THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 3, 250–256
http://dx.doi.org/10.1080/00031305.2015.1109548

Taylor & Francis
Taylor & Francis Group

Calibrated Bayes Factors in Assessing Genetic Association Models

J. G. Liao, Duanping Liao, and Arthur Berg

ABSTRACT

Three competing genetic models—additive, dominant, and recessive—are often considered in genetic association analysis. We propose and develop a calibrated Bayes approach for comparing these competing models that has the desired property of giving equal support to the three models when no genetic association is present. The naïve approach with noncalibrated priors is shown to produce misleading Bayes factors. The method is fully developed with simulation studies, real data analyses, and an efficient algorithm based on an asymptotic approximation. An illuminating connection to the Kullback–Leibler divergence is also established. The proposed calibrated prior can serve as a reference prior for a genetic association study or as a common baseline prior for comparing Bayes analyses of different datasets.

ARTICLE HISTORY
Received April 2015
Revised October 2015

KEYWORDS
Asymptotics; Bayes factor; Calibrated Bayes; Kullback–Leibler; Statistical Genetics

1. Introduction

This article considers the Bayes analysis of a common design of case-control genetic association studies. As an example, consider the case–control dataset in Table 1 from Sasieni (1997) which compares the HLA genotypes in women with and without cervical intraepithelial neoplasia. Typical analyses of this type of data primarily focus on testing for a statistical association between genotype and case–control status. Once an association is established, the next natural question is to determine the mode of association. Here, we present a solution based on Bayes factors to quantify the relative support of the data for three primary modes of genetic association.

We will first set up the notation and describe the three basic modes of association—additive, dominant, and recessive—between the three genotypes and disease (Laird and Lange 2010). For a given individual and biallelic single nucleotide polymorphism (SNP) locus, let x be the number of "a" alleles at the SNP (so $x = 0$, 1, or 2), and let D denote the disease status of the individual where $D = 1$ represents a "case" and $D = 0$ represents a "control." A natural model for the disease probability $P(D = 1|x)$ in a prospective design is (Chatterjee et al. 2009)

$$P(D = 1|x) = \frac{e^{\alpha+\beta g(x)}}{1 + e^{\alpha+\beta g(x)}},$$

where $g(\cdot)$ is chosen with respect to the genetic model of interest: $g(x) = x, g(x) = I(x \geq 1)$, or $g(x) = I(x = 2)$, for the additive, dominant, or recessive model, respectively.

Let ψ_j $(j = 0, 1, 2)$ be the odds ratios of disease for exposure $x = j$ relative to $x = 0$; that is,

$$\psi_0 = 1$$
$$\psi_1 = \frac{P(D = 1|x = 1)}{1 - P(D = 1|x = 1)} \cdot \frac{1 - P(D = 1|x = 0)}{P(D = 1|x = 0)}$$
$$\psi_2 = \frac{P(D = 1|x = 2)}{1 - P(D = 1|x = 2)} \cdot \frac{1 - P(D = 1|x = 0)}{P(D = 1|x = 0)}$$

The different genetic models can be characterized by the conditions in Table 2. It is particularly noted that the odds ratios in the table depend on β in the logistic regression but not on α.

The statistical literature so far has primarily focused on the hypothesis testing of $H_0 : \beta = 0$ with emphasis on achieving higher statistical power under a fixed Type I error. When H_0 is rejected, in the frequentist framework, the Akaike information criterion (AIC) (Akaike 1973) can be used to choose between the three competing models, which in this case is equivalent to picking the model with the largest log-likelihood. But AIC can only order the three competing models, it does not provide a numerical strength for the relative models. For this task, the Bayes factor (Kass and Raftery 1995) is uniquely suited. Gelman et al. (2014) (pp. 182–184) discussed the utility of the Bayes factor as particularly applicable when a discrete set of competing models are compared.

The Bayes factor follows from the fundamental Bayes' theorem and is conceptually superior for assessing the relative support of competing association models. This point will be made clear in the next section. The biggest barrier to its application is the need to specify a prior distribution for (α, β) under each of the three association models as such prior information is often scarce. This task is critical to the analysis as Bayes factors are highly sensitive to the specified prior distribution—much more so than the posterior distribution for parameter estimation (Aitkin 1991; Sinharay and Stern 2002). The noninformative prior that gives reasonable posterior distribution can lead to unreasonable Bayes factors.

This article proposes a new approach to computing the Bayes factors for the three genetic association models by calibrating the priors to have a desired frequentist property. Little (2006, 2011, 2012) provided a general description of a calibrated Bayes

CONTACT J. G. Liao jliao@phs.psu.edu Division of Biostatistics and Bioinformatics, Penn State University, College of Medicine, 500 University Drive, Hershey, PA 17033.

Table 1. Case–control genetic association study studied in Sasieni (1997).

	AA	Aa	aa	Total
Case	40	45	28	113
Control	273	100	43	416
Total	313	145	71	529

Table 2. Odds ratio characterization for different genetic association models.

Model	Characterization
No Association	$\psi_1 = \psi_2 = 1$
Additive (M_1)	$\psi_2 = \psi_1^2 = e^{2\beta}$
Dominant (M_2)	$\psi_1 = \psi_2 = e^{\beta}$
Recessive (M_3)	$\psi_1 = 1, \psi_2 = e^{\beta}$

approach, and the spirit of this approach is used to formulate the calibrated priors. Our calibration is based on the insight that the null H_0 is obtained by having $\beta = 0$ for all three association models, so these three types of associations are effectively deviations from H_0 in different directions. In the absence of prior information, the proposed calibrated approach is superior to the naïve approach of using the same diffuse noninformative priors for each association. The calibrated prior is also useful as a common baseline prior when comparing datasets generated from different sources.

Bayesian analysis of case–control studies was systematically reviewed by Mukherjee, Sinha, and Ghosh (2005). A general discussion of genetic association models is presented in Balding (2006) and connections to Bayesian methodology are described in Stephens and Balding (2009). Wakefield (2009) compared and contrasted the Bayes factor of $H_0 : \beta = 0$ over $H_1 : \beta \neq 0$ with a frequentist p-value approach. Xu, Yuan, and Zheng (2012) used test statistics based on the Bayes factor to enhance statistical power. There have been several approaches that comprehensively incorporate all three association models which address issues associated with multiple comparisons (Li and Gastwirth 2002; González et al. 2008; Hothorn and Hothorn 2009; Talluri, Wang, and Shete 2014).

2. Bayes Factors

A general data structure for a case–control study at a single biallelic SNP is presented in Table 3.

A natural model for these data is

$$\boldsymbol{y}_1 = (y_{10}, y_{11}, y_{12})' \sim \text{multinomial}(n_1, (\pi_{10}, \pi_{11}, \pi_{12}))$$
$$\boldsymbol{y}_0 = (y_{00}, y_{01}, y_{02})' \sim \text{multinomial}(n_0, (\pi_{00}, \pi_{01}, \pi_{02})),$$

where $\pi_{1j} = P(x = j|D = 1)$ and $\pi_{0j} = P(x = j|D = 0)$. It follows that (Satten and Kupper 1993; Mukherjee, Sinha, and Ghosh 2005; Chatterjee et al. 2009)

$$\pi_{1j} = \frac{\pi_{0j}\psi_j}{\sum_{k=0}^{2} \pi_{0k}\psi_k}, \qquad j = 0, 1, 2. \tag{1}$$

Table 3. A general data structure for a case–control genetic association model.

	$x = 0$	$x = 1$	$x = 2$	Total
Case	y_{10}	y_{11}	y_{12}	n_1
Control	y_{00}	y_{01}	y_{02}	n_0
Total	m_0	m_1	m_2	n

Since ψ_1 and ψ_2 only depend on the single parameter β (see Table 2), there are only three free parameters in this model: π_{01}, π_{02}, and β. We have the following decomposition for the likelihood of $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)'$:

$$\begin{aligned} p(\boldsymbol{y}|\pi_{01}, \pi_{02}, \beta, M_j) &= p(\boldsymbol{y}|y_{11} + y_{01} = m_1, y_{12} + y_{02} \\ &= m_2, \beta, M_j) \\ &\quad \times p(y_{11} + y_{01} = m_1, y_{12} + y_{02} \\ &= m_2, \pi_{01}, \pi_{02}, \beta, M_j). \end{aligned}$$

The first factor depends only on β and has the following noncentral hypergeometric distribution (McCullagh and Nelder 1989; Agresti 1992, p. 261)

$$f(\boldsymbol{y}|y_{11} + y_{01} = m_1, y_{12} + y_{02} = m_2, \beta, M_j) = \frac{\binom{n_1}{y_{10}, y_{11}, y_{12}}\binom{n_0}{y_{00}, y_{01}, y_{02}}\psi_1^{y_{11}}\psi_2^{y_{12}}}{\sum_{(z_{10}, z_{11}, z_{21})} \binom{n_1}{z_{10}, z_{11}, z_{12}}\binom{n_0}{m_0 - z_{10}, m_1 - z_{11}, m_2 - z_{12}}\psi_1^{z_{11}}\psi_2^{z_{12}}}, \tag{2}$$

where the denominator sums over all triplets (z_{10}, z_{11}, z_{12}) that form a 2×3 contingency table with the same row and column margins as the data table when replacing (y_{10}, y_{11}, y_{12}) with (z_{10}, z_{11}, z_{12}) (McCullagh and Nelder 1989, p. 261). To simplify notation, from here on we write the conditional distribution $f(\boldsymbol{y}|y_{11} + y_{01} = m_1, y_{12} + y_{02} = m_2, \beta, M_j)$ simply as $f(\boldsymbol{y}|\beta, M_j)$.

A straightforward Bayes factor calculation requires specification of prior distributions for β, π_{01}, and π_{02} under each of the three association models. In this article, we propose to base our Bayes factor on the conditional distribution $f(\boldsymbol{y}|\beta, M_j)$. There is a long tradition of using conditional inference in statistics to simplify analyses (Agresti 1992; Reid 1995). The conditional approach here simplifies the inference by only requiring a prior for β to be specified. This is particularly useful in analyzing thousands of SNPs in a genomewide-association study, as further discussed below, because it can be impractical to specify an informative prior for each SNP. Computationally, the conditional distribution reduces a three-dimensional integration to a one-dimensional integration making the computation more robust (Liao 1999). A Bayes formulation of conditional logistic is given by Rice (2004). However, we need to be concerned if ignoring the second factor in Equation (2) will lead to a substantial loss of information. Little (1989) argued that the marginal sum of a 2×2 table contains little information about the odds ratio. His argument can be extended to the 2×3 table here. Briefly, the value of $y_{11} + y_{01}$ and $y_{12} + y_{02}$ reflect their underlying means, which are $n_1\pi_{11} + n_0\pi_{01}$ and $n_1\pi_{12} + n_0\pi_{02}$. Irrespective of the size of these two means, however, we can always find a β as large as desired or as small as desired to satisfy the two marginal means with an appropriate pair of π_{01}, and π_{02}. Therefore, we cannot glean much information about β from $y_{11} + y_{01}$ and $y_{12} + y_{02}$ unless we can specify a strong prior on π_{01} and π_{02}. In general, basing inference on a conditional likelihood is analogous to throwing away part of the data (or, equivalently, the corresponding experiment) that either contains little information about the parameter of interest or contains information that is hard to untangle without a strong prior on the nuisance

parameters. It is noted that the same conditional distribution $f(\boldsymbol{y}|\beta, M_j)$ is obtained in a prospective study by conditioning on the total number of observed cases n_1 and the total number of observed controls n_0.

Let $f(\beta|M_j)$ be the specified prior distribution of β under model M_j, $j = 1, 2, 3$ (see Table 2). The marginal distribution of y under model M_j is given by

$$f(\boldsymbol{y}|M_j) = \int f(\boldsymbol{y}|\beta, M_j) f(\beta|M_j)\, d\beta. \tag{3}$$

Let $p(M_j)$ be the prior probability for model M_j being true. The posterior probability of model M_j given data $\boldsymbol{y}$ is given by Bayes' theorem:

$$f(M_j|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|M_j)P(M_j)}{\sum_{i=1}^{3} f(\boldsymbol{y}|M_i)P(M_i)}. \tag{4}$$

And the posterior odds of model M_j to M_i can be expressed as prior odds multiplied by the Bayes factor:

$$\frac{f(M_i|\boldsymbol{y})}{f(M_j|\boldsymbol{y})} = \underbrace{\frac{P(M_i)}{P(M_j)}}_{\text{prior odds}} \underbrace{\frac{f(\boldsymbol{y}|M_i)}{f(\boldsymbol{y}|M_j)}}_{\text{Bayes factor}}.$$

The Bayes factor

$$K_{ij} \triangleq \frac{f(\boldsymbol{y}|M_i)}{f(\boldsymbol{y}|M_j)} \tag{5}$$

can be directly interpreted as the support of data y of model M_i over model M_j. In this article, we will focus on Bayes factors. From these Bayes factors, one can easily compute the posterior distributions (4) for any set of priors on M_j.

This Bayes factor approach allows us to compare competing models using the fundamental Bayes' theorem. Specifying the prior $f(\beta|M_j)$ can be difficult as prior information is often not available, especially if the underlying genetic association is unknown. Furthermore, Bayes factors can be very sensitive to the priors (Aitkin 1991; Sinharay and Stern 2002). In the next section, the three genetic models are compared using Bayes factors that are calculated under priors that put them on equal grounds under the null hypothesis of no genetic association.

3. Calibrating The Priors

As in common practice, we shall specify prior $f(\beta|M_j)$ as $N(0, \sigma_j^2)$. This prior says that the unknown β is a quantity around 0, where σ_j should be large enough so that, for example, $|\beta| < \sigma_j$ for all plausible values of β. This guidance, however, is not sufficient in specifying the values of σ_1, σ_2, and σ_3.

We first show that the naïve choice of $\sigma_1 = \sigma_2 = \sigma_3$ can favor M_3 twice as much as M_1 even when there is clearly no association between the disease status and genotype. Consider an artificial dataset in Table 4, in which $y_i/n_i = 1/2$ for $i = 1, 2, 3$.

Table 4. Artificial data clearly demonstrating no association between genotype and case–control status.

	AA	Aa	aa
Case	200	100	25
Control	200	100	25

Table 5. Bayes factors for the data in Table 4 with equal variance priors.

$\sigma_1 = \sigma_2 = \sigma_3$	K_{21}	K_{31}
2	1.30	2.37
8	1.30	2.39

Table 5 gives the Bayes factors under the naïve specification of $\sigma_1 = \sigma_2 = \sigma_3$ evaluated at two different values of 2 and 8. The Bayes factor $K_{31} > 2$ implies $f(M_3|\boldsymbol{y})$ is more than twice as large as $f(M_1|\boldsymbol{y})$ over two different prior variances considered in this simulated example.

Under the null, we would normally expect $K_{21} = K_{32} = K_{31} = 1$ (i.e., no model be favored over another model), but this is not the case as demonstrated in Table 5. Therefore, we propose to calibrate σ_1, $\sigma_2 = k_2\sigma_1$, and $\sigma_3 = k_3\sigma_1$ so that $E_{\boldsymbol{y}\sim H_0}[\log K_{ij}] = 0$. In other words, we wish to determine k_2 and k_3 such that

$$\begin{aligned} E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_1)] &= E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_2)] \\ &= E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_3)]. \end{aligned} \tag{6}$$

The assumption $\boldsymbol{y} \sim H_0$ means that $\boldsymbol{y}$ follows the central hypergeometric distribution with probability mass

$$p(\boldsymbol{y}) = \frac{\binom{n_1}{y_{10}, y_{11}, y_{12}}\binom{n_0}{y_{00}, y_{01}, y_{02}}}{\binom{n}{m_0, m_1, m_2}}. \tag{7}$$

Note that $E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_j)]$ is the Kullback discrepancy of model M_j from H_0. Our calibration therefore requires equal Kullback discrepancy between the null and each M_j.

Given data $\boldsymbol{y}$, each $f(\boldsymbol{y}|M_j)$ is a function of β and $E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_j)]$ can be computed by numerical integration. The parameters k_2 and k_3 can then be solved numerically for any given value of σ_1.

The calibration parameters and calibrated Bayes factors for the artificial data in Table 4 are presented in Table 6.

Here, we observe that substantially different scaling parameters are found for the priors. Furthermore, any Bayes factor that is biased toward one of the three competing models under the null is misleading, but our calibration procedure resolves this issue by equalizing the Bayes factor of each model under H_0.

In Tables 5 and 6, the Bayes factors K_{21} and K_{31} are shown to mostly depend on the choice of k_2 and k_3 and only very weakly on the value of σ_1. We now show this is generally true under some mild conditions. Let $\hat{\beta}_j$ be the MLE of β under model M_j, $j = 1, 2, 3$. For large sample sizes, the likelihood $f(y|\beta, M_j)$ will concentrate in a small interval, which is expressed as $(\hat{\beta}_j - \delta_j, \hat{\beta}_j + \delta_j)$. Now suppose that σ_j is large enough so that the prior $f(\beta|M_j) \sim \frac{1}{\sigma_j}\phi(\beta/\sigma_j)$, with $\phi(\cdot)$ representing the density of a standard normal distribution, is approximately constant on

Table 6. Calculated calibration parameters (k_2, k_3) and resulting Bayes factors showing the calibration is effective.

σ_1	(k_2, k_3)	K_{21}	K_{31}
2	(1.30,2.41)	1.00	0.99
8	(1.30,2.41)	1.00	0.99

this interval. Now further assume that $\frac{\hat{\beta}_j}{\sigma_j}$ is close to 0. This yields the following approximation:

$$\begin{aligned} f(\boldsymbol{y}|M_j) &= \frac{1}{\sigma_j}\int_{-\infty}^{\infty} f(\boldsymbol{y}|\beta, M_j)\phi\left(\frac{\beta}{\sigma_j}\right) d\beta \\ &\approx \frac{1}{\sigma_j}\phi\left(\frac{\hat{\beta}_j}{\sigma_j}\right)\int_{\hat{\beta}_j-\delta_j}^{\hat{\beta}_j+\delta_j} f(\boldsymbol{y}|\beta, M_j)\, d\beta \\ &\approx \frac{1}{\sigma_j}\phi\,(0)\underbrace{\int_{\hat{\beta}_j-\delta_j}^{\hat{\beta}_j+\delta_j} f(\boldsymbol{y}|\beta, M_j)\, d\beta}_{=A_j}. \end{aligned}$$

Note that A_j is free of σ_j. It is now easy to see that K_{ij} mostly depends on the ratio of σ_j to σ_i.

4. Asymptotics

The method for calibrating the priors described in the previous section can be computationally burdensome, especially with large sample sizes, as the calculation of $E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_j)]$ requires summing over thousands or tens of thousands of 2×3 tables of the same margin in the denominator of the noncentral hypergeometric distribution (2). By studying the asymptotics of the proposed procedure, we have derived a closed-form and accurate approximation to k_2 and k_3. Additionally, the asymptotic development has insightful links to Fisher information.

We study the asymptotic properties of the calibration procedure under fixed marginals; that is, we let

$$\begin{aligned} (m_0, m_1, m_2) &= s(m_0^*, m_1^*, m_2^*) \\ (n_0, n_1) &= s(n_0^*, n_1^*), \end{aligned}$$

where $m_0^*, m_1^*, m_2^*, n_0^*, n_1^*$ are all positive and $s \to \infty$ along the positive integers. We start by considering the asymptotic behavior of $f(\boldsymbol{y}|M_j)$ under $s \to \infty$. Applying the Laplace approximation (Kass and Raftery 1995, Section 4) on $f(\boldsymbol{y}|M_j)$ yields

$$\begin{aligned} \log f(\boldsymbol{y}|M_j) = \log\sqrt{2\pi} &+ \underbrace{\log f(\boldsymbol{y}|\hat{\beta}, M_j)}_{*} + \underbrace{\log f(\hat{\beta}|M_j)}_{**} \\ &- \frac{1}{2}\underbrace{\log I(\boldsymbol{y}, \hat{\beta}, M_j)}_{***} + O_p(s^{-1}), \end{aligned} \tag{8}$$

where $\hat{\beta}$ is the maximum likelihood estimate under any of the three association models and $I(\boldsymbol{y}, \hat{\beta}, M_j)$ is the observed Fisher's information; that is,

$$I(\boldsymbol{y}, \hat{\beta}, M_j) = \left.\frac{-d^2 \log f(\boldsymbol{y}|\beta, M_j)}{d\beta^2}\right|_{\beta=\hat{\beta}}.$$

We now proceed to determine the conditions on k_2 and k_3 to approximately obtain (6) based on the asymptotic approximation (8). Note that (6) requires the expected value of the right-hand side of (8) to be equal for each M_j under H_0, and each of the three starred terms in the right-hand side of (8) is considered sequentially.

The first term, $\log f(y|\hat{\beta}, M_j)$, can be ignored in this context since its expected value yields asymptotically equivalent values for each model M_j. More explicitly, we have

$$\begin{aligned} \log f(\boldsymbol{y}|\hat{\beta}, M_j) = &\underbrace{\log f(\boldsymbol{y}|\hat{\beta}, M_j) - \log(f(\boldsymbol{y}|\beta = 0, M_j)}_{\sim \frac{1}{2}\chi_1^2 + O_p(s^{-1})} \\ &+ \log(f(\boldsymbol{y}|\beta = 0, M_j), \end{aligned}$$

where the χ^2 approximation follows from likelihood ratio theory (McCullagh and Nelder 1989, Appendix). Since $\log(f(\boldsymbol{y}|\beta = 0, M_j)$ are identical for each M_j, the expected value of $\log f(\boldsymbol{y}|\hat{\beta}, M_j)$ is asymptotically the same for all M_j.

The second term, $\log f(\hat{\beta}|M_j)$, with the prior specification $\beta|M_j \sim N(0, \sigma_j)$ and the fact that $\hat{\beta} \to 0$ under all M_j, yields

$$\log f(\hat{\beta}|M_j) \to \log f(\beta = 0|M_j) = -\log(\sqrt{2\pi}) - \frac{1}{2}\log\sigma_j^2.$$

Finally, the observed Fisher information in the third term, $I(\boldsymbol{y}, \hat{\beta}, M_j)$, converges to $I(\boldsymbol{y}, 0, M_j)$ and is directly calculated from (2) yielding

$$I(\boldsymbol{y}, M_j) \triangleq I(\boldsymbol{y}, \hat{\beta} = 0, M_j) = \begin{cases} \text{var}\left(y_{11} + 2y_{12}\right) & \text{for } j = 1 \\ \text{var}\left(y_{11} + y_{12}\right) & \text{for } j = 2 \\ \text{var}\left(y_{12}\right) & \text{for } j = 3, \end{cases} \tag{9}$$

where the variance is computed under the central hypergeometric distribution yielding the following expressions:

$$\begin{aligned} \text{var}(y_{1i}) &= \frac{m_i}{n}\left(1 - \frac{m_i}{n}\right)\frac{n_0 n_1}{n-1} \\ \text{cov}(y_{1i}, y_{1j}) &= -\frac{m_i m_j}{n^2}\frac{n_0 n_1}{n-1}. \end{aligned}$$

Therefore, in an effort to satisfy (6), we obtain

$$\begin{aligned} k_2 \approx \tilde{k}_2 &\triangleq \sqrt{\frac{\text{var}\left(y_{11} + 2y_{12}\right)}{\text{var}\left(y_{11} + y_{12}\right)}} \\ &= \sqrt{\frac{-m_1^2 + (n - 4m_2)\, m_1 + 4\,(n - m_2)\, m_2}{m_0\,(m_1 + m_2)}} \\ k_3 \approx \tilde{k}_3 &\triangleq \sqrt{\frac{\text{var}\left(y_{11} + 2y_{12}\right)}{\text{var}\left(y_{12}\right)}} \\ &= \sqrt{\frac{-m_1^2 + (n - 4m_2)\, m_1 + 4\,(n - m_2)\, m_2}{(n - m_2)\, m_2}}. \end{aligned} \tag{10}$$

The above derivations are summarized in the following theorem.

Theorem 1. Let $\beta|M_j \sim N(0, \sigma_j^2)$ with $\sigma_2 = \tilde{k}_2\sigma_1$ and $\sigma_3 = \tilde{k}_3\sigma_1$, then (6) holds asymptotically; that is,

$$\begin{aligned} E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_1)] &= E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_2)] + o(1) \\ &= E_{\boldsymbol{y}\sim H_0}[\log f(\boldsymbol{y}|M_3)] + o(1). \end{aligned}$$

The asymptotic derivation culminating in Theorem 1 yields additional insight into the underlying bias and subsequent correction. For a typical design governed by the Hardy–Weinberg law and $m_0 > m_1 > m_2$, Equation (9) implies

$$I(\boldsymbol{y}, M_1) > I(\boldsymbol{y}, M_2) > I(\boldsymbol{y}, M_3).$$

Table 7. Comparison of k_2 and k_3 to the approximated values $\tilde{k}_2$ and $\tilde{k}_3$ over a broad range of experimental designs.

σ_1	n_1	n_1	n_1	m_1	k_2	$\tilde{k}_2$	k_3	$\tilde{k}_3$
	10	10	10	9	1.80	1.73	1.80	1.73
	40	40	40	36	1.74	1.73	1.74	1.73
2	160	160	160	72	1.74	1.73	1.74	1.73
	40	20	10	21	1.43	1.43	2.45	2.08
	160	80	40	84	1.46	1.47	2.12	2.08
	640	320	160	336	1.47	1.47	2.09	2.08
	10	10	10	9	1.92	1.73	1.92	1.73
	40	40	40	36	1.74	1.73	1.74	1.73
8	160	160	160	72	1.74	1.73	1.74	1.73
	40	20	10	21	1.43	1.47	2.95	2.08
	160	80	40	84	1.46	1.47	2.12	2.08
	640	320	160	336	1.47	1.47	2.09	2.08

So under the naïve prior setting of $\sigma_1 = \sigma_2 = \sigma_3$, the Bayes factor favors M_3 (the recessive model) the most and M_1 (the additive model) the least.

We note that the proposed calibrated prior depends on the Fisher information of a particular model. In Bayesian statistics, previously proposed priors that depend on the Fisher's information include the Jefferys prior and the prior that connects the Schwarz's Bayesian information criterion to a Bayes factor approximation (Kass and Wasserman 1995).

Calculating the true calibration parameters can be computationally burdensome ranging from minutes to hours depending on the sample size, whereas the closed-form approximations can be computed instantaneously. Comparisons of the asymptotic approximations $\tilde{k}_2$ and $\tilde{k}_3$ to the true calibration parameters k_2 and k_3 are made over several experimental designs and presented in Table 7. It is observed that $\tilde{k}_2$ and $\tilde{k}_3$ are accurate approximations to k_2 and k_3. Performance comparisons in terms of the Bayes factors are provided with real data applications in the next section.

The exact calibration in the previous section and the approximate calibration in this section are implemented in R and available at *https://sites.google.com/site/jiangangliao/*.

5. Simulation Study and Applications

Monte Carlo simulation is conducted to study the property of the proposed Bayes factors. We assume the Hardy–Weinberg equilibrium for the control population by making

$$\pi_{00} = p^2, \qquad \pi_{01} = 2p(1-p), \qquad \pi_{02} = (1-p)^2$$

with $p = 0.2$. We simulate under the four models in Table 2. For models M_1, M_2, and M_3, two values of β are considered: $\beta = \log(1.5)$ and $\beta = \log(2)$. The probabilities $\pi_{10}, \pi_{11}, \pi_{12}$ are computed using Equation (1). We then draw a case–control dataset in the form of Table 3 with $n_1 = n_0 = 500$ and $n_0 = n_1 = 1000$. For each dataset, we compute the noncalibrated Bayes factors $\log(K_{21})$ and $\log(K_{31})$ under priors $\sigma_1 = \sigma_2 = \sigma_3 = 2$. These are compared to the calibrated Bayes factors using the approximated calibration parameters $\tilde{k}_2$ and $\tilde{k}_3$. This process is replicated 1000 times and the average of the $\log(K_{21})$ and $\log(K_{31})$ are given in Tables 8 and 9 for the two different sample sizes.

The simulations show a substantial improvement when using calibrated priors over naïve priors. As discussed in Section 4, the noncalibrated Bayes factors have a bias toward the recessive model and to a lesser extent the dominant model. Furthermore, this bias persists even as the sample size grows. The calibrated Bayes factors correct this problem as the average $\log(K_{21})$ and $\log(K_{31})$ are much closer to zero. Under the genetic association models (M_1, M_2, and M_3), the calibrated Bayes factors produce desired results that support the true underlying association model, and the support becomes stronger with increased sample size. It is also noted that the bias that favors the recessive model is still present when the true underlying model is one of M_1, M_2, or M_3, as evidenced by the average $\log(K_{31})$ being always larger for the noncalibrated priors compared to the calibrated counterparts.

We conclude this section with the analysis of two real datasets previously analyzed in the literature. The first dataset, published in *Biometrics* (Sasieni 1997), is the dataset in Table 1. The second dataset is from Lu et al. (2006) that was analyzed in Talluri, Wang, and Shete (2014) and is reproduced in Table 11.

Table 8. Simulation demonstrating reduced bias with calibrated priors ($n_0 = n_1 = 500$).

		Noncalibrated		Calibrated	
Association model	Value of β	avg $\log(K_{21})$	avg $\log(K_{31})$	avg $\log(K_{21})$	avg $\log(K_{31})$
No association	0	0.1677	1.061	0.0057	0.0200
Additive	log(1.5)	− 0.8144	− 3.769	− 1.004	− 4.660
	log(2)	− 3.219	− 12.27	− 3.426	− 13.00
Dominant	log(1.5)	0.7044	− 2.918	0.5426	− 3.921
	log(2)	1.898	− 11.00	1.7452	− 11.96
Recessive	log(1.5)	− 0.07489	1.594	− 0.2660	0.6359
	log(2)	− 0.7766	2.833	− 0.9925	1.960

Table 9. Simulation demonstrating reduced bias with calibrated priors ($n_0 = n_1 = 1000$).

		Noncalibrated		Calibrated	
Association model	Value of β	avg $\log(K_{21})$	avg $\log(K_{31})$	avg $\log(K_{21})$	avg $\log(K_{31})$
No association	0	0.1701	1.0516	0.006793	0.0008285
Additive	log(1.5)	− 1.6752	− 8.0633	− 1.8673	− 8.964
	log(2)	− 6.328	− 25.64	− 6.536	− 26.38
Dominant	log(1.5)	1.3011	− 7.071019	1.1396	− 8.0845
	log(2)	3.787	− 23.09	3.633	− 24.07
Recessive	log(1.5)	− 0.3252	2.1200	− 0.5179	1.1534
	log(2)	− 1.793	4.661	− 2.011	3.781

Table 10. Analysis of dataset in Table 1 from Sasieni (1997).

prior	σ_1	σ_2	σ_3	K_{21}	K_{31}
naïve	2	2	2	2.43	2.25e-4
Calibrated (k_2, k_3)	2	2.92	4.27	1.85	1.15e-4
Approximate ($\tilde{k}_2, \tilde{k}_3$)	2	2.92	4.22	1.84	1.17e-4
naïve	8	8	8	2.70	2.38e-4
Calibrated (k_2, k_3)	8	11.7	17.1	1.87	1.11e-4
Approximate ($\tilde{k}_2, \tilde{k}_3$)	8	11.7	16.9	1.86	1.13e-4

Table 11. Dataset from Lu et al. (2006) considers association of an eNOS polymorphism with breast cancer.

	AA	Aa	aa
Case	203	185	35
Control	167	200	54

Table 12. Analysis of Lu et al. (2006) dataset in Table 11.

prior	σ_1	σ_2	σ_3	K_{21}	K_{31}
naïve	2	2	2	0.445	0.380
Calibrated (k_2, k_3)	2	2.65	4.31	0.339	0.181
Approximate $(\tilde{k}_2, \tilde{k}_3)$	2	2.65	4.23	0.338	0.182
naïve	8	8	8	0.448	0.389
Calibrated (k_2, k_3)	8	10.6	17.2	0.337	0.181
Approximate $(\tilde{k}_2, \tilde{k}_3)$	8	10.6	17.1	0.338	0.181

Results are given in Table 10 and Table 12, respectively. For both datasets, the calibrated prior corrects the inherent bias toward recessive model M_3 with meaningful change in the resulting Bayes factors. For the Lu et al. dataset, for example, the Bayes factor K_{13} changes from $1/0.38 = 2.63$ to $1/0.181 = 5.53$ yielding much stronger support for M_1 after calibration. Additionally, this analysis also demonstrates that the approximations $\tilde{k}_2$ and $\tilde{k}_3$ are very accurate.

6. Discussion

Determining the type of association between disease and a genotype is an important basic problem in biomedical research. The Bayes factor is conceptually a natural approach in quantifying the support of the data for one association model over the other. We show that the naïve choice of equal prior variances can lead to unexpected and undesirable Bayes factors that usually favor the recessive model, however our proposed calibrated prior resolves this inherent bias. This approach led to a comprehensive development of the calibration method including an asymptotic analysis and confirming simulation studies. The proposed calibrated priors can serve as "reference priors" for genetic association models when it is difficult to elucidate a prior for the effect size under each association model. It can also provide a common baseline prior for comparing Bayes analyses of different datasets. More generally, from a methodological perspective, the calibrated Bayes approach proposed in this study takes advantage of a unique structure in competing genetic association models, which can also be used in other applications in which the competing models share a common null. Unfortunately, our approach does not seem to apply to the problem of choosing between nested models as discussed by Kass and Wasserman (1995).

It is now common practices to genotype many variants as in genome-wide association studies (GWAS). The Bayes factors $\log(K_{21})$ and $\log(K_{31})$ can be computed for each of the thousands to millions of SNPs. With so many Bayes factors computed, however, special care is needed in interpreting the results. Following a similar approach as in Efron (2010) for interpreting a large number of p-values in a microarray experiment or GWAS, we can start with the distribution of all $\log(K_{21})$ as a whole and the distribution of all $\log(K_{31})$ as a whole and compare them through simulations against their expected behavior under the global null that no SNP is associated with the disease. Our calibrated Bayes factors make this task easier because all individual $\log(K_{21})$ and individual $\log(K_{31})$ have mean 0 under the global null as a consequence of our calibration. Similarly, we can compare the extreme values of $\log(K_{21})$ and the extreme values $\log(K_{31})$ against their expected distribution under the global null. Substantial future research, however, is needed to fully develop these ideas.

With n_1 and n_0 often as high as 5000, computational time can be a concern. The approximations $\tilde{k}_2$ and $\tilde{k}_3$ only involve the central hypergeometric distribution and can be computed rapidly. The noncentral hypergoemetric in Equation (2) is the only bottleneck for large values of n_1 and n_0. Fortunately, a very accurate normal approximation to (2) is provided in (McCullagh and Nelder 1989, p. 262). With this approximation, Bayes factors for a SNP are no more time consuming than calculating Wilcoxon test. The application of our method to GWAS should therefore be computationally feasible.

Acknowledgment

The authors thank the editor for her careful editing and the associate editor and an anonymous referee for their constructive criticisms.

References

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–153. [251]

Aitkin, M. (1991), "Posterior Bayes Factors," *Journal of the Royal Statistical Society*, Series B, 53, 111–142. [250,252]

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Proceedings of the Second International Symposium of Information Theory*, Akademiai Kiado, Budapest, pp. 267–281. [250]

Balding, D. J. (2006), "A Tutorial on Statistical Methods for Population Association Studies," *Nature Reviews Genetics*, 7, 781–791. [251]

Chatterjee, N., Chen, Y.-H., Luo, S., and Carroll, R. J. (2009), "Analysis of Case–control Association Studies: Snps, Imputation and Haplotypes," *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 24, 489. [250,251]

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Vol. 1), Cambridge, UK: Cambridge University Press. [255]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014), *Bayesian Data Analysis* (Vol. 2), New York: Chapman and Hall/CRC. [250]

González, J. R., Carrasco, J. L., Dudbridge, F., Armengol, L., Estivill, X., and Moreno, V. (2008), "Maximizing Association Statistics Over Genetic Models," *Genetic Epidemiology*, 32, 246–254. [251]

Hothorn, L. A., and Hothorn, T. (2009), "Order-Restricted Scores Test for the Evaluation of Population-Based Case–Control Studies When the Genetic Model is Unknown," *Biometrical Journal*, 51, 659–669. [251]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [250,253]

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934. [254,255]

Laird, N. M., and Lange, C. (2010), *The Fundamentals of Modern Statistical Genetics*, New York: Springer Science & Business Media. [250]

Liao, J. (1999), "A Hierarchical Bayesian Model for Combining Multiple 2×2 Tables Using Conditional Likelihoods," *Biometrics*, 55, 268–272. [251]

Li, B. F. G. Z. Z., and Gastwirth, J. L. (2002), "Trend Tests for Case–control Studies of Genetic Markers: Power, Sample Size and Robustness," *Human Heredity*, 53, 146–152. [251]

Little, R. J. (1989), "Testing the Equality of Two Independent Binomial Proportions," *The American Statistician*, 43, 283–288. [251]

——— (2006), "Calibrated Bayes: A Bayes/Frequentist Roadmap," *The American Statistician*, 60, 213–223. [250]

——— (2011), "Calibrated Bayes, for Statistics in General, and Missing Data in Particular," *Statistical Science*, 26, 162–174. [250]

——— (2012), "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics," *Journal of Official Statistics*, 28, 309. [250]

Lu, J., Wei, Q., Bondy, M. L., Yu, T.-K., Li, D., Brewster, A., Shete, S., Sahin, A., Meric-Bernstam, F., and Wang, L.-E. (2006), "Promoter Polymorphism (−786 $t>c$) in the Endothelial Nitric Oxide Synthase Gene is Associated With Risk of Sporadic Breast Cancer in Non-hispanic White Women Age Younger Than 55 Years," *Cancer*, 107, 2245–2253. [254]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (Vol. 2), London: Chapman and Hall. [251,253,255]

Mukherjee, B., Sinha, S., and Ghosh, M. (2005), "Bayesian Analysis of Case–Control Studies," *Handbook of Statistics*, 25, 793–819. [251]

Reid, N. (1995), "The Roles of Conditioning in Inference," *Statistical Science*, 10, 138–157. [251]

Rice, K. M. (2004), "Equivalence Between Conditional and Mixture Approaches to the Rasch Model and Matched Case–Control Studies, With Applications," *Journal of the American Statistical Association*, 99, 510–522. [251]

Sasieni, P. D. (1997), "From Genotypes to Genes: Doubling the Sample Size," *Biometrics*, 53, 1253–1261. [250]

Satten, G. A., and Kupper, L. L. (1993), "Conditional Regression Analysis of the Exposure-Disease Odds Ratio Using Known Probability-of-Exposure Values," *Biometrics*, 49, 429–440. [251]

Sinharay, S., and Stern, H. S. (2002), "On the Sensitivity of Bayes Factors to the Prior Distributions," *The American Statistician*, 56, 196–201. [250,252]

Stephens, M., and Balding, D. J. (2009), "Bayesian Statistical Methods for Genetic Association Studies," *Nature Reviews Genetics*, 10, 681–690. [251]

Talluri, R., Wang, J., and Shete, S. (2014), "Calculation of Exact *P*-values When SNPS are Tested Using Multiple Genetic Models," *BMC Genetics*, 15, 75. [251,254]

Wakefield, J. (2009), "Bayes Factors for Genome-Wide Association Studies: Comparison With *p*-values," *Genetic Epidemiology*, 33, 79–86. [251]

Xu, J., Yuan, A., and Zheng, G. (2012), "Bayes Factor Based on the Trend Test Incorporating Hardy–Weinberg Disequilibrium: More Power to Detect Genetic Association," *Annals of Human Genetics*, 76, 301–311. [251]