

Quantifying the Fraction of Missing Information for Hypothesis Testing in Statistical and Genetic Studies

Dan L. Nicolae, Xiao-Li Meng and Augustine Kong

Abstract. Many practical studies rely on hypothesis testing procedures applied to data sets with missing information. An important part of the analysis is to determine the impact of the missing data on the performance of the test, and this can be done by properly quantifying the relative (to complete data) amount of available information. The problem is directly motivated by applications to studies, such as linkage analyses and haplotype-based association projects, designed to identify genetic contributions to complex diseases. In the genetic studies the relative information measures are needed for the experimental design, technology comparison, interpretation of the data, and for understanding the behavior of some of the inference tools. The central difficulties in constructing such information measures arise from the multiple, and sometimes conflicting, aims in practice. For large samples, we show that a satisfactory, likelihood-based general solution exists by using appropriate forms of the relative Kullback–Leibler information, and that the proposed measures are computationally inexpensive given the maximized likelihoods with the observed data. Two measures are introduced, under the null and alternative hypothesis respectively. We exemplify the measures on data coming from mapping studies on the inflammatory bowel disease and diabetes. For small-sample problems, which appear rather frequently in practice and sometimes in disguised forms (e.g., measuring individual contributions to a large study), the robust Bayesian approach holds great promise, though the choice of a general-purpose “default prior” is a very challenging problem. We also report several intriguing connections encountered in our investigation, such as the connection with the fundamental identity for the EM algorithm, the connection with the second CR (Chapman–Robbins) lower information bound, the connection with entropy, and connections between likelihood ratios and Bayes factors. We hope that these seemingly unrelated connections, as well as our specific proposals, will stimulate a general discussion and research in this theoretically fascinating and practically needed area.

Key words and phrases: EM algorithm, entropy, Fisher information, genetic linkage studies, haplotype-based association studies, noninformative prior, Kullback–Leibler information, relative information, Cox regression, partial likelihood.

Dan L. Nicolae is Associate Professor, Departments of Medicine and Statistics, The University of Chicago, 5734 S. University Ave, Chicago, Illinois 60637, USA (e-mail: nicolae@galton.uchicago.edu). Xiao-Li Meng is Whipple

V. N. Jones Professor and Chair of Statistics, Harvard University, Cambridge, Massachusetts, USA (e-mail: meng@stat.harvard.edu). Augustine Kong is Vice President

1. MANY CHALLENGES AND AN OVERVIEW

1.1 General Challenges

The central aim of this paper is to establish, in the context of hypothesis testing with incomplete data, a general framework for quantifying the amount of information in the observed data for a specific test being performed, relative to the full amount of information we would have had the data been complete. We do not address the issue of what is the best testing procedure, with or without the complete data, nor the issue of whether a full modeling/estimation strategy should or can be used instead. Rather, we address an increasingly common practical problem where the investigator has chosen the testing procedure, but needs to know the impact of the missing data on the test in terms of the relative loss of information. Such is the case in the genetic studies we briefly review in Sections 2 and 3.

Besides the specific challenges listed in Section 1.2, there are a number of general theoretical and methodological difficulties for establishing this general framework. First, unlike the similar task for estimation, where the notion of “fraction of missing information” is well studied and documented (e.g., Dempster, Laird and Rubin, 1977; Meng and Rubin, 1991), for hypothesis testing, there are two sets of measures to be contemplated, depending on whether the null hypothesis or the posited alternative model can be regarded as approximately adequate. Indeed, this is the very question the hypothesis test aims to provide partial evidence to discriminate.

Second, hypothesis testing procedures, especially those of nonparametric or semiparametric nature, are often constructed without reference to a specific (parametric) model. However, without an explicit model to link the unobserved quantities with the observed data, the very task of measuring how much information we have missed is neither possible in general nor meaningful. It is known, though not widely (e.g., Chernoff, 1979; Meng, 2001), that certain robust statistical procedures for estimation or testing can produce more efficient or powerful results with less data. Consequently, without assuming that our testing procedure is optimal under a specified optimality criterion, we may end up with the seemingly paradoxical situation that additional data may make our procedure less efficient or

powerful. That is, we may declare that more information is available with less data.

Third, in the context of small samples, quantifying information requires going beyond convenient and standard measures such as Fisher information, which is essentially a large-sample measure. Small-sample problems are rather frequent with incomplete data, as missing data reduce effective sample sizes. For the genetic studies we investigate in this paper, the small-sample problems arise even when there appear to be ample amounts of data. For example, we are often interested in measuring information content in individual components (e.g., an individual family in a large linkage study). In haplotype association studies, we often test haplotypes individually—data size may be large enough for testing a common haplotype, but very small for a rare one. In addition, an individual person can be fully informative for one haplotype because we know s/he cannot carry it, but much less so for another when we are uncertain whether s/he carries it or not. All these problems remind us that, in general scientific studies, small-sample problems appear more often than meets the eyes, namely, the numerical value of the sample size, because they sometimes appear in disguised forms.

Given the complex nature of small-sample problems requiring information measures, we literally have spent several years in our quest of finding a general workable approach. Not surprisingly, our conclusion is that robust Bayesian methods hold more promise. As we propose in Section 5, after establishing a likelihood-based large-sample framework in Section 4, this problem can be dealt with by considering posterior measures of the flatness of the entire likelihood surfaces. However, the problem of specifying an appropriate “default” prior is challenging. We report both our promising findings and open problems, hoping to stimulate further development on this practically important and theoretically fascinating topic. We also discuss various interesting theoretical connections (Section 6), as well as further methodological work and applications (Section 7).

1.2 Conflicting Aims in Genetic Studies

The central applied problem that motivated our work was the task to sensibly measure and efficiently compute the amount of information available in a particular genetic data set for a particular hypothesis tested by a particular statistical procedure. All genome-wide linkage screens carried out on qualitative and quantitative traits as well as most of the association studies

extract only part of the underlying information. Missing information can be the result of different sources, such as absence of DNA samples, missing genotypes, spacing between markers, noninformativeness of the markers, or unknown haplotype phase. Investigators want to know how much information is available in the observed data for the purpose of the study *relative* to the amount of information that would have been available if the data were complete. The notion of complete data is problem specific and, in parametric inference, depends on the sufficient statistics; for example, in linkage studies where the IBD (identical by descent) process is sufficient for inference, complete data can be achieved even if genotypes and/or individual samples are missing. Measures of relative information are needed for designing follow-up strategies in linkage studies, for example, using more genetic markers with existing DNA samples versus collecting DNA samples from additional families. Even for situations where we do not intend to recover the missing data, including situations where they cannot possibly be recovered (e.g., DNA samples from deceased ancestors), such measures can still be useful for the interpretation of the data and of the results, and for understanding the behavior of some of the inference tools (e.g., see Section 4.5).

The key methodological challenge is to find a measure that (1) is a reliable index of the relative information specific to a study purpose, (2) conditions on particular data sets, (3) is robust in the sense of general applicability, including to small data sets, (4) is easy to compute and (5) is subject to meaningful combination axioms. The reliability criterion (1) is obvious, and the criterion (2) is necessary because typically an investigator is interested in measuring the relative information in the data set at hand, not with respect to some “average” data set. Criterion (3) is desirable because in a typical genetic linkage study one needs to deal with a large amount of data with a variety of different complex structures (e.g., from a nuclear family to a very complex pedigree), often under time constraints, and thus it is not feasible to design separate measures to suit particular data structures. Criterion (4) is needed for similar reasons—any method without suitable computational efficiency, regardless of its theoretical superiority, will typically be ignored in routine genetic studies given the practical constraints. Criterion (5) ensures certain desirable coherence to prevent paradoxical measure properties (e.g., more informative studies receive less weight in the combined index) when combining studies.

To deal with all these criteria simultaneously requires a careful combination of Bayesian and frequentist perspectives. Some of the criteria [e.g., (1) and (2)] are most easily handled from the Bayesian perspective, and some [e.g., (5)] are easier to satisfy with a frequentist criterion. With large samples, as it is typical, likelihood theory provides a rather satisfactory solution, as we demonstrate in Section 4. For small samples, we have not been able to find a better alternative than to follow a robust Bayesian perspective, which takes full advantage of the Bayesian formulation in deriving information measures with desirable coherent properties, and at the same time it seeks measures that are robust to various misspecifications and are thus more generally applicable. We emphasize, however, that the computational burden associated to these Bayesian measures should not be overlooked, even in this age of the MCMC revolution, for the reasons underlying criterion (4) above. Nevertheless, it is more principled and fruitful to seek ways to increase computational efficiency *after* we establish theoretically sound measures. This is the route we follow.

1.3 Imputing Under the Null or Not—Gaining Insight

For those who have no (direct) interest in genetic studies, the following simple example may provide a stimulus to follow the methods developed in our paper. The example also provides some insights into a somewhat “perplexing” practical question when dealing with hypothesis testing in the presence of missing data: shall we impute under the null or not? We emphasize that the purpose of this example is *not* to illustrate imputation methods. Indeed, neither method discussed below can be recommended in general. Rather, it shows how we can quantify relative information by measuring *how inaccurate is* to erroneously treat imputations as if they were observed data.

Specifically, suppose y_1, \dots, y_n are i.i.d. realizations of Bernoulli(p), but only $n_0 < n$ of them are actually observed. Assuming that the missing data are missing completely at random (Rubin, 1976), we can denote the observed data by y_1, \dots, y_{n_0} . Evidently, a simple large-sample test (assuming n_0 is adequately large) for $H_0: p = p_0$ is to refer the test statistic (where the subscript “ob” stands for “observed data”)

$$(1) \quad T_{\text{ob}} = \frac{\bar{y}_{\text{ob}} - p_0}{\sqrt{p_0(1 - p_0)/n_0}}$$

to the null distribution $N(0, 1)$, where \bar{y}_{ob} is the average of the observed data.

Let us assume that the missing y 's were imputed using two mean-imputation methods. The first method is to impute each missing y by its mean, estimated by \bar{y}_{ob} . The second procedure is to impute each missing y by its mean assuming H_0 is true, that is, by p_0 . Clearly, with either imputation, if we treat the imputed data as if they were observed and apply the test (1) with $n_0 = n$, we will not reach the valid conclusion unless we adjust the null distribution $N(0, 1)$.

For the first method, the average of all data, observed and imputed, is $\bar{y}_1^* = \bar{y}_{ob}$. Therefore, if we erroneously treat the imputed values as real observations, we would compute our test statistics as

$$(2) \quad T_1^* = \frac{\bar{y}_1^* - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{1}{\sqrt{r}} T_{ob},$$

where $r = n_0/n$. In contrast, the second method would lead to

$$(3) \quad T_0^* = \frac{\bar{y}_0^* - p_0}{\sqrt{p_0(1 - p_0)/n}} = \sqrt{r} T_{ob},$$

because the average of all data, observed and imputed, is $\bar{y}_0^* = r\bar{y}_{ob} + (1 - r)p_0$.

Two aspects of the above calculations are important. First, in both cases, the resulting ‘‘completed-data’’ test statistic is proportional to the benchmark given in (1). Consequently, imputing under the null or not leads to the same answer, as long as we adjust the corresponding null distribution accordingly (the generality of this equivalence result obviously needs qualification, but the validity of a test is automatic when its null reference distribution is correctly specified). Second, identities (2) and (3) yield respectively

$$(4) \quad r = \left(\frac{T_{ob}}{T_1^*}\right)^2 \quad \text{and} \quad r = \left(\frac{T_0^*}{T_{ob}}\right)^2.$$

The results in (4) are important because $r = n_0/n$ measures the relative sample sizes, and hence the ‘‘relative information’’ in an i.i.d. setting. These results suggest that we consider measuring the relative information by how liberal the first imputation-based test is, when the imputations under the alternative are treated as real data, or how conservative the second test is, when the imputations under the null are treated as real observations. Our general large-sample results given in Section 4 show that these ideas are in fact general, once we replace the statistics in (4) by their appropriate log-likelihood ratio counterparts (recall the large-sample equivalence between log-likelihood ratio statistics and the Wald statistics in a form similar to T^2). Readers

who are not interested in genetic applications can go directly to Section 4, as Sections 2 and 3 provide the necessary background on the genetic problems to which our methods will be applied.

2. GENETIC LINKAGE ANALYSIS

2.1 Allele-Sharing Methods

Linkage refers to the co-inheritance of two markers or genes because they are located closely on the same chromosome. Allele-sharing methods are part of linkage techniques for locating regions on the genome that are very likely to contain disease susceptibility genes (e.g., Ott, 1991). The data usually consist of genotypes from a large number of markers (polymorphic locations) spread along the genome for individuals from n pedigrees. The allele-sharing methods focus on affected individuals, but genetic data on unaffected relatives are used to infer the inheritance patterns. Alleles at the same locus in two individuals are said to be identical by descent (IBD) if they originate from the same chromosome, and are called identical by state (IBS) if they appear to be the same. For a given location on the genome, the evidence for a disease-susceptibility locus linked to it is given by the sharing of alleles IBD among affected relatives in excess of what is expected if the marker is not linked to a genetic risk factor.

The simplest example of a data structure is the affected sib pair, as shown in Figure 1, where the left diagram shows a family with two affected brothers in which the parental information at a fixed locus is denoted by ‘‘A1’’ and ‘‘A2’’ for the father, and ‘‘A3’’ and ‘‘A4’’ for the mother. The siblings have one allele IBD (A2) which they inherited from their father, and different alleles inherited from their mother. In general, siblings share either two, one or no alleles IBD. Unconditionally, each allele has probability 1/2 to be transmitted; this leads to a probability of 1/4, 1/2, 1/4 for sharing zero, one, two alleles, respectively, identical by descent. Conditioned on the affection status of the sibs,

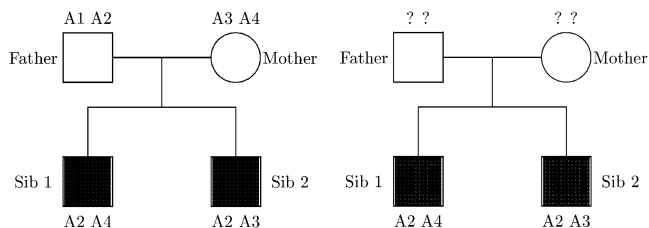


FIG. 1. Pedigree diagrams of an affected sib pair; the IBD sharing is known for the sibs in the left diagram, but only the IBS sharing is known for the sibs in the right diagram.

in the neighborhood of a disease gene, there is an expected increase in the number of alleles IBD across a collection of sib pairs; statistical testing methods are often used to measure the strength of the evidence.

In general, the data are not as simple as in the above example. The pedigree structures can contain far more complicated relations than sib pairs and more than two affected individuals. Most of the data sets extract only part of the underlying IBD information. In general, the information is incomplete at locations between markers. Even at marker locations, a variety of factors can lead to missing information, including any genotype data on deceased or unavailable family members, missing genotypes in the typed individuals, or non-informativeness of the markers. The right diagram of Figure 1 illustrates a family where the parental allele information is missing, so even though the allele sharing among the sib pair appears to be identical in pattern with that of the left diagram, it is not known if the sibs share one or zero alleles IBD as the two “A2” alleles might originate on different parental chromosomes.

In general, the marker information of all the loci on the chromosome is used to calculate a probability distribution on the space of inheritance vectors. For locus t and pedigree i , an inheritance vector, $\omega_i = \omega_i(t)$, is a binary vector that specifies, for all the nonfounding members of the pedigree, which grand-parental alleles are inherited. Under the assumption of no linkage, all inheritance vectors are equally likely, which leads to a uniform prior distribution on their space. For a sib pair, the inheritance vector has four elements, one for each parent-child combination. For example, the first element specifies whether the allele inherited by the first sib from his father originates from the grandfather or grandmother. Assuming no interference (Ott, 1991), a Hidden Markov Model can be used to calculate the inheritance distribution conditional on the genotypes at all marker loci (Lander and Green, 1987). The distribution of the inheritance vectors conditional on the observed data is the basis of the statistical inference, and it is used to determine the conditional distribution of the number of alleles IBD at a given location.

2.2 Hypothesis Testing Using Imputed Sharing Scores

In order to summarize the evidence for linkage in a pedigree, we can use a score S_i (Whittemore and Halpern, 1994; Kruglyak et al., 1996), a measure of IBD sharing among the affected individuals at locus t . In general, S_i is chosen such that it has a higher expected value under linkage than under no linkage. The

standardized form of S_i is $Z_i = (S_i - \mu_i)/\sigma_i$, where $\mu_i = E(S_i|H_0)$ and $\sigma_i^2 = \text{Var}(S_i|H_0)$. The test is typically in the form of linear combination over the n pedigrees,

$$(5) \quad Z = \frac{\sum_{i=1}^n \gamma_i Z_i}{\sqrt{\sum_{i=1}^n \gamma_i^2}},$$

where $\gamma_i \geq 0$ are weights assigned to the individual families. The weights can be chosen according to the number of affecteds and the relationship among them and/or other covariate information. Under the null hypothesis, Z has mean 0 and variance 1. Deviations from the null hypothesis can be tested using a $N(0, 1)$ approximation or the exact distribution of Z .

In general, Z_i 's are not directly observable due to missing information. A common practice is to impute/replace Z_i by $W_i = E(Z_i|\text{data}, H_0)$ to construct a test statistic (Kruglyak et al., 1996),

$$(6) \quad W = \frac{\sum_{i=1}^n \gamma_i W_i}{\sqrt{\sum_{i=1}^n \gamma_i^2}} = E(Z|\text{data}, H_0).$$

The main problem with this test statistic is the difficulty of directly evaluating its statistical significance. A standard $N(0, 1)$ approximation can be very inaccurate when there is a large amount of missing information, as can be seen from the following variance decomposition:

$$(7) \quad \begin{aligned} \text{Var}(Z|H_0) &= \text{Var}(E(Z|\text{data}, H_0)|H_0) \\ &+ E(\text{Var}(Z|\text{data}, H_0)|H_0), \end{aligned}$$

which implies

$$(8) \quad \text{Var}(W|H_0) = 1 - E(\text{Var}(Z|\text{data}, H_0)|H_0) \leq 1.$$

In many cases $\text{Var}(W|H_0)$ can be substantially less than 1, leading to a conservative test when the $N(0, 1)$ approximation is used. A more accurate approach is described in Section 2.3.

It is important to emphasize that, in allele-sharing studies, the amount of missing information can be made arbitrarily low, at least in theory, by increasing the number of markers in the region. That is why, in regions with evidence for linkage, it is important to predict whether by genotyping additional markers one will obtain a more significant deviation from the null. A different strategy for increasing the amount of information is to increase the sample size, that is, to collect DNA samples from additional families. Therefore knowing how much information is missing from the data is important for designing efficient follow-up strategies (see also Nicolae and Kong, 2004).

2.3 Associating a Test With a Model

The linkage methods we described are based on a chosen test statistic. In order to measure the relative information for a test statistic, we need to associate it with a model which specifies the stochastic relationship between the observed data and missing data beyond the null. Otherwise the question of relative information is not well defined, as it is emphasized in Section 1.1. It has been shown by Kong and Cox (1997) that for every test statistic of the form of (5), a class of one-parameter models can be constructed such that the efficient score (Cox and Hinkley, 1974) from each of the models gives asymptotically equivalent results to the given statistic. The inference procedures based on these models can be applied to any pedigree structure and missing data patterns.

As an illustration, we briefly describe the *exponential tilting* model of Kong and Cox (1997) applied to the one-locus allele-sharing statistic. A key assumption underlying this model (and other models for associating tests) is that the distribution of the inheritance vectors satisfies

$$(9) \quad \frac{P(\omega_i | H_A)}{P(\omega_i | H_0)} = \frac{P(Z_i = z(\omega_i) | H_A)}{P(Z_i = z(\omega_i) | H_0)} \quad \text{for all } i,$$

where ω_i is an inheritance vector for pedigree i that leads to a standardized scoring function equal to $z(\omega_i)$, and H_A denotes the alternative hypothesis. Note that any time an investigator employs a test solely based on the Z 's, as far as measuring information concerns, s/he is effectively assuming (9) regardless of whether or not s/he is aware of it.

Under assumption (9), it is sufficient to define the alternative models for Z_i 's. The exponential tilting model has the form

$$(10) \quad P_\theta(Z_i = z) = P_0(Z_i = z)c_i(\theta) \exp(\theta\gamma_i z),$$

where $P_0(Z_i = z)$ is specified by the null (i.e., no linkage) and $c_i(\theta) = [\sum_z P_0(Z_i = z) \exp(\theta\gamma_i z)]^{-1}$ is the renormalization constant. When Z is binary (e.g., as with half-sibs), the model is the same as the logistic regression model

$$(11) \quad \text{logit } P_\theta(Z_i = 1) = \mu_i + \theta\gamma_i,$$

where $\mu_i = \text{logit } P_0(Z_i = 1)$.

Given the exponential tilting model or other similar models (e.g., the linear model of Kong and Cox, 1997), the log-likelihood can be calculated exactly for any missing data patterns under the assumption (9). Similar constructions can be done for multilocus models, as in Nicolae (1999).

3. HAPLOTYPE-BASED ASSOCIATION STUDIES

3.1 Basics of Association Studies

Genetic association studies are designed to study potential associations between genetic variants and phenotypes (i.e., observable traits) on a population scale. The association between the genotype at a given marker and a disease can appear because the genetic variant may be a risk factor for the disease, or because the variant may be strongly correlated, called *in linkage disequilibrium* (LD) in the genetics literature, with a causal locus. The magnitude of the correlation depends on many factors including the distance between the markers and the population history.

For the simplicity of description, we focus here on a simple and popular design, case-control studies, although most results and principles are applicable to other sampling designs including those that incorporate quantitative traits and family-based controls. The simplest genetic variant and a commonly used genetic marker is a single nucleotide polymorphism (SNP) that takes on only two possible alleles. Denoting the two possible alleles as 1 and 2, there are three possible genotypes (1, 1), (1, 2) and (2, 2). The data for a case-control study can then be summarized as a 2-by-3 table where the entries are counts of the three genotype categories for the cases and controls, respectively. The data can be further reduced to a 2-by-2 table, where the entries are counts of the alleles, if a multiplicative model (Terwilliger and Ott, 1992; Falk and Rubinstein, 1987) for allele-risk is assumed. Note that under common assumptions, for a person randomly selected from the population, the two alleles carried are in Hardy-Weinberg equilibrium, that is, they are independent. This might not be true for an affected individual if the genotypes confer different risks, but it is true for the multiplicative model. Since this model is true under the null hypothesis which assumes no difference between the two alleles, assuming the multiplicative model for the purpose of testing does not affect the validity of the p-values. Obviously the power could be reduced if the specified model is different from the true alternative.

When the causal locus genotypes are not part of the data, or when the LD between the markers is strong, it might be more efficient to use more than one marker simultaneously. Most of these multilocus approaches for fine-mapping of disease alleles are based on haplotypes (e.g., McPeck and Strahs, 1999; Pritchard et al., 2000; Lam, Roeder and Devlin, 2000; Morris, Whittaker and

Balding, 2002; Zollner and Pritchard, 2005). Haplotype analyses can be used to investigate untyped genetic variation (Pe'er et al., 2006; Nicolae, 2006a), and can be used to explore which markers could be causal and which are unlikely to be so. A *haplotype* is a sequence of alleles along a chromosome, and hence each person has two haplotypes. The alleles appearing in a haplotype are said to be in *phase*. If the haplotypes are directly observed, then standard methods for analyzing contingency tables could be used to test various models (Gretarsdottir et al., 2003). Possible scenarios range from having a candidate at-risk haplotype to testing the full model (all the haplotypes have different risks) versus the null model (all the haplotypes have the same risk).

3.2 Causes of Incomplete Information

With a case-control study conducted with individual SNPs separately, the sufficient statistic is a 2-by-2 table under the multiplicative model and a likelihood ratio χ^2 test can be used to test the null hypothesis. A common cause of incomplete information is missing genotypes since yield is often less than perfect. The situation becomes more complicated when multiple SNPs are considered jointly. With two SNPs, both having alleles denoted with 1 and 2, there are four possible haplotypes: 1-1 (characterized by allele 1 at both SNPs), 1-2, 2-1 and 2-2. One simple alternative hypothesis is that haplotype 1-1 has risk that is different from the other three haplotypes which are assumed to have the same risk. It could be that we believe the two SNPs are functional and there is interaction between them that leads to increased disease risk for haplotype 1-1, but more common is the hypothesis that the putative, but unobserved, mutation occurred in the 1-1 background and the association between the haplotype and the trait is a result of both being associated with the mutation.

Under the multiplicative model, if haplotypes can be observed directly, then this problem can again be reduced to a 2-by-2 table of haplotype counts where the haplotypes 1-2, 2-1 and 2-2 are collapsed into one. However, for the commonly used technology, SNPs are genotyped separately. For an individual, apart from incomplete information due to missing the genotype for one of the SNPs, there is the issue of uncertainty in phase. Specifically, if the genotypes for the first and second SNP are (1, 2) and (1, 2) respectively, then the two haplotypes could be either (1-1, 2-2) or (1-2, 2-1), that is, the information on phase is missing. In general, there is incomplete information on phase if two or more SNPs that make up the haplotype are heterozygous. In

family-based association studies (e.g., Abecasis, Cardon and Cookson, 2000; Martin et al., 2000; Lange and Laird, 2002a, 2002b), the data on relatives will provide additional information on phase but there will still be uncertainty in inferring the haplotypes. For SNPs that are close together physically, there exist typing technologies that can determine the haplotypes directly, but they are usually much more expensive. Hence, from the design perspective, quantifying loss of information is relevant not only for power/sample-size calculations, but also for the choice of technology.

3.3 Measuring Relative Information Via Test Statistics—a Two-Sample Example

Apart from being relevant for experimental design and the interpretation of data, the amount of missing information is also useful for understanding the behavior of certain testing procedures. While one obvious way to perform testing is to apply a likelihood ratio test based on actual likelihoods computed for the observed incomplete data under the null hypothesis and alternative hypothesis separately, software for such calculations which allows the user to define models in a flexible manner is not readily available. However, available are methods and software based on the EM algorithm that can be applied to one sample to calculate maximum likelihood estimates of haplotype frequencies and expected haplotype counts for individuals or groups assuming the maximum likelihood estimates are the true parameter values (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long, Williams and Urbanek, 1995). Other more sophisticated methods and software to predict haplotype phase and estimate counts also exist (e.g., Stephens, Smith and Donnelly, 2001; Niu et al., 2002). It is very tempting for the user to apply standard testing procedures, such as the likelihood ratio test, by simply treating these expected/predicted counts as the actual observed counts. Doing this is analogous to the example in Section 1.3, except here we are dealing with a two-sample problem.

Specifically, if the original EM computation is applied to the cases and controls jointly as a single group (i.e., as under the null), but with the expectation counts tabulated for the individuals who are then separated into cases and controls, the test is conservative. If, however, the EM computation is applied to the cases and controls separately, then the result is anti-conservative. Moreover, the degree of conservativeness with the first procedure, in large samples, matches the degree of anti-conservativeness of the second procedure. To be more specific, consider the following simple example.

Suppose the observed data consist of 250 patients and 250 controls, or 500 chromosomes each. For a SNP, the patient counts are 300 allele 1 and 200 allele 2, and the control counts are 250 allele 1 and 250 allele 2. Let a and u denote respectively the population frequency of allele 1 in cases and controls. Under the null, the maximum likelihood estimates are $\tilde{a} = \tilde{u} = (300 + 250)/(500 + 500) = 0.55$ and the maximum likelihood estimates under the alternative are $\hat{a} = 300/500 = 0.6$ and $\hat{u} = 250/500 = 0.5$. Simple calculations show that the log-likelihood ratio χ^2 statistic is

$$2[\ell(\hat{a}, \hat{u}) - \ell(\tilde{a}, \tilde{u})] = 10.12.$$

Now suppose there are another 250 cases and 250 controls each with no data yet. Suppose we treat these as missing data and apply the EM computation to the cases and controls jointly. Since $\tilde{a} = \tilde{u} = 0.55$, these extra cases and controls each have expected counts of 275 allele 1 and 225 allele 2. Together with the original counts, this gives 575 allele 1 and 425 allele 2 for the cases, and 525 allele 1 and 475 allele 2 for the controls. The log-likelihood ratio χ^2 statistic computed based on these counts is 5.05, approximately one-half of 10.12.

By contrast, suppose the expected counts for the missing data are computed for the cases and controls separately. In this case, the presumed counts are simply twice the original counts: 600 allele 1 and 400 allele 2 for the cases, and 500 allele 1 and 500 allele 2 for the controls. The log-likelihood ratio χ^2 statistic computed from these counts is 20.24, or exactly double that of 10.12. While this example is extremely simple and unrealistic, the phenomenon seen does extend to real data with haplotypes. Indeed, this is just another example of the relationships given in (4). That is, either ratio will correctly estimate that the relative information is about 50%. The theoretical results in the next section provide a general framework for such estimation.

4. A LARGE-SAMPLE FRAMEWORK

4.1 Variations on the EM Identity

Our large-sample framework is built upon a simple identity involving expected log-likelihood ratios, where the expectation is with respect to the conditional distribution of the missing data given the observed data. Expected lod scores have also been used in the genetics literature to measure the information content of the data (Ott, 2001), and to investigate optimality and validity of analytic strategies (e.g., Cleves and Elston, 1997; Abreu, Greenberg and Hodge, 1999; Daw,

Thompson and Wijsman, 2000). Note that lod stands for logarithm (usually base 10) of the odds, and is used as a statistic for testing whether two loci are linked.

Specifically, let Y_{co} be the complete data and Y_{ob} be the observed data—note that here Y_{ob} is a function of Y_{co} . Let $\ell(\theta|D)$ be the log-likelihood of θ given data D . Then for any θ_1 and θ_2 ,

$$\begin{aligned} & \ell(\theta_1|Y_{co}) - \ell(\theta_2|Y_{co}) \\ &= [\ell(\theta_1|Y_{ob}) - \ell(\theta_2|Y_{ob})] \\ (12) \quad &+ [\log f(Y_{co}|Y_{ob}, \theta_1) \\ & \quad - \log f(Y_{co}|Y_{ob}, \theta_2)]. \end{aligned}$$

By taking conditional expectation with respect to $f(Y_{co}|Y_{ob}, \theta)$, where θ is to be chosen, we have

$$\begin{aligned} & E[\text{lod}(\theta_1, \theta_2|Y_{co})|Y_{ob}, \theta] \\ (13) \quad &= \text{lod}(\theta_1, \theta_2|Y_{ob}) \\ &+ E\left[\log \frac{f(Y_{co}|Y_{ob}, \theta_1)}{f(Y_{co}|Y_{ob}, \theta_2)} \middle| Y_{ob}, \theta\right], \end{aligned}$$

where $\text{lod}(\theta_1, \theta_2|D)$ is the log of odds of θ_1 over θ_2 given data D . Here log can be of any base, and lod is the log of the likelihood ratio, or more generally the log of posterior ratios. Identity (13) is a simple extension of the key identity given in Dempster, Laird and Rubin (1977) for the EM algorithm. Specifically, using the notation of Dempster, Laird and Rubin (1977)

$$\begin{aligned} (14) \quad & Q(\theta|\theta') = E[\ell(\theta|Y_{co})|Y_{ob}, \theta'] \quad \text{and} \\ & H(\theta|\theta') = E[\log f(Y_{co}|Y_{ob}, \theta)|Y_{ob}, \theta'], \end{aligned}$$

identity (13) is the same as

$$\begin{aligned} (15) \quad & Q(\theta_1|\theta) - Q(\theta_2|\theta) \\ &= \ell_{ob}(\theta_1) - \ell_{ob}(\theta_2) + H(\theta_1|\theta) - H(\theta_2|\theta), \end{aligned}$$

where $\ell_{ob}(\theta) \equiv \ell(\theta|Y_{ob})$. In Dempster, Laird and Rubin (1977), (15) was given with $\theta = \theta_2$, and was the basis for establishing the celebrated monotone convergence property of the EM algorithm. As we shall see, this intrinsic connection with the EM algorithm not only helps greatly our theoretical development in Section 6, but more importantly it enables us to compute our information measures directly from quantities that are already used for the EM computation.

Intuitively, if θ_1 is the truth, then if we had more data, which would come from $f(Y_{co}|Y_{ob}, \theta_1)$, we would on

average have a larger lod score than $\text{lod}(\theta_1, \theta_2|Y_{\text{ob}})$. Indeed, by taking $\theta = \theta_1$ in (13) we see

$$(16) \quad \begin{aligned} E[\text{lod}(\theta_1, \theta_2|Y_{\text{co}})|Y_{\text{ob}}, \theta_1] \\ &= \text{lod}(\theta_1, \theta_2|Y_{\text{ob}}) + \text{KL}(\theta_1 : \theta_2) \\ &\geq \text{lod}(\theta_1, \theta_2|Y_{\text{ob}}), \end{aligned}$$

where $\text{KL}(\theta_1 : \theta_2) \geq 0$ is the Kullback–Leibler information—in favor of θ_1 against θ_2 when θ_1 is true—contained in the conditional distribution of Y_{co} given Y_{ob} . The inequality in (16) becomes equality if and only if $\text{KL}(\theta_1 : \theta_2) = 0$, which happens if and only if $f(Y_{\text{co}}|Y_{\text{ob}}, \theta_1) = f(Y_{\text{co}}|Y_{\text{ob}}, \theta_2)$ (a.s.); that is, given Y_{ob} , the additional data would contain no information to discriminate θ_2 from θ_1 . The Kullback–Leibler distance has been used extensively in information theory (e.g., Cover and Thomas, 1991) and mathematical statistics (e.g., Aitchison, 1975). Recent work on using K–L loss includes George, Feng and Xu (2006) and references therein.

Similarly, if θ_2 is the truth, then on average we would expect a smaller $\text{lod}(\theta_1, \theta_2|Y_{\text{co}})$ if we had observed Y_{co} . Mathematically, this is shown by taking $\theta = \theta_2$ in (13), which leads to

$$(17) \quad \begin{aligned} E[\text{lod}(\theta_1, \theta_2|Y_{\text{co}})|Y_{\text{ob}}, \theta_2] \\ &= \text{lod}(\theta_1, \theta_2|Y_{\text{ob}}) - \text{KL}(\theta_2 : \theta_1) \\ &\leq \text{lod}(\theta_1, \theta_2|Y_{\text{ob}}), \end{aligned}$$

and the inequality becomes equality if and only if, as before, $f(Y_{\text{co}}|Y_{\text{ob}}, \theta_1) = f(Y_{\text{co}}|Y_{\text{ob}}, \theta_2)$.

It is important to emphasize that because all the expectations above are conditional upon Y_{ob} , it is legitimate to allow any of the θ 's to depend on Y_{ob} . In particular, the null value θ_0 in the rest of this paper can be either a known fixed value when H_0 is a sharp null, or more generally the constrained MLE of θ from $\ell(\theta|Y_{\text{ob}})$ under the null. It is also important to emphasize that although in this section we focus on large-sample measures primarily because of their reliance on maximum likelihood estimators (MLEs), as discussed below, all the equalities and inequalities discussed above do not involve any approximation, large sample or not. Therefore all measures discussed below can also be very useful for small samples, as long as the MLEs can be trusted (e.g., a small-sample MLE can have good properties, such as under the normal models).

4.2 A Large-Sample Measure of Relative Information Against H_0

Suppose the null value is θ_0 and that the MLE of θ (under H_1) given Y_{ob} is θ_{ob} , and $\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{ob}}) (\geq 0)$ is used to assess the evidence against $H_0 : \theta = \theta_0$. To avoid technical complexity that is not of general interest for our proposals, we will assume (I) θ_{ob} is unique, an assumption typically automatic with large samples, and (II) $\theta_{\text{ob}} \neq \theta_0$, an assumption rarely, if ever, violated in practice. (Nevertheless, for theoretical completeness, we will consider the case of $\theta_{\text{ob}} = \theta_0$ in Section 6 via a limiting argument.) Then, if we intend to measure the information in the unobserved data for discrediting H_0 , under the large-sample assumption, a natural thing to do is to treat θ_{ob} as the “truth,” and measure the expected loss of lod in favor of θ_{ob} relative to the expected complete-data lod score. Namely, we can naturally define

$$(18) \quad \begin{aligned} \mathcal{R}I_1 &= \frac{\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{ob}})}{E[\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{co}})|Y_{\text{ob}}, \theta_{\text{ob}}]} \\ &= \frac{\ell_{\text{ob}}(\theta_{\text{ob}}) - \ell_{\text{ob}}(\theta_0)}{Q(\theta_{\text{ob}}|\theta_{\text{ob}}) - Q(\theta_0|\theta_{\text{ob}})}. \end{aligned}$$

The last expression shows that the computation of $\mathcal{R}I_1$ only requires evaluations, at $\theta = \theta_0$ and $\theta = \theta_{\text{ob}}$, of the observed-data log-likelihood $\ell_{\text{ob}}(\theta)$ and the Q function, where the latter is readily available from the EM algorithm.

Under assumptions (I) and (II), $\mathcal{R}I_1$ is well defined and by (16), $0 < \mathcal{R}I_1 \leq 1$. It is 1 if and only if $\text{KL}(\theta_{\text{ob}} : \theta_0) = 0$, which means that the missing data cannot distinguish between θ_{ob} and θ_0 and thus there is no missing information given Y_{ob} . It approaches 0 if and only if $\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{ob}})/\text{KL}(\theta_{\text{ob}} : \theta_0) \rightarrow 0$, which makes sense because if the observed-data likelihood has diminishing ability, relative to that of the missing-data model [as measured by $\text{KL}(\theta_{\text{ob}} : \theta_0)$], to distinguish between θ_{ob} and θ_0 , then as far as providing evidence against H_0 , the missing information approaches 100%. One very appealing feature of $\mathcal{R}I_1$ is its direct interpretability. As seen in the haplotype example in Section 3.3, a value of $\mathcal{R}I_1 = 0.5$ implies that if we had the complete data, the lod score would be expected to be twice ($\mathcal{R}I_1^{-1} = 2$) as large.

When $\ell(\theta|Y_{\text{co}})$ is linear in a (multidimensional) summary statistics (i.e., a complete-data sufficient statistics) $S(Y_{\text{co}})$, as when the complete-data model is from an exponential family, $\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{co}})$ can be written as $\text{lod}(\theta_{\text{ob}}, \theta_0|S(Y_{\text{co}}))$ and

$$E[\text{lod}(\theta_{\text{ob}}, \theta_0|Y_{\text{co}})|Y_{\text{ob}}, \theta_{\text{ob}}] = \text{lod}(\theta_{\text{ob}}, \theta_0|S^*(Y_{\text{ob}})),$$

where $S^*(Y_{ob}) = E[S(Y_{co})|Y_{ob}, \theta_{ob}]$. That is, $\mathcal{R}I_1$ measures the anti-conservativeness of the completed-data test by pretending that the actual value of the unobserved $S(Y_{co})$ is the same as its imputation under the (estimated) alternative. Therefore, $\mathcal{R}I_1$ is the general version of the first case in (4).

This measure also has the following property when combining data sets. Suppose $Y_{co} = \{Y_{co}^{(1)}, \dots, Y_{co}^{(n)}\}$ are mutually independent and we define $\mathcal{R}I_i$ for each $Y_{co}^{(i)}$ as in (18) but using θ_{ob} instead of individual $\theta_{ob}^{(i)}$ (i.e., an MLE based on $Y_{ob}^{(i)}$); then the overall $\mathcal{R}I$ is a weighted harmonic mean of $\mathcal{R}I_i$'s weighted by the individual lod score, $\text{lod}_i = \text{lod}(\theta_{ob}, \theta_0|Y_{ob}^{(i)})$, namely,

$$(19) \quad \mathcal{R}I_1^{-1} = \frac{\sum_{i=1}^n \text{lod}_i \mathcal{R}I_{1,i}^{-1}}{\sum_{i=1}^n \text{lod}_i}.$$

However, the individual lod score, lod_i , is not necessarily always positive in practice, a problem that is closely related to the problem of defining relative measures for small data sets (e.g., for individual family), as discussed in Section 5. Note that $\mathcal{R}I_1$ can also be expressed as weighted arithmetic mean of $\mathcal{R}I_{1,i}$ if we choose the weights to be proportional to the expected individual complete-data lod score $\text{lod}_i^{(c)} = E[\text{lod}(\theta_{ob}, \theta_0|Y_{co}^{(i)})|Y_{ob}^{(i)}]$:

$$(20) \quad \mathcal{R}I_1 = \frac{\sum_{i=1}^n \text{lod}_i^{(c)} \mathcal{R}I_{1,i}}{\sum_{i=1}^n \text{lod}_i^{(c)}}.$$

Clearly (19) and (20) are equivalent, as long as $\mathcal{R}I_{1,i} > 0$. The harmonic rule (19) is somewhat more appealing because of the direct interpretation of the weight lod_i .

4.3 A Large-Sample Measure of Relative Information Under H_0

Inequality (17) also suggests a large-sample measure of the relative information under H_0 . By taking $\theta_1 = \theta$ and $\theta_2 = \theta_0$ in (17) we obtain that

$$(21) \quad \begin{aligned} & E[\text{lod}(\theta, \theta_0|Y_{co})|Y_{ob}, \theta_0] \\ &= \text{lod}(\theta, \theta_0|Y_{ob}) - \text{KL}(\theta_0 : \theta) \\ &\leq \text{lod}(\theta, \theta_0|Y_{ob}). \end{aligned}$$

Thus, when the additional data are from $f(Y_{co}|Y_{ob}, \theta_0)$, the expected complete lod score cannot exceed the one based on the observed data, for any θ . We can use $\max_{\theta} E[\text{lod}(\theta, \theta_0|Y_{co})|Y_{ob}, \theta_0]$, which cannot exceed $\text{lod}(\theta_{ob}, \theta_0|Y_{ob})$ by (21), as our best estimate of the complete-data lod score; the use of a single point

estimate of the complete-data lod score without considering its uncertainty can be justified under the large-sample assumption. Consequently, we can define

$$(22) \quad \begin{aligned} \mathcal{R}I_0 &= \frac{\max_{\theta} E[\text{lod}(\theta, \theta_0|Y_{co})|Y_{ob}, \theta_0]}{\text{lod}(\theta_{ob}, \theta_0|Y_{ob})} \\ &= \frac{\max_{\theta} [Q(\theta|\theta_0) - Q(\theta_0|\theta_0)]}{\ell_{ob}(\theta_{ob}) - \ell_{ob}(\theta_0)}. \end{aligned}$$

The last expression shows again the computational efficiency of this measure because $\max_{\theta} Q(\theta|\theta_0)$ is the same as carrying out the E-step and M-step of an EM algorithm, by pretending the previous iterated value is $\theta = \theta_0$. However, we emphasize that the use of $\max_{\theta} E[\text{lod}(\theta, \theta_0|Y_{co})|Y_{ob}, \theta_0]$ in our definition of $\mathcal{R}I_0$ instead of $E[\max_{\theta} \text{lod}(\theta, \theta_0|Y_{co})|Y_{ob}, \theta_0]$ is not because this computation is easy, but rather because of the nature of the fundamental identity (13), which requires we maximize the expected complete-data lod score.

Like $\mathcal{R}I_1$, $0 \leq \mathcal{R}I_0 \leq 1$. Unlike $\mathcal{R}I_1$, however, the investigation of when $\mathcal{R}I_0$ approaches one or zero is a more complicated matter, especially when the difference between θ_{ob} and θ_0 is large. This is a partial reflection of the fact that $\mathcal{R}I_0$ is defined under the assumption that the null hypothesis is (approximately) valid, which would be contradicted by a large value of $\delta = \theta_{ob} - \theta_0$, especially under our large-sample assumption. Therefore, it is more sensible to investigate its theoretical properties when δ is small, in which case it is essentially equivalent to $\mathcal{R}I_1$, as we will establish in Section 6. Nevertheless, it is useful to remark here that under the additional assumption that θ_{ob} is the unique stationary point of $\ell_{ob}(\theta)$, the numerator of $\mathcal{R}I_0$ is zero if and only if its denominator is zero, that is, if and only if $\ell_{ob}(\theta_{ob}) = \ell_{ob}(\theta_0)$. [The ‘‘if’’ part of this result is a trivial consequence of (21). The ‘‘only if’’ part follows from the fact that if the numerator is zero, then θ_0 is a maximizer of $Q(\theta|\theta_0)$, which means that θ_0 must also be a stationary point of $\ell_{ob}(\theta)$ by (56) in Appendix A.2.] This demonstrates that in order for $\mathcal{R}I_0$ to be very small, the observed-data likelihood must suffer a diminishing ability to distinguish between θ_{ob} and θ_0 , just as with $\mathcal{R}I_1$.

Also as with $\mathcal{R}I_1$, when $\ell(\theta|Y_{co})$ is linear in $S(Y_{co})$, $\mathcal{R}I_0$ can be computed simply as

$$\mathcal{R}I_0 = \frac{\max_{\theta} \text{lod}(\theta, \theta_0|S_0^*(Y_{ob}))}{\text{lod}(\theta_{ob}, \theta_0|Y_{ob})},$$

where $S_0^*(Y_{ob}) = E(S(Y_{co})|Y_{ob}, \theta_0)$, that is, the mean imputation of the missing $S(Y_{co})$ under the null. Therefore, $\mathcal{R}I_0$ is the general version of the second case in (4), and it measures the conservativeness of our test

when we impute under the null. Its main disadvantage, as previously mentioned, is that it can provide very misleading information when the true θ is far away from the null. On the other hand, because it is computed at the null, it is less sensitive, compared to $\mathcal{R}I_1$, to possible misspecification of the alternative model. We will illustrate this in Section 6.3, where we will discuss further the pros and cons of $\mathcal{R}I_0$.

4.4 Illustration With a Linkage Analysis

In the context of allele-sharing methods, the measures we introduced in the previous sections are implemented in the software ALLEGRO (Gudbjartsson et al., 2000), and are discussed in detail in Nicolae and Kong (2004). In Figure 2, $\mathcal{R}I_1$ and $\mathcal{R}I_0$ are plotted for various locations along chromosome 22 (the unit for the X-axis is CentiMorgans) in a data set consisting of 127 pedigrees used in an inflammatory bowel disease study (Cho et al., 1998). It can be seen that, in this case, the two measures are very close across the entire chromosome. This happens because the sample size is large and the distribution of the family sharing scores is fairly symmetric. Also plotted is an inheritance-vector-based information measure calculated by the software GENEHUNTER (Kruglyak et al., 1996). This measure takes advantage of the fact that the inheritance vectors are equally likely under H_0 and that, for the fixed support of the space of the inheritance vectors, the Shannon entropy (1949) is maximal for the uniform distribution on the support. For the i th pedigree in the study

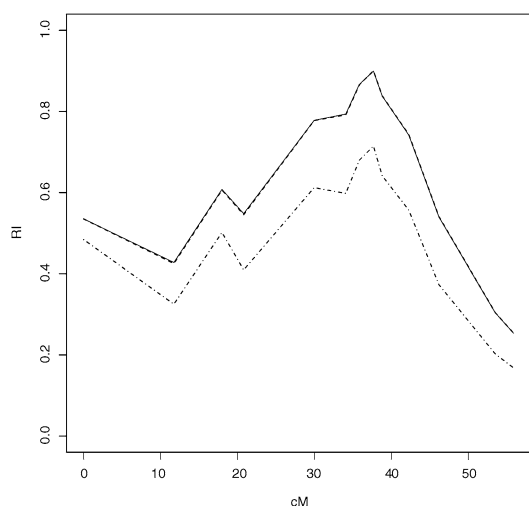


FIG. 2. The large-sample measures of information are plotted against the genetic distance. The top two curves (almost identical) correspond to $\mathcal{R}I_1$ and $\mathcal{R}I_0$; the bottom curve (dot-dashed) corresponds to the entropy-based measure (Kruglyak et al., 1996).

and a given position t , it is defined as

$$1 - \frac{E^i(t)}{E_0^i} \\ \equiv 1 - \frac{-\sum_{\omega_i} P(\omega_i|\text{data}, H_0) \log_2 P(\omega_i|\text{data}, H_0)}{-\sum_{\omega_i} P(\omega_i|H_0) \log_2 P(\omega_i|H_0)},$$

where ω_i was defined in Section 2.1. The definition of the overall information content of a data set is based on the global entropy, which, summed over all n pedigrees, satisfies

$$\mathcal{E}_R = 1 - \frac{E(t)}{E_0} \equiv 1 - \frac{\sum_{i=1}^n E^i(t)}{\sum_{i=1}^n E_0^i}.$$

While \mathcal{E}_R has several desired properties (e.g., it is always between zero and one, and it is one when there is perfect data on the inheritance vectors), it has some deficiencies that make it unsuitable for the linkage application. The most fundamental problem is that it measures the relative information in the whole inheritance vector space, which could be very different from what is available for a particular test statistic that is a function of the inheritance vectors. For example, in the right diagram of Figure 1, we may be nearly certain, and hence suffer very little missing information, that the IBS sharing is actually IBD if we have the knowledge that the allele ‘‘A2’’ has very low population frequency, even though the parental alleles are unknown and therefore \mathcal{E}_R is low (see Nicolae and Kong, 2004, for more details). It is also possible that \mathcal{E}_R is higher than the measures described in this paper (e.g., Thalamuthu et al., 2005), for example in situations where there is a lot of data on unaffected individuals in a family, but little or no data on affected individuals. In these cases, \mathcal{E}_R will capture available information that is not directly of interest when we are performing affecteds-only analyses.

The serious overestimation or underestimation of relative information can have a great impact on the design of follow-up studies. One can decide on increasing the marker density if the relative information is low, as opposed to increasing the sample size. Both strategies are expensive, and therefore deciding what is the most efficient design is of great importance in practice. For example, at the global mode in Figure 2, our measures indicate that we have about 90% relative information, implying that potentially we can increase the lod score by only about 11% ($1/0.9 = 1.11$) if we add markers to make the IBD process approximately known (assuming the value of θ_{ob} remains approximately the same with the additional data). On the other hand,

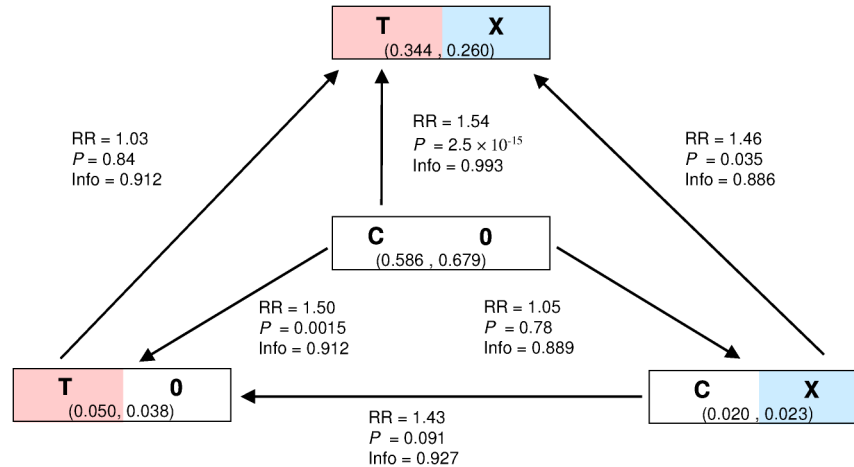


FIG. 3. For each haplotype, estimated frequencies in patients and controls are displayed. RR is estimated risk of the haplotype the arrow is pointing to (h_1) relative to the haplotype the arrow is pointing from (h_2), and is calculated as $[n(h_1)/m(h_1)]/[n(h_2)/m(h_2)]$ where n and m are estimated haplotype counts in patients and controls respectively. P values are calculated based on a likelihood ratio test that properly takes missing information into account. Information shown is $\mathcal{R}I_1$. Very similar numbers are obtained for $\mathcal{R}I_0$.

the entropy-based measure from GENEHUNTER indicates that we have about 70% information, suggesting that a more substantial gain (over 40%) is possible by increasing the density of the markers. Therefore these two approaches of measuring information are likely to lead to different strategies in allocating the resources, but evidently, in this example, it is unlikely the test results will change significantly by adding more markers near the location at the global mode.

4.5 Illustration With a Haplotype-Based Study

In Grant et al. (2006), the gene *TCF7L2* was found to be associated with type-2 diabetes. In particular, allele T of *rs7903146* (SNP402) and allele X of a microsatellite marker DG10S478 are both associated with elevated risk of type-2 diabetes (p -value $< 10^{-10}$). Allele T and allele X are substantially correlated ($r \approx 0.85$) and their effects could not be clearly distinguished from each other in the original study. However, with additional data (Helgason et al., 2007), it became clear that allele T is more strongly associated with diabetes than allele X. SNP402 has alleles T and C, and DG10S478 has alleles X and 0. Jointly there are four haplotypes: TX, CX, T0 and C0. Figure 3 presents pairwise comparisons of these four haplotypes. Data are from 1021 patients ($n = 2042$ chromosomes) and 4273 controls ($m = 8546$ chromosomes). Consistent with the single marker associations, haplotype TX is found to have elevated risk relative to C0. To distinguish between the effects of alleles T and X, haplotype T0 is found to confer risk that is similar to that of TX and has significantly higher risk than C0. By contrast,

haplotype CX is found to have risk similar to that of C0 and significantly lower risk than TX. In other words, given SNP402, DG10S478 does not appear to provide extra information about diabetes, which keeps SNP402 as a strong candidate for being the functional variant.

The yield of the genotypes is not perfect. Each subject has genotypes for at least one of the two markers, but about 3.5% of the genotypes are missing. This together with uncertainty in phase leads to the incomplete information summarized in Figure 3. Interestingly, while the same data are used for the six pairwise comparisons, the fraction of missing information can be quite different. Most striking is that the relative information for the test of TX versus C0 is very close to 100%, while the other tests all have more substantial missing information. We explore some of the reasons below.

Notice that T is highly correlated with X and C highly correlated with 0. As a consequence, TX and C0 are much more common than T0 and CX. Consider a subject whose genotype for D10GS478 is missing. Here we can think of his two alleles for SNP402 one at a time. Given an observed allele T, it is clear that the haplotype is not C0 and quite likely to be TX. Hence, even though incomplete, there is still substantial information provided for the test of TX versus C0. By contrast, we know that this chromosome is useful for the test of TX against T0, but with the allele of DG10S478 missing, that information is completely lost. Even more interesting is that, if the observed allele is C instead, then this haplotype is completely uninformative for the

test of TX versus T0, that is, there is actually no information here whether or not we know the corresponding DG10S478 allele. In effect, the genotype of SNP402 is an ancillary statistic for the test of TX against T0 (or CX against C0). It tells us how much information we can get from this individual assuming that we have no missing data, but by itself does not provide any information for the test. Moreover, if the test of TX versus T0 is of key interest, then effort to fill up missing genotypes for DG10S478 should be focused on those individuals that are T/T homozygous for SNP402.

When genotypes of both markers are observed, uncertainty in phase only exists for those individuals that are doubly heterozygous, that is, having genotypes C/T and O/X. Such an individual either has haplotypes C0/TX (scenario I) or CX/T0 (scenario II). Scenario II provides no information for the test of TX versus C0. Scenario I does contribute something to the test, but by providing a count of 1 to both TX and C0, its impact on the test of TX versus C0 is rather limited. By contrast, for the test of TX versus T0, scenario I adds a count of 1 to TX while scenario II adds a count to T0. Hence, uncertainty in phase has a much bigger impact on the test of TX versus T0 than the test of TX versus C0. This example, therefore, illustrates clearly the importance of measuring *test-specific* relative information.

5. SMALL-SAMPLE EXPLORATORY MEASURES

5.1 A Bayesian Framework

The measures defined in previous sections do not necessarily work with small samples (e.g., data for one family) because they rely on the ability of the MLE to summarize the whole likelihood function. The Bayesian approach becomes a valuable tool in such settings even if we do not necessarily have a reliable prior; we can first construct a coherent measure and then investigate the choice of prior. Since a likelihood quantifies the information in the data through its ability of distinguishing different values of the parameter, it is natural to consider measuring the relative information by comparing how the observed-data likelihood deviates from “flatness” relative to the same deviation in the complete-data likelihood. The Bayesian method is ideal here because we need to assess the change in this deviation due to the joint variability in the missing data and in the parameter. A reasonable measure of this deviation, conditioning on Y_{ob} , is the posterior variance of the likelihood ratio (LR). This measure is appealing because it is naturally scaled via the equality

$$(23) \quad \text{LR}(\theta_0, \theta | Y_{ob}) = \text{E}[\text{LR}(\theta_0, \theta | Y_{co}) | Y_{ob}, \theta],$$

which guarantees that

$$(24) \quad 0 \leq \mathcal{B}I_1^\pi \equiv \frac{\text{Var}[\text{LR}(\theta_0, \theta | Y_{ob}) | Y_{ob}]}{\text{Var}[\text{LR}(\theta_0, \theta | Y_{co}) | Y_{ob}]} \leq 1,$$

where π indexes the underlying prior on θ used by (24), and $\mathcal{B}I$ stands for “Bayes Information.” We assume here that the complete-data likelihood surface is not flat, as otherwise the model is of little interest. The denominator in (24) is therefore positive. We also need to assume that the posterior variances of the two likelihood ratios are finite. This second assumption can be violated in practice, but a second measure we will propose below essentially circumvents this problem.

In the presence of nuisance parameters (under the null), there is also a subtle issue regarding the nuisance part of θ_0 , in the definition of $\mathcal{B}I_1^\pi$. For a full Bayesian calculation, one should leave it unspecified and average it over in the posterior calculation, just as with the θ in $\text{LR}(\theta_0, \theta)$. On the other hand, to be consistent with the large-sample measures as defined in Section 4, we can fix the nuisance parameter part in θ_0 by its observed-data MLE under the null. Identity (23) still holds with such a “fix,” because the calculation there conditions on the observed data. This “fix” may seem to be rather ad hoc from a pure Bayesian point of view. However, it can be viewed as an attempt in capturing the dependence (if any) between the parameter of interest and the nuisance parameter under the null, a dependence that is ignored by a single prior on the nuisance parameter regardless of the null. This subtle issue is related to the difference between “estimation prior” and “hypothesis testing prior,” an issue we will explore in subsequent work. Here we just note that all the Bayesian measures defined in this section can be constructed with either approach for handling the nuisance parameter under the null, although those under shrinking prior toward the null (see Section 5.2) are most easily obtained when the nuisance parameter under the null is fixed at its MLE (or some other known values).

With either approach,

$$\mathcal{B}I_1^\pi = 1 \quad \text{if and only if}$$

$$\text{E}\{\text{Var}[\text{LR}(\theta_0, \theta | Y_{co}) | Y_{ob}, \theta] | Y_{ob}\} = 0,$$

which occurs if and only if for almost all the θ in the support of the posterior, the complete-data likelihood $\text{LR}(\theta_0, \theta | Y_{co})$ is (almost surely) a constant as a function of the missing data, and thus the missing data would offer no additional help in distinguishing θ from θ_0 . On the other hand, $\mathcal{B}I_1^\pi = 0$ if and only if the observed-data likelihood ratio is a constant, and thus

there is no information in the observed data for testing H_0 using $\text{LR}(\theta_0, \theta|Y_{\text{ob}})$. Other characteristics of this measure depend on the choice of the prior π , and they will be discussed in the following sections.

One potential drawback of $\mathcal{B}I_1^\pi$ is that it can be greatly affected by the large variability in the likelihood ratios, as functions of the parameters, for example, when very unlikely parameter values were given nontrivial prior mass. This problem can be circumvented to a large extent by using the posterior variance of the *log-likelihood ratio*,

$$\text{Var}[\text{lod}(\theta, \theta_0|Y_{\text{ob}})|Y_{\text{ob}}].$$

The use of the log scale also makes it much more likely, compared to the ratio scale, that the resulting posterior variances are finite. Evidently, just as with the posterior variance of the likelihood ratio, this is equal to zero if and only if the observed-data likelihood ratio is a constant (on the support of the posterior). Similarly,

$$\text{Var}\left[\log\frac{P(Y_{\text{co}}|Y_{\text{ob}}, \theta)}{P(Y_{\text{co}}|Y_{\text{ob}}, \theta_0)}\middle|Y_{\text{ob}}\right]$$

is equal to zero if and only if there is no additional information in the missing data for testing H_0 . These suggest that we can also measure the relative information by

$$\begin{aligned} \mathcal{B}I_2^\pi &= \text{Var}[\text{lod}(\theta, \theta_0|Y_{\text{ob}})|Y_{\text{ob}}] \\ (25) \quad &\cdot \left(\text{Var}[\text{lod}(\theta, \theta_0|Y_{\text{ob}})|Y_{\text{ob}}] \right. \\ &\left. + \text{Var}\left[\log\frac{P(Y_{\text{co}}|Y_{\text{ob}}, \theta)}{P(Y_{\text{co}}|Y_{\text{ob}}, \theta_0)}\middle|Y_{\text{ob}}\right] \right)^{-1}, \end{aligned}$$

where, as for $\mathcal{B}I_1^\pi$, π indexes the underlying prior on θ .

Although the use of lod is more natural in view of the large-sample measures given in Section 4, it does not admit the nice ‘‘coherence’’ identity for the likelihood ratio as given in (23). Indeed, we had to remove ad hoc a cross term in the denominator of (25) in order to keep the resulting ratio always inside the unit interval. Furthermore, as we show in Section 6, the use of the ratio scale, instead of log ratio, leads to a number of interesting identities between likelihood ratios and Bayes factors, and it is more connected with some finite-sample measure of information in the literature. Whereas such trade-offs need to be explored, our general results in the next section imply that in the neighborhood of θ_0 , the differences between these two measures should be small.

5.2 Limits Under a Shrinking Prior Toward Null

Given their definitions, the immediate question is how to choose π and how to compute $\mathcal{B}I_1^\pi$ and $\mathcal{B}I_2^\pi$ efficiently since, in general, their calculations require integrations that cannot be performed analytically. When the truth is believed to be in a neighborhood of the null value θ_0 , a θ_0 -neighbor approximation to $\mathcal{B}I_1^\pi$ and $\mathcal{B}I_2^\pi$ can be obtained by choosing π to be $U(\theta_0 - \delta, \theta_0 + \delta)$ with $\delta > 0$ small. It is proved in Appendix A.1 that the two Bayesian measures have the same limit as $\delta \rightarrow 0$, denoted by $\mathcal{B}I_0$,

$$\begin{aligned} \mathcal{B}I_0 &= \frac{S^2(\theta_0|Y_{\text{ob}})}{S^2(\theta_0|Y_{\text{ob}}) + \text{Var}(S(\theta_0|Y_{\text{co}})|Y_{\text{ob}}, \theta_0)} \\ (26) \quad &= \frac{S^2(\theta_0|Y_{\text{ob}})}{S^2(\theta_0|Y_{\text{ob}}) + I_{\text{mi}}(\theta_0|Y_{\text{ob}})}, \end{aligned}$$

where $S(\theta|Y_{\text{ob}})$ and $S(\theta|Y_{\text{co}})$ are respectively the observed-data and complete-data score function, and $I_{\text{mi}}(\theta|Y_{\text{ob}})$ is the expected (missing) Fisher information from $f(Y_{\text{co}}|Y_{\text{ob}}, \theta)$. Note that although this result obviously assumes θ is univariate, it can also be applied when only the parameter of interest is univariate, if we fix the nuisance parameter part in θ_0 at its observed-data MLE under the null.

For the exponential tilting linkage model, one can verify that

$$\begin{aligned} \mathcal{B}I_0 &= \frac{W^2}{W^2 + \text{Var}(Z|\text{data}, H_0)} \\ (27) \quad &= 1 - \frac{\text{Var}(Z|\text{data}, H_0)}{W^2 + \text{Var}(Z|\text{data}, H_0)}, \end{aligned}$$

where $W = E(Z|\text{data}, H_0)$, and Z is given in (5). Therefore its computation is straightforward because it only depends on the test statistic and the null hypothesis. Note also that the expectation of the denominator in (27) under the null is simply $\text{Var}(Z|H_0) = 1$. Therefore, if we replace the denominator in (27) by its expected value under the null, we obtain an even simpler approximation $\mathcal{B}I_0 \approx 1 - \text{Var}(Z|\text{data}, H_0)$.

However, $\mathcal{B}I_0$ measures only the relative information in the neighborhood of θ_0 . For example, suppose the data consist of one affected sib-pair like in Figure 1 such that both parents and the sibs are heterozygous with the same pair of alleles at a specific locus (i.e., all individuals have the alleles ‘‘A1’’ and ‘‘A2’’). In this case, the observed-data likelihood from the exponential tilting model is very informative away from θ_0 (see Figure 4), but $\mathcal{B}I_0 = 0$ because the null value $\theta_0 = 0$ turns out to be the *minimizer* of the observed-data likelihood.

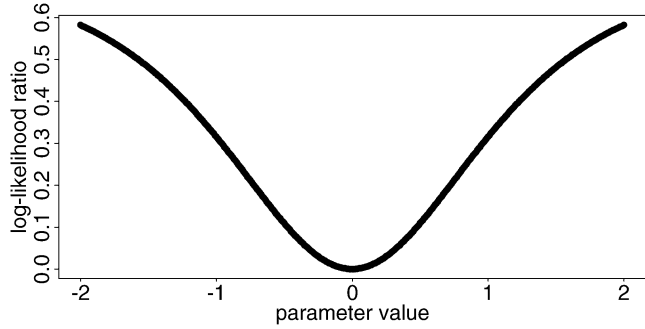


FIG. 4. Log-likelihood ratio for a sib-pair where the parents and sibs are IBS for a heterozygous genotype.

In general, whenever θ_0 is a stationary point of $\ell(\theta|Y_{\text{ob}})$, $\mathcal{B}I_0 = 0$, even if there is almost perfect information. For example, if the data consist of $2n + 1$ sib-pairs such that there is complete information on $2n$ sib-pairs, n sharing 0 alleles IBD and n sharing 2 alleles IBD, and one sib-pair has no information, then $W = 0$ and thus $\mathcal{B}I_0 = 0$. This is clearly a misleading measure. In the next section we propose a remedy for this problem.

5.3 Combining Individual Information Measures

The measures defined in Section 5.1 are inherently small-sample quantities, for the variance terms used in these measures do not naturally admit additivity even for i.i.d. data structures. Whether one can find a satisfying small-sample measure that would automatically admit such additivity is a topic of both theoretical and practical interest, but for our current purposes we can impose such additivity by defining global measures via appropriate combining rules, such as (19). We adopt such rules mainly to maintain the continuity of moving from small-sample to large-sample measures as proposed in Section 4. Whether these are the most sensible rules is a topic that requires further research.

Specifically, suppose our data consist of n independent “small units” (e.g., individual families), $Y_{\text{ob}}^{(i)}$. We apply (24) to each unit and then combine them via the harmonic rule (19) but with weights proportional to $V_i \equiv \text{Var}[\text{LR}(\theta_0, \theta|Y_{\text{ob}}^{(i)})|Y_{\text{ob}}^{(i)}]$. In other words, we define the measure for the aggregated data by first summing up both the numerators and denominators of individual $\mathcal{B}I_{1,i}^\pi$ and then taking the ratio. That is,

$$\begin{aligned} \mathcal{B}I_1^\pi &= \frac{\sum_{i=1}^n \text{Var}[\text{LR}(\theta_0, \theta|Y_{\text{ob}}^{(i)})|Y_{\text{ob}}^{(i)}]}{\sum_{i=1}^n \text{Var}[\text{LR}(\theta_0, \theta|Y_{\text{co}}^{(i)})|Y_{\text{ob}}^{(i)}]} \\ (28) \quad &= \left\{ \frac{\sum_{i=1}^n V_i [\mathcal{B}I_{1,i}^\pi]^{-1}}{\sum_{i=1}^n V_i} \right\}^{-1}. \end{aligned}$$

Similarly, we can define the combined version for $\mathcal{B}I_2^\pi$ from individual $\mathcal{B}I_{2,i}^\pi$, and we can also use the arithmetic combining rule (20). In addition, its limit under the shrinking prior, in analogy to (26), can be expressed as

$$\begin{aligned} \mathcal{B}I_s &= \frac{\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)})}{\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)}) + \sum_{i=1}^n I_{\text{mi}}(\theta_0|Y_{\text{ob}}^{(i)})} \\ (29) \quad &= \frac{\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)})}{\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)}) + I_{\text{mi}}(\theta_0|Y_{\text{ob}})}, \end{aligned}$$

where $I_{\text{mi}}(\theta|Y_{\text{ob}})$ is the expected Fisher information matrix from $f(Y_{\text{co}}|Y_{\text{ob}}, \theta)$, with $Y_{\text{ob}} = \{Y_{\text{ob}}^{(1)}, \dots, Y_{\text{ob}}^{(n)}\}$. We have changed the notation from $\mathcal{B}I_0$ to $\mathcal{B}I_s$ to signify the fact that the latter measure is defined by *summing* up the numerators and denominators of the individual $\mathcal{B}I_0$'s *separately* before forming the combined ratio. The second equation in (29) holds because of the additivity of Fisher information for independent data structures. For the exponential tilting linkage model, this averaging for a shrinking prior leads to

$$\begin{aligned} \mathcal{B}I_s &= \frac{\sum_{i=1}^n W_i^2}{\sum_{i=1}^n W_i^2 + \sum_{i=1}^n \text{Var}(Z_i|\text{data}, H_0)} \\ &= \frac{\sum_{i=1}^n W_i^2/n}{\sum_{i=1}^n W_i^2/n + \text{Var}(Z|\text{data}, H_0)}, \end{aligned}$$

where $W_i = E(Z_i|\text{data}, H_0)$ and $Z = \sum_{i=1}^n Z_i/\sqrt{n}$. This is equal to zero only if all the W_i 's are equal to zero, as opposed to using a global posterior, that is, by applying (26) directly to the whole data set, where $\sum W_i = 0$ is sufficient to cause $\mathcal{B}I_0 = 0$. This difference is an important advantage for $\mathcal{B}I_s$, as we will demonstrate in Section 6.3.

5.4 An Empirical Comparison

To illustrate the proposed Bayesian measures of information, we calculated them for various priors π in a data set containing 21 ulcerative colitis (UC) families (Cho et al., 1998). The choices of priors here were made for investigating the sensitivity to prior specification, so they may not reflect our real knowledge about the problem (e.g., we generally expect θ to be nonnegative in such problems). In Figure 5 the measure of information $\mathcal{B}I_s$ is plotted in comparison with $\mathcal{B}I_2^\pi$, which is calculated as described in the previous section for three different priors. Similar results are obtained using $\mathcal{B}I_1^\pi$. In this example $\mathcal{R}I_1$ and $\mathcal{B}I_s$ are almost identical; $\mathcal{R}I_1$ is therefore not shown. Note that the value of the parameter under the null hypothesis of no

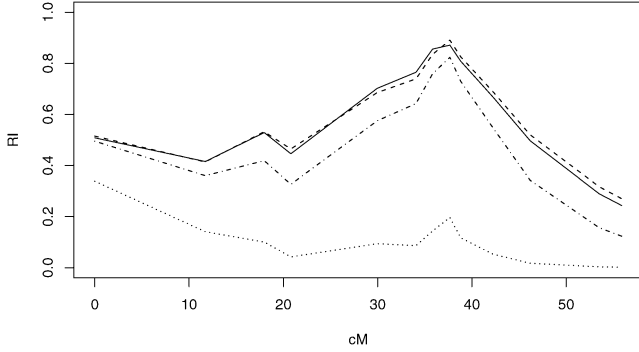


FIG. 5. The Bayesian measures are calculated for a data set containing 21 families. The solid line is $\mathcal{B}I_S$; the dashed line corresponds to $\mathcal{B}I_2^\pi$ calculated using a uniform prior on $(-1, 1)$; the dot-dashed line corresponds to $\mathcal{B}I_2^\pi$ calculated using a uniform prior on $(\min(\theta_{\text{ob}}, \theta_0) - 0.1, \max(\theta_{\text{ob}}, \theta_0) + 0.1)$; the dotted line corresponds to $\mathcal{B}I_2^\pi$ calculated using a uniform prior on $(\theta_{\text{ob}} - 0.1, \theta_{\text{ob}} + 0.1)$.

linkage is equal to zero, and, for this data set, the maximum likelihood estimates for the linkage parameter across the chromosome vary between -0.74 and 0.07 .

We note that the $\mathcal{B}I_2^\pi$ measure calculated using a $\text{Uniform}(-1, 1)$ prior is very close to $\mathcal{B}I_S$, which demonstrates the possibility of having very different priors that result in very similar measures. The Bayesian measure calculated with a prior having a narrower support, that is, uniform on the interval $(\min(\theta_{\text{ob}}, \theta_0) - 0.1, \max(\theta_{\text{ob}}, \theta_0) + 0.1)$, follows the same patterns but is uniformly smaller. Using a prior centered around the maximum likelihood estimate, uniform on the interval $(\theta_{\text{ob}} - 0.1, \theta_{\text{ob}} + 0.1)$, turns out to be very misleading because it gives values that are considerably too small (i.e., in comparison with the large-sample estimates given in Figure 2). We emphasize that symmetric uniform priors were used in Figure 5 simply to demonstrate potential substantial sensitivity to prior specification, as one often expects less erratic behavior from such symmetric and smooth prior specifications. The issue of sensitivity to the choice of prior is further discussed in Section 7.

6. THEORETICAL CONNECTIONS, COMPARISONS AND CURIOSITIES

6.1 The Asymptotic Equivalence to the Estimation Measure

As we discussed previously, a central difficulty in measuring the relative amount of information is that its value will generally depend on the true value of the unknown parameter. One way to explore this dependence is to replace θ_{ob} in the definition of $\mathcal{R}I_0$ or $\mathcal{R}I_1$ by θ in

a suitably defined neighborhood, and to plot it against θ in such a range to check its variability. The use of this type of *relative information function* was proposed in Meng and van Dyk (1996) for the purpose of measuring the rate of convergence of EM-type algorithms, where the function

$$(30) \quad \mathcal{R}I(\theta) = \frac{\ell_{\text{ob}}(\theta_{\text{ob}}) - \ell_{\text{ob}}(\theta)}{Q(\theta_{\text{ob}}|\theta_{\text{ob}}) - Q(\theta|\theta_{\text{ob}})}$$

was termed *relative augmentation function*. Note that $\mathcal{R}I_1$ is simply the value of this function at $\theta = \theta_0$. For simplicity of presentation, we will assume in the following and Section 6.2 that θ is univariate, though all the results are generalizable to multivariate θ by employing appropriate matrix notation and operations. We also assume all the regularity conditions as in Dempster, Laird and Rubin (1977) to guarantee the validity of taking differentiation under integration and for Taylor expansions.

It was shown in Meng and van Dyk (1996) that as $\theta \rightarrow \theta_{\text{ob}}$, $\mathcal{R}I(\theta)$ approaches the so-called *fraction of observed information* for the purpose of estimation:

$$(31) \quad \mathcal{R}I_E = \frac{I_{\text{ob}}}{I_{\text{co}}} \equiv \frac{I_{\text{ob}}}{I_{\text{ob}} + I_{\text{mi}}},$$

where the observed, complete and missing Fisher information are defined, as in Dempster, Laird and Rubin (1977),

$$(32) \quad I_{\text{ob}} \equiv I_{\text{ob}}(\theta_{\text{ob}}) = - \left. \frac{\partial^2 \log f(Y_{\text{ob}}|\theta)}{\partial \theta^2} \right|_{\theta=\theta_{\text{ob}}},$$

$$(33) \quad \begin{aligned} I_{\text{mi}} &\equiv I_{\text{mi}}(\theta_{\text{ob}}) \\ &= \text{E} \left[- \left. \frac{\partial^2 \log f(Y_{\text{co}}|Y_{\text{ob}}; \theta)}{\partial \theta^2} \right| Y_{\text{ob}}; \theta \right] \Big|_{\theta=\theta_{\text{ob}}} \end{aligned}$$

and

$$(34) \quad \begin{aligned} I_{\text{co}} &\equiv I_{\text{co}}(\theta_{\text{ob}}) \\ &= \text{E} \left[- \left. \frac{\partial^2 \log f(Y_{\text{co}}|\theta)}{\partial \theta^2} \right| Y_{\text{ob}}; \theta \right] \Big|_{\theta=\theta_{\text{ob}}} \end{aligned}$$

$$= I_{\text{ob}} + I_{\text{mi}},$$

where the last identity is known as the “missing-data principle,” and is a directed consequence of (15). The $\mathcal{R}I_E$ measure plays a key role in determining the rate of convergence of the EM algorithm and its various extensions (e.g., Dempster, Laird and Rubin, 1977; Meng and Rubin, 1991, 1993; Meng, 1994; Meng and van Dyk, 1997).

The above limiting result suggests that, when $\delta = \theta_0 - \theta_{\text{ob}}$ is small, we can study the behavior of $\mathcal{R}I_1$

via its connection to $\mathcal{R}I_E$, as we demonstrate in the next section. However, among all the measures we proposed, the measure $\mathcal{B}I_s$ of (29) most closely resembles $\mathcal{R}I_E$ of (31). The main differences are the use of $\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)})$ in place of I_{ob} , and the fact that the Fisher information terms in $\mathcal{R}I_E$ are evaluated at $\theta = \theta_{\text{ob}}$, whereas for $\mathcal{B}I_s$ they are evaluated at $\theta = \theta_0$. It is well known that, under regularity conditions, $\sum_{i=1}^n S^2(\theta_0|Y_{\text{ob}}^{(i)})/n$ will converge to the expected Fisher information under the null. Consequently, under the null, $\mathcal{B}I_s$ and $\mathcal{R}I_E$ are asymptotically equivalent. This equivalence may suggest to directly define $\mathcal{B}I_s$ in terms of the ‘‘observed Fisher information at θ_0 .’’ However, although $I_{\text{ob}} \equiv I_{\text{ob}}(\theta_{\text{ob}})$ is guaranteed to be nonnegative (definite) when θ_{ob} is in the interior of the parameter space Θ , this is not necessarily true for $I_{\text{ob}}(\theta_0)$. Therefore, for small-sample problems for which the use of I_{ob} is inadequate (e.g., when the MLE θ_{ob} is on the boundary of Θ), the direct substitution of I_{ob} by $I_{\text{ob}}(\theta_0)$ will not lead, in general, to a nonnegative measure. The $\mathcal{B}I_s$ measure circumvents this problem by using the sum of individual squared scores instead of $I_{\text{ob}}(\theta_0)$, which guarantees that the resulting measure is inside the unit interval, and that it is consistent with $\mathcal{R}I_E$ for large samples. Therefore $\mathcal{B}I_s$ can be viewed as a small-sample extension of $\mathcal{R}I_E$ in the neighborhood of the null.

6.2 Finite-Sample Equivalence in the Neighborhood of the Null

For both $\mathcal{R}I_1$ and $\mathcal{R}I_0$, their equivalence to $\mathcal{R}I_E$ in the neighborhood of θ_0 can be established for finite-sample sizes. (Therefore, $\mathcal{R}I_E$ can also be defined as the value of either $\mathcal{R}I_1$ or $\mathcal{R}I_0$ when $\theta_{\text{ob}} = \theta_0$.) Specifically, denote $\ell_{\text{ob}}^{(k)}$ the k th derivative of $\ell_{\text{ob}}(\theta)$ at $\theta = \theta_{\text{ob}}$, and

$$(35) \quad Q_{\text{ob}}^{(i,j)} = \left. \frac{\partial^{i+j} Q(\theta_1|\theta_2)}{\partial \theta_1^i \partial \theta_2^j} \right|_{\theta_1=\theta_2=\theta_{\text{ob}}}.$$

It is proved in Appendix A.2 that

$$(36) \quad \mathcal{R}I_1 = \mathcal{R}I_E + \frac{Q_{\text{ob}}^{(3,0)} \mathcal{R}I_E - \ell_{\text{ob}}^{(3)}}{3I_{\text{co}}} \delta + O(\delta^2).$$

In deriving this result, we have utilized the following well-known identities in the literature of the EM algorithm (e.g., Dempster, Laird and Rubin, 1977; Meng and Rubin, 1991):

$$(37) \quad Q_{\text{ob}}^{(1,0)} = \ell_{\text{ob}}^{(1)} = 0; \quad Q_{\text{ob}}^{(2,0)} = -I_{\text{co}}.$$

Under the assumption that $Q(\theta|\theta_0)$ has a unique maximizer as a function of θ , an assumption that is

easily satisfied in most of the applications when EM is useful, we also prove in Appendix A.2 that

$$(38) \quad \begin{aligned} \mathcal{R}I_0 &= \mathcal{R}I_E \\ &+ (3I_{\text{ob}}(Q_{\text{ob}}^{(3,0)} + Q_{\text{ob}}^{(2,1)}) \\ &- 2\ell_{\text{ob}}^{(3)} - Q_{\text{ob}}^{(3,0)} \mathcal{R}I_E^2)(3I_{\text{co}})^{-1} \delta \\ &+ O(\delta^2). \end{aligned}$$

These expansions are useful for comparing the first-order (in δ) behavior of $\mathcal{R}I_1$ and $\mathcal{R}I_0$. For example, we suspect that, for many applications, $\mathcal{R}I_0$ is a conservative estimate of the actual relative information, where $\mathcal{R}I_1$ is a more accurate measure. One way to validate this or to identify situations where this conjecture is true is to compare the two coefficients of δ and to determine the appropriate conditions for $\mathcal{R}I_0 < \mathcal{R}I_1$ to the first order in the neighborhood of θ_0 (away from the neighborhood the comparison is not very meaningful because $\mathcal{R}I_0$ can be seriously biased). Due to the complex nature of these two coefficients, we only present in the next section a simple example to illustrate the conservatism of $\mathcal{R}I_1$, and leave the general theoretical investigation to subsequent work.

We also remark here that when the true θ is believed to be close to θ_0 , a measure like $\mathcal{R}I_0$ can be used to construct reasonable bounds. For example, we can expect $\min\{\mathcal{R}I_0, \mathcal{R}I_1\}$ to be a reasonable lower bound and $\max\{\mathcal{R}I_0, \mathcal{R}I_1\}$ an upper bound for relative information, or we can use $\mathcal{R}I_{0.5} \equiv \sqrt{\mathcal{R}I_0 \mathcal{R}I_1}$ as a compromise. In future work, we intend to investigate the reliability and applicability of such bounds and compromise. Here we simply note a computational advantage of $\mathcal{R}I_{0.5}$ that follows from

$$(39) \quad \mathcal{R}I_{0.5} = \left[\frac{\max_{\theta} [Q(\theta|\theta_0) - Q(\theta_0|\theta_0)]}{Q(\theta_{\text{ob}}|\theta_{\text{ob}}) - Q(\theta_0|\theta_{\text{ob}})} \right]^{1/2},$$

which avoids entirely the calculation of the observed-data log-likelihood function $\ell_{\text{ob}}(\theta)$, which is often harder to compute than the expected complete-data log-likelihood $Q(\theta|\theta')$. Furthermore, whenever $\mathcal{R}I_1$ and $\mathcal{R}I_0$ are close to each other, as in our real-data examples, $\mathcal{R}I_{0.5}$ will be practically the same as either $\mathcal{R}I_1$ or $\mathcal{R}I_0$.

6.3 An Illustrative Finite-Sample Comparison

Let $Y_{\text{co}} = \{y_1, \dots, y_n\}$ be i.i.d. samples from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown, and the null hypothesis is $H_0: \mu = \mu_0$. Suppose our observed data Y_{ob} is a size- m random sample of Y_{co} , where $0 < m < n$. Then it should be clear that the relative

information is $r = m/n$ by any reasonable argument. Indeed, straightforward calculation shows $\mathcal{R}I_1 = r$ regardless of the actual value of Y_{ob} . However,

$$\begin{aligned} \mathcal{R}I_0 &= \frac{1}{r} \left[1 - \frac{\log(1 + (1 - r^2)(t_0^2/m))}{\log(1 + t_0^2/m)} \right] \\ (40) \quad &= r - \frac{r(1 - r^2)}{2} \frac{t_0^2}{m} + O\left(\left(\frac{t_0^2}{m}\right)^2\right), \end{aligned}$$

where $t_0 = (\bar{y}_m - \mu_0)/\sqrt{\hat{\sigma}_m^2/m}$, which differs from the usual t -statistic (under the null) only due to the use of MLE for σ^2 , $\hat{\sigma}_m^2 = (1 - 1/m)s_m^2$, instead of the sample variance s_m^2 . From (40), it is clear that $\mathcal{R}I_0$ approaches r whenever t_0^2/m is small, which implies that $\mathcal{R}I_0$ will recover (reasonably) the correct information when the null hypothesis is (approximately) correct.

In contrast, for a fixed sample size m , $\mathcal{R}I_0$ approaches zero if $t_0^2 \rightarrow \infty$ because for large t_0^2 , $\mathcal{R}I_0$ behaves like $-r^{-1} \log(1 - r^2)/\log(1 + t_0^2/m)$. The reason is that the larger t_0^2 is, the stronger is the evidence that the null is false, and thus the more conservative we become when we impute $\text{lod}(\mu, \mu_0|Y_{co})$ using $E[\text{lod}(\mu, \mu_0|Y_{co})|Y_{ob}, \mu_0]$. In other words, whereas $\mathcal{R}I_0$ is a good measure of how conservative the inference is, this example demonstrates that measuring conservatism in general is not necessarily the same as measuring the relative information. However, when the true θ is in a reasonable neighborhood of θ_0 , $\mathcal{R}I_0$ can be a valuable measure, especially because it is more robust to the posited alternative model and thus can serve as a useful diagnostic measure complementing $\mathcal{R}I_1$. We also note the potentially different impacts of nuisance parameter on $\mathcal{R}I_0$ and $\mathcal{R}I_1$. When σ^2 is known, $\mathcal{R}I_0 = \mathcal{R}I_1 = r$. However, whereas $\mathcal{R}I_1$ remains the same when σ^2 is unknown, $\mathcal{R}I_0$ is greatly affected.

It is also informative to see how $\mathcal{B}I_0$ of (26) and $\mathcal{B}I_s$ of (29) compare in this simple problem. For reasons discussed previously, we fix here the nuisance parameter σ^2 at its MLE under the null, $\tilde{\sigma}_{ob}^2 = \sum_{i=1}^m (y_i - \mu_0)^2/m$. We therefore effectively have a single-parameter μ , whose score function given a normal sample $\{y_1, \dots, y_m\}$ is $S_m(\mu) = m(\bar{y}_m - \mu)/\sigma^2$ (where σ^2 is treated as known). Using the fact that $I_{mi}(\mu|Y_{ob}) = (n - m)/\sigma^2$, we have from (26), after setting $\sigma^2 = \tilde{\sigma}_{ob}^2$,

$$\begin{aligned} \mathcal{B}I_0 &= \frac{m^2(\bar{y}_m - \mu_0)^2/\tilde{\sigma}_{ob}^4}{m^2(\bar{y}_m - \mu_0)^2/\tilde{\sigma}_{ob}^4 + (n - m)/\tilde{\sigma}_{ob}^2} \\ (41) \quad &= \frac{rt_0^2}{rt_0^2 + (1 - r)(1 + t_0^2/m)}. \end{aligned}$$

It should not be a surprise to see that $\mathcal{B}I_0 = 0$ when $t_0 = 0$, that is, when μ_0 happens to be the MLE of θ , \bar{y}_m , a phenomenon we previously noted in Section 5.2. However, this simple example provides some clues on why this happens.

Recall that $\mathcal{B}I_0$ was derived by assuming that the prior shrinks to the null. This is very strong prior information, and it inevitably influences our measure of the relative information. Consider the situation when $t_0 = 0$, in which case our observed data are completely consistent with our strong prior that $\theta = \theta_0$. In that sense, the information from the observed data is completely useless because it does not provide anything more than we a priori knew (or rather, assumed). Hence it is not a contradiction for $\mathcal{B}I_0$ to declare zero relative information when clearly the relative information in the observed data should be r . It is not a contradiction because $\mathcal{B}I_0$ has incorporated the prior information, whereas $r = m/n$ measures the relative information in the data under our posited model. This argument appears to be further substantiated when we consider the other extreme, namely, when $t_0^2 \rightarrow \infty$. By the same logic, in this case, the observed data are extremely informative as they provide strong evidence to contradict the prior, and the degree of contradiction is such that, even with more data, it is unlikely to be altered. Consequently, one can expect $\mathcal{B}I_0$ to be close to 1, which indeed follows from (41) when m is large because $\mathcal{B}I_0 \rightarrow [1 + (r^{-1} - 1)m^{-1}]^{-1}$ when $t_0^2 \rightarrow \infty$.

The above discussion indicates a potential problem with any Bayesian measure, as it is inevitable that some prior information will “leak” into our measure of relative information in the data alone (for a specified test). When we have reliable prior information, it is a very interesting issue to investigate/debate whether our relative information should include the prior information (e.g., in the extreme case when we know the null is true for certain, the data become irrelevant, and one can always consider we have 100% information). Nevertheless, in cases where the prior is introduced for convenience, as largely the case for our setting, it is desirable to reduce any unintended influence as much as possible. In this regard, it was a pleasant surprise to see that the $\mathcal{B}I_s$ defined in (29) is able to recover the correct answer in this example. Specifically, letting $\sigma^2 = \tilde{\sigma}_{ob}^2$, (29) becomes

$$\begin{aligned} \mathcal{B}I_s &= \frac{\sum_{i=1}^m (y_i - \mu_0)^2/\tilde{\sigma}_{ob}^4}{\sum_{i=1}^m (y_i - \mu_0)^2/\tilde{\sigma}_{ob}^4 + (n - m)/\tilde{\sigma}_{ob}^2} \\ (42) \quad &= \frac{m}{m + (n - m)} = r. \end{aligned}$$

It is curious that $\mathcal{B}I_S$ has this ability of “removing” the impact of prior information that affected $\mathcal{B}I_0$ in this finite-sample setting; how generally this result holds (even approximately) is a topic for future research.

6.4 Connections to the Two CR Information Lower Bounds

Our large-sample measures have interesting connections with classic measures based on Fisher information, as shown in Section 6.1. Are there similar connections for the small-sample Bayesian measures? The Bayesian measures are based on posterior variances of likelihood ratios or their logarithms. It turns out that there are several interesting connections, or at least analogies, in both frequentist and Bayesian literature. In a frequentist setting, just as the well-known Cramér–Rao lower bound provides a finite-sample information bound that is determined by the Fisher information, there is a more general Chapman–Robbins information bound (Chapman and Robbins, 1951) that is based on sampling variance of the *likelihood ratio*. Specifically, let X have a multivariate pdf/pmf $f(X|\theta)$ with θ taking values in some parameter space Θ . For each θ , let $S_\theta = \{x : f(x|\theta) > 0\}$ be the support of $f(X|\theta)$. Suppose $T(X)$ is an unbiased estimator of a real-valued function $\tau(\theta)$. Let

$$\Phi_\theta = \{\phi \in \Theta : \tau(\phi) \neq \tau(\theta) \text{ and } S_\phi \subset S_\theta\}.$$

Then

$$\text{Var}(T(X)|\theta) \geq \sup_{\phi \in \Phi_\theta} \left[\frac{[\tau(\phi) - \tau(\theta)]^2}{\text{Var}(LR(\phi, \theta|X)|\theta)} \right],$$

where $LR(\phi, \theta|X)$ denotes the likelihood ratio function $f(X|\phi)/f(X|\theta)$.

This “second CR” bound is more general than the first one because it requires neither differentiability of $\tau(\theta)$ nor the existence of Fisher information (e.g., as in the case of discrete parameters). It provides an interesting analogy to the proposed Bayesian measures because it is based also on the variability of the likelihood ratio, where ϕ and θ can be arbitrarily apart. The central connection here is that while our large-sample measures have close ties with Fisher information (as detailed in Section 6.1), which is also intimately connected with the “first CR” bound (i.e., Cramér–Rao bound), our small-sample measures are based on variances of likelihood ratio, which is connected with the “second CR” bound. The fact that the second CR bound is more general than the first

CR bound is also consistent with our expectation that our Bayesian measures ultimately should be more general than the likelihood-based large-sample measures, though currently this is still just an expectation, not a realization.

6.5 Connections Between Likelihood Ratio and Bayes Factors

The variances in our Bayesian measures are more general than the one used by the second CR bound because we average over not only the missing data but also the posterior distribution of θ . Examining the posterior distribution of the entire likelihood ratio might seem a case of “using data twice,” but the following several identities suggest that such a practice is natural from the Bayesian point of view (indeed, the use of posterior distribution of the likelihood ratio has been previously advocated by Dempster, 1997).

First, suppose we have a *proper* prior $\pi(\theta)$; then it is easy to verify that

$$\begin{aligned} & E[LR(\theta_0, \theta|Y_{ob})|Y_{ob}] \\ (43) \quad &= \int \frac{f(Y_{ob}|\theta_0)}{f(Y_{ob}|\theta)} \frac{f(Y_{ob}|\theta)\pi(\theta)}{f_\pi(Y_{ob})} d\theta \\ &= \frac{f(Y_{ob}|\theta_0)}{f_\pi(Y_{ob})} \equiv \text{BF}_{ob}, \end{aligned}$$

where $f_\pi(Y_{ob}) = \int f(Y_{ob}|\theta)\pi(\theta) d\theta$. (Note that here we assume θ_0 is fixed at a known value.)

In other words, the posterior mean of our likelihood ratio is simply the well-known Bayes factor for assessing the probability of the model under $\theta = \theta_0$ relative to the model under $\theta \sim \pi(\theta)$. This shows that the Bayes factor is a very natural generalization of likelihood ratio by taking into account our uncertainty in θ while accessing the evidence in the data against the hypothesized null value $\theta = \theta_0$. It also shows that it is quite natural to consider posterior quantification of the likelihood ratio itself. Incidentally, applying identity (43) first with $Y_{ob} = Y_{co}$ and then averaging the resulting identity over the posterior predictive distribution $f(Y_{co}|Y_{ob})$, we also obtain the following intriguing result:

$$\begin{aligned} & E[\text{BF}_{co}|Y_{ob}] = E[LR(\theta_0, \theta|Y_{co})|Y_{ob}] \\ (44) \quad &= E[LR(\theta_0, \theta|Y_{ob})|Y_{ob}] = \text{BF}_{ob}. \end{aligned}$$

In other words, the observed-data Bayes factor BF_{ob} is the posterior average of any of these three quantities: the observed-data likelihood ratio, the complete-data likelihood ratio, or the complete-data Bayes fac-

tor. Identities (23), (43) and (44) together demonstrate the “coherence” of likelihood ratio and Bayes factor as well as between them. Identity (44) also suggests an easy way of computing BF_{ob} via Monte Carlo averaging of complete-data or observed-data likelihood ratios. We note, however, that the posterior distributions of BF_{co} , $\text{LR}(\theta_0, \theta|Y_{\text{co}})$ and $\text{LR}(\theta_0, \theta|Y_{\text{ob}})$ are generally different. In particular, because of (23) and (43), we have that

$$(45) \quad \begin{aligned} & \max\{\text{Var}[\text{BF}_{\text{co}}|Y_{\text{ob}}], \text{Var}[\text{LR}(\theta_0, \theta|Y_{\text{ob}})|Y_{\text{ob}}]\} \\ & \leq \text{Var}[\text{LR}(\theta_0, \theta|Y_{\text{co}})|Y_{\text{ob}}]. \end{aligned}$$

Given the clear interpretation and utility of the posterior mean of the likelihood ratio, we would naturally consider the posterior variance of the likelihood ratio. That is, we can measure the posterior uncertainty in our likelihood ratio evidence. These are exactly the quantities used in defining $\mathcal{B}I_1^\pi$ in (24), where the numerator and denominator are respectively the posterior variances of the observed-data and complete-data likelihood ratios. The following equivalent expression of $\mathcal{B}I_1^\pi$ further demonstrates how $\mathcal{B}I_1^\pi$ measures relative “flatness” in the likelihood ratio surfaces:

$$(46) \quad \mathcal{B}I_1^\pi = \frac{\text{Cov}_\pi[\text{LR}(\theta_0, \theta|Y_{\text{ob}}), \text{LR}(\theta, \theta_0|Y_{\text{ob}})]}{\text{Cov}_{\pi, \theta_0}[\text{LR}(\theta_0, \theta|Y_{\text{co}}), \text{LR}(\theta, \theta_0|Y_{\text{co}})]},$$

where Cov_π is the covariance operator with respect to the prior $\pi(\theta)$, and $\text{Cov}_{\pi, \theta_0}$ is with respect to $f(Y_{\text{co}}|Y_{\text{ob}}, \theta_0)\pi(\theta)$. In other words, the flatness of the likelihood ratio surfaces is measured by the covariance of the likelihood ratio and its reciprocal. Although this expression itself is intuitive because a positive function is flat if and only if it is proportional to its reciprocal, the equivalence between (24) and (46) is a bit curious because (24) is based on *posterior variance* whereas (46) is based on *prior covariance*.

6.6 Connections to Entropy and R^2

It would be a serious oversight if we do not emphasize the connections of the information measures we discuss in this paper to the vast literature on entropy. Indeed, essentially all measures we presented have an entropy flavor, from the large-sample ones based on Kullback–Leibler information to the small-sample ones involving second-order entropy in the form of $\int (\log p(\theta))^2 p(\theta) d\theta$ (see Zellner, 2003). This is very natural given that the entropy is a fundamental type of information measure (e.g., Akaike, 1985). Indeed, much of the classic results on information measure in optimal sequential designs, which our genetic applica-

tions resemble (i.e., as one needs to decide the next step given what has been observed), are based on entropy-like quantities and their generalizations. This includes both Kullback–Leibler information and Chernoff information (Chernoff, 1979). A central difference between that literature and our current proposals is that the existing literature focuses on quantifying the *absolute* amount of information in an experiment/design, whereas our main objective here is to quantify the *relative* amount of information compared to the absolute amount of information that we would have if there were no missing data (e.g., known IBD sharing in linkage studies). Furthermore, we investigate two sets of relative information, depending on whether we can assume the true parameter is in a neighborhood of the null or not. To the best of our knowledge, our study is the first serious investigation of the roles of null and alternative hypotheses in measuring relative information.

Because our Bayesian measures $\mathcal{B}I_1^\pi$ and $\mathcal{B}I_2^\pi$ are defined as ratios of variances, it is also important to emphasize their connections to the regression R^2 and to other measures of association/correlation such as the linkage disequilibrium measure r^2 (e.g., Devlin and Risch, 1995). These measures are related to Fisher information and can also be used to estimate relative information. The main differences are that ours are defined via the *posterior variability* of the *whole likelihood ratio or log-likelihood ratio*, instead of *sampling variances of individual statistics or variables*. More details on measures of association/correlation used to quantify relative information can be found elsewhere (Nicolae, 2006b).

7. LIMITATIONS AND FURTHER WORK

7.1 Further Theoretical and Methodological Work

Clearly much remains to be done, especially for the small-sample problems. With large samples, we believe the measures we proposed, especially $\mathcal{R}I_1$, satisfy essentially all five criteria as discussed in Section 1.2. For small samples, the various Bayesian measures we proposed, while all satisfy the second criterion, have pros and cons regarding the rest of the criteria. The most pronounced problem, of course, is the choice of a general-purpose “default prior.” Here we emphasize that the desire for “general purpose” is motivated by the observation that in many applications the investigators need to compute the information

measures for many data sets (e.g., different families or pedigrees and different loci in linkage analysis; different tests for different haplotype models in the association studies) under time constraints. Therefore it is typically not feasible to construct specific priors for each data set at hand, nor is it desirable given that the purpose of hypothesis testing, in the genetic applications we are interested in, has more of a screening nature. A requirement for constructing problem-specific priors would be typically viewed as too much of a burden to be practically appealing. On the other hand, standard recipes for constructing “default” priors do not seem to be generally applicable either. For example, the use of Jeffreys’ prior is typically out of the question because the calculation of the expected Fisher information requires us to specify a reliable distribution over the state space of Y_{ob} for arbitrary value of θ , which is typically very hard, if not impossible, to do. Furthermore, the properties of Jeffreys’ prior are not clear when we try to avoid the use of Fisher information in the first place.

Second, whereas $\mathcal{B}I_s$ provides a nice connection between small-sample and large-sample measures in the neighborhood of θ_0 , we currently do not have such a measure when the null is far from the truth. This is of great theoretical and practical concern, at least in the context of genetic studies, because the regions where there is strong evidence against the null are precisely the regions we try to identify. One possible strategy is to start by estimating θ based on the aggregated data (e.g., using data from the other families), and then use a prior that shrinks toward this estimated θ when computing information measure for individual components (e.g., families). In future work we plan to evaluate this strategy, as a part of the general investigation of the sensitivity of our Bayesian measures to prior specifications once we move out the neighborhood of the null.

Third, even for large samples, our measures $\mathcal{R}I_0$ and $\mathcal{R}I_1$ can be sensitive to the posited linkage or association model, which may or may not capture the real biological process that leads to the linkage or association. This would be particularly true for $\mathcal{R}I_1$, which relies more heavily on the model associated with the test than $\mathcal{R}I_0$. Although such sensitivity is inevitable because without a specific alternative model the very notion of relative information may not even be defined, as we emphasized previously, it is important to understand to what degree our information measures can change with our working model. Both theoretical and empirical investigations are needed, especially for classes of problems that are common in practice. Also

needed are investigations of the impact of nuisance parameters on these measures. The haplotype association examples involve nuisance parameters, for example, population genotype risks or population haplotype frequencies, and $\mathcal{R}I_1$ seems to work adequately in practice. Nevertheless, it would be interesting to see if further refinements are possible. The illustrative example of Section 6.3 strongly suggests that further research is necessary to investigate the possible complications caused by the nuisance parameters, especially for $\mathcal{R}I_0$.

7.2 Other Applications

The genetic applications presented in this paper focus on the allele-sharing linkage methods and the haplotype-based association studies, but there are many other areas in genetics where measuring relative information is important. For example, in the past years the markers used in genome-wide searches for susceptibility loci were mostly microsatellites. These are markers that have many alleles, and are generally very informative, but are not very common across the genome. Because the applications focused on small regions of the genome, this lack of abundance of the microsatellites has led to the still increasing popularity of the SNPs as genetic markers. The SNPs are not as informative as the microsatellites, but they are highly abundant. Also new technology platforms such as the Affymetrix GeneChip Mapping 10K, 100K and 500K Arrays (Matsuzaki, Loi and Dong, 2004) are available for SNP genotyping, and they come with a substantial reduction in cost. Given that both the microsatellites and the SNPs are currently used in gene-mapping studies, a fundamental and practical question is how many SNPs we need in order to obtain the same amount of information as obtained by using microsatellites. Differences between SNPs and microsatellites have been investigated for linkage (e.g., Kruglyak, 1997; Schaid et al., 2004; Evans and Cardon, 2004; Middleton et al., 2004; Thalamuthu et al., 2005), and measures of relative information extracted have been proposed (Teng and Siegmund, 1998), but the answers to similar questions will be different for different applications. We plan to further explore the use of the proposed measures of information to other problems of this sort. The comparisons between the relative information of sets of SNPs to that of sets of microsatellites (relative to the underlying complete information) will allow us to make sensible comparisons of the maps for a particular study purpose.

The gene-mapping research has focused recently on genome-wide association studies that are thought to

have better power to localize genes contributing more modestly to disease susceptibility. In these studies, new measures are needed for quantifying the loss in information due to untyped SNPs, or even SNPs that have not been discovered. Also, novel tools for measuring information are necessary in choosing a subset of “tagging” SNPs to type for a disease project based on the data from the HAPMAP project (The International HapMap Consortium, 2003).

Other possible applications are in testing for gene-environment interaction. This can be done in both linkage and association studies, and can increase the power of detecting risk factors. In most of these studies, the environmental and the clinical data are also incomplete. A natural question then arises: “what is the most efficient way to allocate the resources: what percentage should be devoted to collect more genetic information and what percentage should be used to collect more covariate information?” The answer depends again on the specific study, and the problem is more complicated because the environmental and clinical information can be subject to much more complicated missing-data patterns, often due to unknown reasons. Research is clearly needed in this direction to explore to what extent it is possible to sensibly measure the relative information for guiding the allocation of resources, and we hope the general framework we set up in this paper provides a starting point, if not a solution.

APPENDIX

A.1 Proof for Section 5.2

In order to prove the shrinking prior limit results in Section 5.2, we need the following lemma.

LEMMA A.1. *Let t be a fixed real number, and let a_i and b_i , $i = 1, 2, 3, 4$, be real continuous functions defined on an open interval containing t , such that a_i and b_i are three times differentiable in a neighborhood of t . Let $\tilde{a}_i(\delta; t) = \int_{t-\delta}^{t+\delta} a_i(x) dx$, and similarly for $\tilde{b}_i(\delta; t)$, where $i = 1, 2, 3, 4$. If*

$$(47) \quad \begin{aligned} a_1(t)a_2(t) &= a_3(t)a_4(t), \\ b_1(t)b_2(t) &= b_3(t)b_4(t), \end{aligned}$$

but

$$(48) \quad \begin{aligned} b_1''(t)b_2(t) + b_1(t)b_2''(t) \\ - b_3''(t)b_4(t) - b_3(t)b_4''(t) \neq 0, \end{aligned}$$

then

$$\begin{aligned} \lim_{\delta \rightarrow 0} \frac{\tilde{a}_1(\delta; t)\tilde{a}_2(\delta; t) - \tilde{a}_3(\delta; t)\tilde{a}_4(\delta; t)}{\tilde{b}_1(\delta; t)\tilde{b}_2(\delta; t) - \tilde{b}_3(\delta; t)\tilde{b}_4(\delta; t)} \\ = (a_1''(t)a_2(t) + a_1(t)a_2''(t) \\ - a_3''(t)a_4(t) - a_3(t)a_4''(t)) \\ \cdot (b_1''(t)b_2(t) + b_1(t)b_2''(t) \\ - b_3''(t)b_4(t) - b_3(t)b_4''(t))^{-1}. \end{aligned}$$

PROOF. The proof follows from the simple Taylor expansion

$$\tilde{a}_i(\delta; t) = 2a_i(t)\delta + \frac{1}{3}a_i''(t)\delta^3 + O(\delta^5),$$

and conditions (47) and (48). \square

PROPOSITION A.1. *Let π be $U(\theta_0 - \delta, \theta_0 + \delta)$. Then*

$$(49) \quad \lim_{\delta \rightarrow 0} \mathcal{B}I_k^\pi = \frac{S^2(\theta_0|Y_{\text{ob}})}{S^2(\theta_0|Y_{\text{ob}}) + I_{\text{mi}}(\theta_0|Y_{\text{ob}})}, \quad k = 1, 2.$$

PROOF. Let $a_1(\theta) \equiv b_1(\theta) = \exp[\text{lod}(\theta, \theta_0|Y_{\text{ob}})]$, $b_2(\theta) = E[\exp[\text{lod}(\theta, \theta|Y_{\text{co}})]|Y_{\text{ob}}, \theta_0]$ and $a_2(\theta) = a_1^{-1}(\theta)$. Then, as in (46), it is straightforward to verify that

$$(50) \quad \mathcal{B}I_1^\pi = \frac{\int a_1(\theta)\pi(\theta) d\theta \int a_2(\theta)\pi(\theta) d\theta - 1}{\int b_1(\theta)\pi(\theta) d\theta \int b_2(\theta)\pi(\theta) d\theta - 1}.$$

We can then apply Lemma A.1 with $a_3 = a_4 = b_3 = b_4 \equiv 1$. The result for $k = 1$ in (49) then follows because

$$\begin{aligned} a_1''(\theta_0) &= \ell''(\theta_0|Y_{\text{ob}}) + S^2(\theta_0|Y_{\text{ob}}), \\ a_2''(\theta_0) &= -\ell''(\theta_0|Y_{\text{ob}}) + S^2(\theta_0|Y_{\text{ob}}) \end{aligned}$$

and

$$\begin{aligned} b_2''(\theta_0) &= E[-\ell''(\theta_0|Y_{\text{co}}) + \ell^2(\theta_0|Y_{\text{co}})|Y_{\text{ob}}, \theta_0] \\ &= 2I_{\text{mi}}(\theta_0|Y_{\text{ob}}) - \ell''(\theta_0|Y_{\text{ob}}) + S^2(\theta_0|Y_{\text{ob}}). \end{aligned}$$

Note that condition (47) holds because $a_i(\theta_0) = b_i(\theta_0) = 1$ for all i .

For $k = 2$, the limit can be calculated by observing that

$$\begin{aligned} \mathcal{B}I_2^\pi &= \left(1 + \text{Var} \left[\log \frac{P(Y_{\text{co}}|Y_{\text{ob}}, \theta)}{P(Y_{\text{co}}|Y_{\text{ob}}, \theta_0)} \middle| Y_{\text{ob}} \right] \right. \\ &\quad \left. / \text{Var}[\text{lod}(\theta, \theta_0|Y_{\text{ob}})|Y_{\text{ob}}] \right)^{-1} \end{aligned}$$

and then calculating the limit of the ratio in the denominator. A little algebra shows that this ratio can be expressed as

$$(51) \quad \left(\int a_1(\theta)\pi(\theta) d\theta \int a_2(\theta)\pi(\theta) d\theta - \left[\int a_3(\theta)\pi(\theta) d\theta \right]^2 \right) \cdot \left(\int b_1(\theta)\pi(\theta) d\theta \int b_2(\theta)\pi(\theta) d\theta - \left[\int b_3(\theta)\pi(\theta) d\theta \right]^2 \right)^{-1},$$

where $a_1(\theta) = b_1(\theta)$ are the same as in (50), but

$$a_2(\theta) = E[(\text{lod}(\theta, \theta_0|Y_{\text{co}}) - \text{lod}(\theta, \theta_0|Y_{\text{ob}}))^2 \cdot \exp(\text{lod}(\theta, \theta_0|Y_{\text{co}})) | Y_{\text{ob}}, \theta_0],$$

$$a_3(\theta) = E[(\text{lod}(\theta, \theta_0|Y_{\text{co}}) - \text{lod}(\theta, \theta_0|Y_{\text{ob}})) \cdot \exp(\text{lod}(\theta, \theta_0|Y_{\text{co}})) | Y_{\text{ob}}, \theta_0],$$

$$b_2(\theta) = \text{lod}^2(\theta, \theta_0|Y_{\text{ob}})a_1(\theta) \quad \text{and}$$

$$b_3(\theta) = \text{lod}(\theta, \theta_0|Y_{\text{ob}})a_1(\theta).$$

To apply Lemma A.1, we let $a_4 = a_3$ and $b_4 = b_3$. Noting that $a_i(\theta_0) = b_i(\theta_0) = 0$ for all $i = 2, 3, 4$ [and hence condition (47) holds], we only need to compute $a_2''(\theta_0)$ and $b_2''(\theta_0)$ in order to obtain the limit. This calculation is facilitated by the formula

$$\begin{aligned} & \frac{d^2}{dx^2}[g^2(x) \exp(f(x))] \\ &= 2g'^2 \exp(f) + 2gg'' \exp(f) + 4gg'f' \exp(f) \\ & \quad + g^2 f'' \exp(f) + g^2 f'^2 \exp(f). \end{aligned}$$

The result then follows because

$$b_2''(\theta_0) = 2\ell'^2(\theta_0|Y_{\text{ob}}) = 2S^2(\theta_0|Y_{\text{ob}})$$

and

$$\begin{aligned} a_2''(\theta_0) &= 2E[(S(\theta_0|Y_{\text{co}}) - S(\theta_0|Y_{\text{ob}}))^2 | Y_{\text{ob}}, \theta_0] \\ &\equiv 2I_{\text{mi}}(\theta_0|Y_{\text{ob}}). \quad \square \end{aligned}$$

A.2 Derivations for Section 6.2

The derivations are based on the following lemma, which is trivial to verify using the Taylor expansion.

LEMMA A.2. *Let f and g be continuous functions defined on an open interval containing zero, such that $f(\delta) = a_1 + a_2\delta + O(\delta^2)$ and $g(\delta) = b_1 + b_2\delta + O(\delta^2)$ as $\delta \rightarrow 0$. Then*

$$\frac{f(\delta)}{g(\delta)} = \frac{a_1}{b_1} + \frac{a_2 - b_2(a_1/b_1)}{b_1} \delta + O(\delta^2).$$

As in Section 5, we let $\delta = \theta_0 - \theta_{\text{ob}}$. For $\mathcal{R}I_1$, we need to expand both $\ell_{\text{ob}}(\theta_0)$ and $Q(\theta_0|\theta_{\text{ob}})$, as functions of δ . Using the notation given in Section 6.2 and (37), we have

$$(52) \quad \ell_{\text{ob}}(\theta_0) - \ell_{\text{ob}}(\theta_{\text{ob}}) = -\frac{I_{\text{ob}}}{2}\delta^2 + \frac{\ell_{\text{ob}}^{(3)}}{6}\delta^3 + O(\delta^4)$$

and

$$(53) \quad \begin{aligned} & Q(\theta_0|\theta_{\text{ob}}) - Q(\theta_{\text{ob}}|\theta_{\text{ob}}) \\ &= -\frac{I_{\text{co}}}{2}\delta^2 + \frac{Q_{\text{ob}}^{(3,0)}}{6}\delta^3 + O(\delta^4). \end{aligned}$$

Expansion (36) then follows directly from Lemma A.2.

To establish a similar expansion for $\mathcal{R}I_0$, let θ_Q be the maximizer of $Q(\theta|\theta_0)$; recall we assume that θ_Q is unique. Then

$$(54) \quad \mathcal{R}I_0 = \frac{Q(\theta_Q|\theta_0) - Q(\theta_0|\theta_0)}{\ell_{\text{ob}}(\theta_{\text{ob}}) - \ell_{\text{ob}}(\theta_0)}.$$

However, even when $\delta = \theta_0 - \theta_{\text{ob}}$ is small, it is not immediate that θ_Q would be close to θ_{ob} as well. We now show that when δ is small enough, $Q^{(1,0)}(\theta_0|\theta_0)$ and $Q^{(1,0)}(\theta_{\text{ob}}|\theta_0)$ have opposite signs. Consequently, θ_Q , the unique solution of $Q^{(1,0)}(\theta|\theta_0) = 0$, must be between θ_0 and θ_{ob} , and hence $|\theta_Q - \theta_{\text{ob}}| \leq |\delta|$.

To see this, we first expand $g(\theta) \equiv Q^{(1,0)}(\theta|\theta)$ around $g(\theta_{\text{ob}})$ to obtain

$$(55) \quad \begin{aligned} g(\theta_0) - g(\theta_{\text{ob}}) &= g^{(1)}(\theta_{\text{ob}})\delta + O(\delta^2) \\ &= [Q_{\text{ob}}^{(2,0)} + Q_{\text{ob}}^{(1,1)}]\delta + O(\delta^2). \end{aligned}$$

But the following general result, proved in Meng (2000):

$$(56) \quad \ell_{\text{ob}}^{(k+1)}(\theta) = \sum_{j=0}^k \binom{j}{k} Q^{(j+1, k-j)}(\theta|\theta) \quad \text{for any } k \geq 0,$$

implies that $g(\theta_{\text{ob}}) = 0$ and $Q_{\text{ob}}^{(2,0)} + Q_{\text{ob}}^{(1,1)} = \ell_{\text{ob}}^{(2)} = -I_{\text{ob}}$. Consequently,

$$(57) \quad Q^{(1,0)}(\theta_0|\theta_0) = -I_{\text{ob}}\delta + O(\delta^2).$$

For $Q^{(1,0)}(\theta_{\text{ob}}|\theta_0)$, using the notation in (15) and (35), we have

$$(58) \quad \begin{aligned} & Q^{(1,0)}(\theta_{\text{ob}}|\theta_0) \\ &= \ell_{\text{ob}}^{(1)}(\theta_{\text{ob}}) + H^{(1,0)}(\theta_{\text{ob}}|\theta_0) \\ &= H^{(2,0)}(\theta_0|\theta_0)(\theta_{\text{ob}} - \theta_0) + O(\delta^2) \\ &= I_{\text{mi}}(\theta_0)\delta + O(\delta^2), \end{aligned}$$

where $I_{mi}(\theta)$ is as defined in (33). Since both I_{ob} and $I_{mi}(\theta_0)$ are positive, we conclude from (57) and (58) that $Q^{(1,0)}(\theta_{ob}|\theta_0)$ and $Q^{(1,0)}(\theta_0|\theta_0)$ have opposite signs when δ is small enough. Therefore we have established that $\theta_Q - \theta_0 = O(\delta)$, and consequently we can express

$$(59) \quad \theta_Q - \theta_0 = B\delta + C\delta^2 + O(\delta^3),$$

where B and C are $O(1)$ as $\delta \rightarrow 0$ and are to be determined.

To determine B and C , we first note that

$$(60) \quad \begin{aligned} Q^{(1,0)}(\theta_0|\theta_0) &= \ell_{ob}^{(1)}(\theta_0) \\ &= -I_{ob}\delta + \frac{\ell_{ob}^{(3)}}{2}\delta^2 + O(\delta^3) \end{aligned}$$

and

$$(61) \quad \begin{aligned} 0 &= Q^{(1,0)}(\theta_Q|\theta_0) \\ &= \ell_{ob}^{(1)}(\theta_0) + G^{(2)}(\theta_0)(\theta_Q - \theta_0) \\ &\quad + \frac{G^{(3)}(\theta_0)}{2}(\theta_Q - \theta_0)^2 + O(\delta^3), \end{aligned}$$

where $G^{(k)}(\theta) \equiv Q^{(k,0)}(\theta|\theta)$. Substituting (59) and (60) into (61) and solving for B and C , we obtain

$$(62) \quad \begin{aligned} B &= \frac{I_{ob}}{G^{(2)}(\theta_0)} \quad \text{and} \\ C &= -\frac{\ell_{ob}^{(3)} + B^2 G^{(3)}(\theta_0)}{2G^{(2)}(\theta_0)}. \end{aligned}$$

Noting that $G^{(1)}(\theta_0) = \ell^{(1)}(\theta_0)$ and (60), we then obtain

$$\begin{aligned} Q(\theta_Q|\theta_0) - Q(\theta_0|\theta_0) &= G^{(1)}(\theta_0)(\theta_Q - \theta_0) + \frac{G^{(2)}(\theta_0)}{2}(\theta_Q - \theta_0)^2 \\ &\quad + \frac{G^{(3)}(\theta_0)}{6}(\theta_Q - \theta_0)^3 + O(\delta^4) \\ &= \left[-I_{ob}B + \frac{1}{2}B^2 G^{(2)}(\theta_0) \right] \delta^2 \\ &\quad + \left[\frac{1}{2}B\ell_{ob}^{(3)} - CI_{ob} \right. \\ &\quad \left. + BCG^{(2)}(\theta_0) + \frac{1}{6}B^3 G^{(3)}(\theta_0) \right] \delta^3 + O(\delta^4) \\ &= -\frac{I_{ob}^2}{2G^{(2)}(\theta_0)}\delta^2 + \left[\frac{\ell_{ob}^{(3)} I_{ob}}{2G^{(2)}(\theta_0)} + \frac{G^{(3)}(\theta_0) I_{ob}^3}{6[G^{(2)}(\theta_0)]^3} \right] \delta^3 \\ &\quad + O(\delta^4). \end{aligned}$$

Combining this expansion with

$$\begin{aligned} G^{(2)}(\theta_0) &= -I_{co} + [Q_{ob}^{(3,0)} + Q_{ob}^{(2,1)}]\delta + O(\delta^2), \\ G^{(3)}(\theta_0) &= Q_{ob}^{(3,0)} + [Q_{ob}^{(4,0)} + Q_{ob}^{(3,1)}]\delta + O(\delta^2) \end{aligned}$$

and applying Lemma A.2, we obtain

$$\begin{aligned} Q(\theta_Q|\theta_0) - Q(\theta_0|\theta_0) &= \frac{I_{ob}}{2} \mathcal{R}I_E \delta^2 \\ &\quad + \frac{1}{2} \left[\mathcal{R}I_E [I_{ob}(Q_{ob}^{(3,0)} + Q_{ob}^{(2,1)}) - \ell_{ob}^{(3)}] \right. \\ &\quad \left. - \frac{Q_{ob}^{(3,0)}}{3} (\mathcal{R}I_E)^3 \right] \delta^3 \\ &\quad + O(\delta^4). \end{aligned}$$

By Lemma A.2, the above equation and (52) together imply that $\mathcal{R}I_0$ of (54) has the expansion (38).

ACKNOWLEDGMENTS

We thank Daniel Gudbjartsson for many helpful discussions and suggestions, and Judy H. Cho for providing the inflammatory bowel disease data. For the diabetes example illustrated in Figure 3, we thank Daniel Gudbjartsson for providing the software that performed likelihood and information calculations, Gubmar Thorleifsson for constructing the figure, and the diabetes research group at Decode Genetics for generating and providing the data. We also want to thank a number of reviewers for very constructive comments and suggestions. This research was supported in part by several National Science Foundation grants (Nicolae and Meng).

REFERENCES

- ABECASIS, G. R., CARDON, L. R. and COOKSON, W. O. C. (2000). A general test of association for quantitative traits in nuclear families. *Amer. J. Human Genetics* **66** 279–292.
- ABREU, P., GREENBERG, D. and HODGE, S. (1999). Direct power comparisons between simple lod scores and npl scores for linkage analysis in complex diseases. *Amer. J. Human Genetics* **65** 847–857.
- AITCHISON, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353
- AKAIKE, H. (1985). Prediction and entropy. In *Celebration of Statistics: The ISI Centenary Volume* 1–24 (A. Atkinson and S. Fienberg, eds.). Springer, New York. MR0816143
- CHAPMAN, D. C. and ROBBINS, H. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* **22** 581–586. MR0044084

- CHERNOFF, H. (1979). *Sequential Analysis and Optimal Design*. SIAM, Philadelphia, PA.
- CHO, J. H., NICOLAE, D. L., GOLD, L. H. and FIELDS, C. T. ET AL. (1998). Identification of novel susceptibility loci for inflammatory bowel disease. *Proc. Natl. Acad. Sci. USA* **95** 7502–7507.
- CLEVES, M. A. and ELSTON, R. C. (1997). Alternative test for linkage between two loci. *Genetic Epidemiology* **14** 117–131.
- COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York. [MR1122806](#)
- COX, D. R. and HINKLEY, D. (1974). *Theoretical Statistics*. Chapman and Hall, London. [MR0370837](#)
- DAW, E. W., THOMPSON, E. A. and WIJSMAN, E. M. (2000). Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology* **19** 366–380.
- DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Statist. Comput.* **7** 247–252.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–37. [MR0501537](#)
- DEVLIN, B. and RISCH, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29** 311–322.
- EVANS, D. E. and CARDON, L. R. (2004). Guidelines for genotyping in genomewide linkage studies: Single-Nucleotide Polymorphism maps versus microsatellite maps. *Amer. J. Human Genetics* **75** 687–692.
- EXCOFFIER, L. and SLATKIN, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12** 921–927.
- FALK, C. T. and RUBINSTEIN, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Human Genetics* **51** 227–233.
- GEORGE, E., FENG, L. and XU, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.* **34** 78–91. [MR2275235](#)
- GRANT, S. F., THORLEIFSSON, G., REYNISDOTTIR, I., BENEDIKTSSON, R., MANOLESCU, A. and SAINZ, J. ET AL. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics* **38** 320–323.
- GREARS, S., THORLEIFSSON, G. and REYNISDOTTIR, S. T. ET AL. (2003). The gene encoding phosphodiesterase 4d confers risk of ischemic stroke. *Nature Genetics* **35** 131–138.
- GUDBJARTSSON, D. F., JONASSON, K., FRIGGE, M. L. and KONG, A. (2000). Allegro, a new program for multipoint linkage analysis. *Nature Genetics* **25** 12–13.
- HAWLEY, M. and KIDD, K. (1995). HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Heredity* **86** 409–411.
- HELGASON, A., PALSSON, S., THORLEIFSSON, G. and GRANT, S. F. ET AL. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature Genetics* **39** 218–225.
- KONG, A. and COX, N. J. (1997). Allele-sharing models: Lod scores and accurate linkage tests. *Amer. J. Human Genetics* **61** 1179–1188.
- KRUGLYAK, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* **17** 21–24.
- KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. and LANDER, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Amer. J. Human Genetics* **58** 1347–1363.
- LAM, J. C., ROEDER, K. and DEVLIN, B. (2000). Haplotype fine mapping by evolutionary trees. *Amer. J. Human Genetics* **66** 659–673.
- LANDER, E. S. and GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84** 2363–2367.
- LANGE, C. and LAIRD, N. M. (2002a). Analytical sample size and power calculations for a general class of family-based association tests: Dichotomous traits. *Amer. J. Human Genetics* **71** 575–584.
- LANGE, C. and LAIRD, N. M. (2002b). On a general class of conditional tests for family-based association studies in genetics: The asymptotic distribution, the conditional power and optimality considerations. *Genetic Epidemiology* **23** 165–180.
- LONG, J. C., WILLIAMS, R. C. and URBANEK, M. (1995). An E–M algorithm and testing strategy for multiple locus haplotypes. *Amer. J. Human Genetics* **59** 799–810.
- MARTIN, E. R., MONKS, S. A., WARREN, L. L. and KAPLAN, N. L. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Amer. J. Human Genetics* **67** 146–154.
- MATSUZAKI, H., LOI, H. and DONG, S. ET AL. (2004). Parallel genotyping of over 10,000 snps using a one-primer assay on a high-density oligonucleotide array. *Genome Research* **14** 414–425.
- MCPEEK, M. S. and STRAHS, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Amer. J. Human Genetics* **65** 858–875.
- MENG, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22** 326–339. [MR1272086](#)
- MENG, X.-L. (2000). Discussion of “Optimization transfer using surrogate objective functions” by K. Lange, D. Hunter and I. Yang. *J. Comput. Graph. Statist.* **9** 35–43. [MR1819865](#)
- MENG, X.-L. (2001). A congenial overview and investigation of multiple imputation inference under uncongentiality. In *Survey Nonresponse* (R. Groves, D. Dillman, J. Eltinge and R. Little, eds.) 343–356. Wiley, New York.
- MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. [MR1243503](#)
- MENG, X.-L. and VAN DYK, D. (1996). Minimum information ratio and relative augmentation function. In *Proceedings of the Statistical Computing Section of the American Statistical Association* 73–78.
- MENG, X.-L. and VAN DYK, D. A. (1997). The EM algorithm— an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567. [MR1452025](#)
- MIDDLETON, F. A. ET AL. (2004). Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide polymorphism (snp) genotyping assay: A comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Amer. J. Human Genetics* **74** 886–897.

- MORRIS, A. P., WHITTAKER, J. C. and BALDING, D. J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Amer. J. Human Genetics* **70** 686–707.
- NICOLAE, D. L. (1999). Allele sharing models in gene mapping: A likelihood approach. Ph.D. thesis, Dept. Statistics, Univ. Chicago.
- NICOLAE, D. L. (2006a). Testing untyped alleles (TUNA)—applications to genome-wide association studies. *Genetic Epidemiology* **30** 718–727.
- NICOLAE, D. L. (2006b). Quantifying the amount of missing information in genetic association studies. *Genetic Epidemiology* **30** 703–717.
- NICOLAE, D. L. and KONG, A. (2004). Measuring the relative information in allele-sharing linkage studies. *Biometrics* **60** 368–275. [MR2066271](#)
- NIU, T., QIN, Z. S., XU, X. and LIU, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Amer. J. Human Genetics* **71** 1242–1247.
- OTT, J. (1991). *Analysis of Human Genetic Linkage*. Johns Hopkins Univ. Press, Baltimore.
- OTT, J. (2001). Major strengths and weaknesses of the lod score method. *Adv. Genetics* **42** 125–132.
- PE’ER, I., DE BAKKER, P. I., MALLER, J., YELENSKY, R., ALTSHULER, D. and DALY, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics* **38** 663–667.
- PRITCHARD, J. K., STEPHENS, M., ROSNBERG, N. A. and DONNELLY, P. (2000). Association mapping in structured populations. *Amer. J. Human Genetics* **67** 170–181.
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. [MR0455196](#)
- SCHAI, D. J., GUENTHER, J. C., CHRISTENSEN, G. B., HEBBRING, S., ROSENOW, C., HILKER, C. A., MCDONNELL, S. K., CUNNINGHAM, J. M., SLAGER, S., BLUTE, M. L. and THIBODEAU, S. N. (2004). Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancersusceptibility loci. *Amer. J. Human Genetics* **75** 948–965.
- SHANNON, C. E. (1949). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 623–656. [MR0026286](#)
- STEPHENS, M., SMITH, N. J. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *Amer. J. Human Genetics* **68** 978–989.
- TENG, J. and SIEGMUND, D. O. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics* **54** 1247–1265. [MR1671602](#)
- TERWILLIGER, J. D. and OTT, J. (1992). A haplotype-based ‘haplotype relative risk’ approach to detecting allelic associations. *Human Heredity* **42** 337–346.
- THALAMUTHU, A., MUKHOPADHYAY, I., RAY, A. and WEEKS, D. E. (2005). A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content. *BMC Genetics* **6** (Suppl 1) S27.
- THE INTERNATIONAL HAPMAP CONSORTIUM (2003). The international hapmap project. *Nature* **426** 789–796.
- WHITEMORE, A. S. and HALPERN, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50** 118–127.
- ZELLNER, A. (2003). Some aspects of the history of Bayesian information processing. Technical report, The Graduate School of Business, Univ. Chicago.
- ZOLLNER, S. and PRITCHARD, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169** 1071–1092.