The Minimal Belief Principle: A New Method for Parametric Inference

by

C. Liu and J. Zhang
Purdue University

Technical Report #07-13

Department of Statistics
Purdue University

November 2007

THE MINIMAL BELIEF PRINCIPLE: A NEW METHOD FOR PARAMETRIC INFERENCE

Chuanhai Liu and Jianchun Zhang

Purdue University

Abstract: Contemporary very-high-dimensional (VHD) statistical problems call attention more than ever to solving the fundamental problem of scientific inference, that is, to make situation-specific inference with credible evidential support. After scrutinizing the great innovative ideas behind Fisher's fiducial argument and the Dempster-Shafer (DS) theory for scientific inference, we recognize that given a postulated sampling model, reasoning for statistical inference (about a particular realization of random variables) should be different from reasoning for data generation. The classical belief in distributional invariance of pivotal variables does not distinguish these two types of reasoning processes and is thus often too strong to be believable. Intuitively, beliefs with higher credibility can be obtained from the classical belief by making it weaker. This general idea is termed as the "minimal belief" (MB) principle. Technically, the proposed method is built on the DS theory, and provides ways to capture realistically more "don't know" and thereby to build better DS models for solving VHD problems. It is shown that for general single-parameter and certain multiparameter distributions, the MB posteriors are obtained in closed form. The method is illustrated with a variety of examples, including the simple test of significance, the Behrens-Fisher problem, the multinomial model, and the many-normal-means problem. The many-normal-means example offers an MB perspective of often-crude Bayesian and related shrinkage techniques, which have been considered necessary in the last half a century.

Key words and phrases: Bayesian methods, Dempster-Shafer theory, fiducial inference, Likelihood principle, Stein's paradox.

1. Introduction

Modern statistical problems in dealing with massive data with complex structures impose more challenges than ever to statistical inferential methods. New challenges make scientists who care about their situation-specific assessment of uncertainty to think carefully about frequentist and Bayesian methods and the fundamental problem of scientific inference. We take the view shared by Fisher

(1959) and Dempster (2007), among a few others, that the fundamental problem is to make situation-specific inference with credible evidential support. Under this view, scientists would find that (1) frequentist theory somewhat obfuscates by answering questions about a "long run" that is almost always irrelevant to the practitioner's situation, among other criticisms concerning efficiency and multi-parameter/multiassertion problems; and (2) Bayesian argument always requires "probabilities for everything" and hence often leads to *ad hoc* probability assignments that lack credible empirical support, especially in tackling very-high-dimensional (VHD) problems. Fisher's fiducial argument and the Dempster-Shafer (DS) theory were proposed to tackle the fundamental problem of scientific inference. These methods are nowadays rarely visible in the statistical literature, perhaps, due to the criticisms led by Savage (1976). To the authors, the great innovative idea behind Fisher's attempted solution to scientific inference, that is, the fiducial argument, has not been well understood by people, including Fisher himself, in the past century.

1.1. The fiducial argument

The fiducial argument was introduced by R. A. Fisher in his paper "Inverse Probability" in 1930 in an attempt to derive posterior distributions for unknown parameters without use of priors, as an alternative to the Bayesian argument. Since then, the fiducial argument has been a subject full of discussions and controversies. To many contemporary statisticians, for example, "*fiducial inference stands as Fisher's one great failure*" (Zabell (1992)). To a few statisticians, however, Fisher's attitude toward statistical inference from late 1920's to late 1950's is a key to understanding inference. The spirit of *letting data tell all* in fiducial inference has inspired continued efforts in understanding the fiducial argument for scientific inference (*e.g.*, Dempster (1966), Fraser (1966), Dawid and Stone (1982), Wang (2000), to name a few).

The idea behind fiducial argument is as follows. A set of observations is supposed to have been taken from a population distributions $F(x;\theta)$ with unknown parameter θ. Without prior information about the true value of the parameter θ, we want to specify our *a posterior* uncertainty about θ by assigning a posterior distribution for θ. To make this idea more clear, consider the following example. Suppose that a single observation x is a realization of the random

variable X, where $X \sim N(\theta, 1)$ with unknown mean θ and unity variance. We write $X = \theta + Z$, where $Z \sim N(0, 1)$. Unlike the Bayesian argument, the fiducial argument does not assign a distribution to θ *a priori.* Given the observed data x, we specify our posterior uncertainty about θ through *a posteriori* reasoning via *continuing to believe* $Z \sim N(0, 1)$. This leads to the fiducial posterior distribution $\theta|x \sim N(x, 1)$. The variable Z is called the *pivotal* variable, which plays a central role in developing frequentist procedures. It is well known that inference on θ based on the fiducial posterior $\theta|x \sim N(x, 1)$ is well-calibrated, a property that is not necessary but is nice to have. This fact makes fiducial inference appealing. Unfortunately, the fiducial inference is not in general well calibrated for inference about functions of θ such as θ^2 in the present example and about multiparameters in general. Stein's paradox (see, Stein (1956))provides a famous example. This implies that the *a posteriori* belief in Fisher's fiducial argument is typically too strong.

1.2. The DS theory

Dempster's (1966) extensions of the fiducial argument are fundamentally important. First, he extended the fiducial argument to the cases with discrete observable variables. Second, he suggested to approximate continuous observable cases using the multinomial distribution. To some extent, the multinomial model can be thus viewed as the *atomic* component in DS models. Recently, Dempster (2007) introduced the Poisson model as an alternative *atomic* model for DS, which appears easy to work with technically and is practically the same as the multinomial model. As a result, this leads to a new inferential method known as the Dempster-Shafer (DS) theory. A new elegant and powerful calculus that deals with the associated (a)-random sets for DS analysis has been developed in Dempster (1966), Shafer (1976), and Dempster (2007). To the authors, the new idea known as "data fusion" of DS is particularly intriguing because it suggests that one does not have to believe or start with the *likelihood principle* for situation-specific assessment of uncertainty. Dempster's rule of combination plays an important role in DS and leads DS itself to an extension to the Bayesian method (*e.g.*, Dempster (1968, 2007)).

Dempster (2007) introduced a useful and convenient formulation of DS output in terms of (p, q, r), as opposed to Bayesian output (p, q), for assertions. More

specifically, p expresses the probability for the truth of an assertion, q expresses the probability against the truth of the assertion, and r represents a residual probability of the new category of "don't know", an important component that is needed in scientific inference and has been missing in other inferential theories.

1.3. The MB Principle

In building practical DS models, DS users have been adopting traditional *a posteriori* reasoning used in Bayesian, frequentist or confidence-interval, and fiducial arguments, which we refer to as the *classical belief* (CB) principle. To the authors, however, the classical belief principle appears to be somewhat stronger than necessary. Intuitively, beliefs with higher credibility can be obtained from the classical belief by making it weaker. We term such a general idea in this paper as the "minimal belief" (MB) principle. Intuitively, the MB principle admits realistically more "don't know" and, thereby, allows for more credible DS inference. It is in this sense, in terms of "don't know", that MB is weaker than CB. In other words, MB is introduced in an attempt to achieve higher credibility. However, the MB principle serves as a general guidance rather than a precisely defined mathematical term. It may take different mathematical forms for sampling models of different data structures. The intuitive idea of MB is explained by a motivating example in Section 2 and discussed further in Section 3.

Technically, the proposed inferential method is built on the DS theory because the MB principle is intended to be a new way to build alternative DS models. We refer to Dempster (2007) as an important introduction to the DS theory. We believe that MB-DS models, or simply MB models in the sequel, are particularly useful for solving VHD problems. We focus in this article on the fundamental idea of MB.

The remaining part of this article is arranged as follows. A motivating example is given in Section 2. Some basics of minimal belief principle are formulated in Section 3. Single parameter cases are treated in section 4 for both discrete and continuous models. In section 5, multiparameter cases are investigated with several illustrative examples, including the normal distribution with unknown mean and variance and the Behrens-Fisher problem. MB estimation of the multinomial model is treated in section 6. Section 7 concludes with a brief discussion.

2. A Motivating Example: the many-normal-means problem

As a motivating example, an MB perspective of Stein's paradox is presented in this Section. Suppose that X_i's are independent and $X_i \sim N(\mu_i, 1)$ for $i = 1, \cdots, n$, where μ_i's are unknown parameters to be estimated. The least squares estimator (or maximum likelihood estimator) is found to be inadmissible under square loss when $n \geq 3$. This discovery is known as Stein's paradox. James and Stein (1961) showed that the James-Stein estimator, which was originally proposed by Stein (1956), always achieves lower mean square error than the least squares estimator. Efron and Morris (1973) gave a (parametric) empirical Bayes interpretation of the James-Stein estimator. All these methods shrink the least squares estimator towards a common value, such as zero.

We write the sampling model as follows: $X_i = \mu_i + Z_i$, where $Z_i \overset{iid}{\sim} N(0,1)$ for $i = 1, ..., n$. Given the observed data, Z_i's are known to have occurred and remained unknown to us. With CB, we make inference about μ_i by repeatedly "firing a shot at $(Z_1, ..., Z_n)'$ in the n-dimensional Euclidean space $\mathcal{R}^n$ with a random draw $(Y_1, ..., Y_n)'$ from $N^n(0,1) = N_n(0, I_n)$." The deviations of the points in the 2-dimensional (2D) scatter plots from the diagonal line in the upper panel of Figure 2.1 indicate the failure of the attempt of CB to make (joint) inference about $(\mu_1, ..., \mu_n)'$ as n increases. This implies that CB is typically too strong to be believable, meaning that we may actually know less than what is formulated via CB.

A weaker belief that came to mind is to "fire a shot at the ordered Z_i's, denoted by $(Z_{(1)}, ..., Z_{(n)})'$, with an ordered draw $(Y_{(1)}, ..., Y_{(n)})'$." The 2D scatter plots in the lower panel of Figure 2.1, although somewhat misleading in that the surrogates are in $\mathcal{R}^n$ rather than $\mathcal{R}^2$, show the dramatically increased credibility, of course, in the case with the known order or permutation. Note that this weaker belief differs from CB by an unknown permutation. In reality, statistician "doesn't know" how the ordered values correspond to the observations. Thus, introducing the "don't know" component into inference, as in the DS theory, is fundamentally important for credible inference.

For convenience, we term the general idea of using beliefs weaker than CB the *minimal belief* (MB) principle in hopes of achieving *maximal credibility*. For inference, we could live with either vacuous knowledge about the permutation

or credible knowledge, if any, about the permutation, both can be handled in general by using the DS calculus. In this paper, we focus on the case with the vacuous prior on permutations, while leave the general treatment of permutations to future research. The purpose of this paper is to elaborate the MB idea and see how it performs.

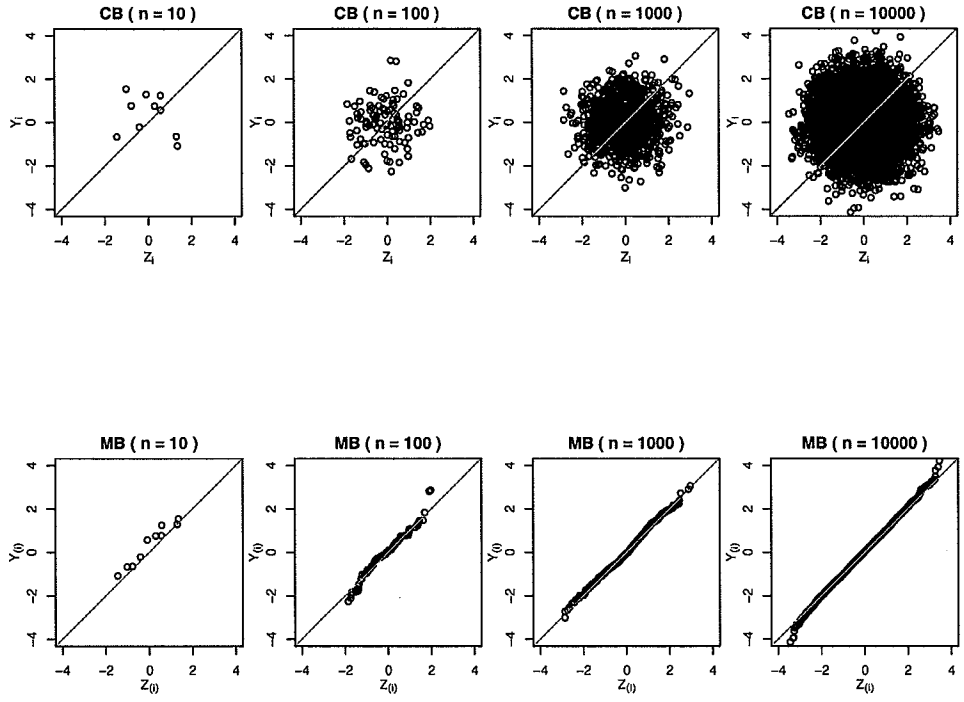

Figure 2.1: Classical Belief (CB) vs. Minimal Belief (MB). The sampling model: $X_i = \mu_i + Z_i$ for $i = 1, \cdots, n$ with $Z_i \overset{iid}{\sim} N(0,1)$. The upper panel is the scatter plots of Y_i's vs. Z_i's. The surrogate variables $\{Y_i\}_1^n$ are used for inference under CB (upper panel). The lower panel is the scatter plots of ordered Y_i's vs. ordered Z_i's. The surrogate variables $\{Y_{(i)}\}_1^n$ are used for inference under MB (lower panel). n is the sample size.

For readers who are not familiar with the DS theory, we demonstrate what MB can offer by presenting an experimental study. For those who are more

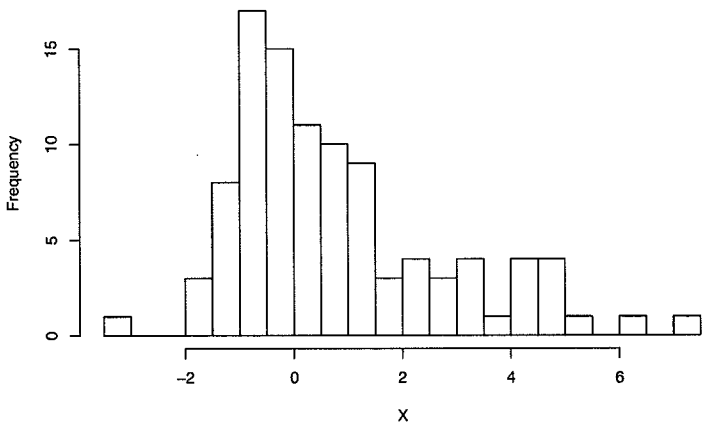

Figure 2.2: The histogram of a sample of size $n = 100$, $\{X_i\}_{i=1}^n$, with X_i generated from $N(\mu_i, 1)$, where $\mu_i \approx 0$ for $i = 1, ..., 80$, and $\mu_i \overset{iid}{\sim} 3 + \text{Expo}(1)$ for $i = 81, ..., 100$.

familiar with the usual probability calculus, such as Bayesian posteriors, than with the DS calculus, we need to specify a usual conditional distribution of the unknown permutation given the observed data X_i's and the surrogate variables Y_i's or simply Z_i's, for both notational simplicity and *a posteriori* thinking about the unknown Z_i's. We warn the reader that forcing "don't know" to go away, that is, $r = 0$, can lead to incredible inference. Nevertheless, the experiment presented below does offer, to some extent, an MB perspective of what is going on in making inference about the surrogate variables that are related to Stein's paradox.

One idea of specifying a usual conditional distribution for sensible *a posteriori* reasoning is to consider the empirical process on the real line:

$$F_n(x) \equiv \frac{|\{X_i : \ X_i \leq x, i = 1, ..., n\}|}{n} \qquad (-\infty < x < \infty) \tag{2.1}$$

where $|\{X_i : \ X_i \leq x, i = 1, ..., n\}|$ stands for the number of the observed X_is that are less than or equal to x. For given permutation $\pi = (\pi_1, ..., \pi_n)'$ of $(1, ..., n)'$, and thereby uniquely defined $\mu_1, ..., \mu_n$, $F_n(x)$ has the following mean

and covariance functions

$$\bar{F}_n(x) \equiv \mathrm{E}(F_n(x)) = \frac{1}{n}\sum_{i=1}^{n} F(x|\mu_i) \qquad (-\infty < x < \infty) \tag{2.2}$$

and

$$\mathrm{E}(F_n(x), F_n(y)) = \frac{1}{n^2}\sum_{i=1}^{n} F(x|\mu_i)[1 - F(y|\mu_i)] \quad (-\infty < x \leq y < \infty) \tag{2.3}$$

where $F(x|\mu)$ denotes the cdf of $N(\mu, 1)$. Our belief for specifying a posterior for the permutation is that the process

$$F_n(x) - \bar{F}_n(x) \tag{2.4}$$

is approximately a Brownian bridge with covariance function given in (2.3), where $F_n(x)$ and $\bar{F}_n(x)$ are defined in (2.1) and (2.2). For producing preliminary results, we considered the usual distribution for the permutation, that is, in the DS context non-zero masses are assigned only to singletons in the set of all the possible permutations.

For a simulation study, $n = 100$ μ_is were chosen as follows. 80 μ_is were set to be approximately zero and the other 20 μ_is were generated from the "shifted" exponential distribution $3+\mathrm{Expo}(1)$. The data $X_{obs} = \{X_i\}_{i=1}^n$ were then generated from $N(\mu_i, 1)$ for $i = 1, ..., n$. The simulated data is shown by the histogram in Figure 2.2. The marginal posterior distributions of μ_i, based on 1,000 posterior draws, are displayed using boxplots in Figure 2.3, where the whiskers represent approximately 99% "confidence intervals". Note that the (normal-distribution based) 99% "confidence intervals" length is about 5.15. We see from Figure 2.3 that implicit shrinkage has taken place. More shrinkage is applied to clustered X_is, for example, those in neighborhoods of 0 and 4. For values approximately from 2 to 3, the marginal posteriors of the corresponding μ_is have wider "confidence intervals", representing more uncertainty than that produced by the fiducial argument. For the isolated observed values, such as those near 6, it does not have a big shrinkage effect.

The procedure for computing the preliminary results is outlined as follows. A finite number of points in $(0, 1)$ covering the observed data range is used to approximate the Brownian bridge. The values of a Brownian bridge at the selected points form a discrete Brownian bridge. For taking a posterior draw of the

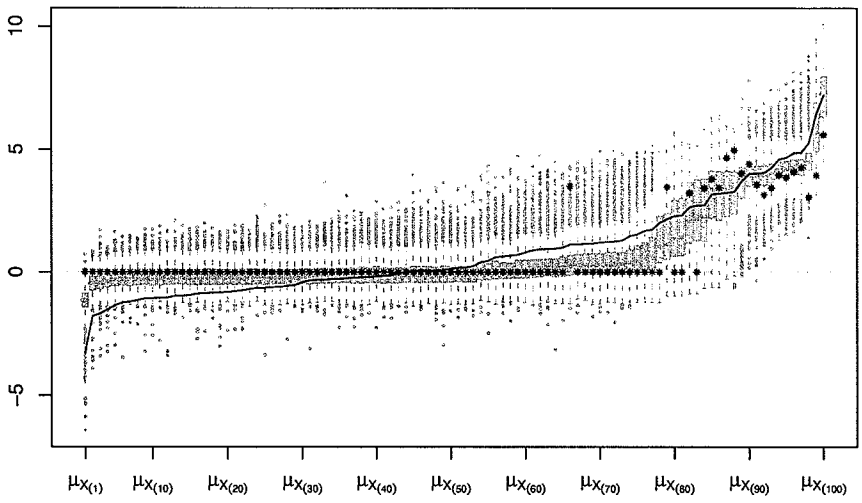

Figure 2.3: The marginal posteriors of μ_i, indexed by the ordered observed data, $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, where blue curve: $X_{(i)}$, box-and-whisker plots: (marginal) posteriors of $\mu_{X_{(i)}}$, and red asterisks: true values of $\mu_{X_{(i)}}$, which are unknown in reality.

permutation for given $\{X_i\}_{i=1}^n$ and $\{Z_{(i)}\}_{i=1}^n$, the basic idea is to simulate a (discrete) Brownian bridge and find a permutation that matches $F_n(x) - \bar{F}_n(x)$ (with the associate variance-covariance matrix) to the simulated Brownian bridge. This idea is in spirit the general idea of "continuing to believe", subject to the "non-conflict"/feasibility constraint in DS sense. That is, a simulated Brownian bridge $B(x)$ is said to be feasible if there exists at least one permutation that makes $B(x)$ match $F_n(x) - \bar{F}_n(x)$. Approximation methods are used to generate non-conflict Brownian bridge. Technical details, together with the general MB-DS approach (*i.e.*, with "don't know"), shall be reported elsewhere.

3. The Basics of MB Principle

The MB principle introduced in the motivating example in Section 2 leads us to the thought that there should be a fundamental difference between "data generation" and parameter inference, which are treated as the same in CB infer-

ence. By "data generation", we mean that the random variables, such as pivotal variables in Fisher's fiducial argument, follow their long-run frequency distribution, before the observable data are obtained. On the other hand, by "parametric inference", we mean inference about unknown quantities that are associated with particular realizations of random variables. We refer to these pivotal variables as *surrogate* variables not to abuse the well established concept of "pivotal". As a matter of fact, the concept of "surrogate" is more general than "pivotal". It makes us reason using random surrogates associated with individual data points rather than using sufficient statistics based on the likelihood principle. As is seen in the motivating example in Section 2, MB leads to realistically or credibly more "don't know" in the DS context than CB. Technically, we make use of the DS concept of random sets and the (p, q, r) formulation to make posterior probability statements about assertions of interest.

The formulation of a general framework for *a posteriori* reasoning towards building MB models can be helpful. An MB model for posterior distribution based inference can be defined as follows. Suppose the data are generated according to the sampling model: $X_{obs} = G(\theta, U)$, where $\theta \in \Theta \subset \mathcal{R}^d$, U has a distribution of $f(u)$, $u \in \mathcal{S}_u$. Without loss of generality, we assume the surrogate variable $U = (U_1, \cdots, U_n)$ with $U_i \overset{iid}{\sim} U(0, 1)$. A continuing-to-believe (CB, a slightly abused notation) variable C is defined via a many-to-one measurable mapping: $C = \mathcal{M}(U)$. For example, $\mathcal{M}(U_1, \cdots, U_n) = (U_{(1)}, \cdots, U_{(n)})$. The *inverse set* consists of all feasible U for given C and X_{obs}:

$$S(C) = \{U : C = \mathcal{M}(U) \text{ and } X_{obs} = G(\theta, U) \text{ for some } \theta \in \Theta\}$$

The MB model is then specified by:

(1) the distribution of $C|X_{obs}$ that is the same as the sampling distribution of C, that is, the CB principle is applied to C; and

(2) a conditional DS-distribution over $S(C)$ for given C and X_{obs}, which is a usual probability mass function on the power set of $S(C)$.

The following two simple examples illustrate how MB inference is performed. More examples are given in Sections 4, 5, and 6.

Example 3.1. *Normal distribution with known variance.* Consider the normal distribution with unknown mean μ and unity variance. We can write $X =$

$\mu + \Phi^{-1}(U)$, where $U \sim U(0,1)$. Assume that only a single observation x is available, then we are sure that there is a corresponding u, which is a realization from $U(0,1)$ but unknown to us. Note that there is no permutation issue involved. The surrogate variable is nothing but the pivotal variable. The MB posterior distribution for $\mu = x - \Phi^{-1}(u)$ can be written as

$$\mu | x \sim N(x, 1)$$

which is correspondingly the same as the fiducial distribution.

Example 3.2. *Two normal samples with known variances.* Consider the simplest case in which we have two independent observations, namely, x_1 from $N(\mu_1, 1)$ and x_2 from $N(\mu_2, 1)$ with unknown parameters (μ_1, μ_2). We are interested in making inference about functions of (μ_1, μ_2), *e.g.*, $\mu_1 - \mu_2$, $\mu_1^2 + \mu_2^2$, etc. MB gives the a-random set for (μ_1, μ_2):

$$\left\{ \Big(x_1 - Z_1, x_2 - Z_2\Big) \text{ or } \Big(x_1 - Z_2, x_2 - Z_1\Big) \right\}$$

where Z_1 and Z_2 are understood as permutation or order statistics of two independent standard normal random variables. For inference about $\mu_1 - \mu_2$, we use the a-random interval

$$\Big[(x_1 - x_2) - \sqrt{2}|Z|, (x_1 - x_2) + \sqrt{2}|Z|\Big]$$

by considering the projection on the line of $\mu_1 - \mu_2$, where Z is a standard normal random variable and $\sqrt{2}|Z|$ has the same distribution as $(Z_{(2)} - Z_{(1)})$, the range of the two order statistics. A justification on the use of above a-random interval can be made along the line of introducing generalized MB below. The probability (p) and plausibility $(p + r)$ about the assertion $\{\mu_1 - \mu_2 \leq \Delta\}$ is plotted as in Figure 3.4. It is noticed that the r (don't know) for the foregoing assertion is close to 1 for $\Delta \approx x_1 - x_2$. For the testing problem, for example, $H_0 : \mu_1 \leq \mu_2$ vs. $H_a : \mu_1 > \mu_2$, one would confirm H_0 if $x_2 > x_1$, and H_a if $x_1 > x_2$.

To end the discussion on the basics of MB, in the spirit of MB we consider an extension of the idea of specifying posterior using weaker belief by replacing a surrogate point with a random set. This can be referred to as the generalized MB (GMB) principle. For example, taking an ordered draw, denoted by $(Y_{(1)}, ..., Y_{(n)})'$ with $Y_i \overset{iid}{\sim} U(0,1)$ and construct the n-dimensional box determined

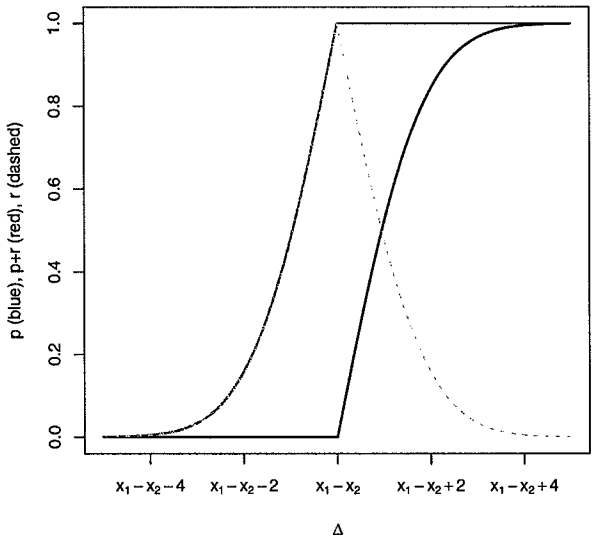

Figure 3.4: The probability p (blue) and the plausibility $(p+r)$ (red) about the assertion $\{\mu_1 - \mu_2 \leq \Delta\}$. The dashed line represents the "don't know" (r) part.

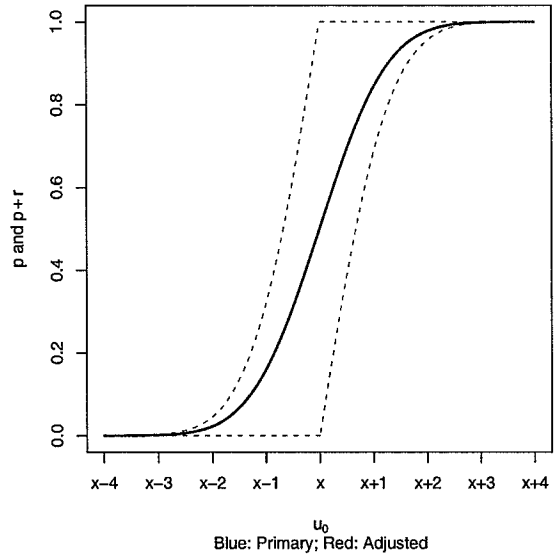

Figure 3.5: The MB posterior CDF (red) and the adjusted posterior CDFs for a single normal variate with known variance, assuming that the surrogate is between $[U/2, (1+U)/2]$.

by

$$(Y_{(i-1)} + Y_{(i)})/2 \leq U_{(i)} \leq (Y_{(i)} + Y_{(i+1)})/2, \qquad (i = 1, ..., n) \qquad (3.1)$$

where $Y_{(0)} = 0$ and $Y_{(n+1)} = 1$, we fire a shot with the n-dimensional box (3.1), instead of the single point $(Y_{(1)}, ..., Y_{(n)})'$, at the target $(U_{(1)}, ..., U_{(n)})'$. This idea is particularly useful for adjusting a-random sets defined by multiple points (see, *e.g.*, Example 5.2). GMB is also useful for solving the problem of hypothesis testing, which is illustrated with the following example.

Example 3.1. (*cont'd*). Suppose that u is known to be $[\frac{U}{2}, \frac{1+U}{2}]$ under the above GMB scenario, where $U \sim U(0,1)$, it is easy to calculate the GMB posterior (cumulative) distributions for the lower bound and upper bound of μ, which are:

$$\underline{H}(\mu_0|x) = \begin{cases} 2\Phi(\mu_0 - x) & \text{for } \mu_0 \leq x, \\ 1 & \text{for } \mu_0 > x \end{cases}$$

and

$$\overline{H}(\mu_0|x) = \begin{cases} 0 & \text{for } \mu_0 \leq x, \\ 2\Phi(\mu_0 - x) - 1 & \text{for } \mu_0 > x \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function for standard normal random variable. The GMB posterior functions are in fact the probability (p) and the plausibility ($p + r$) about the assertion $\{\mu \leq \mu_0\}$ for given x (see Figure 3.5). The problem of significance testing, such as $H_0 : \mu \leq \mu_0$ vs. $H_a : \mu > \mu_0$, is then resolved nicely by using the DS (p, q, r) output to confirm either the *null* or the *alternative*, or fails to confirm either when r is large. In other words, the DS "don't know" component makes it unnecessary Fisher's argument with the force of a logical disjunction for significance testing.

4. The Single Parameter Case

It is relatively easy to perform MB inference when there exists at most a unique feasible permutation. This is typically the case with univariate distributions. In this case, CB and MB are the same, subject to a factor $n!$ associated with the feasibility condition. The difference is that in MB inference, the infeasible permutations are excluded before making inference. While in CB inference, a similar treatment is conducted implicitly when applying Dempster's rule of combination. In most cases, MB inference has a closed form solution.

4.1. Continuous Distribution

For the continuous sampling distributions, we have the following result:

Theorem 1 *Suppose that* $\{x_1, \cdots, x_n\}$ *is a sample of size* n *generated by* $U_i = F(X_i|\theta)$, *where* $U_i \overset{iid}{\sim} U(0,1)$ *for* $i = 1, \cdots, n$. $F(\cdot|\theta)$ *is the cumulative distribution function for* X_i*'s with parameter* $\theta \in \Theta \subseteq (-\infty, \infty)$. *Assume that:*

(1) $U = F(X|\theta)$ *is one-to-one mapping for any two variables given the third variable being fixed;*

(2) For any x *in sampling space* $\mathcal{X}$, $\Theta_x \equiv \{\theta : F(x|\theta) \in (0,1)\}$ *is a non-empty interval;*

(3) $F(x|\theta)$ *as a function of* θ *is strictly monotone (increasing or decreasing) and differentiable over* Θ_x*; and*

(4) $F(x|\theta)$ *as a function of* x *is strictly increasing and differentiable over* $\mathcal{X}$*; For any* $\delta > 0$, *let the intervals* $[x_i, x_i + \delta]$ *be a coarsened version of the sample, then the a-random sets for MB inference about* θ *are all singletons with the MB posterior distribution proportional to*

$$\left[\sum_{j=1}^{n} \frac{\left|\frac{\partial F(x_j|\theta)}{\partial \theta}\right|}{\frac{\partial F(x_j|\theta)}{\partial x}}\right] \prod_{i=1}^{n} \frac{\partial F(x_i|\theta)}{\partial x} \tag{4.1}$$

in the limit of $\delta \to 0$.

Proof: Noticing that the permutation is unique, we shall use DS argument to prove the result. Without loss of generality, we assume $F(x|\theta)$ as a function of θ is strictly decreasing over the interval Θ_x. For any $x \in [x_i, x_i + \delta]$, Θ_x is an interval, and therefore $\Theta_{i,\delta} \equiv \cup_{x_i \le x \le x_i + \delta} \Theta_x$ is an interval due to the continuity assumptions. Hence, given the observed interval $[x_i, x_i + \delta]$, U_i is uniform over the interval $\mathcal{U}_{i,\delta} \equiv \{u : F(x|\theta) = u; x_i \le x \le x_i + \delta\}$. Thus, the coarsened version of the sample obtained by replacing each x_i with the interval $[x_i, x_i + \delta]$ induces an a-random interval for θ (see Figure 4.6):

$$[\theta(U_i, x_i), \theta(U_i, x_i + \delta)],$$

where U_i is uniformly distributed over $\mathcal{U}_{i,\delta}$. For any real numbers $a, b \in \cap_i \Theta_{x_i}$

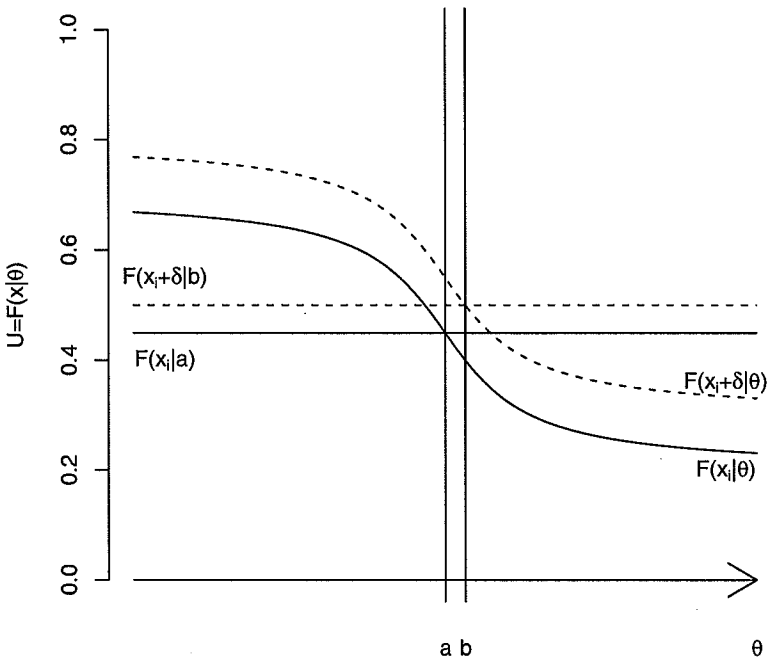

Figure 4.6: Curves of $F(x_i|\theta)$ and $F(x_i+\delta|\theta)$ as functions of θ. Note that the event $\{\theta(U_i, x_i) \leq a\}$ is equivalent to $\{U_i \geq F(x_i|a)\}$ and that $\{\theta(U_i, x_i+\delta) \geq b\}$ is equivalent to $\{U_i \leq F(x_i+\delta|b)\}$.

and $a \leq b$, we have:

$$\begin{aligned} & \Pr\big(\theta(U_i, x_i) \leq a, \theta(U_i, x_i+\delta) \geq b\big) \\ = \; & \Pr\big(F(x_i|a) \leq U_i \leq F(x_i+\delta|b)\big) \\ \propto \; & \max\{F(x_i+\delta|b) - F(x_i|a), 0\} \end{aligned}$$

Note that any non-empty intersection of all the individual a-random intervals is still an interval. This yields the combined probability function for

$$\cap_{i=1}^{n} [\theta(U_i, x_i), \theta(U_i, x_i+\delta)], \tag{4.2}$$

which is proportional to $\prod_{i=1}^{n} \max\{F(x_i+\delta|b) - F(x_i|a), 0\}$. Denote by $m(x, y)$

the associated density function for the interval in (4.2), we have

$$c([a,b]) \equiv \int_{\infty}^{a}\int_{b}^{\infty} m(x,y)dydx$$

It follows that the density of the lower end of the combined a-random interval is

$$\begin{aligned} m_\delta(\theta) &\equiv \left.\frac{\partial c([a,b])}{\partial a}\right|_{a=\theta,b=\theta} \\ &\propto \sum_{j=1}^{n}\left|\frac{\partial F(x_j|\theta)}{\partial\theta}\right|\prod_{i\neq j}[F(x_i+\delta|\theta)-F(x_i|\theta)] \\ &= \left[\sum_{j=1}^{n}\frac{\left|\frac{\partial F(x_j|\theta)}{\partial\theta}\right|}{F(x_j+\delta|\theta)-F(x_j|\theta)}\right]\prod_{i=1}^{n}[F(x_i+\delta|\theta)-F(x_i|\theta)] \end{aligned}$$

Hence, for any $\theta_1,\theta_2\in\cap_i\Theta_{x_i}$ as $\delta\to 0$, we have

$$\frac{m_\delta(\theta_1)}{m_\delta(\theta_2)}\to\frac{\left[\sum_{i=1}^{n}\frac{\left|\frac{\partial F(x_j|\theta_1)}{\partial\theta}\right|}{\frac{\partial F(x_j|\theta_1)}{\partial x}}\right]\prod_{i=1}^{n}\frac{\partial F(x_i|\theta_1)}{x}}{\left[\sum_{i=1}^{n}\frac{\left|\frac{\partial F(x_j|\theta_2)}{\partial\theta}\right|}{\frac{\partial F(x_j|\theta_2)}{\partial x}}\right]\prod_{i=1}^{n}\frac{\partial F(x_i|\theta_2)}{x}}.$$

That is, as $\delta\to 0$, we obtain

$$m(\theta)\propto\left[\sum_{j=1}^{n}\frac{\left|\frac{\partial F(x_j|\theta)}{\partial\theta}\right|}{\frac{\partial F(x_j|\theta)}{\partial x}}\right]\prod_{i=1}^{n}\frac{\partial F(x_i|\theta)}{\partial x}. \tag{4.3}$$

Similarly, the density of the upper end of the combined a-random interval can be obtained and is the same as $m(\theta)$ in equation (4.3). Thus, equation (4.3) is the MB posterior density function for θ.

Remark 4.1. For location parameters, MB posteriors coincide with the Bayesian posteriors obtained with the use of the flat prior. In other words, if the family of the sampling distributions, indexed by the unknown parameter θ, can be transformed to a location family, MB inference supports the likelihood principle (see Lindley (1958) for an interesting discussion on the relationship between fiducial and Bayesian arguments). It is seen from Theorem 1 and the following examples that in general, MB inference does not support the likelihood principle

for statistical inference. This implies that in general, Fisher's sufficient statistics are no longer "sufficient" for MB inference.

Remark 4.2. Although the DS calculus is used in the proof of Theorem 1, MB posteriors are different from conventional DS posteriors, that is, those given by the DS multinomial approximation (Dempster (1966)). The multinomial-based and Poisson-based DS posteriors for univariate distributions with single parameters can also be obtained in closed form (Dempster (1969) and (Liu (2007b)). It is worth noting that although multinomial approximation is used to introduce "don't know" in DS inference, conventional DS posteriors are CB based and thus different from MB posteriors.

4.2. Discrete Distribution

MB posteriors for univariate discrete distributions can also be obtained in closed form. Here we consider, as an example, the Poisson distribution. The Poisson model is treated as an *atomic* model in DS calculus (2007). Here, we adopt the MB approach to the Poisson model.

Example 4.1. *Poisson distribution.* Suppose $X_i = k(k = 0, 1, \cdots)$ if $0 \leq U_i - F(X_i - 1|\theta) < F(X_i|\theta) - F(X_i - 1|\theta) = f(X_i|\theta)(i = 1, \cdots, n)$ are *i.i.d.* sample of size n from Poisson distribution with parameter θ, where $F(\cdot)$ and $f(\cdot)$ are the cumulative distribution function and probability mass function of the Poisson distribution, respectively. The a-random interval for θ is $\cap_{i=1}^{n}[\theta(U_i, x_i - 1), \theta(U_i, x_i)]$. Using an approach similar to the one in the proof of Theorem 1, we can obtain the joint posterior distribution for the a-random interval, which is proportional to $\prod_{i=1}^{n}[F(x_i|b) - F(x_i - 1|a)]$. This leads to the MB posterior densities of lower and upper endpoints:

$$\begin{aligned}
\underline{h}(\theta|x) &\propto \left[-\sum_{i=1}^{n}\frac{\frac{\partial F(x_i-1|\theta)}{\partial\theta}}{f(x_i|\theta)}\right]\left[\prod_{i=1}^{n}f(x_i|\theta)\right] \propto \frac{1}{\theta}\prod_{i=1}^{n}f(x_i|\theta) \\
&\propto \text{dGamma}(\sum_{i=1}^{n}x_i, n) \\
\text{and} & \\
\overline{h}(\theta|x) &\propto \left[-\sum_{i=1}^{n}\frac{\frac{\partial F(x_i|\theta)}{\partial\theta}}{f(x_i|\theta)}\right]\left[\prod_{i=1}^{n}f(x_i|\theta)\right] \propto \prod_{i=1}^{n}f(x_i|\theta) \\
&\propto \text{dGamma}(1+\sum_{i=1}^{n}x_i, n),
\end{aligned}$$

where dGamma denotes the density of Gamma distribution. These marginal cdf results are the same as those of Dempster (2007). However, the distribution of the MB a-random intervals is different from that of the DS a-random intervals. Furthermore, the DS posterior obtained via the DS approximation through discretization is dramatically different from the MB posterior (see Liu (2007b)).

5. Multiparameter Cases

Fiducial inference for the multiparameter case has been a challenging problem. It would be a natural way to tackle the problem by taking the DS approach to the single parameter case. For example, Dempster (1966,1972) considered the sampling class of structures of the second kind and applying DS rule of combination for the multinomial model. However, the a-random set is difficult to compute in the multinomial case. The MB results appear attractive and easy to calculate in many cases. We consider in this section two examples and in the next section the multinomial model.

Example 5.1. *Normal distribution with unknown variance.* Suppose that $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ for $i = 1, \cdots, n$ with unknown μ and σ^2 to be estimated. There is a unique feasible permutation. The a-random sets are necessarily singletons. The distribution for (μ, σ^2) can then be obtained by the trick of considering the conditionals via arguments made with the surrogate variables. By Theorem 1, it is easy to obtain:

$$\mu|(x_1, \cdots, x_n, \sigma^2) \sim N(\overline{x}, \sigma^2/n)$$

$$\sigma^2|(x_1, \cdots, x_n, \mu) \sim \text{invGamma}\left(\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}\right)$$

The joint posterior distribution for (μ, σ^2) does exist since the conditionals are determined by the common surrogate variables. Of course, it can be shown that the two conditionals are compatible based on the well-known theory (see, Arnold *et al.* (2001)): *the conditionals $f(\theta_1|\theta_2)$ and $f(\theta_2|\theta_1)$ are compatible iff the following two conditional hold: (1) $\Theta \equiv \{(\theta_1, \theta_2) : f(\theta_1|\theta_2) > 0\} = \{(\theta_1, \theta_2) : f(\theta_2|\theta_1) > 0\}$; and (2) there exists function $u_1(\theta_1)$ and $u_2(\theta_2)$ such that $\frac{f(\theta_1|\theta_2)}{f(\theta_2|\theta_1)} = u_1(\theta_1)u_2(\theta_2)$ for all $(\theta_1, \theta_2) \in \Theta$.*

The joint MB posterior distribution can be obtained from these conditionals

as follows (see, *e.g.*, Gelman and Speed (1993,1999)):

$$f(\mu,\sigma^2|x_1,\cdots,x_n) \propto \frac{f(\mu|x_1,\cdots,x_n,\sigma^2)f(\sigma^2|x_1,\cdots,x_n,\mu_0)}{f(\mu_0|x_1,\cdots,x_n,\sigma^2)}$$
$$\propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\overline{x}-\mu)^2]\right)$$

where $\overline{x}$ and s^2 are the sample mean and sample variance, respectively, μ_0 is arbitrarily fixed value. It is interesting to see that the MB result agrees with the commonly used Bayesian posterior (see, *e.g.*, Gelman *et al.* (2004), p.74), which is obtained by using the prior $f(\mu,\sigma^2)\propto\sigma^{-2}$ rather than the Jeffreys prior.

It is worth noting that the *a posteriori* surrogate variables in this case reduce to two independent surrogate variables given the observed data $\{x_i\}_{i=1}^n$. More specifically, we have the a-random point/variable $\overline{x}+\frac{\sigma}{\sqrt{n}}Z$ for μ conditioning on σ^2, and the a-random point/variable $(n-1)s^2V$ for σ^2, where Z and V are independent, Z is a standard normal random variable, V is an inverse-Gamma random variable with degree-of-freedom $(n-1)/2$ and rate $1/2$.

Example 5.2. *The Behrens-Fisher problem.* The Behrens-Fisher problem is about comparing the means of two normal distributions with unknown means and variances. This problem has long been of interest in the theory of statistical inference. Although many methods have been proposed, no definitive solutions appear to exist. A comprehensive review of this problem can be found in Kim and Cohen (1998). Here, we consider the simplest case, where we have two observations for each population, namely, $n_1=n_2=n=2$. For each population, the a-random sets are singletons as in the case of $N(\mu,\sigma^2)$ (see Example 5.4), because feasible permutations are unique. However, the permutation is not unique between the two populations. The resulted a-random set in fact consists of doubletons, corresponding to the unknown permutations between the two populations. Denote the sample means and sample variances by $\overline{x}_i$ and s_i^2, for $i=1,2$. We then have the a-random sets for $(\mu_1|\sigma_1^2,\sigma_1^2,\mu_2|\sigma_2^2,\sigma_2^2)$:

$$\left(\overline{x}_1+\frac{\sigma_1}{\sqrt{n}}Z_1,(n-1)s_1^2V_1,\overline{x}_2+\frac{\sigma_2}{\sqrt{n}}Z_2,(n-1)s_2^2V_2\right)$$

or

$$\left(\overline{x}_1+\frac{\sigma_1}{\sqrt{n}}Z_2,(n-1)s_1^2V_2,\overline{x}_2+\frac{\sigma_2}{\sqrt{n}}Z_1,(n-1)s_2^2V_1\right),$$

where "|" means conditioning. Note that (Z_1, V_1) and (Z_2, V_2) as a-random variables are no longer independent *a posteriori*. However, the two variables in each pair are still independent.

The MB approach for inference about functions of (μ_1, μ_2) with the vacuous prior for the permutation is to consider the marginal a-random sets for (μ_1, μ_2), that is,

$$\left\{\left(\overline{x}_1 + \sqrt{\frac{s_1^2}{2}}C_1, \overline{x}_2 + \sqrt{\frac{s_2^2}{2}}C_2\right) \text{ or } \left(\overline{x}_1 + \sqrt{\frac{s_1^2}{2}}C_2, \overline{x}_2 + \sqrt{\frac{s_2^2}{2}}C_1\right)\right\}$$

where C_1, C_2 (with a slightly abuse of notation) are understood as permutation or order statistics of two independent Cauchy random variables.

The a-random interval for $\mu_1 - \mu_2$ can then be obtained as:

$$\left[\overline{x}_1 - \overline{x}_2 + (\sqrt{\frac{s_2^2}{2}}C_{(1)} - \sqrt{\frac{s_1^2}{2}}C_{(2)}), \overline{x}_1 - \overline{x}_2 + (\sqrt{\frac{s_2^2}{2}}C_{(2)} - \sqrt{\frac{s_1^2}{2}}C_{(1)})\right]$$

by considering the projection along the line of $\mu_1 - \mu_2$. The framework of GMB introduced in the end of Section 3 provides a justification for using this a-random interval. The posterior density functions for the two endpoints can be calculated analytically as follows:

$$\underline{h}(\mu_0|x) = \begin{cases} \frac{1}{\pi}\frac{a+b}{(\mu_0-\delta)^2+(a+b)^2}[1 - \frac{2}{\pi}\arctan\frac{\mu_0-\delta}{|a-b|}] & \text{for } a \neq b, \\ \frac{1}{\pi}\frac{4a}{(\mu_0-\delta)^2+4a^2}I(\mu_0 \leq \delta) & \text{for } a = b \end{cases}$$

and

$$\overline{h}(\mu_0|x) = \begin{cases} \frac{1}{\pi}\frac{a+b}{(\mu_0-\delta)^2+(a+b)^2}[1 + \frac{2}{\pi}\arctan\frac{\mu_0-\delta}{|a-b|}] & \text{for } a \neq b, \\ \frac{1}{\pi}\frac{4a}{(\mu_0-\delta)^2+4a^2}I(\mu_0 \geq \delta) & \text{for } a = b \end{cases}$$

where $a = \sqrt{s_1^2/2}, b = \sqrt{s_2^2/2}, \delta = \overline{x}_1 - \overline{x}_2$, $I(\cdot)$ is the indicator function.

6. The multinomial Distribution

The multinomial distribution

$$\text{Multinomia}(P_1, ..., P_K), \quad (P_k \geq 0 \text{ for } k = 1, ..., K \text{ and } \textstyle\sum_{k=1}^{K} P_k = 1)$$

over the sample space $\mathcal{X} \equiv \{1, ..., K\}$ is a practically useful model for analyzing categorical data as well as a theoretically interesting model for developing inferential methods. In practice, the presence of zero and small counts in some

categories is problematic. In this case, a fake small count, say 1/2, is often added to each cell before analysis. Serious attempts in literature do not appear satisfactory, especially when the number of categories with observed small counts is large. In Bayesian analysis, when the Jeffreys or flat prior is used, undesirable marginal posterior occurs when pooling is considered. The reference prior proposed by Berger and Bernardo (1992) has not completely solved this problem. Dempster (1966,1968,1972) studied the binomial and trinomial cases under the Dempster-Shafer framework, however, the general multinomial distribution is analytically intractable. The approach proposed by Walley (1996) using "imprecise Dirichlet" priors is interesting. Since it involves "don't know", Walley's method is closer to DS than to Bayes.

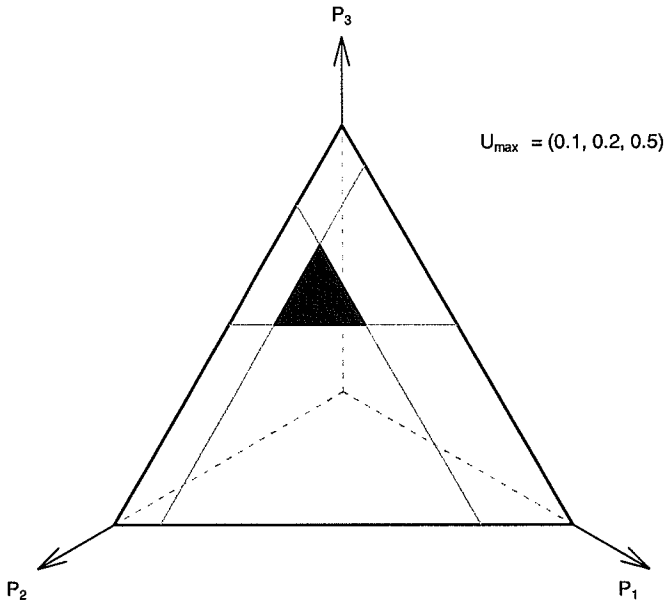

Figure 6.7: The 3-dimensional simplex (the outer large triangle) representing the parameter space of (P_1, P_2, P_3) in the trinomial distribution. The inner shaded triangle similar to the outer triangle represents the a-random region with $U^{(1)}_{\max} = 0.1, U^{(2)}_{\max} = 0.2$, and $U^{(3)}_{\max} = 0.5$, where it is assumed the observed counts $N_j > 0$ for $j = 1, 2$, and 3.

Here, we consider the MB posterior for $(P_1, ..., P_K)'$. Suppose that the data is thought to have been generated according to the following mechanism: draw $U_i \overset{iid}{\sim} U(0,1)$ and set $X_i = k$ if $0 \leq U_i - \sum_{j=1}^{k-1} P_j < P_k$ for $i = 1, \cdots, n$, $k = 1, \cdots, K$. Note that this is nothing but the sampling class of structures of the first kind (see Dempster (1966)). Denote by N_k the number of counts in the k-th category for $k = 1, \cdots, K$, it is evident that $N_k = \sum_{i=1}^{n} I(X_i = k)$ and $\sum_{k=1}^{K} N_k = n$, where $I(\cdot)$ denotes the indicator function.

We consider $V_i \equiv U_i - \sum_{j=1}^{X_i - 1} P_j$ (for $i = 1, \cdots, n$) instead of U_is, as the surrogate variables. The restrictions or permutations for U_is thus disappear when we consider V_is. Thus, the permutations are irrelevant and V_is are i.i.d. uniform random variables on $[0,1]$, subject to the feasibility constraints: $V_i \leq P_{X_i}$ for $i = 1, ..., n$. Aggregating V_is, we obtain for $k = 1, \cdots, K$:

$$V_j^{(k)} \leq P_k \ (j = 1, \cdots, N_k) \iff \max_j V_j^{(k)} \leq P_k$$

where $\{V_j^{(k)}\}_{j=1}^{N_k}$ is the collection of the V_i's with the corresponding X_i value equal to k. Let $U_{\max} = (\max_{1 \leq j \leq N_1} V_j^{(1)}, ..., \max_{1 \leq j \leq N_K} V_j^{(K)})'$, the vector of the lower bounds for the cell probabilities P_1, ..., P_K.

Routine algebraic operation leads to that

$$U_{\max} \overset{d}{=} (D_1, ..., D_K)$$

where $(D_0, D_1, ..., D_K)' \sim \text{Dirichlet}_{K+1}(1, N_1, \cdots, N_K)$. The "expanded" a-random vector $(D_0, D_1, ..., D_K)'$, with D_0 corresponding to "don't know", lies in the $(K+1)$-dimensional simplex. Thus, the a-random set for $(P_1, ..., P_K)'$ is a shrunk version of the K-dimensional simplex, the space of $(P_1, \cdots, P_k)$. Figure 6.7 displays such an a-random set in the case of $K = 3$.

Remark 6.1. The MB posterior, given in terms of the DS calculus, defines a DS-Dirichlet process. The DS-Dirichlet allows for simple and credible solutions to a class of non-parametric problems. For its frequentist and Bayesian counterpart, see Ferguson (1973).

Remark 6.2. In the case of $K = 2$, the MB posterior is the same as the DS posterior (see, for example, Dempster (1966) and Liu (2007a)).

7. Conclusion

In this article, we considered a new principle, called MB, for parametric inference. MB is proposed in the framework of the DS theory. Thus, the MB

principle can be viewed as a new way to build alternative DS models. The intuitive idea of MB is that the weaker the belief, the higher the credibility. It is typical that unlike the fiducial argument and DS theory, MB inference has no *a posteriori* independence assumptions and therefore is different from the existing inferential methods. It is shown that MB leads to simple and intuitively attractive results for the multinomial model and it does shrinkage automatically for the many-normal-mean problem. The MB principle also raises questions on "sufficiency" of the likelihood principle, which is critical to frequentist, Bayesian, and fiducial methods.

Like every new method, MB needs further intensive investigations, especially for multiparameter cases for which we have only touched a few examples. While the vacuous prior on the unknown permutation in MB allows for conservative inference, specification of credible DS-distribution for the unknown permutation for possible sharper inferential results without losing credibility deserves further study. Limited experimental results on a variety of statistical problems, including deconvolution, multiple testing, and variable selection, based on work with our collaborators are encouraging. We expect that future research on MB will be fruitful, leading us to a new era of the DS theory, especially for VHD statistical problems.

Acknowledgment

The authors are grateful to Professor Arthur P. Dempster for sharing his deep understanding of scientific inference as well as for his thoughtful comments and suggestions. The authorship order is alphabetic. The second author's research is supported by Purdue Research Foundation.

References

ARNOLD, B. C., CASTILLO, E., and SARABIA, J.M. (2001). Conditionally specified distributions: an introduction (with discussion). *Statist. Sci.* **16** 249-274.

BERGER, J. O. and BERNARDO, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79** 25-37

DAVID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference. *Annals of Statist.* **10** 1054-1067.

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355-374.

DEMPSTER, A. P. (1968). A generalization of Bayesian inference. *J. R. Statist. Soc. B* **30** 205-247.

DEMPSTER, A. P. (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Annals of Math. Statist.* **40** 953-969.

DEMPSTER, A. P. (1972). A class of random convex polytopes. *Ann. Math. Statist.* **43** 260-272.

DEMPSTER, A. P. (2007). The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning.* To appear.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117-130.

EFRON, B. (1992). Introduction to James and Stein (1961) estimation with quadratic loss. *Breakthroughs in Statistics* **1** 437-442.

FRASER, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53** 1-9.

FERGUSON, T. S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statist.*, **1** 209-230.

FISHER, R. A. (1959). *Statistical methods for scientific induction*, 2nd Ed. Hafner Publishing Company, New York.

GELMAN, A. and SPEED, T.P. (1993). Characterizing a joint probability distribution by conditionals. *J. R. Statist. Soc. B* **55** 185-188.

GELMAN, A. and SPEED, T.P. (1999). Corrigendum: Characterizing a joint probability distribution by conditionals. *J. R. Statist. Soc. B* **61** 483.

GELMAN, A., CARLIN, J. B., STERN, H. S., and RUBIN, D. B. (2004). *Bayesian Data Analysis.* 2nd edition. CRC press.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.* Univ. of California Press, 361-379.

KIM, S. and COHEN, A. S. (1998). One the Behrens-Fisher problem: a review. *J. Educational and Behavioral Statist.* **23** 356-377.

LINDLEY, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc. B* **20** 102-107.

LIU, C. (2007a). Discussion on "Modeling and Predicting and Predicting Probabilities with Outlooks" by M. Fygenson, *Statistica Sinica*, to appear.

LIU, C. (2007b). The Dempster-Shafer theory for single parameter univariate distributions. Technical Report, TR07-12, Department of Statistics, Purdue University.

SAVAGE, L. J. (1976, 1970 Fisher lecture). On rereading R. A. Fisher. *Annals of Statist.*, **4** 441-500.

SHAFER, G. (1976). *A mathematical theory of evidence.* Princeton University Press, Princeton, New Jersey.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the third Berkeley symposium on mathematical statistics and probability.* Univ. of California Press, 197-206.

STEIN, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30** 877-880.

WALLEY, P. (1996). Inferences from multinomial data: Learning about a bag of marbles (with discussion). *J. Roy. Statist. Soc. B.*, **58** 3-57.

WANG, Y. H. (2000). Fiducial interval: what are they? *American Statistician* **54** 105-111.

ZABELL, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369-387.

Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN, 47907-2067

E-mail: chuanhai@stat.purdue.edu

Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN, 47907-2067

E-mail: zhang10@stat.purdue.edu