Derivation of Mixture Distributions and Weighted Likelihood Function [1]

Xiaogang Wang [2] and James V. Zidek [3]

York University and University of British Columbia.

Abbreviated title:

Derivation of Mixtures and Weighted Likelihood

Abstract

In this article the widely used mixture distribution and a new device for inference, the weighted likelihood function, are derived within an information-theoretic framework. Mixture distributions are proved to be optima that minimize the entropy loss under certain constraints. The weighted likelihood is derived to use information from different samples from populations other than that under study to trade bias for precision and thereby yield inferential procedures, in particular estimates, that are more reliable than their classical counterparts.

AMS 2000 Classification: 62B10; 62E17; 62H12

Keywords: Euler-Lagrange equations; Relative entropy; Mixture distributions; Weighted likelihood.

1 Introduction

Origins of this paper can be found in Stein (1956) who showed that bias could be traded for precision. Moreover he showed that "strength" could be "borrowed" from data drawn independently from populations other than that about which inferences were to be made. Specifically, under certain reasonable conditions, if normal population means are to be estimated simultaneously from independent samples, then the sample averages can be outperformed in terms of expected combined squared-errors of estimation. Moreover, each of the improved mean estimators relies on the data from all other populations.

Stein's result had dramatic impact, in as much as it challenged conventional paradigms that supported use of the sample averages. Moreover, since the likelihood method had produced the

[1]This research is supported in part by the Natural Sciences and Engineering Research Council of Canada

[2]Xiaogang(Steven) Wang, Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada, M3J 1P3. *E-mail address: stevenw@mathstat.yorku.ca* .

[3]James V. Zidek, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2. *Email address: jim@stat.yorku.ca*

sample averages in the first place, while failing to produce Stein's superior alternative, it cast some doubt on the method itself. Can the likelihood be extended to yield Stein's result, more specifically the estimator of James and Stein (1961). That is the subject of this paper.

An answer to this question might seem unnecessary as it has long been known that a hierarchical empirical Bayes approach using the conventional likelihood explicates the James-Stein estimator. However, not all practitioners embrace the Bayesian approach. Moreover, its use may be impractical or even infeasible in applications where, not uncommonly, ten's of thousands of parameters may be encountered, making impossible, the problems of eliciting genuine (as opposed to ad-hoc or non-Bayesian e.g. improper) priors and carrying out the necessary computations, the speed of modern computers notwithstanding. Thus, deriving a likelihood based alternative seems worthwhile.

To derive an appropriate likelihood in Section 2, we take an approach suggested by Hu and Zidek (2002) based on the maximum entropy approach of Akaike (1977). More precisely, we seek a predictive distribution that minimizes the relative entropy subject to certain constraints. The latter are meant to capture the supposed "resemblance" of the population of inferential interest and others from which independent samples are available. Estimating the unknown population distributions in the spirit of Akaike leads us to the weighted likelihood.

The legitimacy of Akaike's approach has been amply demonstrated through such things as the derivation of the much-used AIC criterion and of Bayes rule, a rule that interestingly enough does not obtain from the classical rationality axioms for subjective probability itself. Most importantly for us, Akaike uses his approach to derive the classical likelihood function. Since that method points ineluctably to the version of the likelihood described next, we are confident that it is the correct choice among many for the role we wish it to play. Discovering the right choice (that is much discussed and applied elsewhere) is the central contribution of this paper.

To describe the likelihood we obtain, suppose from each population $i = 1, \ldots, m$ we independently observe identical, identically distributed random responses, $X_{i1}, \ldots, X_{in_i}$. Each of these responses may be a vector, all having the same dimension (1 in the univariate case). Each $X_{ij},\ j = 1, \ldots, n_i$ is assumed to have a density function $f_i(.; \theta_i),\ i = 1, \ldots, m$. Moreover, we assume the samples from the different populations are independent of each other. Finally, let $\mathbf{X_i} = (X_{i1}, ..., X_{in_i})^t$.

Suppose that only θ_1, an unknown vector of population 1 parameters, is of inferential interest.

Then for fixed $\mathbf{X} = \mathbf{x}$, the weighted likelihood (WL) turns out to be of the form,

$$\mathrm{WL}(\mathbf{x}; \theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{\lambda_i}, \tag{1}$$

where $\lambda = (\lambda_1, ..., \lambda_m)$ is the "weight vector" whose values are not implied by our implementation of Akaike's approach and must be specified in the context of specific applications.
A "maximum weighted likelihood estimator (WLE)", $\tilde{\theta}_1$, for θ_1 may be characterized as

$$\tilde{\theta}_1 = arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\mathbf{x}; \boldsymbol{\theta_1}).$$

To find the WLE, we may compute

$$log\mathrm{WL}(\mathbf{x}; \theta_1) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i \; log \; f_1(x_{ij}; \theta_1).$$

In turn, we may solve the *weighted likelihood equations:*

$$(\partial / \partial \theta_1) \; log \; \mathrm{WL}(\mathbf{x}; \theta_1) = 0.$$

(Note that the uniqueness of the WLE is not assumed.) So we see that weighted likelihood theory closely resembles (and formally includes) classical likelihood theory.

Although we derive the weighted likelihood in this paper, it has appeared elsewhere and we briefly summarize its history. It has been developed for a variety of purposes. Moreover, simple examples are easily constructed where it arises naturally. The multinomial likelihood is one such example, where the (adaptive or sample-based) weights arise naturally. One can also find examples in a Bayesian framework where it is a (classical) integrated likelihood. In spite of the WL's long history, it seems to have been suggested on an ad hoc basis. We are not aware of a "normative" argument like that given here (and in a special case by Hu and Zidek 2001, 2002), assuring that it is the correct choice.

2 Basics Elements

In this section, we introduce the framework from which the weighted likelihood will derive. Before doing so, we recall that for any density functions, $g_1(x)$ and $g_2(x)$, with respect to a sigma finite measure ν, define the relative entropy (in other words, the *Kullback-Leibler divergence*) as:

$$K(g_1, g_2) = E_1 \left(log \frac{g_1(X)}{g_2(X)} \right) = \int log \frac{g_1(x)}{g_2(x)} g_1(x) d\nu(x).$$

In this expression, $log(g_1(X)/g_2(X))$ is defined as $+\infty$ if $g_1(x) > 0$ and $g_2(x) = 0$, so the expectation could be $+\infty$. Although $log(g_1(x)/g_2(x))$ is defined as $-\infty$ when $g_1(x) = 0$ and $g_2(x) > 0$, the integrand, $log(g_1(x)/g_2(x))g_1(x)$ is defined as zero in this case. The properties of the entropy can be found in Csiszar(1975). In particular, the relative entropy is not symmetric and therefore not a distance.

To implement the Akaike approach, we assume the existence of the m population density functions introduced in Section 1 that are unknown and unknowable, playing purely conceptual roles. More specifically, assume σ-finite probability spaces $(\mathcal{X}, \mathcal{F}, \mu_i), i = 1, 2, ..., m$, with probability measures $\{\mu_i\}$ that are *absolutely continuous* with respect to one another. The existence of a σ-finite measure ν that dominates the μ_i follows. We take the f_i to be the Radon-Nikodym derivatives of $\{\mu_i\}$ with respect to ν.

The density functions $f_1, \ldots, f_m \in V$ are all assumed to be continuous where V is a reflexive Banach space. Although V can be quite arbitrary, we take $V = L^p = L^p(\mathcal{X}, \nu)$. It is known that the L^p spaces $(1 < p < \infty)$ are reflexive but that L^1 is not (cf. Royden 1988).

For $i = 1, 2, ..., m$, define

$$\mathcal{E}_i = \{g \in L^p : ||g - f_i||_p < C_i, \int f_i(x) log \frac{f_i(x)}{g(x)} d\nu(x) \leq a_i, \int g(x) d\nu(x) = 1, g(x) > 0.\} \qquad (2)$$

where $a_i \geq 0$ and C_i are constants. Furthermore, let

$$\mathcal{E} = \cap_{i=1}^{m-1} \mathcal{E}_i. \qquad (3)$$

We remark that the set $\mathcal{E}$ will be bounded with respect to the L^p norm and non-empty if the constraints are not too restrictive. The latter is assumed throughout.

According to the maximum entropy principle of Akaike (1977), the goodness of a particular model, q, as the predictive distribution of a random response, X, with true density p, is measured by the relative entropy,

$$B(p; q) = -I(p, q) = -\int p(x) log \frac{p(x)}{q(x)} d\nu(x).$$

We shall not be concerned with the information theoretic significance of the relative entropy; rather, we simply view it as a measure of the discrepancy between the two distributions.

We apply this measure by taking $f = f_1$, the population density of inferential interest. Were it known, we would take $p = f_1$, the best possible choice available. However, it is not known and this measure of performance has only a conceptual device to play.

The other population densities, $\{f_i\ , i = 2, \ldots, m\}$, are also unknown. Yet we believe them to "resemble" f_1 and that knowledge needs to be incorporated in selecting a predictive density. We interpret this to mean that any proposed model, g, must not diverge excessively from each of these other densities even as we pursue the ideal of minimizing that between g and f_1. More specifically, to fit into our relative entropy framework, we ask that $I(g, f_i) \leq a_i$ for constants $a_i,\ i = 1, 2, 3..., m$. The $\{a_i\}$ need not be in fact be known. Their role like that of the $\{f_i\}$ is purely conceptual and the assumption of their existence alone is enough to lead us to a form for the appropriate likelihood.

Thus, for a given set of density functions, $f_1(x)$ being primary, we seek a probability density function $g \in \mathcal{E}$ which minimizes $I(f_1, g) = \int f_1(x) log\ \frac{f_1(x)}{g(x)} d\nu(x)$ over all probability densities, g, satisfying

$$I(f_i, g) \leq a_i, \quad i = 1, 2, ..., m, \tag{4}$$

where a_i, $i = 2, 3, ..., m$, are non-negative constants.

3 Derivation of the Mixture Distribution and the Weighted Likelihood Function

To prove the existence of the optimal solution to the problem posed in the last section, we use the following result. Let $\mathcal{D}$ be a non-empty closed convex subset of L^p, $1 < p < \infty$. Let $g \in L^p$. We define $I(g) : L^p \rightarrow \mathcal{R}$. We are concerned with the minimization problem:

$$\inf_{g \in \mathcal{D}} I(g). \tag{5}$$

To avoid trivial cases, we assume that the function $I(g)$ is proper, *i.e.* it does not take the value $-\infty$ and is not identically equal to $+\infty$. We then have the following known result.

Theorem 3.1 *Assume that $I(g)$ is convex, lower semi-continuous and proper with respect to g. In addition, assume that the set $\mathcal{D}$ is bounded,*

so that there exist a constant M, say, such that

$$\sup_{g \in \mathcal{D}} I(g) < M. \tag{6}$$

*Then the minimization problem (**??**) has at least one solution in $\mathcal{D}$. The solution is unique if the function $I(g)$ is strictly convex on $\mathcal{D}$.*

Proof: (See, for example, Ekeland and Temam 1976, p 35.) $\diamond$

Let $I(g) = I(f_1, g) = \int f_1(x) log \frac{f_1(x)}{g(x)} d\nu(x)$ for some given density f. We minimize $I(g)$ on $\mathcal{E}$ which is defined by (**??**). It can be seen that $I(g)$ is a bounded non-negative strictly convex function with respect to g. It follows that $I(g)$ is continuous with respect to g (*c.f.* Lemma 2.1, Ekeland and Temam 1976). In fact, $I(g)$ is weakly lower semicontinous over L^p, $1 < p < \infty$ (cf. Theorem 1.2, Chapter 3, Dacorogna 1989). Finally, we may conclude from Theorem **??** that I attains its minimum value at a unique point in $\mathcal{E}$. We state this formally in the next Corollary.

Corollary 3.1 *For a given set of density functions $f_1, f_2, ..., f_m$, the minimization problem (**??**) has a unique solution.*

We now establish a necessary property of the optimal solution to the minimization problem.

Theorem 3.2 *For g^* to be the optimal solution to the minimization problem (**??**), it is necessary that it be a mixture distribution, i.e., that there exist non-negative constants $t_1^*, t_2^*, \ldots, t_m^*$ such that $\sum_{i=1}^{m} t_i^* = 1$, and*

$$g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x) \geq 0. \tag{7}$$

The previous theorem implies, in particular, that l_0 must have the same sign as every one of the multipliers l_j as well as 1 implying that all these multipliers are nonnegative, a fact that will play a role in ensuing developments. As well, note that the celebrated *Shannon-Kolmogorov Information Inequality* is a special case of this last result. To see this consider the minimization problem without any constraints, where we seek the optimal density function g^* that minimizes $I(f_1, g)$ for any given $f_1(x)$. According to Theorem **??**, the necessary condition for g^* to be the optimal solution is that

$$g^*(x) = t_1^* f_1(x).$$

Since $t_i^* = 0,\ i = 2, 3, .., m,\ t_1^* = 1$. It then follows that $g^*(x) = f_1(x),\ \ a.e.$.

By the proof of Theorem **??** and the Lagrange theorem, we may find the optimal density function, g^*, by minimizing

$$\begin{aligned} & \int f_1(x) log \frac{f_1(x)}{g(x)} d\nu(x) + l_0 \left(\int g(x) d\nu(x) - 1 \right) + \sum_{i=2}^{m} l_i \left(\int f_i(x) log \frac{f_i(x)}{g(x)} d\nu(x) - a_i \right) \\ = & -\left(\int f_1(x)\ log\ g(x) d\nu(x) + \sum_{i=2}^{m} l_i \int f_i(x)\ log\ g(x) d\nu(x) \right) + l_0 \int g(x) d\nu(x) \\ & + \left(\int f_1(x)\ log\ f_1(x) d\nu(x) + \sum_{i=2}^{m} l_i \left(\int f_i(x)\ log\ f_i(x) d\nu(x) - a_i \right) - l_0 \right). \end{aligned}$$

The last term in previous equation does not depend on the choice of g. Thus, the minimization problem considered is equivalent to maximizing over $\mathcal{E}$

$$\begin{aligned}&\int f_1(x)\ log\ g(x)d\nu(x) + \sum_{i=2}^{m} l_i \int f_i(x)\ log\ g(x)d\nu(x)\\ &- l_0 \int g(x)d\nu(x)\\ &= \sum_{i=1}^{m} d_i \int f_i(x)\ log\ g(x)d\nu(x) - l_0 \int g(x)d\nu(x)\end{aligned}$$

where $d_1 = 1,\ d_i = l_i,\ i = 2, 3, ..., m..$ Since the $\{d_i\}$ are non-negative, it follows by reasoning as in the proofs of Theorems **??** and **??** that the optimum may be found by maximizing

$$\sum_{i=1}^{m} d_i \int f_i(x)\ log\ g(x)d\nu(x)$$

over $\mathcal{E}$.

However, f_i's are unknown and we obtain the WL in the non-parametric case, by heuristic reasoning like Akaike has employed.

To that end, observe that any terms in the objective function that involve them must be estimated, the obvious estimator being

$$\sum_{i=1}^{m} d_i \int\ \log g(x)d\hat{F}_i(x),$$

where $\hat{F}_i$ denotes the empirical distribution function for population i=1,. . . ,m. Now we may argue as in the classical case of $i.i.d$ observables where the non-parametric MLE is shown to be the sample empirical distribution. Thus, we see that the optimum is degenerate and puts all of its unit mass on the sample points themselves. In other words the optimum is obtained by maximizing over $g_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$ with $\sum\sum g_{ij} = 1$ and $g_{ij} \geq 0$ the quantity,

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} g_{ij}^{d_i/n_i}.$$

We thus obtain the WL estimator of F_1 as a generalization of what Hu and Zidek (2001) called the relevance weighted empirical distribution, namely

$$\hat{F}_1 = \sum_{1=1}^{m} w_i \hat{F}_i$$

where $\hat{F}_i$ denotes the empirical distribution of the i-th sample and $w_i \propto d_i,\ i = 1, \ldots, m$ are non-negative weights that sum to 1. Thus, by this heuristic reasoning we obtain not only the non-

parametric WL but the WL estimator as well in explicit form. Although this estimate is rather "rough", it is the best that can be obtained without further restrictions.

A natural such restriction brings us to the parametric case where $g(\cdot) = f_1(\cdot|\theta_1)$, and we regard $\theta_1 \in \Theta$ represents a vector of population 1 parameters. Following the lines of the conventional Lagrangian argument given by Hu and Zidek (2002), we may express the optimization problem in Equation (4) differently, at least subject to the regularity conditions. To that end we make the following assumptions.

Assumptions.

1. Subject to the constraints imposed on the optimization problem in Equation (4), $\theta_1 \to I(g, f_1)$ has a unique maximum, θ_1^* in Θ.

2. For each $i = 2, \ldots, m$ the gradient of $g(\cdot) = \log f_1(\cdot|\theta_1)$ with respect to $|\theta_1$ exists a.e. $[\nu]$ and can be taken under the integral sign in $I(g, f_i)$.

Applying a the Lagrange argument (*c.f.* Beavis and Dobbs 1990) we obtain the following result.

Theorem 3.3 *Assume the* $\{\partial \log g/\partial\theta_{1i}\}$ *do not all lie in the hyperplane of functions orthogonal to some non-null element of the space spanned by the* $\{f_i,\ i = 2, \ldots, m\}$ *with respect to the inner product* $(f, h) = \int f h d\nu$. *Then*

$$\theta_1^* = arg \max_{\theta_1 \in \Theta} \sum_{i=1}^{m} d_i \int \log g(x; \theta) dF_i(x),$$

where the $\{d_i\}$ *represent Lagrange multipliers.*

However, as in Akaike's theory, the population distributions for the m populations are unknown and merely play a conceptual role. Thus, the previous theorem's value may primarily be qualitative, yielding some conceptual basis for the choice of the the family of acceptable parametric functions if the Lagrange result is to hold. Furthermore, to obtain a usable form of the objective function in the previous theorem, we need to proceed as in the nonparametric case above and estimate the unknown population distribution. This then gives us the parametric version of the likelihood obtained earlier. The estimate of the parameter of the optimal distribution would be found as

$$\arg \max_{\theta \in \Theta} \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta)^{d_i/n_i}.$$

This implies that the estimate of parameter of the optimal density is equivalent to finding the WLE derived from the weighted likelihood function if the functional form of the optimal density function is known.

The next theorem describes the relationships

between t_i and a_i, for any mixture density function $\sum_{i=1}^{m} t_i f_i(x)$ which satisfies the constraints (**??**) .

Theorem 3.4 *Suppose there exists* $\mathbf{a}^0 = (a_1^0, a_2^0, ..., a_m^0)^t$ *and* $\delta^0 = (\delta_1, \delta_2, ..., \delta_m)^t$ *such that there exists* $g_0(x) = \sum_{i=1}^{m} t_i f_i(x)$ *with* t_i *chosen as a function of* $\mathbf{a}$ *so that* g_0 *achieves equalities in the constraints (**??**) and* $\sum_{i=1}^{m} t_i = 1$ *for any* $\mathbf{a}$ *such that* $|a_i - a_i^0| < \delta_i^0$. *Then the* $\{t_i\}$ *are monotone functions of* a_i, *more precisely,*

$$\begin{aligned} \frac{\partial t_i}{\partial a_i} &\leq 0, \quad i = 2, ..., m, \\ \frac{\partial}{\partial a_i} \sum_{k \neq i} t_k &\geq 0, \ i = 2, ..., m. \end{aligned}$$

Moreover, the weights t_i *are all between* 0 *and* 1.

4 Related Works

In this section we describe some earlier works that although not directly related to the central topic of this paper, is nonetheless quite relevant.

The Kullback-Leibler divergence is also known as the entropy loss. James and Stein (1961)introduce it as a performance criterion in estimating the multinormal variance-covariance matrix. Brown (1968) and Haff (1980) used it to index the losses incurred in estimating both the multinomial variance-covariance matrix and its inverse. Ghosh and Yang (1982) introduced that loss when simultaneously estimating p-independent Binomial and multinomial proportions. Parsian and Nematollahi (1996) consider the estimation of scale parameter under entropy loss function. Trottini and Spezzaferri (2002) show that the criterion based on by logarithmic utility function for estimating the density function by San-Martitni and Spezzaferri (1984) is equivalent to the generalized predictive criterion using the relative entropy. It should also be noted that Bernardo (1979) shows the entropy is a loss function in a Bayesian framework.

The idea of finding a optimal solution with respect to relative entropy under constraints is related to the hypothesis testing for divergence outlined in Kullback (1959, Chapter 3). For any given true density f, the practitioner seeks a probability distribution that is "nearest to the true density with respect to the relative entropy or divergence. The true density, however, is unknown. Kullback (1956) proposes to find the optimal density function which achieved minimum relative

entropy subject to $\int T(x)g(x)dx = C$, where C is usually a multidimensional parameter. The constraint is employed to force $f_1(x)$ to satisfy some other desired characteristics. Although the true density function $f(x)$ is in fact unknown, we suppose in the spirit of configural polysmaplnig by Tukey (1987), Morgenthaler and Tukey (1991) and Easton (1991) that a set of density functions, $f_1, f_2, ..., f_m$ "span" a reasonable range of possible true densities for the observables. They are introduced in the context of robustness in order to find inferential procedures that can work well in the face of a wide variety of stochastic behaviors. Thus it is reasonable to impose the constraints as $T_i(x) = log\,(f_i(x)/g(x))$ so that the degree of resemblance of the optimal density to each of the density in the candidate set is reflected by relative entropy. Therefore in order to find the optimal predictive distribution, the desired density function should not only be associated with only one density but also with other candidate densities to a varying degree. Morgenthaler and Tukey (1991) have argued that this kind of approach is more realistic than the usual method of choosing one distribution, say normal, and then estimating the parameters. We also remark that the optimal solution in our analysis differs from the maximum entropy distribution, otherwise known as the Maxwell-Boltzmann distribution. The goal of using maximum entropy is the construction of a density function so that it possesses some desired properties. These two questions might seem to be identical on the surface. But they are quite different in nature. The proof and detailed discussions of the maximum entropy distribution can be found in Cover and Thomas (1991).

The WL extends the local likelihood of Tibshirani and Hastie (1987) since the restriction of the weights to indicators or more generally kernel functions in the local likelihood is relaxed in the WL setting. A detailed discussion of the local likelihood and associated properties can be found in Eguchi and Copas (1998). Versions of the weighted likelihood can be seen in a variety of contexts (*c.f.* Brillinger (1977), Rao (1991), Field and Smith (1994), Newton and Raftery (1994), Markatou, Basu and Lindsay (1998) and Hu and Rosenberger (2000)).

Following Hu (1997), Hu and Zidek (1995, 2001, 2002) extend the local likelihood to a more general setting but with the same aim, that of combining relevant information in samples from other populations thought to resemble that whose parameters are of interest. They call their WL the relevance weighted likelihood estimator (REWL). In other words, referring to the definition above, $f_2(\cdot;\theta_2), ..., f_m(\cdot;\theta_m)$ are thought to be "similar to" $f_1(.;\theta_1)$. We should add that in their extension of the REWL, Hu and Zidek (1995) also consider simultaneous inference for all the θ's.

The classical maximum likelihood estimator (MLE) has asymptotic properties that have not only pointed to good performance but as well, provided useful items for the statistical toolbox, for

example, approximate confidence intervals. Such features carry over to the WLE although these are proved elsewhere. Hu (1997) provides asymptotic theory under a paradigm resembling that of non-parametric regression and function estimation. There, information about θ_1 builds up because the number of populations grows with increasingly many in close proximity to that of θ_1. However, this paradigm does not seem natural in many contexts. So in contrast to Hu, Wang, van Eeden and Zidek (2001) suppose a fixed number of populations with an increasingly large number of observations from each. Under this paradigm, they derive an alternative large sample theory for the WLE. In practice, the weights in the WLE may need to estimated using the data. The asymptotic properties for this case are given in Wang (2001).

Applications of the WLE readily found elsewhere (Hu and Rosenberger 2000; Hu, Rosenberger and Zidek 2000). Hu and Zidek (2001) show how the WLE can be used to predict the number of goals scored in ice hockey when Vancouver's NHL team plays Calgary's. In fact, the outcomes of sports competitions provide a particularly apt domain of application for the WLE, since typically a particular pair of teams will meet only seldom during a single season. However, much "relevant" information comes from their encounters with other teams and that is the

sort of information that makes the WLE work so well in the Hu-Zidek application. Of course, other methods may well be available, but the simplicity of the WLE makes it very attractive. To demonstrate use of the WLE in this paper, we give, as an example in Section 4, the estimation of the success probability of the negative binomial distribution. Concluding remarks appear in Section 5.

The empirical likelihood (Owen 2001) points to an approach intermediate between non-parametric and parametric ones we have adopted in this paper. One such approach may be found in an undated working paper of Kitamura, Tripath and Ahn found at the URL: www.ssc.wisc.edu/ gtripath/working-papers/cmmel-web.pdf. This seems an interesting direction for future work.

5 Appendix

Proof Theorem ??: For the optimal density g^* whose existence is assured, we may without loss of generality assume that the constraints are binding, *i.e* that $I(f_i, g^*) = a_i$, $i = 2, 3, ..., m$ since by reducing the non-binding a's if necessary we obtain the same optimum. Thus the optimization problem with solution g^* can be re-formulated in the context of calculus of variations as follows

$$\min_{g \in \mathcal{E}} I(g) = \min_{g} \int f_1(x)\, log \frac{f_1(x)}{g(x)}\, d\nu(x)$$

where $g \in \mathcal{E}$ means:

$$\int f_i(x)\ log\frac{f_i(x)}{g(x)}\ d\nu(x) = a_i,\ \ i = 2, ..., m;$$
$$\int g(x)d\nu(x) = 1\ \ and\ \ g(x) \geq 0.$$

Define $\psi(x, g) = f_1(x)log\frac{f_1(x)}{g(x)} + l_0 g(x) + \sum_{k=2}^{m} l_k f_k(x) log\frac{f_k(x)}{g(x)}$. Since $\psi(x, g)$ is continuous with respect to g, by an elementary theorem in the calculus of variations (see, for example, Giaquinta and Hildebrandt 1996) it follows that a necessary condition for g^* to be the optimal solution is that it satisfies the *Euler-Lagrange* equation, i.e.

$$\nabla_g \psi - \frac{\partial}{\partial x}(\nabla_{g'} \psi) = 0, \tag{8}$$

where ∇_g and $\nabla_{g'}$ are the derivative operators with respect to g and $g^{'}$ respectively and l_k suitably chosen constants, the so-called "Lagrange multipliers". Notice that $\psi(x, g)$ is not a function of $g^{'}$. That implies $\nabla_{g'} \psi = 0$. Thus *Euler-Lagrange* equation becomes $\nabla_g \psi = 0$. It follows that

$$-\frac{f_1}{g} + l_0 - \sum_{k=2}^{m} l_k \frac{f_k}{g} = 0.$$

We then have

$$g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x),$$

where $t_1^* = 1/l_0, t_i^* = l_k/l_0, k = 2, ..., m$.

The sum of the t_i^*'s must be 1 since $g^* \in \mathcal{E}$ and hence $1 = \int g^*(x)d\nu(x) = \sum_{k=1}^{m} t_k^*$. Likewise,

$$g^*(x) = \sum_{k=1}^{m} t_k^* f_k(x) \geq 0$$

since g^* must be in $\mathcal{E}$ by Corollary **??**.

Finally, we observe that the $\{t_i^*\}$ must be nonnegative for if not we could make $\sum_{k=1}^{m} t_k^* f_k(x)$ uniformly larger by truncating any negative weights to zero and renormalzing the remaining weights so that they sum to 1. The result would satisfy the constraints while reducing the objective function. Hence the original solution could not have been optimal, a contradiction. This completes the proof.

$\diamond$

Proof of Theorem ??:

Let $\phi_i(x) = f_i(x) - f_1(x), i = 2, ..., m$. Then,

$$g_0(x) = f_1(x) + \sum_{k=2}^{m} t_k \phi_k(x)$$

$$and \int \phi_i(x) d\nu(x) = 0, \qquad i = 2, ..., m.$$

It follows that,

$$f_i(x) = g_0(x) + \phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x) \geq 0.$$

This implies that

$$-[\phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)] \leq g_0(x). \tag{9}$$

Since g_0 satisfies the constraints (**??**), it follows that, for $2 \leq i \leq m$

$$\begin{aligned}
\frac{\partial a_i}{\partial t_i} &= \frac{\partial}{\partial t_i}\Big[\int f_i(x)\ log\ \frac{f_i(x)}{g_0(x)}\ d\nu(x)\Big] \\
&= \frac{\partial}{\partial t_i}\Big[\int f_i(x)\ log\ \frac{f_i(x)}{\sum_{k=1}^{m} t_k f_k(x)}\ d\nu(x)\ \Big] \\
&= -\int f_i(x) \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) \\
&= -\int [g_0(x) + \phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)] \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) \\
&= -\int g_0(x)\ \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) - \int\ [\phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)]\ \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) \\
&\leq -\int \phi_i(x) d\nu(x) - \int\ [\phi_i(x) - \sum_{k=2}^{m} t_k \phi_k(x)]\ \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) \\
&\leq -\int \phi_i(x) d\nu(x) + \int g_0(x)\ \frac{\phi_i(x)}{g_0(x)}\ d\nu(x) \qquad by \quad (\mathbf{??}) \\
&= 0.
\end{aligned}$$

Therefore, it follows that, for $i = 2, ..., m$,

$$\frac{\partial t_i}{\partial a_i} = \frac{1}{\frac{\partial a_i}{\partial t_i}} \leq 0.$$

It also follows that

$$\frac{\partial}{\partial a_i} \sum_{k \neq i} t_k \geq 0$$

since $t_1 + t_2 + ... + t_m = 1$.

Note that if we set $a_i = 0$, then $t_i = 1$; if $a_i = \infty$, then $t_i = 0$. Since t_i is a monotone function of a_i for any fixed $a_j, i \neq j$, it follows that $0 \leq t_i \leq 1$, $i = 1, 2, ..., m$. $\diamond$

References

Akaike, H. (1977). On entropy maximization principle. In: P.R. Krishnaiah (Ed.), *Applications of Statistics*. Amsterdam: North-Holland, 27-41.

Brillinger, D.R. (1977). Discussion of Stone (1977). *Ann Statist*, 5, 622-623.

Bernardo, J.M. (1979). Expected information as expected utility, *Ann Statist*, 7, 686-690.

Brown, L.D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann Math Statist*, 37, 1087-1136.

Cover, T.M., Thomas, J.A. (1991). *Elements of Information Theory*. New York: Wiley.

Csiszar, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann Statist*, 3, 146-158.

Dacorogna, B. (1989). *Direct methods in the calculus of variations*. New York: Springer-Verlag.

Easton, G.S. (1991). Compromise maximum likelihood estimators for Location. *JASA*, 83, 1051-1073.

Ekeland, I., Temam, R. (1976). *Convex analysis and variational problems*. New York: American Elsevier Publishing Company, Inc..

Eguchi, S., Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *JRSS, Ser B*, 60, 709-724.

Field, C., Smith, B. (1994). Robust estimation - a Weighted maximum likelihood approach, *International Statist Rev*, 62, 405-424.

Giaquinta, M., Hildebrandt, S. (1996). *Calculus of variations*, New York: Springer-Verlag.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann Statist,* 8, 586-597.

Ghosh, M., Yang, M.C. (1988). Simultaneous estimation of the multivariate precision matrix. *Ann Statist*, 16, 278-191.

Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators, *Can J Statist*, 25, 45-59.

Hu, F., Rosenberger, W.F. (2000). Analysis of time trends in adaptive designs with applications to a neurophysiology experiments. *Statist in Medicine*, 19, 2067-2075.

Hu, F., Zidek, J.V. (1995). Incorporating relevant sample information using the likelihood. Technical Report No. 161 Dept. of Statistics, The University of British Columbia, Vancouver, B.C., Canada.

Hu, F., Zidek, J.V. (2001). The relevance weighted likelihood with applications. In: Ahmed, S.E., Reid, N. (Eds.), *Empirical bayes and likelihood inference*, New York: Springer-Verlag.

Hu, F. and Zidek, J. V.(2002) The weighted likelihood. *Can J Statist*,30, 347-371.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc 4th Berkeley symp math statist prob,* 1, 361-379, Berkely: University of California Press.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Lindsay, B.G. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS regional conference series in probability and statistics*, Vol 5, Hayward.

Markatou, M, Basu A., Lindsay, B.G. (1998). Weighted likelihood equations with bootstrap root search. *JASA*, 93, 740-750.

Morgenthaler, S., Tukey, J.W. (1991). *Configural polysampling: a route to practical robustness*. New York: Wiley.

Newton, M.A., Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *JRSS, Ser B*, 56, 3-48.

Rao, P.B.L.S. (1991). Asymptotic theory of weighted maximum likelihood estimation for growth models. In: Prabhu, N., Vasawa, I.V. (Eds.), *Statistical inference for stochastic processes*, New York: Decker, 183-208.

Royden,H. L. (1988). *Real analysis*. New York: Prentice Hall.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *JRSS, Ser B,* 57, 99-138.

Stein, C. (1956). Inadmissibilityof the usual estimator for the mean of a multivariate normal distribution. *Proc 3rd Berkeley symp on math statist prob*, 1, 197–206.

Tibshirani, R. and Hastie, T. (1987). Local likelihood of statistical predictions. *JRSS, Ser B*, 36 111-147.

Trottini, M. and Spezzaferri, F. (2002). A generalized predictive criterion for model selection. *Can J Statist*, 30, 79-96.

Wang, X. (2001). *Weighted likelihood estimation*. Ph.D. Dissertation. Dept Statistics, U British Columbia.

Wang, X., van Eeden, C., Zidek, J.V. (2003). Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*. To appear.