOR were genotyped on different SNP arrays, and there remained 301, 143 autosome SNPs genotyped in both datasets that passed QC. We then performed principal component analysis to remove the outliers and extracted 3,226 subjects (740 from OHTS and 2486 from NEIGHBOR) that clustered around the European samples in HapMap3 [The International HapMap Consortium, 2010]. We regressed out age, sex, and 6 leading principal components from the raw phenotypes, quantile normalized the residuals and used them as the phenotypes for single SNP analysis. We computed BF, sBF and p_B using prior $\sigma_b = 0.2$.

We first compared BF and sBF in different minor allele frequency (MAF) bins. Different MAF bins correspond to different bins of the informativeness (λ_1) of SNPs. Figure 4 shows that in each bin $\log_{10}$ sBF ~ $\log_{10}$ BF is roughly parallel to the line $y = x$, and more importantly the larger the MAF, the further are the points away from $y = x$, or in other words, $\log_{10}$ sBF − $\log_{10}$ BF is larger, which agrees well with the definition of sBF (Fig. 3), and fits the theoretical predictions (Propositions 3 and 5). Another noticeable feature in Fig. 4 is that the minimum value of sBF is larger than that of BF, which is consistent with the Proposition 2(a).

Next we examined the ranking of SNPs by different test statistics. Table 1 contains the top 20 SNPs in the ranking by BF. Rows were then sorted according to SNP’s chromosome and position. Incidentally, the top 2 hits (*rs*7518099 and *rs*4656461) are the same for all the three test statistics. The rankings are largely similar to one another among the three test statistics: BF, sBF, and p_B, particularly so between BF and p_B. There is, however, the noticeable exception of SNP *rs*7696626; with a ranking by sBF that is much worse than its rankings by BF and p_B. Not surprisingly, this SNP has the smallest MAF (0.023) among the 20 SNPs included in Table 1. This example fits the theoretical observations made in Proposition 3. We permuted the phenotypes once, and recomputed the test statistics. The permutation is to simulate under the null to confirm that $\mathbb{E}_0[\log \mathrm{sBF}] = 0$. Apparently sBF is more invariant against permutation compared to BF in the sense that $\log(\mathrm{sBF}(\boldsymbol{y})/\mathrm{sBF}(\tilde{\boldsymbol{y}})) \approx \log \mathrm{sBF}(\boldsymbol{y})$.

Our choice of the $\sigma_b = 0.2$ represents the prior belief of small but noticeable effect size in the context of GWAS [c.f. Burton et al., 2007]. To check how sensitive our results with respect to the choice of $\sigma_b = 0.2$, we redid the analysis using $\sigma_b = 0.5$. As predicted by our theory (Fig. 3), we observed that with $\sigma_b = 0.5$ BF tends to be smaller and sBF tends to be larger, and p_B remains unchanged. Moreover, the rankings of the SNPs remained mostly unchanged between the two choices of σ_b (Table S1 in Supplementary).

Lastly, although it was not our main objective, we examined the top hits in the association result. Our analysis reproduced three known genetic associations for IOP. Namely, the *TMCO1* gene on chromosome 1 (163.9M-164.0M) which was reported in [van Koolwijk et al., 2012]; a single hit *rs*2025751 in the *PKHD1* gene on chromosome 6 [Hysi et al., 2014]; and a single hit *rs*12150284 in the *GAS7* gene on chromosom 17 [*Ozel et al., 2014*]. A noticeable potentially novel finding is the gene *PEX14* on chrosome 1. Two SNPs, *rs*12120962 and *rs*12127400, have modest association signals from BF and p_B, but their scaled Bayes factors are noteworthy. *PEX14* encodes an essential component of the peroxisomal import machinery. The protein interacts with the cytosolic receptor for proteins containing a *PTS1* peroxisomal targeting signal. Incidentally, *PTS1* is known to elevate the intraocular pressure [Shepard et al., 2007]. In addition, a mutation in *PEX14* results in one form of Zellweger syndrome, and for children who suffer from Zellweger syndrome, congenital glaucoma is a typical neonatal-infantile presentation [Klouwer et al., 2015].

6 Discussion

In this paper, we quantify the null distribution of Bayes factors in the context of multilinear regression. We showed that under the null, 2 log BF is distributed as a weighted sum of chi-squared random variables. The null distribution allows us to compute the p-value associated with Bayes factor analytically, and we have developed a software package to do so efficiently. The software package can be used in a wide range of applications such as ridge regression, variance component model, and genetic association studies. The null distribution of Bayes factors also allows us to study the properties of Bayes factors, and we identified the dominant term in Bayes factor that leads to its excessive prior-dependence and Bartlett's paradox. We proposed the scaled Bayes factor, which depends less on the prior and is immune to Bartlett's paradox. We then studied the properties of the sBF under the null and the local alternatives. Compared to BF, sBF respects more to the informativeness of data.

Very often the covariates $\boldsymbol{L}$ are inferred from a statistical model, for example, imputed allele dosages in Guan and Stephens [2008] and the haplotype loading matrix in Xu and Guan [2014]. One would like to take into account the uncertainty of the inferred $\boldsymbol{L}$. In imputation-based association studies, one may compute the posterior mean of $\boldsymbol{L}$ and then perform the test. But in haplotype association analysis [Xu and Guan, 2014], using the posterior mean of $\boldsymbol{L}$ is impractical as one realization of $\boldsymbol{L}$ may be a column switching of the other, to say the least. A natural solution is to compute a Bayes factor for each realization of $\boldsymbol{L}$ and use the averaged Bayes factor as the test statistic. Then how to evaluate the associated p-value for the averaged Bayes factor? The same question also arises after obtaining an averaged Bayes factor from multiple choices of σ_b's. Two commonly used methods to combine p-values are Fisher's [1948] method and Stouffer et al.'s [1949] method. Fisher's method uses $-2\sum_{i=1}^{k}\log(p_i) \sim \chi^2_{2k}$; and Stouffer's method first obtains a Z-score for each p-value and then uses $\sum_{i=1}^{k} Z_i/\sqrt{k} \sim N(0, 1)$. Both methods assume the p-values to be combined are independent, while in our situations the p-values are obviously dependent. Motivated by Theorem 1, we propose to combine p-values using $2\log(\sum_{i=1}^{k}\exp(W_i/2)/k) \sim \chi^2_1$, where $W_i = \Psi^{-1}(1-p_i)$ and Ψ is the cumulative distribution function of χ^2_1. In essence, we converted each p-value to its associated Bayes factor, averaged Bayes factors, and computed p_B of the average Bayes factor. Since averaging over Bayes factors is always valid and Theorem 1 provides the necessary connection between p-values and Bayes factors, our approach to combining the correlated p-values appears to work well, at least for the afore-mentioned two examples, where the existing methods surely fail.

By definition, $\mathbb{E}_0[\text{BF}] = 1$ which is a *nice* property because it suggests that BF does not favor either the null or the alternative when the data are simulated under the null. A careful investigation into Proposition 3, however, revealed that this seemingly nice property effectively results in a greater penalty on more informative covariates. The scaled Bayes factor, on the other hand, satisfies $\mathbb{E}_0[\log \text{sBF}] = 0$, trading the property $\mathbb{E}_0[\text{BF}] = 1$ of the (unscaled) Bayes factor. Immediately, this trade suggests that sBF favors the alternative over the null (by Jensen's inequality or simply Proposition 2(a)). We argue that this trade brings several benefits to sBF: it depends less on prior; it becomes immune to Bartlett's paradox; and, more importantly, sBF becomes well calibrated with respect to permutation. Suppose we permute the response variable $\boldsymbol{y}$ once to obtain $\tilde{\boldsymbol{y}}$ and compute the test statistic (either BF or sBF) with $\tilde{\boldsymbol{y}}$, treating it as the test statistic under the null. Obviously, $2\log(\text{sBF}(\boldsymbol{y})/\text{sBF}(\tilde{\boldsymbol{y}}))$ is expected to have the same mean as that of $2\log\text{sBF}(\boldsymbol{y})$ (Proposition 2(c)). On the other hand $2\log(\text{BF}(\boldsymbol{y})/\text{BF}(\tilde{\boldsymbol{y}}))$ is expected to have a different (shifted) mean from $2\log\text{BF}(\boldsymbol{y})$. We believe this better calibration of sBF with respect to permutation will make it a better test statistic for Bayesian variable selection regression [Guan and Stephens, 2011].

In genetic association studies one routinely performs millions of simple linear regression to test for the association between each genetic variant and the phenotype. In general, sBF and p_B would produce different rankings for the variants, because their corresponding λ_1's differ. When a special prior, $\sigma_b \propto 1/\delta_1$, is used, however, sBF (and BF) will produce the same ranking as p_B [Wakefield, 2008, Guan and Stephens, 2008]. This prior, which produces the same λ_1 for all covariates, somewhat defeats the purpose of specifying a prior, because it practically eliminates the effect of a variant's variance to its test statistic. In multi-linear regression, such a special prior is the g-prior which sets every λ_i to $g/(g+1)$. At some sense ranking variants using sBF is more "informative" than using p_B, and we would like to control false discovery rate (FDR) for sBF. One approach is specifying the prior odds, multiplying the prior odds with BF or sBF to obtain the posterior odds, and then the posterior probability of association (PPA) for each variant. But specifying the prior odds is somewhat arbitrary, which unfortunately has a strong influence on PPA. An alternative approach is perhaps to develop a procedure that is similar to that of Benjamini and Hochberg [1995]. The Benjamini-Hochberg procedure relies on the null distribution of the p-values (which is uniform) to control FDR, but it is noted that the p-value may not be the optimal statistic for controlling FDR [Sun and Cai, 2009]. Since now we know the null distribution of sBF (and BF), we can estimate the expected FDR for sBF (and BF). Such a FDR controlling procedure will provide an alternative solution to "calibrating" Bayes factors (either scaled or unscaled), and it will strengthen the "Bayes/non-Bayes compromise," which is likely to attract more practitioners to apply Bayesian methods in their studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by United States Department of Agriculture/Agriculture Research Service under contract number 6250-51000-057 and National Institutes of Health under award number R01HG008157. The authors would like to thank Mark Meyer and Dennis Bier at Baylor College of Medicine for editorial assistance. The review comments from the editors and two anonymous reviewers greatly improved the clarity of our presentation.

References

Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics. 2006; 7(10):781–791.

Bartlett MS. A comment on D. V. Lindley's statistical paradox. Biometrika. 1957; 44(1-2):533–534.

Bausch J. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. Journal of Physics A: Mathematical and Theoretical. 2013; 46(50):505202.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995:289–300.

Box GEP. Sampling and bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General). 1980:383–430.

R Burton P, Clayton DG, R Cardon L, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–678. [PubMed: 17554300]

Epstein MP, Duncan R, Ware EB, et al. A statistical approach for rare-variant association testing in affected sibships. The American Journal of Human Genetics. 2015; 96 (4):543–554. [PubMed: 25799106]

Ferguson TS. A course in large sample theory, volume 49. Chapman & Hall London. 1996.

Fisher RA. Questions and answers #14. The American Statistician. 1948; 2(5):30–31.

Good IJ. Saddle-point methods for the multinomial distribution. The Annals of Mathematical Statistics. 1957:861–881.

Good IJ. The bayes/non-bayes compromise: A brief review. Journal of the American Statistical Association. 1992; 87(419):597–606.

Guan Y, Stephens M. Practical issues in imputation-based association mapping. PLoS Genetics. 2008; 4(12):e1000279. [PubMed: 19057666]

Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. Ann Appl Stat. 2011; 5(3):1780–1815.

Hysi PG, Cheng C, Springelkamp H, et al. Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. Nature genetics. 2014; 46(10):1126–1130. [PubMed: 25173106]

Ibrahim JG, Laud PW. On bayesian analysis of generalized linear models using jeffreys's prior. Journal of the American Statistical Association. 1991; 86(416):981–986.

Jeffreys H. The theory of probability. OUP Oxford; 1961.

Kass MA, Heuer DK, Higginbotham EJ, et al. The ocular hypertension treatment study: A randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. Archives of Ophthalmology. 2002; 120 (6):701–713. [PubMed: 12049574]

Kennedy FE. Randomization tests in econometrics. Journal of Business & Economic Statistics. 1995; 13(1):85–94.

Klouwer FCC, Berendse K, Ferdinandusse S, et al. Zellweger spectrum disorders: clinical overview and management approach. Orphanet journal of rare diseases. 2015; 10(1):1. [PubMed: 25603901]

Liang F, Paulo R, Molina G, et al. Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association. 2008; 103(481)

Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007; 39(7):906–913. [PubMed: 17572673]

Meng X. Posterior predictive p-values. The Annals of Statistics. 1994:1142–1160.

Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012; 2(1):86–97.

Nuzzo R. Scientific method: Statistical errors. Nature. 2014; 506:150–152. [PubMed: 24522584]

O'Hagan A, Forster JJ. Kendall's advanced theory of statistics, volume 2B: Bayesian inference, volume 2. Arnold; 2004.

Ozel AB, Moroi SE, Reed DM, et al. Genome-wide association study and meta-analysis of intraocular pressure. Human genetics. 2014; 133(1):41–57. [PubMed: 24002674]

Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of p values in composite null models. Journal of the American Statistical Association. 2000; 95(452):1143–1156.

Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics. 1984; 12(4):1151–1172.

Sawcer S. Bayes factors in complex genetics. European Journal of Human Genetics. 2010; 18 (7):746–750. [PubMed: 20179745]

Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. The American Statistician. 2001; 55(1):62–71.

Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics. 2007; 3(7):e114. [PubMed: 17676998]

Shepard AR, Jacobson N, Millar JC, et al. Glaucoma-causing myocilin mutants require the peroxisomal targeting signal-1 receptor (pts1r) to elevate intraocular pressure. Human Molecular Genetics. 2007; 16(6):609–617. [PubMed: 17317787]

Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009; 10(10):681–690.

Stouffer SA, Suchman EA, DeVinney LC. , et al. The American Soldier, Vol.1: Adjustment during Army Life. Princeton University Press; Princeton: 1949.

Sun W, Cai TT. Large-scale multiple testing under dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(2):393–424.

The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–58. [PubMed: 20811451]

Ulmer M, Li J, Yaspan BL, et al. Genome-wide analysis of central corneal thickness in primary open-angle glaucoma cases in the neighbor and glaugen consortiathe effects of cct-associated variants on poag risk. Investigative Ophthalmology & Visual Science. 2012; 53 (8):4468. [PubMed: 22661486]

van Koolwijk LME, D Ramdas W, Ikram MK, et al. Common genetic determinants of intraocular pressure and primary open-angle glaucoma. PLoS Genet. 2012; 8(5):e1002611. [PubMed: 22570627]

Wacholder S, Chanock S, Garcia-Closas M, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. Journal of the National Cancer Institute. 2004; 96(6):434–442. [PubMed: 15026468]

Wakefield J. Reporting and interpretation in genome-wide association studies. International Journal of Epidemiology. 2008; 37(3):641–653. [PubMed: 18270206]

Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: A review. JAMA. 2014; 311(18):1901–1911. [PubMed: 24825645]

Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics. 1938; 9(1):60–62.

Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011; 89(1):82–93. [PubMed: 21737059]

Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014; 197(3): 823–838. [PubMed: 24812308]

Zellner A. On assessing prior distributions and bayesian regression analysis with g-prior distributions. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti. 1986; 6:233–243.

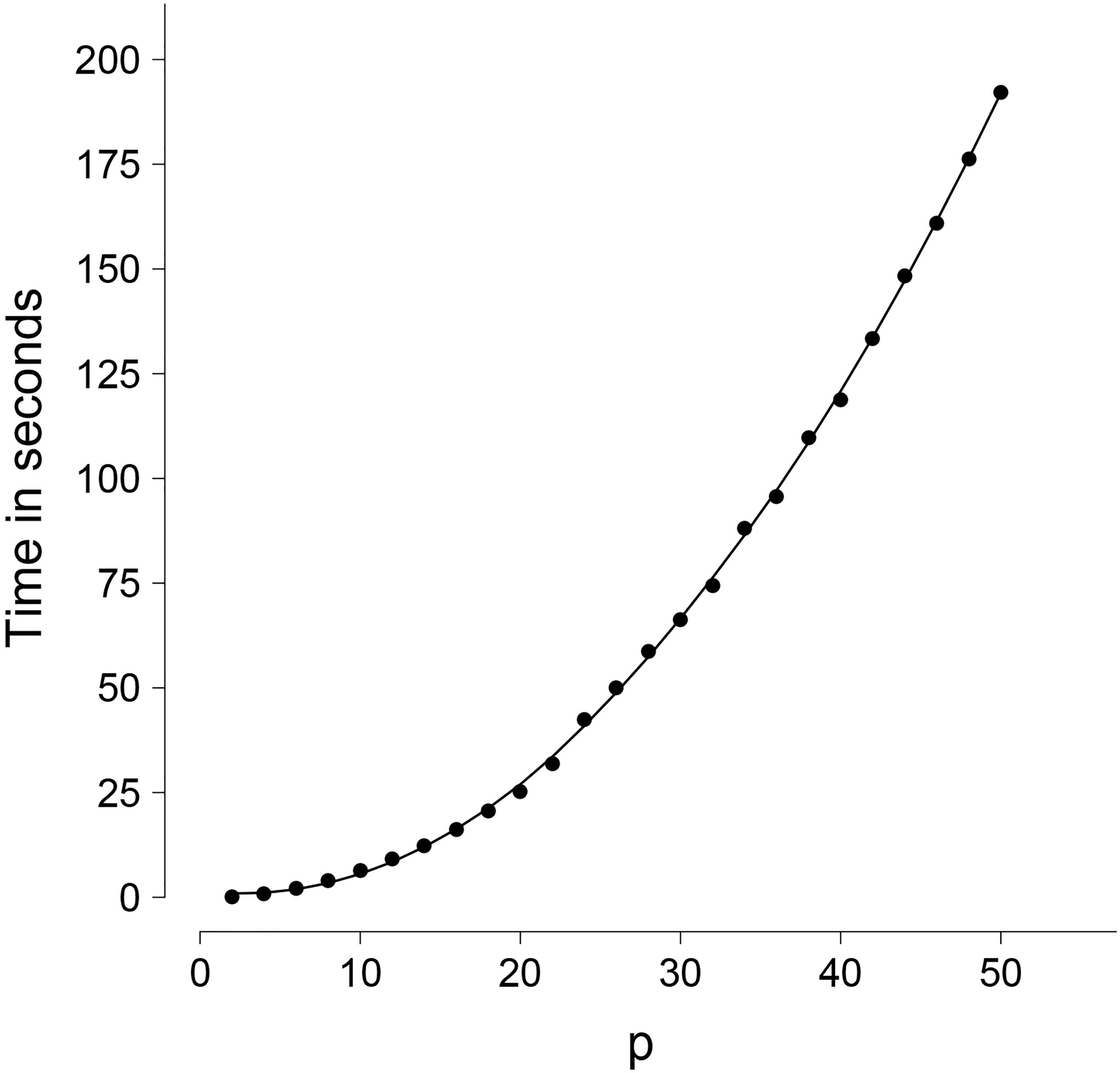

Figure 1.

Speed of evaluating p_B. The plot shows the time spent (y-axis) evaluating 1, 000 density functions to obtain 1, 000 p-values for different number of χ_1^2 components (2–50 on x-axis). The p-values were evaluated with relative error < 10^{-5}.

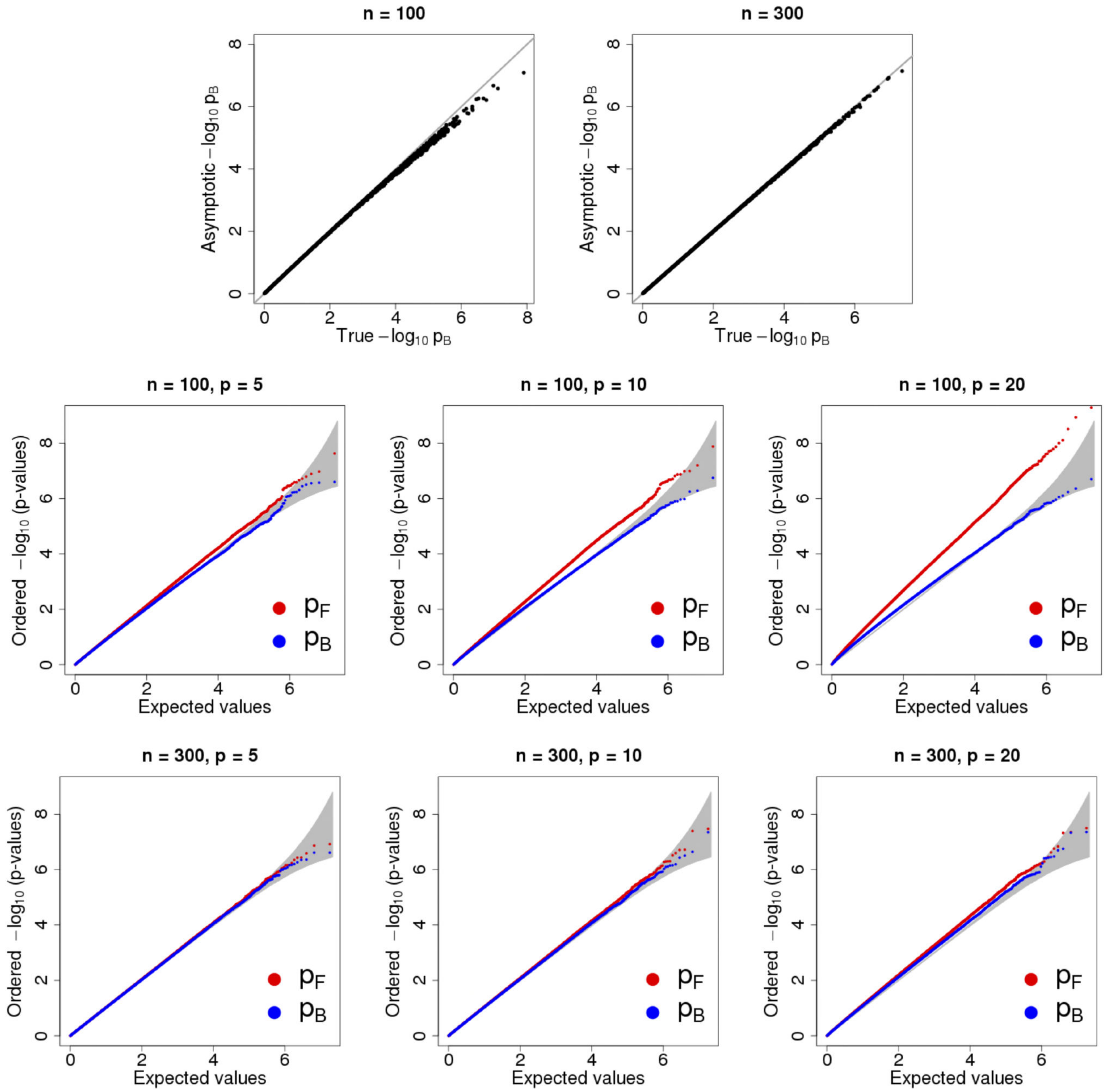

Figure 2.

Accuracy and calibration of p_B. The top row is for simple linear regression. The "true values" for p_B are obtained from F-tests. The y-axis is the asymptotic p_B. Line $y = x$ is marked in grey. Two bottom rows are for multi-linear regression. Red dots represent p_F (from likelihood ratio tests) and blue p_B. The grey region represents a 95% prediction intervals for uniformly distributed p-values.

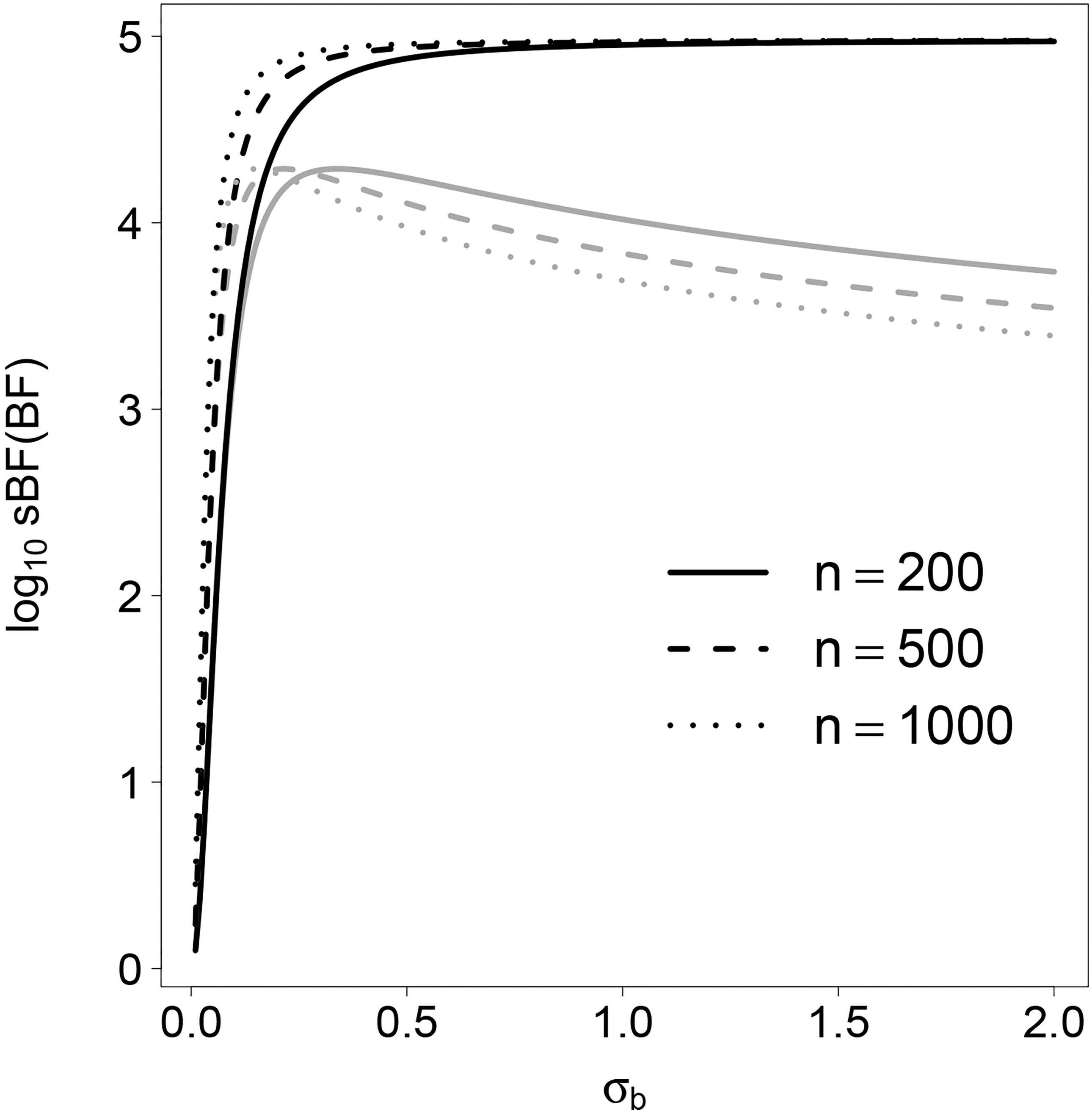

Figure 3.

BF and sBF as functions of σ_b. The plot is for simple linear regression of various sample sizes. BF is in gray and sBF black. BF and sBF are computed assuming the covariate has unit variance.

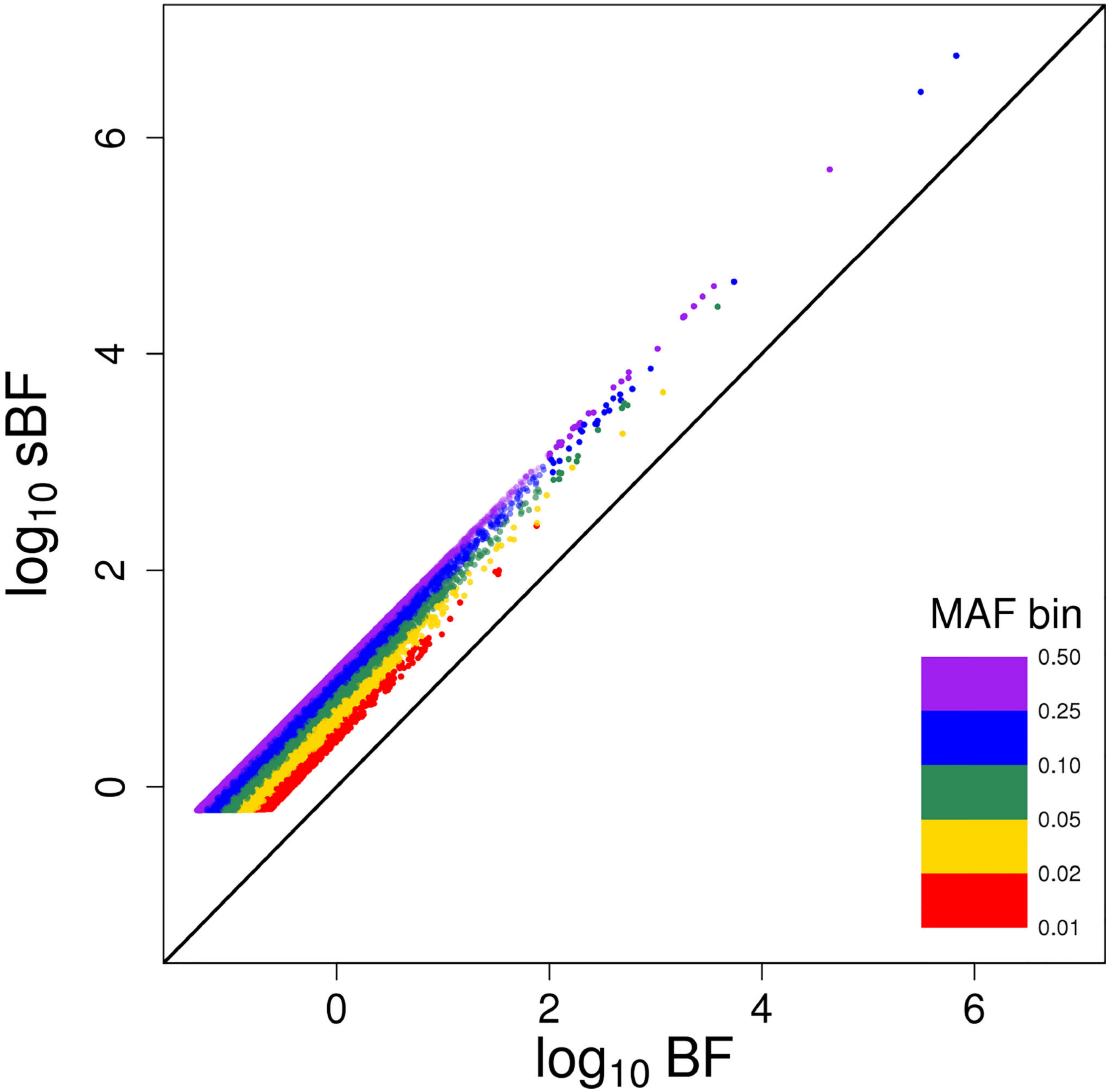

Figure 4.

The distributions of $\log_{10}$ BF and $\log_{10}$ sBF by different bins of minor allele frequency (MAF). The bins are marked by color and the diagonal line is $y = x$.

Table 1

Top 20 single SNP associations.

SNP	Chr	Pos	MAF	*bf*(*y*)	*bf*($\tilde{y}$)	*sbf*(*y*)	*sbf*($\tilde{y}$)	p(y)	p($\tilde{y}$)
rs12120962	1	10.53	0.384	3.88 (5)	−0.90	4.56 (4)	−0.21	5.63 (5)	0.01
rs12127400	1	10.54	0.384	3.61 (9)	−0.90	4.29 (8)	−0.21	5.34 (9)	0.01
rs4656461	1	163.95	0.140	5.71 (2)	−0.57	6.26 (2)	−0.03	7.51 (2)	0.46
rs7411708	1	163.99	0.428	3.69 (8)	−0.68	4.38 (7)	0.01	5.43 (8)	0.52
rs10918276	1	163.99	0.427	3.59 (10)	−0.66	4.28 (9)	0.03	5.33 (10)	0.54
rs7518099	1	164.00	0.140	6.04 (1)	−0.61	6.58 (1)	−0.07	7.85 (1)	0.38
rs972237	2	125.89	0.119	3.05 (15)	−0.62	3.56 (17)	−0.11	4.65 (18)	0.31
rs2728034	3	2.72	0.090	3.80 (6)	−0.62	4.27 (10)	−0.15	5.45 (7)	0.22
rs7645716	3	46.31	0.254	3.34 (11)	−0.88	3.98 (11)	−0.16	5.03 (11)	0.21
rs7696626	4	8.73	0.023	2.96 (18)	−0.33	3.20 (42)	−0.01	4.70 (16)	0.31
rs301088	4	53.53	0.473	2.95 (20)	−0.81	3.64 (16)	−0.11	4.65 (17)	0.31
rs2025751	6	51.73	0.466	3.78 (7)	−0.75	4.47 (6)	−0.06	5.53 (6)	0.41
rs10757601	9	26.18	0.443	3.09 (13)	−0.79	3.78 (12)	−0.10	4.80 (13)	0.33
rs10506464	12	62.50	0.164	2.97 (17)	−0.75	3.54 (18)	−0.18	4.59 (19)	0.15
rs10778292	12	102.78	0.140	4.00 (4)	−0.75	4.54 (5)	−0.21	5.68 (4)	0.02
rs2576969	12	102.80	0.271	3.07 (14)	−0.85	3.71 (14)	−0.20	4.75 (14)	0.08
rs17034938	12	102.85	0.127	3.23 (12)	−0.71	3.75 (13)	−0.19	4.85 (12)	0.13
rs1288861	15	43.50	0.120	2.95 (19)	−0.45	3.46 (20)	0.06	4.54 (21)	0.59
rs4984577	15	93.76	0.367	3.02 (16)	−0.64	3.69 (15)	0.04	4.71 (15)	0.56
rs12150284	17	9.97	0.353	4.95 (3)	−0.75	5.63 (3)	−0.07	6.75 (3)	0.38

The SNPs chosen have top 20 BF values using $\sigma_b = 0.2$. The rankings by three test statistics are given in the parentheses. $\tilde{\boldsymbol{y}}$ is obtained by permuting $\boldsymbol{y}$ once. SNP IDs are in bold if they are mentioned specifically in the main text. The column names are explained as following. Pos: genomic position in megabase pair according to HG18; *bf*(*y*): $\log_{10}$ BF(*y*); *bf*($\tilde{y}$): $\log_{10}$ BF($\tilde{y}$); *sbf*(*y*): $\log_{10}$ sBF(*y*); *sbf*($\tilde{y}$): $\log_{10}$ sBF($\tilde{y}$); *p*(**y**): $-\log_{10} p_B(\mathbf{y})$; and *p*($\tilde{y}$): $-\log_{10} p_B(\tilde{y})$.