Large Scale Multinomial Inferences and Its Applications in Genome Wide Association Studies

Chuanhai Liu and Jun Xie

Abstract Statistical analysis of multinomial counts with a large number K of categories and a small number n of sample size is challenging to both frequentist and Bayesian methods and requires thinking about statistical inference at a very fundamental level. Following the framework of Dempster-Shafer theory of belief functions, a probabilistic inferential model is proposed for this "large K and small n" problem. The inferential model produces a probability triplet (p,q,r) for an assertion conditional on observed data. The probabilities p and q are *for* and *against* the truth of the assertion, whereas $r = 1-p-q$ is the remaining probability called the probability of "don't know". The new inference method is applied in a genome-wide association study with very high dimensional count data, to identify association between genetic variants to the disease Rheumatoid Arthritis.

1 Introduction

We consider statistical inference for the comparison of two large-scale multinomial distributions. This problem is motivated by genome-wide association studies (GWAS) with very high dimensional count data, *i.e.*, single nucleotide polymorphism (SNP) data. SNPs are major genetic variants that may associate with common diseases such as cancer and heart disease. A SNP has three possible genotypes, wild type homozygous, heterozygous, and mutation (rare) homozygous. In genome-wide association studies, genotypes of over 500,000 SNPs are typically measured for disease (case) and normal (control) subjects, resulting in a large amount of count data. In most situations, statistical analysis of genome-wide association data has been on a single SNP at a time, using simple logistic regression or χ^2 tests of association for 2×3 contingency tables [19], where we compare differences in genotype frequency of a SNP between cases and controls to identify association with a disease. However, these methods cannot detect associations of combinatorial SNP effects. If we consider a block of SNPs, for example 10 SNPs, it results in $3^{10} = 59,049$ possible genotypes. This number of categories is much larger than a typical study size of a few thousands subjects. Therefore, most categories will have zero or one observa-

Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN 47907, e-mail: chuanhai@purdue.edu, junxie@purdue.edu

tion. The familiar logistic regression and χ^2 tests are not appropriate any more in such a situation.

Statistical analysis of multinomial counts with a large number K of categories and a small number n of sample size is a challenging problem for both frequentist and Bayesian methods. For Bayesian methods, it is known that in this situation Bayesian priors have tremendous effects on the final inferential results, which is known as the non-scalability phenomena of the non-informative and flat prior distributions. Gelman [11] discussed this problem in real applications of a different context. For frequentist methods, χ^2 tests for contingency tables suffer from the problem of small or zero counts. It is not uncommon that frequentist methods are applied to modified contingency tables obtained by either deleting or combining categories with small observed counts (*e.g.*, zeros and ones).

As a mathematical tool for scientific discovery from observed data, prior-free probabilistic inference becomes increasingly important nowadays when scientists are facing large-scale inferential problems, such as the GWAS problem considered in this paper. Earlier attempts of prior-free probabilistic inference include Fisher's fiducial [9, 12, 13, 26, 25], Fraser's structural inference [10], Dempster-Shafer theory of belief functions [6, 21], generalized p-values and generalized confidence regions [22, 24], "objective" Bayesian inference using intentionally non-informative priors, and the algorithmic inference [1, 2, 3]. The algorithmic inference and generalized fiducial appear to be particular interesting because they expand the scope of the applications of fiducial-type inference dramatically by taking the advantage of modern computing power. However, all these methods have not proved to be effective to produce probabilistic inferential results that have the desirable long-run frequency calibration property, or in other words, have the same interpretation from people to people and from experiments to experiments. Most noticeable is that all fiducial-type inference and "objective" Bayes suffer from what is known as *marginalization paradoxes* or more generally *integration paradoxes*, a fundamental problem in fiducial and "objective" Bayesian inference.

The method recently proposed by Balch [4] is also related to but somewhat different from above fiducial-type methods. Without producing posterior-type of probabilities, it creates confidence regions with declared coverage. The problem is that the concept of Neyman-Pearson confidence intervals does not have the desirable probabilistic interpretation. That is, the confidence or coverage of a Neyman-Pearson confidence interval measures the property of confidence intervals rather than providing an assessment of uncertainty in an unknown parameter. For example, a confidence interval does not claim that values in the confidence interval are more plausible or likely than those outside of the interval.

Recently, Martin, Zhang, and Liu [18, 27] proposed what they called weak beliefs or inferential models (IMs) for probabilistic inference, based on an extension of the framework of Dempster-Shafer theory of belief functions [21, 6]. Following Dempster-Shafer and weak beliefs, we view that probabilistic inference for the truth of an assertion, which is denoted as $\mathscr{A}$, amounts to producing a probability p for the truth of $\mathscr{A}$, a probability q against the truth of $\mathscr{A}$, and a residual probability r, called the probability of "don't know". That is, the triplet (p,q,r) is our uncertainty

assessment of $\mathscr{A}$. In addition, unlike fiducial-type and objective Bayesian inference, the inferential model produces the triplet (p,q,r) that has desirable properties in terms of long-run frequency. Compared to the notions of *Belief* and *Plausibility* in belief functions, p equals *Belief* and $p+r$ is *Plausibility* of $\mathscr{A}$. This probability inferential framework is called Inferential Model (IM).

The rest of the paper is arranged as follows. Section 2 provides a brief review of the IM framework formulated recently in Martin and Liu [15]. A generalized inference model for comparing two multinomials is presented in Section 3. The method is applied in a genome-wide association study to identify SNPs that are potentially associated with a given disease in Section 4. Section 5 concludes with a few remarks.

2 Inferential models: an overview

2.1 The basic framework of IMs

Like the familiar Bayesian and frequentist frameworks, the framework of IMs starts with a given sampling model $f(X;\theta)$ for the observed data $X \in \mathbb{X}$ and some unknown parameter $\theta \in \Theta$. We are interested in an assertion of the parameter $A \subseteq \Theta$. IM consists of three basic steps that we introduce in the following subsections.

2.1.1 The A-step

To assess uncertainty of an assertion on θ based on X, we introduce an unobserved random variable $U \in \mathbb{U}$ with known probability measure P_U to associate X and θ. The unobserved random variable U is called the auxiliary a-variable. This association is defined in such a way that the resulting distribution of X given θ is the postulated sampling distribution $f(X;\theta)$. This association can be intuitively understood as a data generative model and can be written as

$$X = a(U,\theta) \qquad (U \sim \mathrm{P}_U). \tag{1}$$

This process is termed as the Association (A)-step. The A-step is illustrated with the following example.

Example 1 (The Binomial model). Suppose that an observed count X is considered to have come from the binomial distribution

$$X \sim \mathrm{Bin}(n,\theta)$$

with known size parameter n but unknown proportion $\theta \in [0,1]$. Denote by

$$\text{pBin}(x;\theta,n)=\sum_{k=0}^{x}\binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad (x=0,1,...,n;\ 0\le\theta\le 1)$$

the cdf of $\text{Bin}(n,\theta)$. In theory, X can be simulated by first drawing U from the standard uniform distribution $\text{Unif}(0,1)$ and then computing X as the U-quantile of $\text{Bin}(n,\theta)$. This leads to the natural association

$$\text{pBin}(X-1;\theta,n)<U\le\text{pBin}(X;\theta,n). \tag{2}$$

■

In general, the A-step defines the following two inferentially relevant subsets. The first is the collection of all possible candidate values of the parameter corresponding to a value of the a-variable U and the observed data X. This subset is obtained by the multi-valued inverse mapping of (1) as

$$\Theta_X(U)=\{\theta:\theta\in\Theta,\ X=a(U,\theta)\}$$

for any value of U and the observed data X. The second subset is the subset in the auxiliary space $\mathbb{U}$

$$\mathbb{U}_A(X)=\{u:\Theta_X(u)\subseteq A\} \tag{3}$$

that supports for the truth of the assertion $A\subseteq\Theta$ after seeing X. In what follows, we assume that this subset is measurable with respect to the probability measure P_U. Thus, $\mathbb{U}_A(X)$ is an event in the auxiliary space, indexed by $X\in\mathbb{X}$. The event $\mathbb{U}_A(X)$ is called the a-event for A given X.

The candidate sets and a-events are illustrated below with the binomial example.

Example 1 (The Binomial model {Cont'd}). Let $\text{pBeta}(.;\alpha,\beta)$ be the cdf of the Beta distribution with shape parameters α and β. Then the cdf $\text{pBin}(x;\theta,n)$ can be expressed as

$$\text{pBin}(x;\theta,n)=\text{pBeta}(1-\theta;n-x,x+1)=1-\text{pBeta}(\theta;x+1,n-x).$$

This gives an alternative representation of the association (2):

$$1-\text{pBeta}(\theta;X,n-X+1)<U\le 1-\text{pBeta}(\theta;X+1,n-X)$$

or simply

$$\text{pBeta}(\theta;X+1,n-X)<U\le\text{pBeta}(\theta;X,n-X+1)$$

because $1-U$ and U have the same distribution $U\sim\text{Unif}(0,1)$. It follows that given $U=u$ and X, the candidate set is a non-singleton interval,

$$\Theta_X(u)=\{\theta:\text{qBeta}(u;X,n-X+1)\le\theta<\text{qBeta}(u;X+1,n-X)\},$$

where qBeta is the inverse of the cdf $\text{pBeta}(.;\alpha,\beta)$, or quantile of the Beta distribution. Since the candidate set $\Theta_X(u)$ for all $u\in[0,1]$ is an interval, the expression for the a-event $\mathbb{U}_A(X)$ can be tedious, depending on the structure of A. When $A=[a,b]$

is a subinterval in $\Theta = [0,1]$, we have

$$\mathbb{U}_A(X) = \mathbb{U}_{[a,b]}(X) = \{u : \text{ qBeta}(u; X, n-X+1) \geq a, \text{qBeta}(u; X+1, n-X) \leq b\}.$$

It should be noted that $\mathbb{U}_{[a,b]}(X) = \emptyset$ if $\text{pBeta}(a; x, n-x+1) > \text{pBeta}(b; x+1, n-x)$. This can happen when a and b are close to each other. In the case where $A = \{\theta_0\}$ is a singleton, we have $\mathbb{U}_A(X) = \emptyset$ and

$$\begin{aligned}\mathbb{U}_{\{\theta_0\}^c}(X) &= \mathbb{U}_{[0,\theta_0)}(X) \cup \mathbb{U}_{(\theta_0,1]}(X) \\ &= [0, \text{pBeta}(\theta_0; X+1, n-X)) \cup (\text{pBeta}(\theta_0; X, n-X+1), 1].\end{aligned}$$

It can be seen that this is consistent with the candidate set $\Theta_X(u)$. That is, the candidate set $\Theta_X(u)$ is the complement of the a-event supporting $\{\theta_0\}^c$ or against $\{\theta_0\}$. ■

2.1.2 The P-step

Valid or meaningful probabilistic inference about the unknown parameter can be generated with the usual probability calculations in the probability space of the a-variable U. The associated a-event $\mathbb{U}_A(X)$ appears to be a natural choice, and is indeed used in almost all existing fiducial-type inferential methods. Care has to be taken, however, to avoid potential selection bias due to the fact that the associated a-event $\mathbb{U}_A(X)$ is effectively chosen by the observed data X. The IM framework is made bias-free by requiring the use of an appropriate predictive random sets (PRS), denoted by $\mathscr{S}$. A PRS $\mathscr{S}$ is defined by (*i*) a pre-specified and possibly assertion-specific collection of measurable subsets, $\mathbb{S}_A$, in the auxiliary space $\mathbb{U}$ and (*ii*) an appropriate probability measure on $\mathbb{S}_A$.

For simplicity without loss of efficiency [15], $\mathbb{S}_A$ is taken to be a nested sequence of subsets in $\mathbb{U}$, i.e., for all two elements S_1 and S_2 in $\mathbb{S}_A$, it holds that $S_1 \subseteq S_2$ or $S_2 \subseteq S_1$. In this case, the needed probability mass can be specified as

$$\text{P}(\mathscr{S} \subseteq S) = \text{P}_U(S) \qquad (S \in \mathbb{S}_A). \tag{4}$$

The elements of $\mathbb{S}_A$ are called focal elements, and the PRS $\mathscr{S}$ is called a *consonant* PRS. In what follows, PRSs are all consonant and with the probability mass defined by (4). The IM framework makes use of PRS $\mathscr{S}$ to predict the unobserved a-variable associated with the observed data X through the association (1). This is referred as the Prediction (P)-step.

Example 2 (A centered PRS). A simple PRS $\mathscr{S}$ for predicting the a-variable U in the binomial example is the centered PRS

$$\mathscr{S} = [U/2, 1-U/2] \qquad (U \sim \text{Unif}(0,1)).$$

■

2.1.3 The C-step

The last of the three steps of IMs combines the prediction of U with the observed data X to compute the bias-adjusted probability, called the *belief* of A given X,

$$\text{Bel}_X(A;\mathscr{S}) = \text{P}(\mathscr{S} \subseteq \mathbb{U}_A(X)) = \text{P}(\Theta_X(\mathscr{S}) \subseteq A) \tag{5}$$

as evidence in the probability scale supporting the truth of the assertion A. Let

$$\Theta_X(\mathscr{S}) = \cup_{u\in\mathscr{S}}\Theta_X(u). \tag{6}$$

Then it is easy to see that

$$\text{Bel}_X(A;\mathscr{S}) = \text{P}(\Theta_X(\mathscr{S}) \subseteq A). \tag{7}$$

Similarly, the belief of the negation of A, denoted by A^c, can be evaluated using the same PRS $\mathscr{S}$ or a different PRS. As in the Dempster-Schafer theory of belief functions, the probability

$$\text{Pl}_X(A;\mathscr{S}) = 1 - \text{Bel}_X(A^c;\mathscr{S})$$

is called the plausibility of A given X. The pair of belief and plausibility probabilities $(\text{Bel}_X(A;\mathscr{S}), \text{Pl}_X(A;\mathscr{S}))$ provides the IM output as the IM uncertainty assessment of the assertion A. These two probabilities $\text{Bel}_X(A;\mathscr{S})$ and $\text{Pl}_X(A;\mathscr{S})$ are also known as lower and upper probabilities.

Realizations of $\mathscr{S}$ may make the subset (6) empty. In the context of Dempster-Shafer, this is known as a conflict case. The IM method called the elastic belief is proposed by Ermini Leaf and Liu [8] for handling this case. The intuition is that PRSs are designed to be efficient by making use of small focal elements as much as possible. Thus, a conflict realization can be replaced by some larger focal element. To maintain the efficiency, any conflict realizations should be replaced by the smallest focal element that is just big enough to give non-empty $\Theta_X(\mathscr{S})$ defined in (6). Since we use consonant PRSs, the needed modification is straightforward. That is, we define

$$\mathscr{S}_X^\star = \cap_{S\in\mathbb{S}_A, \Theta_X(S)\neq\emptyset} S \tag{8}$$

and call it the data-dependent or elastic version of the PRS $\mathscr{S}$. The belief is defined accordingly as

$$\text{Bel}_X(A;\mathscr{S}_X^\star) = \text{P}(\Theta_X(\mathscr{S}_X^\star) \subseteq A). \tag{9}$$

This modification is taken automatically in the IM framework. For convenience, we write $\text{Bel}_X(A;\mathscr{S})$ for $\text{Bel}_X(A;\mathscr{S}_X^\star)$ and say that the PRS $\mathscr{S}$ is elastic or equipped with elasticity.

To summarize, we have the following three-step IM framework to produce probabilities as uncertainty assessment of an assertion $A \subseteq \Theta$.

A-STEP. Associate the observed data $X \in \mathbb{X}$ and the parameter $\theta \in \Theta$ through an association $X = a(U, \theta)$, which produces the sampling distribution $f(x;\theta)$. This association defines the candidates $\mathbb{U}_X(U)$ and the a-event $\mathbb{U}_A(X)$.

P-STEP. Predict U with a consonant and elastic PRS $\mathscr{S}_A$ for inferring A (and another consonant and elastic PRS $\mathscr{S}_{A^c}$ for inferring A^c).

C-STEP. Combine the prediction and the observed data through the association to evaluate the belief and plausibility of A:

$$\mathrm{Bel}_X(A) \equiv \mathrm{Bel}_X(A;\mathscr{S}_A) \text{ and } \mathrm{Pl}_X(A) \equiv \mathrm{Pl}_X(A;\mathscr{S}_{A^c}) = 1 - \mathrm{Bel}_X(A^c;\mathscr{S}_{A^c}).$$

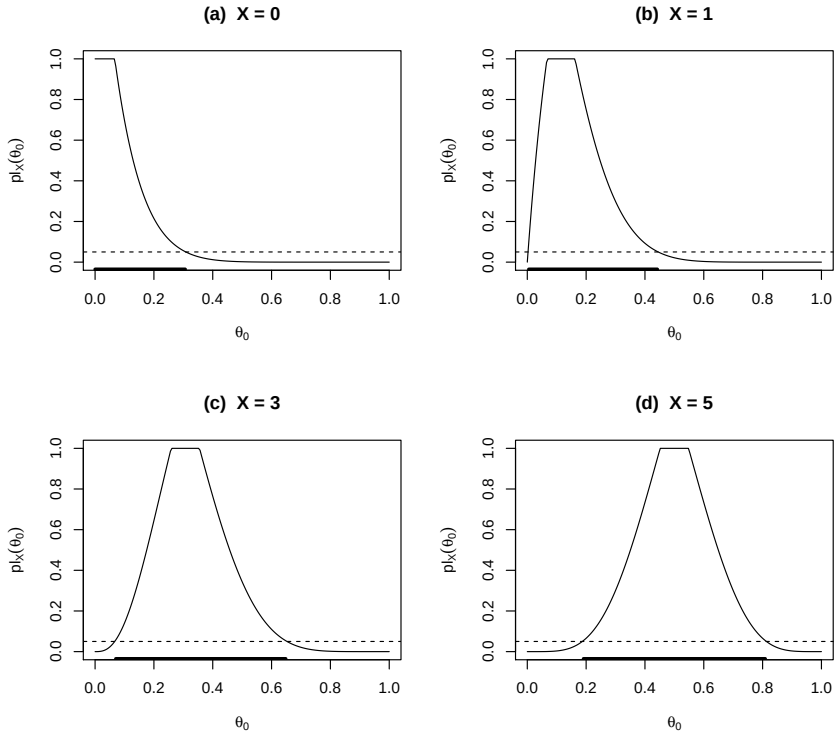

Fig. 1 The plausibility of $\mathscr{A}_{\theta_0} = \{\theta : \theta = \theta_0\}$ given the observed $X = 0, 1, 3, 5$, indicated by the vertical lines, in the binomial example with size $n = 10$.

This is illustrated below with the binomial example.

Example 1 (The Binomial model {Cont'd}). Take the centered PRS and compute the belief and plausibility of all single assertions $\{\theta_0\} \subset [0,1]$ conditional on the observed count X. The belief is zero, and the plausibility can be computed as follows:

$$\begin{aligned}\mathrm{Pl}_X(\{\theta_0\}) &= 1-\mathrm{Bel}_X(\{\theta_0\}^c)\\ &= \begin{cases} 1-2[\mathrm{pBeta}(\theta_0, X+1, n-X)-.5] & \text{if } \mathrm{pBeta}(\theta_0, X+1, n-X) > .5,\\ 1-2[.5-\mathrm{pBeta}(\theta_0, X, n-X+1)] & \text{if } \mathrm{pBeta}(\theta_0, X, n-X+1) < .5,\\ 1 & \text{otherwise}\end{cases}\\ &= 2\min\{1-\mathrm{pBeta}(\theta_0, X+1, n-X), \mathrm{pBeta}(\theta_0, X, n-X+1), .5\}.\end{aligned}$$

The plausibility curves

$$\mathrm{pl}_X(\theta_0) \equiv \mathrm{Pl}_X(\{\theta_0\})$$

for $(n=10, X=0)$, $(n=10, X=1)$, $(n=10, X=3)$, and $(n=10, X=5)$ are shown in Figure 1. The so-called 5% plausibility interval, an IM counterpart of the frequentist 95% confidence interval, is defined as

$$\{\theta : \mathrm{pl}_X(\theta) \geq 0.05\}$$

and shown in Figure 1. We note that constrained problems, such as $\theta \geq 0$, are automatically handled by the use of elastic PRS in the P-step. ■

IM belief and plausibility probabilities are conceptually similar to those in Dempster-Shafer. Using the new notations recommended by Dempster [6], we have

$$p_X(A) = \mathrm{Bel}_X(A),\ q_X(A) = \mathrm{Bel}_X(A^c), \tag{10}$$

and

$$r_X(A) = 1 - p_X(A) - q_X(A),$$

which are referred to as the probabilities for the truth of A, against the truth of A, and of "don't know". In what follows, we also use this probability triplet (p, q, r) in place of the lower and upper probability pair $(\mathrm{Bel}_X(A), \mathrm{Pl}_X(A))$. Unlike Dempster-Shafer and other existing inferential methods such as Fisher's fiducial and "objective" Bayes, IM belief and plausibility probabilities are frequency calibrated. This attractive property along with further developments of IMs is briefly reviewed next.

2.2 Theoretical results and further developments

In scientific inference where experience is converted to knowledge, it is critically important for inferential results to be interpretable from people to people and from experiments to experiments. The IM framework is committed to produce inferential results to have such properties. For this, the concept of validity is defined formally as follows.

Definition 1 (Validity of IMs). The inferential model is valid for assertion A if for every α in $(0,1)$, both

$$P_\theta(\{X : p_X(A) \geq \alpha\}) \leq 1-\alpha \quad \text{and}$$
$$P_\theta(\{X : q_X(A) \geq \alpha\}) \leq 1-\alpha$$

hold respectively for every $\theta \in A^c = \Theta \setminus A$ and for every $\theta \in A$. The probabilities in $P_\theta(\cdot)$ are defined with respect to the random variable X following $f(X;\theta)$.

In other words, validity requires $p_X(A)$ and $q_X(A)$, as functions of the random variable X, to be stochastically bounded by the uniform distribution over the unit interval $(0,1)$ in repeated experiments. Thus, the triplet $(p_X(A), q_X(A), r_X(A))$ provides strength of evidence for both A and A^c in terms of long-run frequency probability. Thresholds for $p_X(A)$ and $q_X(A)$ can be used to confirm the truth and falsity of A. When applying the technique in practices, we report all three probabilities, where a large value of $p_X(A)$ supports A (e.g., the null hypothesis), a large value of $q_X(A)$ supports A^c (e.g., the alternative hypothesis), and a large value of $r_X(A)$ does not support either. For those familiar with the Fisher framework of significance testing, the value $1 - q_X(A)$ is consistent with the notion of p-value. Given a significance level α, when we have

$$p_X(A) + r_X(A) = 1 - q_X(A) < \alpha,$$

the plausibility of the null hypothesis is smaller than the significance level and hence it leads to rejection of $H_0 : \theta \in A$.

Martin and Liu (2012) show that IMs produce valid inferential results. This is due to the use of natural PRSs that satisfy the following credibility.

Definition 2 (Validity of PRS). Let $U \sim \mathrm{P}_U$ with the sample space $\mathbb{U}$. A PRS $\mathscr{S}$ independent of U and with focal elements in $\mathbb{U}$ is said to be valid iff $\mathrm{P}(\mathscr{S} \not\ni U)$, as a function of U, is stochastically smaller than the uniform random variable $\mathrm{Unif}(0,1)$.

Two undergoing further developments of IMs concern efficient inference and applications to challenging statistical problems. This paper provides an example of the latter. The work on efficient IM inference is briefly reviewed here. Martin and Liu [15] discuss the use of assertion-specific PRS for optimal inference. They also argued that the consideration of optimal inference helps resolve the uniqueness issue regarding both the choice of the association for representing the sample model and the choice of the PRS for generating valid inference.

Inference on marginal assertions (i.e., assertions on lower dimensional quantities of the parameters) is both theoretical interesting and practically useful. For example, this is the typical case where "objective" Bayesian and Fisher fiducial are seen to be paradoxical in many models, e.g., the Fieller-Creasy controversy and the Stein paradox in inference on many-normal-means. Constructing assertion-specific PRSs for efficient IM inference is important. An early work on this can be found in Martin and Liu [17].

In the case where the number of observed data points is larger than the number of unknown parameters, efficient inference amounts to combining information. This is discussed in Martin and Liu [15] under the name of conditional IMs (CIMs). The key idea is to condition on components or (data-free) functions of a-variables that are fully observed. This is illustrated by the following simple example concerning the inference about the unit Gaussian mean from a sample.

Example 3 (The unit Gaussian mean). Suppose that $X_1,...,X_n$ form a sample from $N(\theta,1)$. Consider the natural association

$$X_i = \theta + Z_i \qquad (Z_1,..,Z_n \overset{iid}{\sim} N(0,1)).$$

Let $\bar{Z} = n^{-1}\sum_{i=1}^n Z_i$ and let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Then

$$Z_i - \bar{Z} = X_i - \bar{X} \qquad (i = 1,...,n)$$

are fully observed, whereas its component, say,

$$\bar{X} = \theta + \bar{Z}$$

is not. Note that $\{Z_i - \bar{Z} : i = 1,...,n\}$ falls into a $n-1$ dimensional subspace $\mathbb{R}^{n-1}$. Thus valid inference can be made by predicting $\bar{Z}$ conditional on $\{Z_i - \bar{Z} = X_i - \bar{X} : i = 1,...,n\}$, which is equivalent to $\{Z_{i+1} - Z_i = X_{i+1} - X_i : i = 2,...n-1\}$. Routine algebraic operations lead to the combined IM for inference about θ:

$$\bar{X} = \theta + \frac{1}{\sqrt{n}}Z \qquad (Z \sim N(0,1)).$$

■

Martin and Liu [15] show that sufficient statistics can be used as an initial step of CIMs. The argument for the use of sufficient statistics is, however, not the same as that Fisher used (i.e., likelihood based) for establishing the concept of sufficient statistics. It is interesting that it is the observed functions of a-variables that play a fundamental role in the IM framework. This motivated what they call the local CIMs. Local CIMs provide a new way of conducting both valid and efficient probabilistic inference when the dimension of the minimal sufficient statistics is larger than that of the parameters, for which no satisfactory solutions are available based upon other schools of thought on inference.

There are many other situations where combining information is necessary. For example, consider three typical scenarios. The first is known as the constrained parameter inference. In this situation, constraints on parameters are information about parameters in addition to those specified by the sampling model. This type of problems has been challenging to existing methods but handled automatically in IMs with elastic PRSs. The second is known as Bayesian inference with scientifically meaningful priors. Martin and Liu [15] show that this type of Bayesian inference can be viewed as a special case of IMs. The third is the case where prior information is experience based, that is, the prior cannot be represented by a Bayesian prior. While more research is needed, here we consider a sensible way of combining information: to approximate the prior knowledge using a working sampling model with some fake data. This is illustrated by the following example.

Example 4 (The binomial model with imprecise prior knowledge). Suppose that prior (imprecise) knowledge about a binomial proportion θ is available and that

a new count X is obtained from $\text{Bin}(n,\theta)$. If the prior knowledge can be satisfactorily approximated as that obtained from a count X_0 from $\text{Bin}(n_0,\theta)$ with some known n_0 and X_0, with CIMs, the combined inference about θ is obtained from the count $Y = X_0 + X$ from $\text{Bin}(n_0+n,\theta)$.

The work on exact IMs shows that the IM framework provides a promising alternative to existing methods for scientific inference. Compared to existing methods that have been intensively investigated by the large research community in past decades or centuries, there is too much to do. While optimal inference is yet available for challenging statistical problems, it is expected that the way of reasoning with unobserved a-variables helps to develop methods for producing valid and efficient inferential results. The method referred to as Generalized IMs (GIMs) provides such an example. The idea is to gain simplicity while maintaining validity by making use part of an association that represents the sampling model. This idea is used in the next section for our large-scale multinomial inference.

3 A generalized inferential model for comparing two large-scale multinomial models

Now we develop an inferential model for uncertainty assessment of the assertion that two large-scale multinomial models are the same. In the following, the probabilistic inference of multinomial models is valid for data with both small and large number of categories. We start with a motivating example of genome-wide association studies, where we compare SNPs frequencies of control samples and case samples. We scan the whole genome sequence using blocks of SNPs, for example, with a block size of 10 SNPs. For a given block, there are two independent multinomial distributions corresponding to distributions of SNP genotypes of the control and case populations. These two multinomial distributions can be derived by a $2\times K$ table of independent Poisson counts, where K is the total number of SNP genotypes in the block. More specificially, let $N_j^{(i)}$ denote a Poisson count with unknown rates $\lambda_j^{(i)} \geq 0$ for $i = 0,1$ and $j = 1,...,K$. The count $N_j^{(0)}$ represents the number of subjects (or occurrences) of genotyp j in the control group and $N_j^{(1)}$ is the number of subjects (occurrences) of genotype j in the disease group. They are modeled as two independent Poisson counts with the respective rates $\lambda_j^{(0)}$ and $\lambda_j^{(1)}$. The rates $\lambda_j^{(0)}$ and $\lambda_j^{(1)}$ can be interpreted as the expected numbers of occurrences of genotype j in the corresponding populations.

It is well known that conditioning on $m_i = \sum_{j=1}^K N_j^{(i)}$ for $i = 0$ and 1, the observed data $N_j^{(i)}$ follow two independent multinomial models with

$$(N_1^{(i)},...,N_K^{(i)}) \sim \text{Multinomial}\left(m_i, \theta_1^{(i)},...,\theta_K^{(i)}\right) \qquad (i = 0,1)$$

where $\theta_j^{(i)} = \lambda_j^{(i)} / \sum_{j=1}^K \lambda_j^{(i)}$ is the SNPs frequencies of the control ($i = 0$) and case ($i = 1$) populations. The problem of interest here is inference about the assertion that $\theta_j^{(0)} = \theta_j^{(1)}$ for $j = 1, ..., K$. In terms of $\lambda_j^{(i)}$, this assertion can be written as

$$\lambda_j^{(0)} \propto \lambda_j^{(1)} \qquad (j = 1, ..., K).$$

Alternatively, conditional on each column the Poisson counts of the $2 \times K$ table lead to K binomial distributions. Let $\phi_j = \lambda_j^{(1)} / (\lambda_j^{(0)} + \lambda_j^{(1)})$ and write $n_j = N_j^{(0)} + N_j^{(1)}$ and $X_j = N_j^{(1)}$ for $j = 1, ..., K$. Then

$$X_j | n_j \sim \text{Binomial}(n_j, \phi_j) \qquad (j = 1, ..., K). \tag{11}$$

The inference about equal multinomial frequency parameters is the same as inference about

$$\mathscr{A} = \{\phi_j = \phi_0 : j = 1, ..., K \text{ for some } \phi_0 \in [0, 1]\}. \tag{12}$$

For probabilistic inference of (12), we take the generalized inferential model (GIM) approach [16] (See also [21, 23, 7] for examples of belief approaches based on likelihood functions). As mentioned at the end of Section 2, the key idea is to construct a simple IM that produces valid inference. More specifically, inference can be made from a function of the observed data, *e.g.*, $Y = h(\mathbf{X})$ for some specific function $h(.)$ of $\mathbf{X} = (X_1, ..., X_K)$. It is the state of the art to define or choose $Y = h(\mathbf{X})$ for efficient inference. In the following, we define Y based on normal approximation to the binomial problem (11), with the consideration that the resulting GIM output is efficient.

Denote $N = \sum_{j=1}^K n_j$, which is the total sample size of both control and case groups. We introduce a statistic

$$Y = \sum_{j=1}^K w_j \frac{\left(X_j - n_j \frac{\sum_{j=1}^K X_j}{N} \right)^2}{n_j (N - n_j)}$$

where $w_j = (n_j - 1)/(n_j + 1)$ down-weights observations with small column size n_j. The weights are proposed based on simulations. Note that we only consider counts with the column total of $n_j \geq 2$ when calculating Y. On the other hand, the weights such defined do not affect columns with a large number of counts. Let $\phi = (\phi_1, ..., \phi_K)$ denote the parameter of the assertion of interest (12) and $F_\phi(y)$ be the cdf of Y conditioning on $\sum_{j=1}^K X_j$. The conditional distribution $F_\phi(y)$ may be derived using the fact that X_j's follow a (multivariate) hypergeometric distribution conditioning on $\sum_{j=1}^K X_j$. In addition, $F_\phi(y)$ depends on ϕ only through their relative values, say, $\phi / \sum_{j=1}^K \phi_j$. For a data-generating device of the observable quantity Y, we know that Y can be generated by taking the inverse of $F_\phi(y)$ on a uniform random variable U.

The above statistic Y and its cdf $F_\phi(y)$ defines a GIM that associates the observed data Y and the unknown parameter ϕ through the uniform a-variable $U \sim \text{Unif}(0, 1)$.

For simplicity, in what follows we shall take continuous cdf to approximate $F_\phi(y)$. Thus, the A-step is determined by the association

$$F_\phi(Y) = U \qquad (U \sim \text{Unif}(0,1)). \tag{13}$$

Note that the distribution of Y under $\mathscr{A}^c$ is stochastically larger than that under $\mathscr{A}$. Efficient inference is obtained when the PRS for predicting the a-variable U is taken to be the one-sided PRS:

$$\mathscr{S}_U = [0,U] \qquad (U \sim \text{Unif}(0,1)). \tag{14}$$

We use this one-sided PRS for the P-step of the GIM for inference of the assertion (12).

Denote by Ω the space of the parameter ϕ. The C-step combines the A-step and the P-step to induce a random set in Ω:

$$\Omega_Y(U) \equiv \Omega_Y(\mathscr{S}_U) = \{\phi : F_\phi(Y) \le U\}$$

and compute uncertainty assessment on the assertion that the two large-scale multinomial models are the same. The computational details are given below.

The probability for $\mathscr{A}$, $p_Y(\mathscr{A})$, is necessarily zero, as the assertion represents a lower-dimensional space, where all components of ϕ are equal. The probability against the assertion, $q_Y(\mathscr{A})$, is computed by using the fact that

$$\begin{aligned} q(\mathscr{A}) &= \text{P}\left(\Omega_Y(U) \subseteq \mathscr{A}^c\right) \;=\; \text{P}\left(\Omega_Y^c(U) \supseteq \mathscr{A}\right) \\ &= \text{P}\left(\{\phi : F_\phi(Y) > U\} \supseteq \mathscr{A}\right) \\ &= \text{P}\left(U < F_\phi(Y) \text{ for all } \phi \in \mathscr{A}\right) \\ &= \text{P}\left(U < \min_{\phi\in\mathscr{A}} F_\phi(Y)\right) \\ &= \min_{\phi\in\mathscr{A}} F_\phi(Y). \end{aligned}$$

Under $\mathscr{A}$, all components of ϕ are the same. Because the distribution $F_\phi(Y)$ only depends on relative values of the components of ϕ, there is only one quantity of $F_\phi(Y)$ over $\phi \in \mathscr{A}$. The minimization is in fact not necessary. We compute the distribution of Y using a scaled χ^2 distribution with the scale and degrees of freedom estimated from a Monte Carlo sample by the method of moments. More precisely, the Monte Carlo-based method consists of the following four steps:

1. Simulate a Monte Carlo sample of size M, denoted by $\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(M)}$, from the sampling distribution of $\mathbf{X}$, conditional on $\sum_{j=1}^K X_j$.
2. Compute $Y^{(i)} = h(\mathbf{X}^{(i)})$ for $i = 1,\ldots,M$ to obtain a sample of size M: $Y^{(1)},\ldots,Y^{(M)}$, from the sampling distribution of Y.
3. Calculate the sample mean and variance of the sample $Y^{(1)},\ldots,Y^{(M)}$, that is,

$$\bar{Y} = \frac{1}{M}\sum_{i=1}^{M} Y^{(i)} \quad \text{and} \quad S_Y^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(Y^{(i)} - \bar{Y}\right)^2.$$

4. Find the degrees of freedom ν and the scale parameter γ of the χ^2 approximation with its first two moments matching the sample mean and variance calculated in Step 3, i.e.,
$$\gamma = \frac{S_Y^2}{2Y} \quad \text{and} \quad \nu = \frac{\bar{Y}}{\gamma}.$$

Steps 2-4 in the above proposed method are straightforward. The details of Step 1 are as follows. Note that conditional on $\tau = \sum_{j=1}^{K} X_j$ with $\phi \propto \mathbf{1}_K = (1,...,1)$, the sampling distribution of $\mathbf{X} = (X_1,...,X_K)$ is the well-known multivariate hypergeometric distribution with known parameters K, $n = (n_1,...,n_K)$, and $\tau = \sum_{j=1}^{K} X_j$. The multivariate hypergeometric distribution has attractive properties including that the marginal distributions and conditional distributions are also hypergeometric. This allows for a simple way of generating multivariate hypergeometric distributions with methods of simulating univariate hypergeometric distributions. For example, the marginal distribution of the first component X_1 is the (univariate) hypergeometric distribution with the parameters $\tau = \sum_{j=1}^{K} X_j$, $\sum_{j=1}^{K} n_j - \tau$, and $\sum_{j=1}^{K} n_j$, which stand for the number of "white balls" in the urn, the number of "black balls" in the urn, and the number of balls drawn from the urn respectively in the familiar context of drawing without replacement from an urn consisting of both black and white balls. Methods for pseudo random generation of the (univariate) hypergeometric distribution are available [14]. An algorithm is implemented as the function `rhyper` in R [20]. Using `rhyper` we generate samples of multivariate hypergeometric from multiple simulations of univariate hypergeometric distributions to create random samples of $\mathbf{X}$ from its joint distribution.

4 Application in genome-wide association study

We apply the methodology on the GAW16 (Genetic Analysis Workshop 16) data from the North American Rheumatoid Arthritis Consortium. This genome-wide association study aims at identifying genetic variants, more specifically single nucleotide polymorphisms, which are associated with the Rheumatoid Arthritis disease. The data consists of 2062 samples, where 868 are cases and 1194 are controls. For each sample, whole genome SNPs are observed with a total coverage of 545,080 SNPs.

We partition the entire SNP sequence on each chromosome into a sequence of m blocks of consecutive SNPs, each block consisting of, for example, 10 SNPs. For each block, indexed by $b = 1,...,m$, our proposed analysis of the two-sample multinomial counts produces (p_b, q_b, r_b) output for the assertion that "the two samples, cases versus controls, are from the same population". The (p_b, q_b, r_b) output has $p_b = 0$ and q_b providing evidence against the assertion.

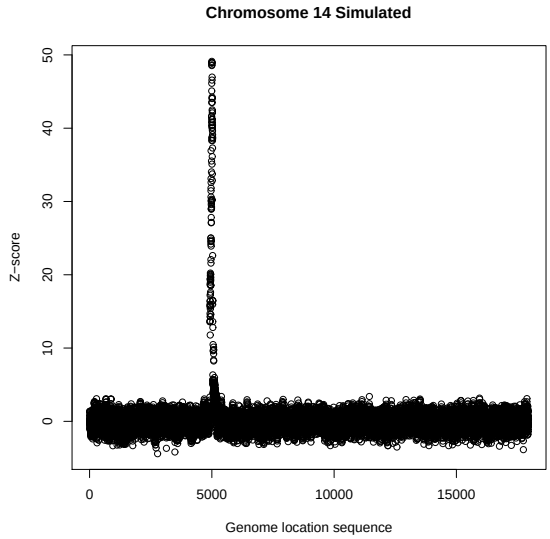

Fig. 2 The time-series plots of the Z-scores of the probabilities for the assertion that control and case populations are different, computed based on the inferential model for the simulated data from chromosome 14.

To assess our IM method for this "large K and small n" problem, we first consider a simulated study, using the real SNP genotype data to randomly simulate disease phenotypes. A phenotype variable y is generated from a simple additive model $y_i = \sum_j x_{ij} b_{ij} + e_i$, where x_{ij} denotes the SNP genotype of subject i at SNP j, $x_{ij} = 0$, 1, or 2 for wild type homozygous, heterozygous, and mutation homozygous, respectively, $i = 1, ..., 2062$, and j from 1 to the number of SNPs for a chromosome. We consider chromosome 14, which contains 17,947 SNPs in the Rheumatoid Arthristis genotype data. The coefficient b_{ij} is the effect of the j-th SNP for the i-th subject and is set equal to zero except for five SNPs at positions $j = 5000, 5001, ..., 5004$, where b_{ij} for these five SNPs are simulated from independent normal distributions with means of 5 and standard deviation of 1. In addition e_i is the residual effect generated from a normal distribution with mean of 0 and standard deviation of 1. At the end, disease subjects are sampled from the individuals with phenotypes y exceeding a threshold, which is the normal quantile corresponding to the proportion of $868/2062 = 0.42$, and controls are sampled from the remaining individuals. This simulation creates a new case-control data set with disease causal SNPs at positions $j = 5000, 5001, ..., 5004$ in chromosome 14.

Figure 2 displays a sequence of the q-value for chromosomes 14 in terms of Z-score, $Z = \Phi^{-1}(q_b)$, where $\Phi^{-1}(.)$ stands for the cdf of the standard normal distribution. Positions around 5000 have very large q_b values hence show strong evidences against the assertion that "the two samples, cases versus controls, are from the same population". In other words, IM provides a probability inference that

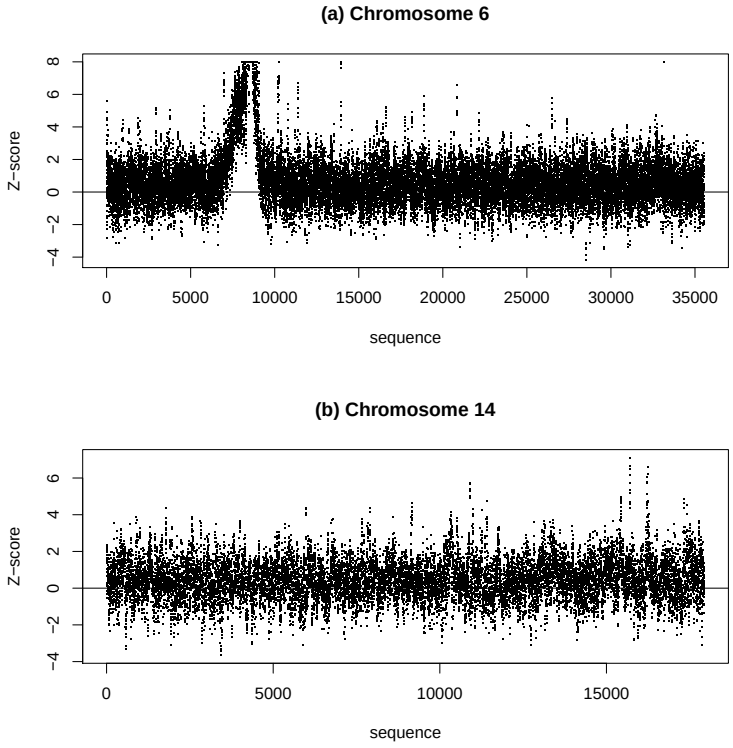

Fig. 3 The time-series plots of the Z-scores of the probabilities for the assertion that control and case populations are different, computed based on the two-multinomial model for SNPs in blocks of 10 in (a) chromosome 6 and (b) chromosome 14.

SNPs around 5000 have different frequencies between disease and control subjects and hence are associated with the disease.

We now apply the IM method in the real data of 868 cases and 1194 controls and compare SNP genotype frequencies between the two groups. A block of 10 SNPs are studied at a time, with the scale of multinomial up to $3^{10} = 59,049$ categories. Figure 3 displays sequences of the q-value for chromosomes 6 and 14 in terms of Z-score, $Z = \Phi^{-1}(q_b)$. When larger than 8, the values of the Z-scores are replaced with 8 in the plots. Figure 4 displays the histograms of the q-value for chromosomes 6 and 14. Large values in Figure 3 (a) correspond to those on the right tail in Figure 4 (a). They indicate that there are some blocks on chromosome 6 potentially associated with Rheumatoid Arthritis. This result is consistent with the known fact that the HLA (human leukocyte antigen) region on chromosome 6 contributes to disease risk. On the other hand, Figures 3 (b) and 4 (b) shows that there are very few blocks on chromosome 14 that have Z-scores larger than 6 and are considered to associate with Rheumatoid Arthritis. Except for large values, the q-value in Figures 4 (a) and 4 (b) have very smooth distributions. This implies that we can specify a null

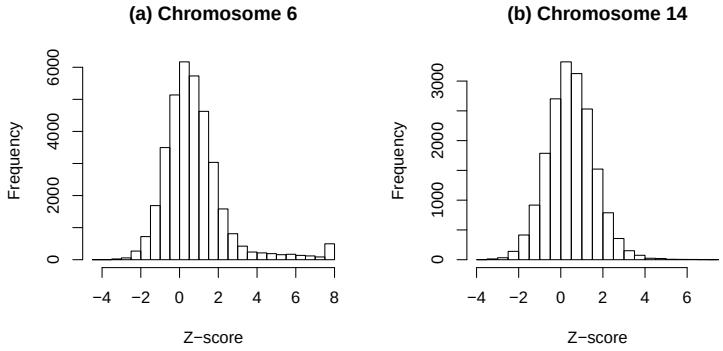

Fig. 4 Histograms of the Z-scores of the probabilities for the assertion that control and case populations are different, computed based on the two-multinomial model for SNPs in blocks of 10 in (a) chromosome 6 and (b) chromosome 14.

distribution so that SNPs or blocks potentially associated with Rheumatoid Arthritis can be identified.

For this large K and small n problem, with K up to $3^{10} = 59,049$, it is difficult to apply standard frequentist or Bayesian approaches and no such analysis has been done for a block of SNPs. Instead, we conduct a standard approach of χ^2 tests for 2×3 contingenecy tables for a single SNP at a time and compare the results with our previous analysis. The simple χ^2 tests of one SNP at a time identify the same HLA region on chromosome 6 with significant association to the disease. However, the simple χ^2 tests also produce many extremely significant SNPs, corresponding to Z-scores larger than 10, on all other chromosomes. This result indicates that the standard method tends to make falsely significant associations whereas our IM method is more accurate in assessing uncertainty for this challenging problem.

5 Conclusion

The difficulty of existing statistical methods for large-scale multinomial counts requires thinking about statistical inference at a very fundamental level and demands novel ideas beyond the current two dominant schools of thought, the frequentist and Bayesian. We propose a probabilistic inferential model, which uses auxiliary random variables for reasoning towards inference rather than constructing fiducial probabilities in the attempt to replace Bayesian posterior probabilities. The proposed method works for data of both small and large sample sizes. It produces inferential results that have desirable frequency properties. Our future research includes further investigation of the arbitrariness of the unobserved auxiliary random variable, specification of the predictive random sets, and choice of partial sampling model in generalized inferential models. We have discussed these issues in Section 2.2. More

discussion of these problems can be found in the on-going work [15, 17]. We believe that the proposed method will 1) generates useful tools for applied statisticians who are challenged by very high dimensional count data, and 2) call attention to fundamental research on statistical inference and problems considered by founding fathers such as Ronald Fisher and Jerzy Neyman.

Acknowledgements This work is supported by the National Science Foundation Grant DMS-1007678. The authors thank the Editor and the anonymous referees for their very helpful comments and suggestions that helped enhance the paper. The authors also thank Kelvin Ma for assistance in the simulation studies in Section 4.

References

1. Apolloni, B.; Malchiodi, D.; Gaito, S. : Algorithmic Inference in Machine Learning, International Series on Advanced Intelligence, 5 (2nd ed.), Adelaide: Magill, *Advanced Knowledge International* (2006).
2. Apolloni, B. and Bassis, S. : Algorithmic inference of two-parameter gamma distribution. Communications in Statistics–Simulation and Computation , 38(9), 1950-1968 (2009).
3. Apolloni, B. and Bassis, S. : Confidence About Possible Explanations, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, VOL. 41, NO. 6 (2011).
4. Balch, M. S. : Mathematical foundations for a theory of confidence structures International Journal of Approximate Reasoning, Volume 53, 1003-1019 (2012).
5. Dempster, A. P. : Further examples of inconsistencies in the fiducial argument. Annual of Mathematical Statistics, 34, 884-891 (1963).
6. Dempster, A. P. : The Dempster-Shafer calculus for statisticians. International Journal of Approximate Reasoning, 48, 265-277 (2008).
7. Denoeux, T. : Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions. International Journal of Approximate Reasoning, 42(3), 228-252 (2006).
8. Ermini Leaf, D. and Liu, C. : Inference about constrained parameters using the elastic belief method International Journal of Approximate Reasoning, Volume 53, Issue 5, 709-727 (2012).
9. Fisher, R. A. : *Statistical Methods and Scientific Inference*. London: Oliver & Boyd (1956).
10. Fraser, D. A. S. : *The Structure of Inference*. New York: Wiley (1968).
11. Gelman, A. : Bayesian model-building by pure thought: some principles and examples. *Statistica Sinica*, **6**, 215-232 (1996).
12. Hannig, J. : On generalized fiducial inference. Statist. Sinica, 19, 491-544 (2009).
13. Hannig, J. and Lee, T. C. M. : Generalized fiducial inference for wavelet regression. Biometrika, 96, 847-860 (2009).
14. Kachitvichyanukul, V. and Schmeiser, B. : Computer generation of hypergeometric random variates. Journal of Statistical Computation and Simulation, 22, 127-145 (1985).
15. Martin, R. and Liu, C. : Inferential models: A framework for prior-free posterior probabilistic inference. Journal of American Statistical Association, to appear (2013).
16. Martin, R. and Liu, C. : Generalized inferential models. Technical Report, Department of Statistics, Purdue University (2011), http://www.stat.purdue.edu/ chuanhai/docs/imlik-4.pdf.
17. Martin, R. and Liu, C. : *Inferential Models: Reasoning with uncertainty*. Chapman & Hall/CRC (2013).
18. Martin, R., Zhang, J., and Liu, C. : Dempster-Shafer theory and statistical inference with weak beliefs. Statistical Science, 25, 72-87 (2010).

19. McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. : Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews, Genetics, 9, 356-369 (2008).
20. R Development Core Team : *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (2011).
21. Shafer, G. : *A mathematical theory of evidence.* Princeton University Press, Princeton, New Jersey (1976).
22. Tsui, K. W. and Weerahandi, S. : Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J. Amer. Statist. Assoc., 84, 602-607 (1989).
23. Wasserman, L. A. : Belief functions and statistical evidence. The Canadian Journal of Statistics, 18(3), 183-196 (1990).
24. Weerahandi, S. : Generalized confidence intervals. J. Amer. Statist. Assoc., 88, 899-905 (1993).
25. Xie, M., Singh, K., and Strawderman, W. E. : Confidence distributions and a unifying framework for meta-analysis. J. Amer. Statist. Assoc., 106, 320-333 (2011).
26. Xie, M. and Singh, K. : Confidence distribution, the frequentist distribution of a parameter - a review. Int. Statist. Rev., to appear (2012).
27. Zhang, J. and Liu, C. : Dempster-Shafer inference with weak beliefs. Statistica Sinica, 21, 475-494 (2011).