9

Bayesian Methods in Fisher's Statistical Genetics World

Radu V. Craiu and Lei Sun

University of Toronto, Toronto, ON

9.1 Background and Introduction

Statistical genetics is a scientific discipline that covers any statistical analysis of genetic data. The interplay between statistics and genetics has a long history, dating back to the seminal work by Fisher almost a century ago to confirm the genetic theory of chromosomal inheritance (Piegorsch, 1990). Recent advancements in genotyping (i.e., collecting genetic data) technologies have produced vast amounts of data, offering statisticians great opportunities in methodological development, implementation and application. For example, the high-dimensional genome-wide association studies conducted in the last few years have led to the development of a catalog of novel statistical methods for dissecting the genetic architecture of complex human traits; see, e.g., Thomas et al. (2009) and Begum et al. (2012).

The types of genetic data available are quite diverse; they include microsatellites, single-nucleotide polymorphisms, copy number variations, DNA methylation, and gene expression. The corresponding statistical methodologies are equally diverse, as illustrated by Bull et al. in Chapter 8. For clarity and a more focused discussion, this expository piece is centered around studies of genetic association between single-nucleotide polymorphisms and heritable human traits. In the following, we first provide relevant genetic terminology. We then formulate genetic association studies in terms of regression models in which inferences on the regression coefficients are of interest. Using a published genome-wide association study as an example, we first describe the commonly used frequentist approaches to achieve testing and estimation objectives, and we then discuss alternative Bayesian methods and associated advantages as well as challenges. We conclude with discussions of other recent developments in Bayesian statistical genetics, focusing on contributions made by Canadian statisticians, and comment on future directions.

9.1.1 Basic Genetic Terminology

The building blocks of the human genome are base pairs, which are part of the DNA in each person's chromosomes. There are about three billion base pairs and over 99% are identical between individuals. A base pair that varies in a population and has two variants (alleles) is called a single-nucleotide polymorphism (SNP).

Let A and a denote the two alleles, and without loss of generality, let a be the minor allele for which the population allele frequency, $p(a)$, is such that $p(a) \leq .5$. This frequency is called the minor allele frequency (MAF) and some terminology is based on this: a common SNP has MAF $\geq .05$, a low frequency SNP has $.01 <$ MAF $< .05$, and a rare SNP or rare variant has MAF $\leq .01$.

Although the exact number is a moving target, it is believed that there are over 10 million common SNPs and even more rare variants (1000GenomesProjectConsortium, 2012). Human chromosomes are paired, with one inherited from the mother and the other from the father. Therefore, at each SNP location along the genome, there are two alleles forming a person's genotype, which can be AA, Aa, or aa.

9.1.2 Statistical Set-Up of Genetic Association Studies: Two Intertwined Issues

Let Y be the trait (i.e., phenotype) of interest, e.g., the presence of Type 1 diabetes (T1D) or blood glucose level, and let X represent the genotype of a SNP under study. Genetic association studies assess whether the response variable Y varies between different levels of X.

For example, if individuals with genotype aa tend to have a higher risk of developing T1D than individuals with genotype AA, then there is an association between X and Y. To statistically assess the relationship between Y and X, a so-called simple linear regression model can be considered if the trait is (approximately) normally distributed (e.g., blood glucose level), viz.

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{9.1}$$

where Y_i is the trait value for individual i, X_i is the SNP genotype for this individual, and $\epsilon_1, \ldots, \epsilon_n$ represent independent random variations, which are assumed to follow a Normal distribution with mean 0 and variance σ^2.

In many applications, the three genotypes, AA, Aa and aa, are coded numerically (additively) as $X = 0$, 1 and 2 to represent the number of copies of the minor allele a. In that case, model (9.1) assumes that increasing X by one unit will increase the Y value, on average, by β. This is also called the linear additive model because the effect of two copies of the minor allele (genotype aa, $X = 2$) is assumed to be twice the effect of 1 copy (genotype Aa, $X = 1$). In some genetic settings, other genotype models may be used, which we discuss in Section 9.1.3.

If the trait of interest is binary (e.g., the presence of Type 1 diabetes, T1D), the classical logistic regression can be used (Agresti, 2002),

$$\text{logit}\{\text{E}(Y_i)\} = \alpha + \beta X_i, \tag{9.2}$$

where Y_i indicates whether individual i has the disease ($Y_i = 1$) or not ($Y_i = 0$), $\text{E}(Y_i) = \Pr(Y = 1)$ and

$$\text{logit}\{\Pr(Y = 1)\} = \log\left\{\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)}\right\}$$

is called the log odds. Model (9.2) implies that when increasing X by 1 unit we expect an increase in the log odds value of Y by β, and the interpretation of β is the well known log odds ratio (log OR). The log odds, rather than just $\text{E}(Y_i)$, is used in (9.2) because it tends to give a better description of the relationship between Y_i and X_i.

In either regression setting, the primary objective of an association study is to identify which SNPs are related to the trait; this is equivalent to comparing the two hypotheses for each SNP, viz.

$$\mathcal{H}_0 : \beta = 0, \quad \mathcal{H}_1 : \beta \neq 0.$$

It is in this context that we will describe in Section 9.2 the commonly used frequentist and Bayesian approaches to assess whether the evidence provided by the data supports $\mathcal{H}_0$ or $\mathcal{H}_1$.

To a statistician, this may seem like an exceedingly simple problem. However, significant complications arise in applications. For example:

a) additional variables (also known as covariates) such as sex and age may need to be included in the regression models (9.1) or (9.2);

b) measurement errors occur in both X and Y;

c) individuals' phenotypes and genotypes can be correlated;

d) individuals may come from different populations;

e) multiple (common or rare) SNPs need to be jointly analyzed to increase power; and

f) multiple (binary, continuous or both) traits can be of interest.

Proper statistical treatment for any of these issues requires experience in statistical genetics and genetic epidemiology. Since we do not assume here that the reader has such prior knowledge, we will focus on explaining some of the basic statistical techniques used for genetic association studies. We briefly discuss some more complex issues in Section 9.4.

Once a trait-associated SNP has been identified, it is of interest to report the corresponding genetic effect size such as log OR, β in (9.2), and to plan follow-up studies that seek to replicate the initial finding. An interesting statistical question arises here. Let us assume that β is the true log OR of an associated SNP, and without loss of generality, assume that the minor allele of the SNP increases the risk of the disease under study (i.e., OR > 1 and $\beta > 0$). Let $\hat{\beta}$ be an estimated value reported from a study based on (9.2). It is known that $\hat{\beta}$ is unbiased, which means that if we were to repeat the experiment over and over again by collecting a different sample from the same population and applying the same statistical analysis, then the average of all the values of $\hat{\beta}$ would be β. However, not all studies are successful in identifying the association and only studies with sufficiently large $\hat{\beta}$ values are significant (i.e., result in rejecting $\mathcal{H}_0$). Therefore, $\hat{\beta}$ reported from a significant study is on average bigger than the true value. How to correct for this upward bias requires non-trivial statistical remedies, which we discuss in Section 9.3.

9.1.3 GWAS and an Example

Many current "high-throughput" genetic studies, such as the genome-wide association studies (GWAS) or next generation sequencing (NGS) studies, comprehensively investigate the whole genome to identify trait-associated SNPs. In that case, the scientific objective is to look for association between a trait and genotypes of hundreds of thousands or millions of SNPs.

Let X_{ij} be the genotype for individual i at SNP j with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. One of the key features of GWAS and NGS is $n \ll p$, where n is in the range of 1,000 to 10,000 while p is in the range of 1 million for GWAS (or 10 millions for NGS). Although joint analyses of multiple SNPs have recently been tackled using more advanced regression methods such as the Lasso (see Chapter 5 by Tibshirani), single-SNP association analysis is still predominant due to its simplicity and interpretability. In that case, regression models (9.1) or (9.2) are fitted repeatedly for each SNP $j \in \{1, \ldots, p\}$, and the corresponding (frequentist) decision rule concerning $\mathcal{H}_0$ is based on a p-value, which is the probability, computed assuming that $\mathcal{H}_0$ is true, that the test statistic is as extreme as the one observed.

Intuitively, the smaller the p-value, the smaller the evidence in favor of $\mathcal{H}_0$ provided by the data. Traditionally, if the p-value is less than .05 we reject the null hypothesis $\mathcal{H}_0$. However, even if the SNP is not associated with the trait, there is a 5% probability of wrongly rejecting $\mathcal{H}_0$; this is known as the Type 1 error. In the GWAS setting, we perform millions of tests in one analysis and errors accumulate. In order to control the overall (i.e., genome-wide or family-wise) Type 1 error rate at 5%, a stringent criterion such as p-value $< 5 \times 10^{-8}$ is typically used for each SNP included in the study (Dudbridge and Gusnanto, 2008).

WTCCC (2007) by the Wellcome Trust Case Control Consortium is a landmark work in which a GWAS was conducted for each of the seven common diseases, including coronary heart disease, Type 1 diabetes (T1D), Type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension, with a total sample size of 16,179. Here we focus on the T1D study for which the sample consisted of 1,926 individuals with T1D (cases) and 2,872 subjects without T1D (controls). Genotype data were collected for these individuals at $469,557$ SNPs (after data quality control). To assess the association evidence between T1D and each of the SNPs, WTCCC (2007) calculated p-values for testing $\mathcal{H}_0 : \beta_j = 0$ for $1 \leq j \leq 469{,}557$, and reported the findings in their Table 2 (five SNPs with p-values $< 5 \times 10^{-7}$) and Table 3 (seven additional SNPs with p-values $< 10^{-5}$). They also provided Bayes factors for each of the 12 SNPs (Section 9.2.2).

Section 9.2 describes the details of their frequentist and Bayesian association analyses. Section 9.3 discusses the upward bias inherent in their reported β_j estimates and reviews different approaches for reducing the estimation bias using only the summary statistics.

9.2 Identification of Trait-Associated SNPs

9.2.1 *p*-value

To identify SNPs associated with T1D using as an example the WTCCC (2007) study, first consider the logistic regression model

$$\text{logit}\{\text{E}(Y_i)\} = \alpha + \beta_j X_{ij},$$

where Y_i indicates whether individual i in the sample has T1D ($Y_i = 1$) or not ($Y_i = 0$), and $X_{ij} = 0, 1$ or 2 if the genotype of SNP j for individual i is, respectively, AA, Aa or aa. This is the additive model as discussed in Section 9.1.2. The p-value for testing $\mathcal{H}_0 : \beta_j = 0$ is denoted as $p_{j,\text{add}}$. Covariates such as age at onset of T1D were not included in the WTCCC analysis.

To discover SNPs with non-additive genetic effects, it is also important to consider the more flexible genotypic model, where the genotype of a SNP is treated as a categorical variable with three levels. For the genotypic model,

$$\text{logit}\{\text{E}(Y_i)\} = \alpha + \beta_{1j} X1_{ij} + \beta_{2j} X2_{ij}$$

can be used, where $X1$ and $X2$ are the standard dummy variables: $(X1_{ij} = 0, X2_{ij} = 0)$ for genotype AA, $(X1_{ij} = 1, X2_{ij} = 0)$ for Aa and $(X1_{ij} = 0, X2_{ij} = 1)$ for aa. The p-value for testing $\mathcal{H}_0 : \beta_{1j} = \beta_{2j} = 0$ is denoted as $p_{j,\text{geno}}$. WTCCC (2007) then defined SNPs with strong evidence of association with T1D as SNPs with $\min(p_{j,\text{add}}, p_{j,\text{geno}}) < 5 \times 10^{-7}$, and moderate evidence of association as $\min(p_{j,\text{add}}, p_{j,\text{geno}}) < 10^{-5}$. We summarize their key results in Table 9.1. Although the two models generally give similar results, SNP rs17166496 clearly did not follow the additive assumption and would have been missed without the consideration of the alternative genotypic model.

TABLE 9.1: Results of the WTCCC genome-wide association study (GWAS) of Type 1 diabetes (T1D). All values except the ranks are restated here based on Tables 2 and 3 of WTCCC (2007). Note that the ranks in the table are within these 12 SNPs not at the genome-wide level involving all 469,557 SNPs.

			Additive Model			
Chromosome	SNP	MAF	Frequentist		Bayesian	
			p-value	Rank	$\log_{10}$BF	Rank
6	rs9272346	.387	2.42 E-134	1	141.93	1
1	rs6679677	.096	1.17 E-26	2	23.07	2
12	rs17696736	.424	2.17 E-15	3	12.53	3
12	rs11171739	.423	1.14 E-11	4	8.89	4
16	rs12708716	.350	9.24 E-08	5	5.15	5
4	rs17388568	.260	5.00 E-07	6	4.42	6
18	rs2542151	.163	1.89 E-06	7	3.91	7
10	rs2104286	.286	7.97 E-06	8	3.31	9
5	rs2544677	.242	8.23 E-06	9	3.32	8
1	rs2639703	.276	8.46 E-06	10	3.25	10
12	rs11052552	.486	1.02 E-04	11	2.22	11
5	rs17166496	.391	6.06 E-01	12	−.97	12

			Genotypic Model			
Chromosome	SNP	MAF	Frequentist		Bayesian	
			p-value	Rank	$\log_{10}$BF	Rank
6	rs9272346	.387	5.47 E-134	1	139.77	1
1	rs6679677	.096	5.43 E-26	2	22.83	2
12	rs17696736	.424	1.51 E-14	3	11.56	3
12	rs11171739	.423	9.71 E-11	4	8.24	4
16	rs12708716	.350	4.92 E-07	5	4.71	5
4	rs17388568	.260	3.27 E-06	7	3.89	6
18	rs2542151	.163	1.16 E-05	9	3.52	8
10	rs2104286	.286	4.32 E-05	11	2.88	11
5	rs2544677	.242	4.43 E-05	12	2.70	12
1	rs2639703	.276	1.74 E-05	10	3.06	10
12	rs11052552	.486	7.24 E-07	6	3.80	7
5	rs17166496	.391	5.20 E-06	8	3.25	9

9.2.2 Bayes Factor

The previous section showed that alternative models may lead to different conclusions regarding the association of a SNP with a given trait. Combining analyses produced by different models is desirable when there is no clear evidence supporting a specific model. However, this is difficult under the frequentist framework and leads us to consider Bayesian alternatives. WTCCC (2007) were among the first who considered a Bayesian framework for GWAS, and in particular they relied on Bayes factors (Kass and Raftery, 1995) to rank SNPs. Other early Bayesian work in the GWAS context includes Marchini et al. (2007) and Wakefield (2007), and Stephens and Balding (2009) gave an excellent review on this topic.

The Bayesian approach treats the parameter of interest, θ, as a random variable for which a prior distribution, $p(\theta|M)$, is first defined under assumed model M. The prior distribution can be interpreted as a summary of our information about the parameter θ before the data is collected. Inference is developed based on the corresponding posterior density which is conditional on the observed data D and is calculated using

$$p(\theta|D,M) = \frac{f(D|\theta,M)p(\theta|M)}{p(D|M)} = \frac{f(D|\theta,M)p(\theta|M)}{\int_\Theta f(D|\theta,M)p(\theta|M)\mathrm{d}\theta}. \tag{9.3}$$

The denominator of (9.3) is the probability of data under model M. We can interpret $p(D|M)$ as the weighted average probability of observing the data under the assumed model M, over all possible values of the parameter θ.

Let us revisit the logistic additive model (9.2). For a given SNP, let M_0 ($\mathcal{H}_0 : \beta = 0$) denote the null model of no association and M_1 ($\mathcal{H}_1 : \beta \neq 0$) the alternative model. To decide which of the two competing models, M_0 and M_1, is more suitable, a natural choice is to rely on the Bayes factor of M_1 against M_0, viz.

$$\mathrm{BF}_{10} = \frac{p(D|M_1)}{p(D|M_0)}\,.$$

In the WTCCC (2007) application, $\theta_0 = (\alpha, 0)$, $\theta_1 = (\alpha, \beta)$, $\alpha \sim \mathcal{N}(0,1)$ and $\beta \sim \mathcal{N}(0,.2)$. The prior for β reflects the belief that genetic effects (expressed as the odds ratio, OR = $\exp(\beta)$) of SNPs associated with complex human traits are in the range of .5 to 2 and most likely between .67 and 1.5; see the Supplementary Figure 1 of WTCCC (2007). The choice of prior is study specific and subjective, which contributes to scientists' reluctance to using Bayesian methods; see Chapter 10 by Gustafson on Bayesian methods in observational epidemiology studies. However, techniques such as model-averaging (see below) can alleviate some of the concerns that different prior specifications might lead to different conclusions.

Let BF_{10}(add) denote the Bayes factor for SNP j under the additive model and BF_{10}(geno) under the genotypic model. In the latter case, $\theta_0 = (\alpha, 0, 0)$ and $\theta_1 = (\alpha, \beta_1, \beta_2)$, and prior distributions are specified for both β_1 and β_2; see WTCCC (2007) for more details on the choice of prior. For each of the 12 SNPs selected based on the p-value criterion as described in Section 9.2.1, the corresponding Bayes factors, BF_{10}(add) and BF_{10}(geno) were also calculated and are restated (on the $\log_{10}$ scale) in Table 9.1.

9.2.3 Additional Considerations

The immediate conclusion from Table 9.1 is that rankings of the SNPs are remarkably consistent between p-value and Bayes factor. This consistency has been steadily reported in the GWAS literature; see, e.g., Strömberg et al. (2009). It has been theoretically justified by Wakefield (2009) for case-control GWAS. However, such conclusions depend on several assumptions, an important one being that the MAFs are not too small. Although SNPs analyzed by GWAS are usually common SNPs, this is not the case for the emerging NGS

TABLE 9.2: Results of WTCCC GWAS of T1D after "Model Averaging." Here, min(p-value) = min(p-value$_{\text{add}}$, p-value$_{\text{geno}}$), $\log_{10}\text{BF}(.8, .2) = \log_{10}\{.8\,\text{BF(add)} + .2\,\text{BF(geno)}\}$, p-value$_{\text{add}}$, p-value$_{\text{geno}}$, BF(add) and BF(geno) are from Table 9.1.

Chro-			Frequentist		Bayesian	
mosome	SNP	MAF	min(p-value)	Rank	$\log_{10}$ BF(.8, .2)	Rank
6	rs9272346	.387	2.42 E-134	1	141.83	1
1	rs6679677	.096	1.17 E-26	2	23.03	2
12	rs17696736	.424	2.17 E-15	3	12.44	3
12	rs11171739	.423	1.14 E-11	4	8.82	4
16	rs12708716	.350	9.24 E-08	5	5.09	5
4	rs17388568	.260	5.00 E-07	6	4.35	6
18	rs2542151	.163	1.89 E-06	8	3.86	7
10	rs2104286	.286	7.97 E-06	10	3.25	8
5	rs2544677	.242	8.23 E-06	11	3.25	9
1	rs2639703	.276	8.46 E-06	12	3.22	10
12	rs11052552	.486	7.24 E-07	7	3.14	11
5	rs17166496	.391	5.20 E-06	9	2.55	12

area. For analysis of rare variants, it is unclear if the traditional frequentist approaches (e.g., Lee et al., 2012; Derkach et al., 2014) and Bayesian methods (e.g., Yi and Zhi, 2011) lead to similar rankings.

Let us reconsider SNP rs17166496 in Table 9.1 for which different genetic model assumptions (additive vs genotypic) resulted in strikingly different results. Let $p(M)$ denote the prior distribution for genetic model M, and for simplicity consider only two models, $M \in \{\text{add}, \text{geno}\}$, with user-specified prior probabilities

$$\begin{aligned} \Pr(M = \text{add}) &= p(\text{add}), \\ \Pr(M = \text{geno}) &= p(\text{geno}) = 1 - p(\text{add}). \end{aligned}$$

Under the Bayesian framework, we can define an overall Bayes factor as a weighted average of the individual Bayes factors, viz.

$$\text{BF}_{10} = p(\text{add}) \times \text{BF}_{10}(\text{add}) + p(\text{geno}) \times \text{BF}_{10}(\text{geno}).$$

For a more detailed discussion we refer the readers to Stephens and Balding (2009). Therefore, instead of selecting one "best" model and reporting the corresponding finding, the Bayesian analysis provides a principled way to combine the evidence from each model, weighted by the prior belief in that model. For the SNPs in Table 9.1, Table 9.2 provides the Bayes factors after model averaging (see Section 9.3.1), with more weight ($p(\text{add}) = .8$) given to the additive genetic model reflecting the common belief that most SNPs act in an (approximately) additive manner (Hill et al., 2008).

9.3 Replication of a Significant Finding

In most scientific studies, a negative result is usually not published and a positive finding needs to be replicated in an independent sample. For example, other investigators might want to validate the association between SNP rs2542151 and T1D found by WTCCC (2007), and they would typically use the reported genetic effect size (i.e., $\exp(\hat{\beta})$) to calculate the sample size needed for a replication study (e.g., with 80% power at the .05 level). However, as discussed in Section 9.1.2, the use of the same data first for testing and then for estimation leads to an upward-biased estimate of the effect size and, therefore, an under-powered replication study. This phenomenon is also known as the winner's curse and is ubiquitous in GWAS. In the following, we discuss the frequentist and Bayesian approaches to correcting such bias using only the summary statistics that are typically reported.

9.3.1 Conditional MLE vs. Bayesian Model Averaging

Consider $\hat{\beta}$, the estimator of β, $\widehat{SE}(\hat{\beta})$, the standard error of $\hat{\beta}$, and

$$T = \hat{\beta}/\widehat{SE}(\hat{\beta}) \,,$$

the test statistic for testing $\mathcal{H}_0$ as defined in Section 9.1.2. To simplify the discussion, assume that $\beta > 0$ for a truly associated SNP. Since we focus attention on estimates for which the associated test was significant (i.e., $T > c$, where c is some specified value), $\hat{\beta}$ is not an (approximately) unbiased estimator of β; in fact, $\mathrm{E}(\hat{\beta}|T > c) \geq \beta$.

Ghosh et al. (2008) proposed to obtain the maximum likelihood estimate (MLE) based on the conditional likelihood that incorporates the fact that the observed test statistic exceeded the significance criterion. Specifically, under the Normal assumption that T follows the $\mathcal{N}[\beta/SE(\hat{\beta}), 1]$ distribution, the conditional MLE of β is the parameter value that maximizes the likelihood

$$L(\beta) = p(T|T > c) = \frac{\phi\{T - \beta/\widehat{SE}(\hat{\beta})\}}{\Phi\{-c + \beta/\widehat{SE}(\hat{\beta})\}} \,, \tag{9.4}$$

where ϕ and Φ are respectively the density and cumulative distribution function of $\mathcal{N}(0, 1)$ (Ghosh et al., 2008). The $L(\beta)$ in (9.4) is a conditional likelihood because it quantifies the conditional probability of the observed data given that they yielded a significant finding for β. Conditioning on the significant result creates a more realistic statistical framework. Indeed, compared with the unconditional $\hat{\beta}$, the estimate obtained from this conditional MLE approach is, on average, closer to the true value β. However, it is difficult to incorporate prior information on β into (9.4) in settings where it might be available.

As an alternative, Xu et al. (2011) considered the Bayesian approach that specifies a prior distribution for the parameter of interest, $\theta = \beta$ (i.e., the log OR), and infers the posterior as in (9.3). In the same vein as the conditional likelihood approach, the Bayesian inference is performed conditional on the significant finding for β. Therefore, the prior specification can incorporate this information. In this setting, Xu et al. (2011) specified the prior distribution for β as a mixture of a discrete distribution that places all probability mass at 0 and a continuous distribution g with support on $\mathbb{R}$,

$$p(\beta) = \xi\delta_{\{0\}}(\beta) + (1-\xi)g(\beta). \tag{9.5}$$

Priors such as (9.5) are known as spike-and-slab priors and have been used repeatedly in Bayesian variable selection and shrinkage estimation; see, e.g., Kuo and Mallick (1998). In this application, the mixture parameter ξ allows the possibility that the observed significance is due to chance (i.e., $\beta = 0$); the $g(\beta)$ density function reflects the prior belief about the range and distribution of the effect of a truly associated SNP. The Bayesian estimator of β is the posterior mean $\mathrm{E}(\beta|D)$. The spike at 0 will shrink the posterior mean and reduce the posterior variance of β. The prior distribution $p(\xi)$ assigned to ξ allows us to quantify the uncertainty about a discovery being true or false and influence the amount of shrinkage in the posterior. For instance, if we use a Beta prior for ξ, denoted $\mathcal{B}(a, b)$, and if the sample size of the discovery study was small, we may choose large a and small b values (e.g., $a = 8$ and $b = .5$) to represent our initial skeptical view on the significance of any finding from such a study.

In practice, a researcher may have a difficult time deciding which particular prior to use for the analysis. In many cases, the analysis can be better served by considering several priors rather than choosing just one of them. Suppose M_1 is the model with $p(\xi) = \mathcal{B}(8, .5)$ and M_2 with $p(\xi) = \mathcal{B}(.5, 8)$, and let $p(M_1)$ and $p(M_2)$ be the prior probabilities for the two models; see Xu et al. (2011) for the specification of $p(M_1)$ and $p(M_2)$. Then the Bayesian model averaging (BMA) paradigm offers a consistent method to combine the two models by using

$$p(\theta|D) = \sum_{k=1}^{2} p(\theta|D, M_k)p(M_k|D), \tag{9.6}$$

where $p(M_k|D)$ is the posterior probability of model M_k and is proportional to $p(M_k|D) \propto p(D|M_k)p(M_k)$. Formally, the Bayesian BMA estimator is the mean of the distribution $p(\theta|D)$, but in practice, the theoretical mean is not available in closed form and it is approximated via the Monte Carlo method as discussed in Section 9.3.2.

Using SNP rs2542151 as an example, the genetic effect reported by WTCCC (2007) was

$$\hat{\beta}_{\text{naive}} = .285, \quad \widehat{OR}_{\text{naive}} = 1.33.$$

However, estimates were substantially smaller after bias correction by both the frequentist approach,

$$\hat{\beta}_{\text{freq}} = .140, \quad \widehat{OR}_{\text{freq}} = 1.15,$$

and the Bayesian method,

$$\widehat{\beta}_{\text{Bayes}} = .117, \quad \widehat{OR}_{\text{Bayes}} = 1.12.$$

This reduction in genetic effect estimate (OR from 1.33 to just over 1.1) is practically meaningful and important because a) the sample size needed for a successful replication study would not be underestimated; and b) the potential clinical importance of this SNP would not be overestimated. Between the two correction methods, Xu et al. (2011) have shown that the Bayesian estimator has better accuracy (as measured by the mean squared error) than the conditional MLE in studies with low power, where the winner's curse is more likely to impact the conclusions of the analyses.

9.3.2 Computational Considerations

Realistic Bayesian modeling and analysis have been made possible thanks to significant advances in computational algorithms, particularly Markov Chain Monte Carlo (MCMC) samplers. Note that the integral that appears in the denominator of (9.3) is often intractable. For example, the model considered here along with the prior (9.5) and a uniform prior on β (i.e., θ) yields a posterior distribution that cannot be studied analytically, e.g., one cannot compute in closed form the posterior mean, viz.

$$\mathrm{E}(\theta|D,M) = \int \theta\, p(\theta|D,M)\mathrm{d}\theta.$$

However, we can still approximate $\mathrm{E}(\theta|D,M)$ as long as we can draw from $p(\theta|D,M)$. For instance, if we can generate a sample $\theta_1,\ldots,\theta_K$ from the posterior $p(\theta|D,M)$ we can then estimate $\mathrm{E}(\theta|D,M)$ by

$$\mathrm{E}(\widehat{\theta|D,M}) = \frac{1}{K}\sum_{i=1}^{K}\theta_i.$$

Due to space constraints, we cannot get into the details of constructing the algorithms used to sample from posterior distributions, but we refer the reader to the review article of Craiu and Rosenthal (2014) and Rosenthal's review of Metropolis algorithms in Chapter 6.

The tight connection between Bayesian inference and computational algorithms has also been exploited in situations in which the likelihood cannot be calculated in closed form (e.g., Tavaré et al., 1997) or when it may be too expensive to compute; see, e.g., Wegmann et al. (2009) and Row et al. (2011).

9.4 Conclusion and Discussion

The collection of vast amounts of genetic data in recent years has provided statisticians with numerous challenges and opportunities. In the effort to understand the genetic architecture of complex human traits, one important area of research has focused on genetic association studies which are the central theme of the discussion here. In addition to the frequentist framework we have emphasized the value of the alternative Bayesian paradigm. For many, a major hurdle in a Bayesian analysis comes from the computational and methodological challenges involved in the study of the posterior distribution via MCMC sampling. However, the increase in computational expertise among statistical geneticists and the advent of user-friendly software has spurred renewed interest in Bayesian methods.

The frequentist and Bayesian bias-correction methods discussed in Section 9.3 assume that only the summary statistics are available and only for the reported significant SNPs. When genome-wide results are available, it might be beneficial to let the empirical distribution of the estimated effects influence the specification of the prior. Built upon the work by Efron (2011), this empirical Bayes approach has been used by Ferguson et al. (2013) to tackle the winner's curse in a setting where multiple genetic effects are dealt with jointly. When the original data are available, one could use the alternative bootstrap-based method reviewed by Bull et al. in Chapter 8.

Bayesian methods have also been successfully used in other genetic settings. For example, Lo and Gottardo (2007) coupled a hierarchical Bayesian model with an empirical Bayes specification of the prior to produce inference about differential expression in microarray studies; Gottardo et al. (2008) proposed a Bayesian analysis of Chromatin-immunoprecipitation microarrays in which the hierarchical structure of the model accounts for the spatial correlation present between neighboring probes; Wu et al. (2009) proposed a Bayesian segmentation approach that identifies copy number variations (DNA segments that exhibit duplications and deletions when compared to a reference genome); and Scott-Boyer et al. (2012) introduced a Bayesian hierarchical model that can combine genotypic and gene expression data to detect the so-called "expression quantitative trait loci (eQTL)."

Recent progress in genetic association studies has taught us some lessons. Although simple statistical techniques can identify many trait-associated genetic variants, these "low hanging fruits" are only a few small pieces of a much bigger puzzle. To solve the remaining puzzle, more sophisticated methodology is needed to mine the already available data, to analyze new kinds of data, and to combine different sources of data. We believe that the Bayesian methodology can play an important role in solving some of these issues. Dialog between statisticians and other scientists is becoming a critical component of the analytical process for these emerging studies.

Acknowledgments

The authors thank the editor, Jerald F. Lawless, and two external reviewers for helpful comments. The authors' research is supported by the Natural Sciences and Engineering Research Council of Canada and by the Canadian Institutes of Health Research.

About the Authors

Radu V. Craiu is a professor of statistics at the University of Toronto. He studied mathematics at the Universitatea din Bucureşti (MS, 1996) and statistics at the University of Chicago (PhD, 2001). His research interests include computational methods for Bayesian inference, especially Markov Chain Monte Carlo sampling algorithms, copula models, model selection and statistical genetics. He is an associate editor for *The Canadian Journal of Statistics*.

Lei Sun is an associate professor of biostatistics and statistics at the University of Toronto. She studied mathematics at Fudan University and obtained her PhD in statistics from the University of Chicago in 2001. Her primary research interest is statistical genetics, developing statistical methods and computational tools for high-dimensional studies of complex human traits.

Bibliography

1000GenomesProjectConsortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65.

Agresti, A. (2002). *Categorical Data Analysis.* Wiley, New York.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*, 40:3777–3784.

Craiu, R. V. and Rosenthal, J. S. (2014). Bayesian computation via Markov Chain Monte Carlo. *Annual Review of Statistics and its Applications*, 1:7.1–7.23.

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: A review and some new results. *Statistical Science*, 29:in press.

Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32:227–234.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106:1602–1614.

Ferguson, J. P., Cho, J. H., Yang, C., and Zhao, H. (2013). Empirical Bayes correction for the Winner's Curse in genetic association studies. *Genetic Epidemiology*, 37:60–68.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *American Journal of Human Genetics*, 82:1064–1074.

Gottardo, R., Li, W., Johnson, W. E., and Liu, X. S. (2008). A flexible and powerful Bayesian hierarchical model for ChIP-Chip experiments. *Biometrics*, 64:468–478.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, 4(2):e1000008.

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

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā, Series B*, 60:65–81.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13:762–775.

Lo, K. and Gottardo, R. (2007). Flexible empirical Bayes models for differential gene expression. *Bioinformatics*, 23:328–335.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913.

Piegorsch, W. W. (1990). Fisher's contributions to genetics and heredity, with special emphasis on the Gregor Mendel controversy. *Biometrics*, 46:915–924.

Row, J. R., Brooks, R. J., Mackinnon, C. A., Lawson, A., Crother, B. I., White, M., and Lougheed, S. C. (2011). Approximate Bayesian computation reveals the factors that influence genetic diversity and population structure of foxsnakes. *Journal of Evolutionary Biology*, 24:2364–2377.

Scott-Boyer, M. P., Imholte, G. C., Tayeb, A., Labbe, A., Deschepper, C. F., and Gottardo, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11:1–30.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Review Genetics*, 10:681–690.

Strömberg, U., Björk, J., Vineis, P., Broberg, K., and Zeggini, E. (2009). Ranking of genome-wide association scan signals by different measures. *International Journal of Epidemiology*, 38:1364–1373.

Tavaré, S., Balding, D. J., Griffith, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518.

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P., and Stram, D. O. (2009). Methodological issues in multistage genome-wide association studies. *Statistical Science*, 24:414–429.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, 81:208–227.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, 33:79–86.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov Chain Monte Carlo without likelihood. *Genetics*, 182:1207–1218.

WTCCC, W. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678.

Wu, L. Y., Chipman, H. A., Bull, S. B., and Briollais, L. (2009). A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics*, 25:1669–1679.

Xu, L., Craiu, R. V., and Sun, L. (2011). Bayesian methods to overcome the Winner's Curse in genetic studies. *The Annals of Applied Statistics*, 5:201–231.

Yi, N. and Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. *Genetic Epidemiology*, 35:57–69.