Ann. Inst. Statist. Math.
Vol. 57, No. 4, 687–701 (2005)
©2005 The Institute of Statistical Mathematics

# DERIVATION OF MIXTURE DISTRIBUTIONS AND WEIGHTED LIKELIHOOD FUNCTION AS MINIMIZERS OF KL-DIVERGENCE SUBJECT TO CONSTRAINTS

XIAOGANG WANG[1] AND JAMES V. ZIDEK[2]

[1] *Department of Mathematics and Statistics, York University, 4700 Keele Street, ON, Canada M3J 1P3*
[2] *Department of Statistics, University of British Columbia, 333-6356 Agriculture Road, BC, Canada V6T 1Z2*

(Received May 26, 2003; revised July 7, 2004)

**Abstract.** In this article, mixture distributions and weighted likelihoods are derived within an information-theoretic framework and shown to be closely related. This surprising relationship obtains in spite of the arithmetic form of the former and the geometric form of the latter. Mixture distributions are shown to be optima that minimize the entropy loss under certain constraints. The same framework implies the weighted likelihood when the distributions in the mixture are unknown and information from independent samples generated by them have to be used instead. Thus the likelihood weights trade bias for precision and yield inferential procedures such as estimates that can be more reliable than their classical counterparts.

*Key words and phrases*: Euler-Lagrange equations, relative entropy, mixture distributions, weighted likelihood.

## 1. Introduction

To introduce the results we present in this paper, suppose a statistician, S, is required to predict $y$, the realized value of a random variable $Y$ with unknown probability density function (PDF), say $f$. Instead of a point prediction, S is allowed to state the prediction as a predictive PDF, say $g$. Finally, S receives a reward, $\log g(y)$. Thus, by selecting $g$, S expects a return of $\int f(y) \log g(y) dy$. How should $g$ by selected?

Although it is not the framework employed in this paper, we now state an approach taken by both Frequentists (Fs) and Bayesians (Bs) for different reasons. Suppose $f = f_j$ with probability $\pi_j$, $j = 1, \dots, m$, $\pi_j \geq 0$, $\sum \pi_j = 1$, where for simplicity we suppose $m = 2$. The expected return then becomes $\int (\pi_1 f_1(y) + \pi_2 f_2(y)) \log g(y) dy$ and a familiar calculation then leads to an optimal $g$,

$$g^*(y) = \pi_1 f_1(y) + \pi_2 f_2(y). \tag{1.1}$$

The Fs call $g^*$ a mixture model, the Bs a model average. In the aforementioned context leading to equation (1.1) the $f_1$ and $f_2$ are usually viewed as competitors. For example, $f_1$ and $f_2$, respectively, could represent the PDF of a surficial geological measurement distribution given the presence or absence of specified sub-surface deposits.

In contrast, in this paper's context, $f_1$ and $f_2$ play the roles of complementary models thought to resemble each other and the true density function of interest. For example, they could be the unknown population PDF's for a response measured on two recent surveys while $f$ represents that of the population to be measured in the near future. Although the population would have changed somewhat from one survey to the next, S would view the density functions as being quite similar. S's quest for an optimum $g$ leads us in the Section 2 to formalize the notion of "similarity" in a new paradigm rooted in the celebrated maximum entropy criterion of Akaike. If the true density function $f$ is unknown, the resulting optimum predictor is shown in that section to be another mixture model

$$g^* = \pi_1 f_1 + \pi_2 f_2, \tag{1.2}$$

where $\pi_i \geq 0$, $i = 1, 2$, and $\pi_1 + \pi_2 = 1$. The above mixture is obtained from quite a different rationale than the one in equation (1.1). Here, the $\pi_i$'s are derived from the assumptions made in Section 2 and they can only be determined if $f_i$'s and their relationships to the true density function $f$ are both known.

If, more realistically, the $f_i$'s were unknown, then little progress could be made without having data obtained by repeatedly measuring random variables having these PDF's. Heuristically, one might then expect an estimate of $g^*$ to be found by using the PDFs obtained by differentiating the empirical cumulative distribution functions (CDFs) associated with the respective samples. In fact, that is precisely the estimate found in Section 3 in the non-parametric case where $g^*$ does not have a specified parametric form. Moreover, as we will see in Section 3, that natural estimator turns out to be none other than the weighted likelihood estimator of the mixture model. In that section, we present our principal result, a fundamental relationship between the arithmetically averaged predictor in equation (1.2) and the geometrically averaged weighted likelihood described below.

However, before getting into a more detailed description of our results, we would note that the origin of this paper lies in Stein (1956) who showed that bias could be traded for precision. Moreover he showed that "strengths" could be "borrowed" from data drawn independently from populations other than the population of inferential interest. 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 populations.

Stein's result challenged conventional paradigms which supported the use of the sample averages. Moreover, since the likelihood method produces the sample average in the first place, while failing to produce Stein's superior alternative, it casts some doubts 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.

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). The legitimacy of Akaike's approach has been amply demonstrated through, for example, its generation of the celebrated AIC criterion. Akaike also used his approach to derive the classical likelihood function and thereby provided us with a blueprint for our construction of the weighted likelihood, a central contribution of this paper.

To describe that likelihood, we suppose for simplicity that $f = f_1$. Moreover, assume that from each of the populations associated with the $f_i$'s, we observe independent and identically distributed random variables, $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, m$. Each of these random variables may be a vector, all having the same dimension. Each $X_{ij}$, $j = 1, \ldots, n_i$, is assumed to have a density function $f_i$, $i = 1, \ldots, m$. Moreover, we assume the samples from the different populations are independent of each other. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in_i})^t$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m)$.

Suppose only $\theta_1$, an unknown vector of parameters of the first population, is of inferential interest. Moreover, we initially limit our search for an optimal predictive PDF to a parametric class, $g(\cdot) = f_1(\cdot \mid \theta_1)$ (although we briefly consider the non-parametric alternative as well). Then for fixed $\boldsymbol{X} = \boldsymbol{x}$, we derive in Section 3, the weighted likelihood (WL) as

$$\mathrm{WL}(\boldsymbol{x}; \theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{\lambda_i}, \tag{1.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ 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 $\theta_1$ is defined as

$$\tilde{\theta}_1 = \arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\boldsymbol{x}; \theta_1). \tag{1.4}$$

To find the WLE, we may compute

$$\log \mathrm{WL}(\boldsymbol{x}; \theta_1) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i \log f_1(x_{ij}; \theta_1). \tag{1.5}$$

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

$$(\partial / \partial \theta_1) \log \mathrm{WL}(\boldsymbol{x}; \theta_1) = 0. \tag{1.6}$$

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

The weighted likelihood (WL) has been developed for a variety of purposes and it can have a variety of forms, as seen in our summary of Section 4. One example: the multinomial likelihood where sample-based weights (we call "adaptive") arise naturally. In a Bayesian framework it can arise as an integrated likelihood. In spite of the WL's long history, it seems to have been suggested in specific instances on an *ad hoc* basis. We are not aware of any "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

For density functions, $g_1(x)$ and $g_2(x)$, with respect to a $\sigma$−finite measure $\nu$, *Kullback-Leibler* divergence is defined as:

$$KL(g_1, g_2) = E_1 \left( \log \frac{g_1(X)}{g_2(X)} \right) = \int \log \frac{g_1(x)}{g_2(x)} g_1(x) d\nu(x). \tag{2.1}$$

In this expression, $\log(g_1(x)/g_2(x))$ is defined as $+\infty$, if $g_1(x) > 0$ and $g_2(x) = 0$. Therefore 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. 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.

The properties of the entropy can be found in Csiszar (1975) and Cover and Thomas (1991). In particular, the relative entropy is not symmetric and therefore not a distance. The Kullback-Leibler divergence has also been known as the entropy loss. James and Stein (1961) introduced it as a performance criterion in estimating the multinormal variance-covariance matrix. Brown (1966) and Haff (1980) used it to index the losses incurred in estimating both the multinomial variance-covariance matrix and its inverse. Ghosh and Yang (1988) introduced this as a loss function for simultaneously estimating $p$-independent binomial and multinomial proportions. Parsian and Nematollahi (1996) considered the estimation of scale parameter under entropy loss function. Trottini and Spezzaferri (2002) showed that the criterion based on logarithmic utility function for estimating the density function by San Martini and Spezzaferri (1984) is equivalent to the generalized predictive criterion using the relative entropy. Bernardo (1979) showed the entropy is a loss function in a Bayesian framework.

According to the maximum entropy principle of Akaike (1977), the goodness of a particular model $g$, as the predictive distribution of a random response, $X$, with true density $f$, is measured by the *Kullback-Leibler* divergence (relative entropy),

$$I(f,g) = KL(f;g) = \int \log \frac{f(x)}{g(x)} f(x) d\nu(x). \tag{2.2}$$

The above distance is minimized if we set $g(x) = f(x)$ for all $x$.

It is rarely possible to assume that an underlying distribution is exactly characterized by a proposed statistical model $g(x;\theta)$. It is more reasonable to assume that the proposed statistical model lies in a close proximity to the true underlying distribution. One neighborhood enveloping the model is defined by Eguchi and Copas (1998) as

$$N_f(\epsilon) = \cup_{\theta\in\Theta}\{g : I(f,g) \leq \epsilon\}, \tag{2.3}$$

where $\epsilon \geq 0$.

We further assume the existence of the $m$ population density functions that are unknown and they play purely conceptual roles. More specifically, assume $\sigma$-finite probability spaces $(\Omega, \mathcal{F}, \mu_i)$, $i = 1, 2, \ldots, m$, with probability measures $\mu_i$'s that are *absolutely continuous* with respect to one another. The existence of a $\sigma$-finite measure $\nu$ that dominates the $\mu_i$'s then follows. We take the $f_i$ to be the Radon-Nikodym derivatives of $\mu_i$ with respect to $\nu$ for $i = 1, 2, \ldots, m$.

We apply this measure by taking $f = f_1$, the population density of inferential interest. If it were known, we should set $g = f_1$, the best choice available, assuming the problem is well-posed so that this choice satisfies the constraints. If, more realistically, it were not known as we now suppose, this measure of performance would play only a conceptual role. The other population densities, $f_i$, $i = 2, \ldots, m$, are also assumed to be unknown. However, suppose we believe they "resemble" the density function of interest $f_1$, then this knowledge should to be incorporated in selecting a predictive density. If $f_1$ is considered as the primary and "closest" statistical model to the true underlying distribution, the incorporation of other available distributions would allow the proposed model $g$ to capture some important characteristics of the true underlying distribution that might be missed by the primary model $f_1$. We interpret this to mean that any proposed model, $g$, must not diverge excessively from each of these other densities even as we minimize the difference between $g$ and $f_1$.

More specifically, to fit into our relative entropy framework, we require that $I(f_i, g) \leq a_i$ for constants $a_i$, $i = 2, 3, \ldots, m$. The $a_i$ represents the magnitude of resemblance between the "best" model and a candidate model $f_i$, $i = 2, \ldots, m$. In fact, the $a_i$'s might not be known. Their roles like that of the $f_i$'s are 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 $g$ which minimizes $I(f_1, g) = \int f_1(x) \log \frac{f_1(x)}{g(x)} d\nu(x)$ over all probability densities satisfying

$$I(f_i, g) \leq a_i, \qquad i = 2, \ldots, m, \tag{2.4}$$

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

## 3. Derivation of the mixtures and weighted likelihoods

### 3.1 *Derivation of the mixture distributions*

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 $V = L^p = L^p(\Omega, \nu)$. It is known that the $L^p$ spaces $(1 < p < \infty)$ are reflexive but that $L^1$ is not (see Royden (1988) for example).

For $i = 2, \ldots, m$, we define

$$\mathcal{E}_i = \left\{ g \in L^p : \|g - f_i\|_p < C_i, \int f_i(x) \log \frac{f_i(x)}{g(x)} d\nu(x) \leq a_i, \right. \\ \left. \int g(x) d\nu(x) = 1, g(x) > 0 \right\}, \tag{3.1}$$

where $a_i \geq 0$ and $C_i$, $i = 2, 3, \ldots, m$, are non-negative constants. Furthermore, we define

$$\mathcal{E} = \cap_{i=2}^{m} \mathcal{E}_i. \tag{3.2}$$

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.

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 $g \in L^p$ and $J(g) : L^p \to \mathcal{R}$ denote a general mapping. We are interested in the following minimization problem:

$$\inf_{g \in \mathcal{D}} J(g). \tag{3.3}$$

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

THEOREM 3.1. *Assume that $J(g)$ is convex, lower semi-continuous and proper with respect to $g$. In addition, assume that the set $\mathcal{D}$ is bounded, so that there exists a constant $M$ say, such that*

$$\sup_{g\in\mathcal{D}} J(g) < M, \tag{3.4}$$

*then the minimization problem defined by equation* (3.3) *has at least one solution. Furthermore, the solution is unique if the function $J(g)$ is strictly convex on $\mathcal{D}$.*

The proof of the above theorem can be found in Ekeland and Temam (1976).

Let $J(g) = I(f_1, g)$ for a given density $f_1$. We minimize $I(f_1, g)$ with respect to $g$ on $\mathcal{E}$ as defined by equation (3.2). It can be seen that $I(f_1, g)$ is a bounded non-negative strictly convex function with respect to $g$. It follows that $I(f_1, g)$ is continuous with respect to $g$ (Lemma 2.1, Ekeland and Temam (1976)). In fact, $I(f_1, g)$ is weakly lower semi-continous with respect to $g$ over $L^p$, $1 < p < \infty$ (Theorem 1.2, Dacorogna (1989) for example). We conclude from Theorem 3.1 that $I(f_1, g)$ attains its minimum value at a unique point in $\mathcal{E}$ for any given density $f_1$.

COROLLARY 3.1. *For a given set of density functions $f_1, f_2, \ldots, f_m$, the minimization problem defined by equation* (2.4) *has a unique solution.*

We now establish a necessary property of the optimal solution to the minimization problem defined by equation (2.4) and thereby obtains equation (1.2).

THEOREM 3.2. *For $g^*$ to be the optimal solution to the minimization problem defined by equation* (2.4), *it is necessary that it be a mixture distribution, i.e., that there exist non-negative constants $\pi_1, \ldots, \pi_m$ such that $\sum_{i=1}^m \pi_i = 1$, and*

$$g^*(x) = \sum_{i=1}^{m} \pi_i f_i(x) \geq 0. \tag{3.5}$$

Note that the celebrated *Shannon-Kolmogorov Information Inequality* is a special case of this last result. To see this, consider the minimization problem defined by equation (2.4) without any constraint. Thus we seek the optimal density function $g^*$ that minimizes $I(f_1, g)$ for any given $f_1$. According to Theorem 3.2, the necessary condition for $g^*$ to be the optimal solution is $g^*(x) = \pi_1 f_1(x)$. Since $\pi_1 = 1$ and $\pi_i = 0$, $i = 2, 3, \ldots, m$, it then follows that $g^*(x) = f_1(x)$ for all $x$.

The previous theorem also states that the optimal density is actually a mixture of all the available densities, $f_1, f_2, \ldots, f_m$. The weight or proportion $\pi_i$ should reflect the importance of each density. The nature of relative entropy means that a smaller value of $a_i$ for a particular density corresponds to greater importance or resemblance to the true density.

In our framework, the importance of each density is expressed or controlled by $a_i$, although the relationship between the $a$'s and the $\pi$'s is neither simple nor of much practical value. However, the next theorem gives us some qualitative understanding of that relationship. In fact, it describes the relationships between the weight and $a_i$ for any mixture density function which satisfies the constraints defined by equation (2.4).

THEOREM 3.3. *Suppose there exists* $\boldsymbol{a}^0 = (a_2^0, \ldots, a_m^0)^t$ *such that there exists* $g_0(x) = \sum_{i=1}^m t_i f_i(x)$ *with* $t_i$ *chosen as a function of* $\boldsymbol{a}^0$ *so that* $g_0$ *achieves equalities in the constraints defined by equation* (2.4) *and* $\sum_{i=1}^m t_i = 1$. *In addition, there exists* $\boldsymbol{\delta}^0 = (\delta_2, \ldots, \delta_m)^t$ *such that* $|a_i - a_i^0| < \delta_i^0$, $i = 2, \ldots, m$, *for any* $\boldsymbol{a}$. *Then the* $t_i$*'s are monotone functions of* $a_i$, *more precisely,*

$$\frac{\partial t_i}{\partial a_i} \leq 0, \qquad i = 2, \ldots, m,$$

$$\frac{\partial}{\partial a_i} \sum_{k \neq i} t_k \geq 0, \qquad i = 2, \ldots, m.$$

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

### 3.2 *Derivation of weighted likelihood functions*

By the Lagrange theorem, the minimization problem defined by equation (2.4) is equivalent to seeking the optimal density $g^*$ which minimizes the following:

$$S(g) = \int f_1(x) \log \frac{f_1(x)}{g(x)} d\nu(x) + l_0 \left( \int g(x) d\nu(x) - 1 \right) + \sum_{i=2}^m l_i \left( \int f_i(x) \log \frac{f_i(x)}{g(x)} d\nu(x) - a_i \right). \tag{3.6}$$

Rewrite $S(g)$ as follows:

$$\begin{aligned} S(g) = &- \left( \int f_1(x) \log g(x) d\nu(x) + \sum_{i=2}^m l_i \int f_i(x) \log g(x) d\nu(x) \right) \\ &+ l_0 \int g(x) d\nu(x) \\ &+ \left( \int f_1(x) \log f_1(x) d\nu(x) \right. \\ &+ \left. \sum_{i=2}^m l_i \left( \int f_i(x) \log f_i(x) d\nu(x) - a_i \right) - l_0 \right). \end{aligned} \tag{3.7}$$

Thus, the minimization problem considered is equivalent to maximizing the following

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

where $d_1 = 1$, $d_i = l_i$, $i = 2, 3, \ldots, m$. Theorem 3.2 implies, in particular, that $l_0$ must have the same sign as every one of the multipliers $l_j$'s as well as 1 implying that all these multipliers are nonnegative. Since the $d_i$'s are non-negative, thuse the optimum can be found by maximizing

$$\sum_{i=1}^m d_i \int f_i(x) \log g(x) d\nu(x). \tag{3.8}$$

Suppose a functional form for $g$ can be prescribed, i.e. assume $g(\cdot) = f_1(\cdot \mid \theta_1)$ where $\theta_1 \in \Theta$ represents a parameter vector for the population of interest. To that end, we make the following assumptions:

(i) Subject to the constraints imposed on the optimization problem in equation (2.4), the function $\theta_1 \rightarrow I(f_1, g)$ has a unique maximum, $\theta_1^*$ in $\Theta$.

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

By applying the Lagrange argument (c.f. Beavis and Dobbs (1990)), we then obtain the following result.

THEOREM 3.4. *Assume the* $\frac{\partial \log g(x|\theta_1)}{\partial \theta_{1k}}$, $k = 1, 2, \ldots, d$, *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* $\langle f, h \rangle = \int f(x)h(x)d\nu(x)$. *Then the unique optimum satisfies the following*:

$$\theta_1^* = \arg\max_{\theta_1 \in \Theta} \sum_{i=1}^{m} d_i \int \log f_1(x; \theta) dF_i(x), \tag{3.9}$$

*where the* $d_i$*'s represent Lagrange multipliers.*

However, as in Akaike's theory, the distributions for the $m$ populations are unknown and merely play a conceptual role. Thus, the role of the previous theorem is qualitative and it limits our choice of the family of acceptable parametric functions if the Lagrange result is to hold. We can then obtain the WL in the parametric case by heuristic reasoning like Akaike employed. By replacing the distribution functions by their empirical counterparts, we then seek $\theta_1$ which maximizes the following:

$$\sum_{i=1}^{m} d_i \int \log f_1(x; \theta_1) d\hat{F}_i(x), \tag{3.10}$$

where $\hat{F}_i$ denotes the empirical distribution function for population $i = 1, \ldots, m$. This then gives us the parametric version of the likelihood defined earlier. The estimate of the parameter of the optimal distribution can then be derived from

$$\tilde{\theta}_1 = \arg\max_{\theta \in \Theta} \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(X_{ij}; \theta)^{d_i/n_i}. \tag{3.11}$$

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. Finally, $g^*(\cdot) = f_1(\cdot \mid \tilde{\theta}_1)$ provides the required predictive PDF derived from all the samples. It, unlike its non-parametric counterpart below, is not in mixture model form.

If there were no constraints imposed, then $d_i = 0$ for $i = 2, 3, \ldots, m$. Thus the goal is minimize $I(f_1, g)$. It then follows that the weighted likelihood can be simplified to

$$\prod_{j=1}^{n_1} f_1(X_{1j}; \theta)^{1/n_1}. \tag{3.12}$$

For this special case, we then have

$$\log WL(\theta) = \frac{1}{n_1}\sum_{j=1}^{n_1} \log f_1(X_{1j};\theta). \tag{3.13}$$

It follows that the WLE for $\theta_1$ coincides with the classical MLE when there is no connection among the samples. This shows that our framework could reproduce the derivation of MLE presented by Akaike (1973).

Furthermore we can also obtain the WL in the non-parametric case, by heuristic reasoning like Akaike has employed. Observe that any unknown term in the objective function defined by equation (3.8) must be estimated. Now we may argue as in the classical case in which the non-parametric MLE is shown to be the sample empirical distribution. Thus, we see that the optimum density is degenerate and puts all of its unit mass on the sample points. In other words the optimum is obtained by maximizing the following quantity over $g_{ij}$, $i = 1,\dots,m$, $j = 1,\dots,n_i$ with $\sum\sum g_{ij} = 1$ and $g_{ij} \geq 0$,

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} g_{ij}^{d_i/n_i}. \tag{3.14}$$

We also obtain the WL estimator of $F_1$ as a generalization of what Hu and Zidek (1993, 2001) called the relevance weighted empirical distribution, namely

$$\hat{F}_1 = \sum_{1=1}^{m} \pi_i \hat{F}_i \tag{3.15}$$

where $\hat{F}_i$ denotes the empirical distribution of the $i$-th sample and $\pi_i \propto d_i$, $i = 1,\dots,m$ are non-negative weights that sum to 1. Thus, by this heuristic reasoning we obtain not only the non-parametric WL in explicit form but the WL estimator as well. Although this estimate is rather "rough", it is the best that can be obtained without parametric restrictions.

We remark that the weights $d_i$'s, although dictated in principle by our constraints, are not easily specified in practice. This is because $a_i$'s which represent the relationships between $f_1$ and $f_i$, $i = 2,\dots,m$ are not known. Thus the likelihood weights must be either specified or estimated. Indeed, it may well be preferable to choose $d_i$'s adaptively. Wang and Zidek (2004) suggested selecting them adaptively by using cross-validation. The adaptive weights proposed there are designed to control any possible bias. The asymptotic properties of WLE are shown in Wang *et al.* (2004).

## 4. Related work and discussion

In this section, we describe related work including some not directly connected to the central topic of this paper.

### 4.1 *James-Stein estimator and the WLE*

Let $Y_i \sim N(\theta_i, 1)$, $i = 1, 2, \dots, m$, where $m \geq 3$. Assume $Y_1, Y_2, \dots, Y_m$ are independent. However they might not be identically distributed as they come from normal distributions with possibly different means. The *James-Stein* estimator is defined as

$$\theta_1^{JS} = \bar{y} + (1 - B^{JS})(y_1 - \bar{y}), \tag{4.1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $B^{JS} = (m-3)/\sum_{k=2}^{m}(y_k - \bar{y})^2$.

The weighted likelihood function for this case is

$$WL(\theta_1) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\lambda_i(y_i - \theta_1)^2. \tag{4.2}$$

It then follows that

$$\hat{\theta}^{WLE} = \sum_{i=1}^{m}\lambda_i y_i. \tag{4.3}$$

Choosing $\lambda_1^{JS} = 1 - \frac{m-1}{m}B^{JS}$ and $\lambda_i^{JS} = \frac{B^{JS}}{m}$, $i = 2, 3, \ldots, m$ makes the WLE coincide with the James-Stein estimator. So, just as the James-Stein estimator can be explained by the empirical Bayes paradigm (Efron and Morris (1973)) so can our extension to the classical likelihood paradigm.

### 4.2 *Weighted likelihood functions*

The local likelihood of Tibshirani and Hastie (1987) extended the idea of local fitting to likelihood-based regression models. Staniswalis (1989) generalized the theory of Tibshirani and Hastie (1987) through non-parametric kernel estimation of a regression function. In the context of non-parametric regression, she defined the *weighted likelihood* as

$$W(\theta) = \sum_{i=1}^{n} W\left(\frac{z_0 - z_i}{b}\right)\log f(y_i;\theta), \tag{4.4}$$

where $z_i$ are fixed and $b$ is a single unknown parameter. Other versions of the local likelihood have been proposed or discussed by Copas (1995), Hjort and Jones (1996) and Loader (1996). Eguchi and Copas (1998) gave a general form of the local likelihood. Still other weighted likelihoods resembling those in this article have been studied. Markatou *et al.* (1997, 1998) proposed one for robust estimation. Hunsberger (1994) also adopted the term "weighted likelihood" when using kernel estimators for the parametric and non-parametric components of semi-parametric regression models. We should add that versions of the weighted likelihood can also be seen in a variety of contexts (c.f. Brillinger (1977), Rao (1991), Field and Smith (1994), Newton and Raftery (1994)).

Hu (1997) proposed the relevance weighted likelihood which is closely related the weighted likelihood presented in this article. The properties of relevance weighted likelihood can be seen in Hu and Zidek (1993, 1995, 2001, 2002). The relevance weighted likelihood generalizes the core of the local likelihood as the weights are not restricted to kernel weights. But it differs from the local likelihood since it does not have a bias correction term. Hu and Rosenberger (2000) discussed various choices for the weights in the context of relevance weighted likelihood.

### 4.3 *Discussion and future work*

The idea of finding an 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 that satisfies certain constraints. 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 polysampling* by Morgenthaler and Tukey (1991) and Easton (1991) that a set of density functions, $f_1, f_2, \ldots, f_m$ "span" a reasonable range of possible true densities for the observations. 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.

The empirical likelihood, a non-parametric procedure with likelihood foundations, seems natural and appealing. Moreover, many desirable features of the likelihood carry over to its empirical counterpart (Owen (2001)). LeBlanc and Crowley (1995) proposed the use of local empirical likelihood to estimate a "conditional functional". Ren (2001) used a weighted empirical likelihood ratio confidence interval for the mean with censored data. That likelihood strongly resembled the one Hu and Zidek (1993) proposed, although they did not consider the impact of censoring. We will investigate this direction further in future work.

Finally, we would note that the approach taken in this paper will seem unnecessary since Bayesian methods could well be used instead. In particular, it has long been known that a hierarchical empirical Bayes approach using the conventional likelihood also generates the James-Stein estimator. However, not all practitioners embrace the Bayesian approach. Moreover, the use of Bayesian approach may be impractical or even infeasible in applications where, not uncommonly, ten's of thousands of parameters may be encountered. This makes it almost impossible to eliciting genuine (as opposed to ad-hoc or non-Bayesian e.g. improper) prior distributions and carry out the necessary computations. Thus, deriving a likelihood-based alternative seems worthwhile.

## Acknowledgements

We thank the two anonymous referees for their excellent comments and suggestions that greatly improved the quality of our presentation. This research is supported in part by the Natural Sciences and Engineering Research Council of Canada.

## Appendix: Proof of theorems

Proof of Theorem 3.2. For the optimal density $g^*$ whose existence is assured, we may without loss of generality assume that the constraints are binding, i.e. that $I(f_i, g^*) = a_i$, $i = 2, 3, \ldots, m$ 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

$$\min_{g \in \mathcal{E}} I(f_1, g) = \min_{g \in \mathcal{E}} \int \log \frac{f_1(x)}{g(x)} f_1(x) d\nu(x), \tag{A.1}$$

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

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

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.

$$\nabla_g \psi - \frac{\partial}{\partial x}(\nabla_{g'}\psi) = 0, \tag{A.2}$$

where $\nabla_g$ 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

$$-\frac{f_1}{g} + l_0 - \sum_{k=2}^m l_k \frac{f_k}{g} = 0. \tag{A.3}$$

We then have

$$g^*(x) = \sum_{k=1}^m \pi_k f_k(x), \tag{A.4}$$

where $\pi_1 = 1/l_0$, $\pi_i = l_k/l_0$, $k = 2,\ldots,m$.

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

$$g^*(x) = \sum_{k=1}^m \pi_k f_k(x) \geq 0, \tag{A.5}$$

since $g^*$ must be in $\mathcal{E}$ by Corollary 3.1.

Finally, we observe that the $\{\pi_i\}$ must be nonnegative for if not we could make $\sum_{k=1}^m \pi_k f_k(x)$ uniformly larger by truncating any negative weights to zero and renormalizing 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. □

PROOF OF THEOREM 3.3. Let $\phi_i(x) = f_i(x) - f_1(x)$, $i = 2,\ldots,m$. Then,

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

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

It follows that,

$$f_i(x) = g_0(x) + \phi_i(x) - \sum_{k=2}^m t_k\phi_k(x) \geq 0. \tag{A.6}$$

This implies that

$$-\left[\phi_i(x) - \sum_{k=2}^m t_k\phi_k(x)\right] \leq g_0(x). \tag{A.7}$$

Since $g_0$ satisfies the constraints by equation (2.4), it follows that, for $2 \le i \le m$

$$\begin{aligned}
\frac{\partial a_i}{\partial t_i} &= \frac{\partial}{\partial t_i}\left[\int f_i(x)\log\frac{f_i(x)}{g_0(x)}d\nu(x)\right] \\
&= \frac{\partial}{\partial t_i}\left[\int f_i(x)\log\frac{f_i(x)}{\sum_{k=1}^m t_k f_k(x)}d\nu(x)\right] \\
&= -\int f_i(x)\frac{\phi_i(x)}{g_0(x)}d\nu(x) \\
&= -\int\left[g_0(x)+\phi_i(x)-\sum_{k=2}^m t_k\phi_k(x)\right]\frac{\phi_i(x)}{g_0(x)}d\nu(x) \\
&= -\int g_0(x)\frac{\phi_i(x)}{g_0(x)}d\nu(x)-\int\left[\phi_i(x)-\sum_{k=2}^m t_k\phi_k(x)\right]\frac{\phi_i(x)}{g_0(x)}d\nu(x) \\
&\le -\int \phi_i(x)d\nu(x)-\int\left[\phi_i(x)-\sum_{k=2}^m t_k\phi_k(x)\right]\frac{\phi_i(x)}{g_0(x)}d\nu(x) \\
&\le -\int \phi_i(x)d\nu(x)+\int g_0(x)\frac{\phi_i(x)}{g_0(x)}d\nu(x) \qquad \text{by (A.7)} \\
&= 0.
\end{aligned}$$

Therefore, it follows that, for $i = 2, \ldots, m$,

$$\frac{\partial t_i}{\partial a_i} = \frac{1}{\dfrac{\partial a_i}{\partial t_i}} \le 0. \tag{A.8}$$

We then have

$$\frac{\partial}{\partial a_i}\sum_{k\ne i} t_k \ge 0 \tag{A.9}$$

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

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 $a_j$, $i \ne j$, it follows that $0 \le t_i \le 1$, $i = 1, 2, \ldots, m$. □

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proceedings of the Second International Symposium on Information Theory* (eds. B. N. Petrov and F. Caski), 267–281, Akademiai Kiado, Budapest.

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

Beavis, B. and Dobbs, I. (1990). *Optimization and Stability Theory for Economic Analysis*, Cambridge University Press, New York.

Bernardo, J. M. (1979). Expected information as expected utility, *The Annals of Statistics*, **7**, 686–690.

Brillinger, D. R. (1977). Discussion of Stone (1977), *The Annals of Statistics*, **5**, 622–623.

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters, *The Annals of Mathematical Statistics*, **37**, 1087–1136.

Copas, J. B. (1995). Local likelihood based on kernel censoring, *Journal of the Royal Statistical Society Series B*, **57**, 221–235.

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

Csiszar, I. (1975). *I*-divergence geometry of probability distributions and minimization problems, *The Annals of Probability*, **3**, 146–158.

Dacorogna, B. (1989). *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York.

Easton, G. S. (1991). Compromise maximum likelihood estimators for location, *Journal of the American Statistical Association*, **83**, 1051–1073.

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.

Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics, *Journal of the Royal Statistical Society Series B*, **60**, 709–724.

Ekeland, I. and Temam, R. (1976). *Convex Analysis and Variational Problems*, American Elsevier Publishing Company, New York.

Field, C. and Smith, B. (1994). Robust estimation: A weighted maximum likelihood approach, *International Statistics Review*, **62**, 405–424.

Ghosh, M. and Yang, M. C. (1988). Simultaneous estimation of the multivariate precision matrix, *The Annals of Statistics*, **16**, 278–291.

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

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix, *The Annals of Statistics*, **8**, 586–597.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric non-parametric density estimation, *The Annals of Statistics*, **24**, 1619–1647.

Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators, *The Canadian Journal of Statistics*, **25**, 45–59.

Hu, F. and Rosenberger, W. F. (2000). Analysis of time trends in adaptive designs with applications to neurophysiology experiment, *Statistics in Medicine*, **19**, 2067–2075.

Hu, F. and Zidek, J. V. (1993). A relevance weighted nonparametric quantile estimator, Tech. Report No. 134, Department of Statistics, University of British Columbia, Vancouver, Canada.

Hu, F. and Zidek, J. V. (1995). Incorporating relevant sample information using the likelihood, Tech. Report No. 161, Department of Statistics, University of British Columbia, Vancouver, Canada.

Hu, F. and Zidek, J. V. (2001). The relevance weighted likelihood with applications, *Empirical Bayes and Likelihood Inference* (eds. S. E. Ahmed and N. Reid), Springer-Verlag, New York.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

Hunsberger, S. (1994). Semiparametric regression in likelihood-based models, *Journal of the American Statistical Association*, **89**, 1354–1365.

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 361–379, University of California Press, Berkeley, California.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

LeBlanc, M. and Crowley, M. (1995). Semiparametric regression functionals, *Journal of the American Statistical Association*, **90**, 95–105.

Loader, C. R. (1996). Local likelihood density estimation, *The Annals of Statistics*, **24**, 1602–1618.

Markatou, M., Basu, A. and Lindsay, B. G. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression, *Journal of Statistical Planning and Inference*, **57**, 215–232.

Markatou, M., Basu, A. and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search, *Journal of the American Statistical Association*, **93**, 740–750.

Morgenthaler, S. and Tukey, J. W. (1991). *Configural Polysampling: A Route to Practical Robustness*, Wiley, New York.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society Series B*, **56**, 3–48.

Owen, A. B. (2001). *Empirical Likelihood*, Chapman and Hall, New York.

Parsian, A. and Nematollahi, N. (1996). Estimation of scale parameter under entropy loss, *Journal of Statistical Planning and Inference*, **52**, 77–91.

Rao, P. B. L. S. (1991). Asymptotic theory of weighted maximum likelihood estimation for growth models, *Statistical Inference for Stochastic Processes* (eds. N. U. Prabhu and I. V. Vasawa), 183–208, Marcel Dekker.

Ren, J. (2001). Weighted empirical likelihood ratio confidence interval for the mean with censored data, *Annals of the Institute of Statistical Mathematics*, **53**, 498–516.

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

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion, *Journal of the Royal Statistical Society Series B*, **57**, 99–138.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in a likelihood-based models, *Journal of the American Statistical Association*, **84**, 276–283.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceeding of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 197–206, University of California Press, Berkeley.

Tibshirani, R. and Hastie, T. (1987). Local likelihood of statistical predictions, *Journal of the Royal Statistical Society Series B*, **36**, 111–147.

Trottini, M. and Spezzaferri, F. (2002). A generalized predictive criterion for model selection, *The Canadian Journal of Statistics*, **30**, 79–96.

Wang, X. and Zidek J. V. (2004). Selecting likelihood weights by cross-validation, *The Annals of Statistics* (in press).

Wang, X., van Eeden, C. and Zidek, J. V. (2004). Asymptotic properties of maximum weighted likelihood estimators, *Journal of Statistical Inference and Planning*, **119**, 37–54.