The Canadian Journal of Statistics
Vol. 30, No. 3, 2002, Pages 347–371
La revue canadienne de statistique

The weighted likelihood[1]

Feifang HU and James V. ZIDEK

Key words and phrases: Entropy maximization principle; function estimation; information; James–Stein estimation; kernel; likelihood; maximum likelihood; nonparametric regression; normal mean estimation; relevance; smoothing.

MSC 2000: Primary 62A01; secondary 62F10.

Abstract: The authors consider a weighted version of the classical likelihood that applies when the need is felt to diminish the role of some of the data in order to trade bias for precision. They propose an axiomatic derivation of the weighted likelihood, for which they show that aspects of classical theory continue to obtain. They suggest a data-based method of selecting the weights and show that it leads to the James–Stein estimator in various contexts. They also provide applications.

La vraisemblance pondérée

Résumé : Les auteurs considèrent une version pondérée de la vraisemblance classique qui s'impose lorsqu'il apparaît opportun d'amenuiser l'influence de certaines observations dans le but d'atteindre un équilibre entre biais et précision. Ils décrivent une axiomatique qui conduit à la vraisemblance pondérée, pour laquelle ils montrent que certains pans de la théorie classique continuent de s'appliquer. Ils suggèrent une méthode adaptative de sélection des poids et montrent qu'elle mène à l'estimateur de James–Stein dans différents contextes. Ils présentent en outre quelques applications.

1. INTRODUCTION

In this paper, we describe our version of the weighted likelihood (WL) and demonstrate its use in a number of applications. We also describe other versions, some being special cases and some not.

To put our work in perspective, consider the problem of selecting a predictive distribution for goals in any designated future ice-hockey game of a given season between the National Hockey League's Vancouver Canucks and Calgary Flames. A traditional approach would take the sequence of Canucks home games played against the Flames as an independent and identically distributed sequence of observations. The classical likelihood method could be used to find estimates of the model parameters. A similar analysis could be made for the Flames's home games against the Canucks. Finally, the outcome of future games between these two teams could be predicted using the estimated model.

However, this approach, which uses only the direct information, would be poor, especially toward the beginning of the season when little or no data were available. Our approach brings in additional, relevant data from the games these teams played against other opponents. These other data would bias the estimates of the sampling model parameters for Canucks–Flames games. However, they would also provide an indication of the relative strengths of these two teams. Thus, we would trade bias for information. In Section 7, we return to this example to show that trade-off pays off. Our predictor is substantially more accurate than the one based on the classical estimator alone.

The WL proposed here has its roots in the celebrated paper of Stein (1956). That paper, arguably the most important in the context of the Wald paradigm (i.e., decision theory), had a substantial impact on the course of the development of statistical theory. It addressed the problem

[1]This paper was the basis for the second author's Gold Medal address given at the 29th Annual Meeting of the Statistical Society of Canada held in Burnaby, British Columbia, in June 2001. / Cet article a inspiré l'allocution du médaillé d'or prononcée par le second auteur à l'occasion du 29ième congrès annuel de la Société statistique du Canada tenu à Burnaby (Colombie-Britannique) en juin 2001.

of simultaneously estimating the means of $p \geq 3$ normal populations using independent samples that had been drawn from each.

Stein's paper includes an heuristic derivation of what has become known as the James–Stein estimator (James & Stein 1961). The latter was shown to have a uniformly smaller combined mean square error than the maximum likelihood estimator (MLE), whenever the number of populations (assumed to have a known variance) exceeded 2. In contrast, the heuristic argument merely makes a case for shrinking the MLE towards zero when p becomes large. However, although the result is less precise, the simplicity of that argument makes the case seem compelling and the result quite general as well as fundamental. In particular, the argument calls into question the validity of the "large n (sample size)" paradigm that is often used to justify the MLE, in a "large p (number of parameters)" situation.

Also Stein's argument shows that in terms of the sum of the mean-square-error-of-estimation criterion, the sample averages could be improved on by borrowing information from the other samples. That somewhat unintuitive result has been variously called the "Stein paradox," the "Stein effect" or the "Stein phenomenon." In any case the result: (i) cast doubt on the large sample paradigm inasmuch as it justifies the use of sample averages by demonstrating their good asymptotic qualities; (ii) cast doubt on unbiasedness and minimum variance unbiased estimation, criteria that in many cases supported the use in applications of the sample average when Stein's result made this seem naive; (iii) showed that bias could be traded for increases in precision; (iv) showed that information could profitably be borrowed from samples independently drawn from populations different from that under study; (v) paved the way to the widespread use of hierarchical Bayesian methods in practice by Bayesians and non-Bayesians alike, the latter through the notion of "random effects." Taken together, (i)–(v) cast doubt on the classical likelihood that gave the sample averages in the first place.

We show in this paper that when suitably extended, the likelihood can yield estimators that do not have the difficulties mentioned above, that are encountered by sample averages. Further, we show how that extension allows us to trade bias for precision in a likelihood context. We show how the celebrated estimator of James & Stein (1961) that flowed out of the seminal paper of Stein (1956) can be derived there.

In particular, we further extend the likelihood that Hu (1994) calls the "relevance weighted likelihood" (REWL). The REWL arises when in addition to (or instead of) the sample from the study population, relevant but independent samples from other populations are available. By downweighting these other samples according to their relevance, the REWL incorporates their information while downplaying their bias.

The idea of relevance weighting has been well recognized in the special case of nonparametric regression where estimates of the regression function $m(x)$ at any given point x use suitably downweighted data obtained at nearby points. In particular, the relevance weighted likelihood has already been proposed in that context. (References are given in Section 8.) However, as far as we know, the first general formalization of the REWL is in Hu (1994).

For reasons given below, the weights in our extension of the relevance weighted likelihood may in fact be negative. We indicate situations in which the bias-precision trade-off can be made without merely exploiting relevant information as it does in the James–Stein estimator.

The problem of combining information addressed here can be solved, at least conceptually, using a hierarchical Bayesian approach. But as Efron (1996, p. 540) notes: "In practice, however, it is difficult to apply hierarchical Bayes methods because of the substantial Bayesian inputs that they require." Whereas Efron retains that framework but substitutes empirical Bayesian steps for those inputs, we start at a more basic level and appeal to Akaike's entropy maximization principle to obtain and motivate our approach and in particular, the WL. We do this first in the case where the underlying densities do not have a parametric form (Section 3) and then in the parametric case (Section 4). An advantage of our approach is its consistency with approaches taken within the well-developed field of nonparametric regression referred to above. We discover that many of the

classical properties of the likelihood carry over to the WL (Section 3). In particular, as Hu (1997) as well as Wang, van Eeden & Zidek (2001) have shown, the asymptotic theory of Wald for the maximum likelihood estimator (MLE) can be extended to its counterpart for the WL.

We address the problems of selecting the weights in Section 4. A general method for doing so is presented, which is based on Akaike's maximum entropy criterion. We recognize, however, that these weights may best be chosen in particular applications where context-specific information is available. Then in Section 5, we look at the problem of estimating the means of populations that are normally distributed. We show how the WL can generate better estimates than the MLE when the means lie in restricted parameter spaces. When the means are unrestricted but must be estimated simultaneously, we see that the James–Stein estimator obtains from use of the WL in conjunction with the method of Section 4.

To further demonstrate the scope of our theory, we show in Section 6 how an extension of the classical likelihood ratio test leads to a generalized Shewhart quality control chart.

Our work is linked to a substantial body of existing theory. Rather than attempt to survey that literature in this section, we give a brief summary of the most relevant links in Section 8. We give our concluding remarks in Section 9.

2. NONPARAMETRIC WEIGHTED LIKELIHOOD

In Section 2.1, we derive in a special case the nonparametric weighted likelihood, or NP-WL, for short. See Hu & Zidek (2001). That derivation, which has been extended to the general case by Wang (2001), is based on an approach due to Akaike (1973, 1977, 1978, 1982, 1983, 1985). Then in Section 2.2, we describe the "bias-precision" trade-off involved in the estimation of a population distribution.

2.1. Deriving the likelihood.

We consider n populations labelled $i = 1, \ldots, n$. Suppose for each i, Y_i represents a measurable attribute or a vector of such attributes. These Y_i are assumed to be independently distributed. The unknown population distribution of Y_i has a probability density function, f_i, with respect to a σ-finite measure μ. Although the population distributions are not necessarily identical, assume that they are thought to resemble one another to varying degrees. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the vector or matrix of these measurable attributes.

Suppose that from each population i, $n_i \geq 0$ items are randomly and independently sampled, yielding $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, Y_{ij} representing the Y_i measured on the jth item sampled from the ith population, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$; by convention, $\mathbf{Y}_i$ is taken to be the null vector when $n_j = 0$. Assume that the Y_{ij}, $j = 1, \ldots, n_i$ are independent as well as identically distributed, each having its associated population distribution. Denote the realization of $\mathbf{Y}_i$ by $\mathbf{y}_i$, $i = 1, \ldots, n$.

Inferential interest concerns the joint population distribution of the n populations, i.e., of $\mathbf{Y}$. Taking Akaike's approach, we seek a predictive density based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$, say $\hat{f}_{\mathbf{Y}}$ for $\mathbf{Y}$. Akaike measures the accuracy of that predictive distribution by the negative of Boltzman's entropy, or equivalently by

$$\int f_{\mathbf{Y}}(y) \ln\{\hat{f}_{\mathbf{Y}}(y)\}\, d\mu(y). \tag{1}$$

That measure of divergence is, in fact, a special case of a large family of such measures (Trottini & Spezzaferri 2002). In any case, the optimum choice of $f_{\mathbf{Y}}$ would maximize the quantity in (1). Indeed, it would just be $f_{\mathbf{Y}}$ if the latter were known. However, the key to Akaike's approach is the idea that $f_{\mathbf{Y}}$ exists but is unknowable. It plays a purely conceptual role in the analysis.

Since the n populations are sampled independently, we require $\hat{f}_{\mathbf{Y}}$, like $f_{\mathbf{Y}}$, to be a product of its marginal counterparts, so that

$$\hat{f}_{\mathbf{Y}} = \hat{f}_1 \times \cdots \times f_n, \tag{2}$$

the $\hat{f}_i$ being predictive densities for the individual populations. It then follows that we may find the optimum $\hat{f}_Y$ by finding the optimum $\hat{f}_i$ for each i, the one that maximizes

$$\int f_i(y) \ln\{\hat{f}_i(y)\}\, d\mu(y), \quad 1 \leq i \leq n. \tag{3}$$

Since the population distributions are thought to resemble each other, we might heuristically conclude that the $\mathbf{y}_j$ with $j \neq i$ provide information of value in constructing $\hat{f}_i(y)$. With "resemble" appropriately defined, that heuristic can be rigorously demonstrated as shown below.

To be consistent in our use of Boltzman's entropy, we adopt it in interpreting "resemble" and define that term to mean

$$\int f_j(y) \ln\{f_j(y)\}\, d\mu(y) \geq \int f_j(y) \ln\{f_i(y)\}\, d\mu(y) \geq c_{ji} \tag{4}$$

for all $i \neq j$. The c_{ji} reflect the degree to which the distribution of population i resembles that of j. If

$$c_{jj} = \int f_j(y) \ln\{f_j(y)\}\, d\mu(y),$$

the negative entropy for the distribution of population j, then the restriction in (4) would entail the requirement that $f_i = f_j$ since in general the mapping $g \mapsto \int h(y) \ln\{g(y)\}\, d\mu(y)$ is uniquely maximized by $g = h$ for any two densities g and h.

An anonymous referee asked if $c_{ij} = c_{ji}$, noting that it would be difficult to distinguish between them in practice. In fact, these two bounds need not be the same since the measure of divergence is asymmetrical. However, their role is purely conceptual. In the sequel, they will be absorbed into weights derived from them whose specification in practice will be discussed.

To help interpret the conditions in (4), suppose the f_i are normal probability density functions with means μ_i and unit variances. Then these conditions could equivalently be expressed as

$$|\mu_i - \mu_j| \leq \sqrt{-2c_{ij} - 1 - \log(2\pi)}$$

for all i and j.

Given the conditions in (4), we restrict $\hat{f}_i$ to the class of densities g for which

$$\int f_j(y) \ln\{g(y)\}\, d\mu(y) \geq c_{ji}, \quad j \neq i.$$

That class is convex, and the functional

$$g \to \int f_i(y) \ln\{g(y)\}\, d\mu(y)$$

is strictly concave.

Although the population response distribution is unknown, we can nevertheless easily characterize the optimum solution in a mathematical sense in the simple case where the Y_i are discrete. To do so, we use Lagrange's method. Subject to $\int g_i\, d\mu = 1$, g_i maximizes

$$\sum_{j=1}^{n} \lambda_{ij} \int f_j(y) \ln\{g(y)\}\, d\mu(y), \tag{5}$$

the multipliers $\lambda_{i1}, \ldots, \lambda_{in}$ being chosen to ensure the validity of the conditions (4). (Note that the inequalities in (4) must first be converted to equalities, as we may without loss of generality by appropriately increasing the values of the c_{ij} as necessary.) Thus, ideally we obtain

$$\hat{f}_i = \sum_{j=1}^{n} \omega_{ij} f_j \tag{6}$$

with $\omega_{ij} = \lambda_{ij} / \sum_\ell \lambda_{i\ell}$. It can be shown that the ω_{ij} are nonnegative. A proof of this fact is given by Wang & Zidek (2002) and, for brevity, will not be included here.

Using (6), we obtain the optimal cumulative distribution function (CDF) for population $i = 1, \ldots, n$, namely

$$\hat{F}_i = \bar{F}_i = \sum_{j=1}^{n} \omega_{ij} F_j, \tag{7}$$

where F_j denotes the CDF of the jth population. (In the discussion below, this quantity is considered in relation to the bias-information trade-off that is fundamental to this paper.)

Since these distributions are unknown, we are forced to estimate the quantity in (5). Let us begin by re-expressing it as

$$\sum_{j=1}^{n} \lambda_{ij} \int \ln\{g(y)\}\, dF_j(y).$$

Its natural estimate in our nonparametric context would replace F_j by its empirical distribution F_j^{emp} for those $j = 1, \ldots, n$ such that $n_j > 0$. The terms in equation (7) for which $n_j = 0$ cannot be so estimated since no data are available. [That could include term i corresponding to the population of central interest.] We therefore exclude them by setting $\lambda_{ij} = 0$ when $n_j = 0$ in the summands of (5) as a practical concession.

After estimating the objective function in (5) in the manner described in the previous paragraph, we see that g_i and hence f_i would be found by maximizing, over densities g, the NP–WL

$$\prod_{j=1}^{n} \prod_{\ell=1}^{n_j} g^{\lambda_{ij}/n_j}(y_{j\ell}), \tag{8}$$

where we take $\lambda_{ij}/n_j = 0$ when $n_j = 0$.

Applying Lagrange's method once more, we obtain

$$f_i = \sum_{j=1}^{n} \omega_{ij} f_j^{\text{emp}}, \tag{9}$$

where f_i^{emp} denotes the discrete empirical density function for the jth population obtained as the derivative of F_j^{emp} with respect to counting measure and $\omega_{ij} = 0$ if $n_j = 0$. Thus the maximum NP-WL estimate of f_i proves to be the obvious estimate obtained from (6) when the population densities are unknown and we set $\lambda_{ij} = 0$ if $n_j = 0$. From (9), we obtain an estimate of the CDF for population i, namely

$$\hat{F}_i = \sum_{j=1}^{n} \omega_{ij} F_j^{\text{emp}}.$$

We call this the "relevance weighted empirical distribution" (REWED) after Hu & Zidek (1993a), who introduced that terminology in the special case where $n_j \equiv 1$.

The joint population predictive density can now be found from (2), and this completes our analysis for the case where the Y_i are discrete. Using a more complex analysis, Wang (2001) has shown that the derivation given above can be extended to the general case where μ is an arbitrary sigma-finite measure. In the sequel that extension is assumed. As well, we formally drop the unnecessary restriction to nonnegative weights so that our weighted likelihood will formally generate the James–Stein estimator, among others. With these extensions, we now define the NP-WL.

DEFINITION 1. *Suppose the sample vectors (or matrices when the $y_{j\ell}$ represent response vectors)* $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn_j})$, $i = 1, \ldots, n$, *with* $n_j \geq 0$, *have been independently observed from* n *populations. For population* i, *define the* nonparametric weighted likelihood (NP-WL) *as that given in* (8). *The joint all-population likelihood is the product over* i *of that above.*

2.2. Discussion.

We find a counterpart of the familiar "bias-variance" trade-off in this nonparametric distribution estimation context. For population i, bias might be characterized by $F_i - \overline{F}_i$; see (7). That function expresses, for each point in its domain, the overall degree to which the family of population CDFs resemble i when averaged with the weights ω_{ij}. Ensuring that this bias converges to 0 as the number of populations $n \to \infty$ imposes a restriction on the weights (which depend on n).

Of course, when $n_i > 0$, we can eliminate the bias altogether with the then permissible choice $\omega_{ii} = 1$. Indeed this would yield the classical estimator $\widehat{F}_i = F_i^{\text{emp}}$ for F_i. However in reducing the bias this way we lose the information about population i in the samples from populations, $j \neq i$, information that flows through the known resemblance of their distributions to it. That in turn would reduce the precision of our estimator which might be expressed through $\overline{F}_i - \widehat{F}_i$. Intuitively speaking, that latter would increase while the former decreases. In summary, we see a bias-precision trade-off.

In unpublished work summarized in Hu & Zidek (2001), Hu & Zidek (1993a) give conditions on the weights which ensure, asymptotically at least, that precision will improve as bias is reduced. More precisely, suppose the weights are chosen so that $|F_i(y) - \overline{F}_i(y)| \to 0$ as $n \to \infty$. Then $|\overline{F}_i(y) - \widehat{F}_i(y)| \to 0$ almost surely for each y as $n \to \infty$ provided that

$$\sum_{n=1}^{\infty} \exp\left(-\varepsilon^2 \Big/ \sum_{j=1}^{n} \omega_{nij}^2\right) < \infty$$

for all $\varepsilon > 0$, where we have added the suffix n to the weights to emphasize their dependence on n. In particular, this almost sure convergence to 0 of the estimation error at each y will be assured by the sufficient condition $\max_j \omega_{nij} = o(1/\ln n)$.

Under appropriate conditions, Hu & Zidek (1993a, 2001) also prove strong consistency and asymptotic normality for the estimators of the quantiles of F_i derived from those of $\widehat{F}_i$. Their simulation studies (see Hu & Zidek 2001) confirm the value of these quantile estimators.

While the derivation given in this section points to the form of the NP-WL, the problem of selecting the λ_{ij} remains. Although in principle they are determined by the c_{ij}, they are not determined in explicit form. More importantly, the difficult practical problem of specifying the c_{ij} remains. We address the problem of selecting the weights in Section 4.

3. THE PARAMETRIC WEIGHTED LIKELIHOOD

In Section 3.1, we present the parametric alternative to the nonparametric weighted likelihood of the last section and call it the "parametric-weighted likelihood" (P-WL). Examples in Section 3.2 show the broad range of potential applications as well as the potential value of our likelihood-like approach to inference. Asymptotic theory is described in Section 3.3.

3.1. The likelihood.

Reconsider equations (1)–(3). We now impose the additional condition that $\hat{f}_i$ be constrained to a parametric class of hypothetical alternatives of the form $g(y) = f_i(y \mid \theta_i)$, $\theta_i \in \Theta$, where Θ represents an open subset of Euclidean space. In this context, the f_i represent specified functions so that only the θ_i need to be selected.

Let us now add the constraints in (4) and the assumptions needed for the analysis of the previous section as extended by Wang (2001). Specifically, we assume that subject to the constraints, a unique maximum θ_i^* of $\theta_i \mapsto \int \ln\{f_i(y \mid \theta_i)\} f_i(y)\, d\mu(y)$ is attained at an interior point

of Θ. Assume also that this function together with $\theta_i \mapsto \int \ln\{f_i(y \mid \theta_i)\} f_j(y)\, d\mu(y)$, $j \neq i$, are continuously differentiable on Θ. Then θ_i^* maximizes

$$\sum_{j=1}^{n} \omega_{ij} \int \ln\{f_i(y \mid \theta_i)\} f_j(y)\, d\mu(y)$$

for certain constants ω_{ij}.

However, the f_j are unknown and we are forced to estimate them. To do so we use the empirical distribution as in the previous section. Thus we would maximize instead

$$\sum_{j=1}^{n} \omega_{ij} \int \ln\{f_i(y \mid \theta_i)\}\, dF_j^{\text{emp}}(y) = \sum_{j=1}^{n} \frac{\omega_{ij}}{n_j} \sum_{\ell=1}^{n_j} \ln\{f_i(y_{j\ell} \mid \theta_i)\}, \tag{10}$$

for $i = 1, \ldots, n$. Combining over i, the estimates in (10) gives us the P-WL in the next definition, where we let $\lambda_{ij}/n_j = 0$ when $n_j = 0$ as before.

DEFINITION 2. *Suppose the sample vectors (or matrices when the $y_{j\ell}$ represent response vectors) $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn_j})$, $i = 1, \ldots, n$ with $n_j \geq 0$ have been independently observed from n populations. Define the joint all-population* parametric weighted likelihood (P-WL) *as*

$$\theta = (\theta_1, \ldots, \theta_n) \mapsto \prod_{i=1}^{n} \prod_{j=1}^{n} \prod_{\ell=1}^{n_j} f_i^{\lambda_{ij}/n_j}(y_{j\ell} \mid \theta_i).$$

Using the P-WL, we may define the maximum weighted likelihood estimate (WLE) of θ in the obvious way.

3.2. Examples.

Example 1 (Parametric nonparametric regression). Let $n_j \boxminus 1$ and

$$f_j = f\{y_j - m(x_j)\}, \quad x_j \in [a, b], \quad j = 1, \ldots, n,$$

$m(x)$ being a smooth function and the density, f being a known density and the y_j being independently observed, conditional on the x_j.

Here and in general, choosing the λ_{ij} is analogous to choosing a kernel and bandwidth in nonparametric regression theory. Indeed, in the domain of that theory, we can find the λ_{ij} directly from the corresponding kernels (and their bandwidths), making our task easy in that case.

When f represents a normal distribution centered at 0, the WLE for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ with $\mu_i = m(x_i)$, $i = 1, \ldots, n$, is easily seen to be a linear function of $\mathbf{y} = (y_1, \ldots, y_n)$. Its coefficients will depend on the normal variances σ_j^2 that are assumed known. Thus, it resembles the members of the various classes of estimators that have been proposed in nonparametric regression theory even when normal error distributions are not assumed; see Eubank (1988), Fan & Gijbels (1996), Härdle (1990) as well as Wand & Jones (1995).

However, such linear estimators would not be WLEs if f were not the density of a normal distribution. And if the form of f were unknown save for symmetry about 0, the results of Hu & Zidek (1993a, 2001) suggest the use of a more robust alternative to the linear estimator obtained from the nonparametric WLE of f. In summary, our WL theory points to alternatives to the common nonparametric regression estimators except possibly in the normal case.

The next example differs from the last one in that we allow the weights to depend on the data themselves.

Example 2 (Robustness). Here the observations are assumed to be scalars, $n_j \equiv 1, \theta_i \equiv \theta_1$ and $f_i \equiv f_1$ except that some of the y_j are thought to be outliers coming from some population other than 1.

In this example, that is related to the weighted estimating equations approach of Field & Smith (1994), the outliers need to be downweighted. To do this, we may order the data as $y_{(1)}, \ldots, y_{(n)}$ and assign weights λ_{1j} depending on the degree to which we regard the associated data as outlying. The appropriate P-WL becomes

$$\prod_{j=1}^{n} f_1^{\lambda_{1j}}(y_{(i)} \mid \theta_1).$$

In the extreme case, when a fraction 2ε are deemed to be outliers, we could choose $\lambda_{1j} = 0$, when $i \leq [n\varepsilon]$ or $i \geq [n(1-\varepsilon)]$. (Here $[x]$ denotes the greatest integer less than x.) Then the P-WL becomes a trimmed likelihood. The trimmed mean would then be obtained in certain cases as a WLE.

Example 3 (Generalized smoothing). Suppose $n_j \equiv 1$ and $(Y_1, X_1), \ldots, (Y_n, X_n)$ are n data pairs, Y_j having PDF $f\{y, \theta(X_j)\}$ with parameter $\theta(X_j)$. Interest lies in the study population corresponding to a fixed value $X = x_i$ (that need not be one of the x_j). The weights λ_{ij} enable us to represent the degree to which the information from the populations corresponding to X_j should be used in fitting $f\{y, \theta(x_i)\}$. The WL becomes

$$\prod_{j=1}^{n} f^{\lambda_{ij}}\{y_j, \theta(x_j)\}.$$

This example extends substantially the domain of classical linear nonparameter regression theory.

3.3. Discussion.

Let us now refer to the P-WL more simply as the WL. The likelihood is obtained from the WL in the special case when all the data are independently drawn from the study population. However, even here there may be a role for the WL as Example 2 demonstrates.

The likelihood usually derives from the sampling density by inverting what is fixed and what varies. In particular, conditional on f (or the parameter of f), the likelihood integrates to unity over the sample space. This duality between likelihood and sampling density may be useful for determining the likelihood. However, it does not seem intrinsic. In the usual case of iid sampling, we could take the nth root of the inverted sampling density without apparent loss and without preserving the aforementioned property. Moreover, in the Bayesian framework the sample space need not even be specified, although likelihood can certainly be defined. So we do not see the lack of duality with sampling as a shortcoming of our proposed extension of the likelihood. The usual asymptotic theory, appropriately modified still obtains, as shown elsewhere (Hu 1997; Wang 2001).

We can easily show that the WL is preserved under arbitrary differentiable data-transformations (with nonvanishing Jacobians) when the sampling densities are absolutely continuous with respect to the Lebesgue measure. So the WL inherits this important property of the likelihood.

The very important likelihood principle of classical statistical inference tells us that the likelihood embraces all relevant information about the parameter. Indeed, according to the factorization theorem, sufficiency may be defined through the likelihood. Standard constructions of minimally sufficient statistics rely on the likelihood.

The counterpart of the likelihood theory in our setting would be the WL principle. Lacking the invertibility of likelihood and sampling density, we might resort to a WL-based definition of sufficiency and call it "WL-based sufficiency." That notion enables us to reduce the dimension of the observation vector to that of any (WL-based) sufficient vector-valued function of the data, while retaining all information in the WL.

DEFINITION 3. *Call* WL-based sufficient *any real or vector-valued statistic that determines the WL up to an arbitrary multiplicative factor which does not depend on* f. WL-based minimal-sufficient statistics *are functions of every other WL-based sufficient statistic.*

A WL-based minimal-sufficient statistic yields maximal data reduction. Such a statistic need not be unique. The *factorization theorem* remains true for our notion of sufficiency.

THEOREM 1. *A necessary and sufficient condition that* S *be WL-based sufficient for the parametric family* $\mathcal{F}$ *indexed by* θ *is that there exist functions* $m_1(s,\theta)$ *and* $m_2(\mathbf{y})$ *such that for all* $\theta \in \Omega$, *the WL may be represented as* $m_1(s,\theta)m_2(\mathbf{y})$.

Accepting the WL as the basis for inference makes reliance on WL-based sufficient statistics inevitable. The seemingly reasonable estimators we obtain in applications, such as those of this paper, depend on the data only through such a statistic offering some support for our principle. Just as the conventional likelihood (regarded as a function) is sufficient, the WL is WL-based sufficient. (This fact follows from the factorization theorem.) However, WL-based sufficient statistics lack the property of conventional sufficiency which renders the conditional sampling distribution of the data given a sufficient statistic independent of θ. Finding an analogous interpretation of the principle of WL-based sufficiency remains an open problem.

4. SELECTING THE WEIGHTS

We believe the best choice of the weights will depend on the context. For example, the particular context of quality control in the next section suggests a way to select those weights for control charting. The general problem of weight selection resembles that of choosing a kernel and bandwidth in nonparametric regression. Indeed, in that context, we can choose the weights directly from the corresponding kernels and their bandwidths in an ad hoc fashion (Hu & Zidek 1997).

To gain further insight into the role of these weights, we concentrate on population i alone, assume $n_j \equiv 1$ and focus on the case where "nearness" can be interpreted as the difference between population parameters. Consider the ln-WL ratio

$$\Lambda = \ln\left\{\prod_{j=1}^{n} f_i^{\lambda_{ij}}(Y_{ij} \mid \theta_i)\right\} \Big/ \left\{\prod_{j=1}^{n} f_i^{\lambda_{ij}}(Y_{ij} \mid \theta_i + \Delta)\right\}$$

and its capability to discriminate against alternatives near θ_i (assumed for this discussion to be real-valued). We have approximately

$$\mathrm{E}(\Lambda) \approx -\Delta \sum_{j=1}^{n} \lambda_{ij} \mathrm{E}\left\{\frac{\partial \ln f_i(Y_{ij};\theta_i)}{\partial \theta_i}\right\} + \frac{\Delta^2}{2} \sum_{j=1}^{n} \lambda_{ij} \mathrm{E}\left\{-\frac{\partial^2 \ln f_i(Y_{ij};\theta_i)}{\partial \theta_i^2}\right\}.$$

When the Y_{ij} are identically distributed, each with density function $f_i(y;\theta_i)$, we see that this capability to discriminate against local alternatives depends on Fisher's information, namely

$$\sum_{j=1}^{n} \lambda_{ij} \mathrm{E}\left\{-\frac{\partial^2 \ln f_i(Y_{ij};\theta_i)}{\partial \theta_i^2}\right\}.$$

In this case, we must choose $\lambda_{ij} = 1$ for all j to optimize this capability by maximizing the expected size of the ln-WL ratio. In this way, we make maximal use of the information in all of the observations.

However, when not all of the observations come from the same population i, we see the effect of bias in the dominant, nonzero first-order term. The choice of the λ_{ij} determines the trade-off between the bias in the first term and the information in the lesser, second term. Clearly, the optimal λ_{ij} must depend somewhat sensitively on the bias expressed by the term

$$\mathrm{E}\left\{-\frac{\partial \ln f_i(Y_{ij};\theta_i)}{\partial \theta_i}\right\}.$$

The approach of Stigler (1990) and our derivation in Section 2 suggest a general approach to the selection of weights. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ be a given vector of weights and $\hat{\theta}_{\boldsymbol{\lambda}} = (\hat{\theta}_{1\lambda_1}, \ldots, \hat{\theta}_{n\lambda_n})$ denote the WLE for $\theta = (\theta_1, \ldots, \theta_n)$ based on those weights. Define

$$\boldsymbol{\lambda}(\theta) = \arg\max_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \mathrm{E}\left[\int f_i(y \mid \theta_i) \ln\{f_i(y \mid \hat{\theta}_{i\lambda_i})\}\, d\mu(y) \Big| \theta\right]. \tag{11}$$

Finally, let $\hat{\theta}$ be any reasonable estimator of θ. Then the resulting $\hat{\lambda}$ would constitute a data-based choice of the weights. For example, these might be found by substituting the MLE of θ in equation (11). We refer to the WL with data-dependent weights as the *adaptive WL*.

Generally, $\boldsymbol{\lambda}(\theta)$ would need to be found by computational methods. However, we may obtain an approximation to it by adopting a second order Taylor approximation for $\ln f_i(y \mid \hat{\theta}_{i\lambda_i})$ (see Edwards 1984, p. 145):

$$\boldsymbol{\lambda}(\theta) \approx \arg\min_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \mathrm{E}\left\{(\hat{\theta}_{i\lambda_i} - \theta_i)' I_i(\theta_i)(\hat{\theta}_{i\lambda_i} - \theta_i) \Big| \theta\right\},$$

where I_i represents Fisher's information matrix corresponding to the density $f_i(\cdot \mid \theta_i)$.

The value of this approach to specifying weights will be the subject of further study. Meanwhile, we illustrate it with results in succeeding sections. Those results also demonstrate the generality of the WL-based approach for inference which goes well beyond nonparametric regression, one of its important sub-domains of applicability. They show in fact how the well-known James–Stein estimator can be viewed as a relative of nonparametric regression estimators.

Before leaving this section, we should remark that Wang (2001) has developed a cross-validatory approach to the determination of the adaptive WL. His approach can be applied in cases where all the n_j are strictly positive.

5. ESTIMATING NORMAL MEANS

In this section, we suppose $n_j \equiv 1$ and let $Y_{i1} = Y_i$. Assume $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, are independent random variables, the μ_i being unknown parameters while the σ_i are known. Given $Y_i = y_i$, $i = 1, \ldots, n$, we require an estimate of μ_1, thought to resemble the other μ_i.

Classical likelihood-based estimation theory suggests the MLE

$$\hat{\mu}_1 = y_1,$$

an estimator that fails to use our knowledge of the similarity amongst the population means.

An alternative which uses the data more fully would be

$$\hat{\mu}_1^* = \sum_{i=1}^{n} w_i y_i,$$

where $w_1 + \cdots + w_n = 1$. This last estimator (the WLE) can be found by maximizing the appropriate WL, namely

$$\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right\}\right]^{w_i}.$$

Contrast the latter with the usual likelihood function for this problem:

$$\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right\},$$

which unlike the P-WL involves all the population means. This last likelihood does not reflect our prior information about the resemblance of the μ_i to μ_1 and in particular, does not let us use the information in y_i about μ_1.

When might the WLE be preferable to μ_1 or the maximum likelihood estimator? In this section, we answer this question in two different contexts and find that the former can have advantages over the latter. We thereby demonstrate the value of the relevant information.

5.1. Restricted parameter spaces.

Jointly restricting the population parameters in some way can naturally lead to the need to combine the information in their independent samples. Consider for example the case where $|\mu_i - \mu_1| \le c_i$ for specified constants $c_i \ge 0$ and $c_1 = 0$. In this case, the MLE would become a truncated version of $\hat{\mu}_1$. However, for squared error loss such truncated estimators are inadmissible under very general conditions; see Charras & van Eeden (1991). Moreover, except in the case $n = 2$ they cannot be found in explicit form. Hence we compare $\hat{\mu}_1^*$ with $\hat{\mu}_1$ in the next theorem, where $\mathbf{w}' = (w_1, \ldots, w_n)$, $\mu' = (\mu_1, \ldots, \mu_n)$, $C' = (c_1, \ldots, c_n)$, $\mathbf{1}' : 1 \times n = (1, \ldots, 1)$, and $D = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

THEOREM 2. *For all μ such that $|\mu_i - \mu_1| \le c_i$, $i = 1, \ldots, n$, we have* $\mathrm{E}\{(\hat{\mu}_1^* - \mu_1)^2 \mid \mu\} < \mathrm{E}\{(\hat{\mu}_1 - \mu_1)^2 \mid \mu\}$ *when*

$$\mathbf{w} = \mathbf{w}^0 \equiv \lambda(D + C'C)^{-1}\mathbf{1}, \tag{12}$$

where $\lambda^{-1} = \mathbf{1}'(D + C'C)^{-1}\mathbf{1}$.

Proof. We can easily show that

$$\mathrm{E}\{(\hat{\mu}_1^* - \mu_1)^2 \mid \mu\} \le \mathbf{w}'(D + C'C)\mathbf{w}. \tag{13}$$

The bound in (13) is a strictly convex function on the convex set

$$\left\{\mathbf{w} : \sum_{i=1}^{n} w_i = 1\right\}.$$

Moreover, its minimum must be attained at an interior point of $\mathbb{R}^n$. Hence it has a unique global minimum at such a point and we may invoke Lagrange's necessary condition to characterize the optimal weights $\mathbf{w}^0$ such as those given in equation (12). When these optimal weights are chosen, the upper bound in equation (13) becomes λ. Thus

$$\begin{aligned}\mathrm{E}\{(\hat{\mu}_1 - \mu_1)^2 \mid \mu\} - \mathrm{E}\{(\hat{\mu}_1^* - \mu_1)^2 \mid \mu\} &\ge 1/\sigma_1^{-2} - \lambda \\ &\propto \sum_{i=1}^{n}\sigma_i^{-2} - \sigma_1^{-2}\end{aligned}$$

$$
\begin{aligned}
& - \left(\sum_{i=1}^{n} c_i \sigma_i^{-2}\right)^2 \left(1 + \sum_{i=1}^{n} c_i^2 \sigma_i^{-2}\right)^{-1} \\
\geq & \sum_{i=2}^{n} \sigma_i^{-2} - \left(\sum_{i=2}^{n} c_i \sigma_i^{-2}\right)^2 \left(1 + \sum_{i=2}^{n} c_i^2 \sigma_i^{-2}\right)^{-1} \\
\geq & \sum_{i=2}^{n} \sigma_i^{-2} \left(1 + \sum_{i=2}^{n} c_i^2 \sigma_i^{-2}\right)^{-1} \\
> & \ 0,
\end{aligned}
$$

which proves our assertion. □

We now turn to the case $n = 2$, where a direct comparison of μ_1^* with the MLE becomes feasible. With the optimal weights given in (12), we obtain with $c = c_2$ for simplicity

$$\hat{\mu}_1^* = y_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + c^2}\, z,$$

where $z = y_2 - y_1$.

The MLE, $\hat{\mu}_1^{\text{mle}}$ would still use just y_1 unless the condition $|y_1 - y_2| \leq c$ is violated. If that condition fails,

$$\hat{\mu}_1^{\text{mle}} = \frac{\sigma_1^{-2} y_1 + \sigma_2^{-2}[y_2 - c]}{2} \quad \text{or} \quad \frac{\sigma_1^{-2} y_1 + \sigma_2^{-2}[y_2 + c]}{2}$$

according as $y_1 < [y_2 - c]$ or $y_1 > [y_2 + c]$. So the MLE does then bring y_2 into the estimation of μ_1, but only crudely through truncation. In Figure 1, we compare the performance of the WLE and the MLE through their mean square errors of estimation when $\sigma_1 = \sigma_2 = 1$ for simplicity.

In agreement with intuition, the WLE loses its advantage over the MLE as $\sigma_2^2 \to \infty$ or $c \to \infty$. The extra information in y_2 becomes useless in these extreme cases because of the uncontrolled noise or bias, respectively, in the second sample. When $c = 0$, the WLE becomes the MLE for the full data set.

When $\sigma_2^2 \to 0$, the problem under consideration becomes that of estimating a bounded normal mean, a much studied problem; see Bickel (1981) as well as Casella & Strawderman (1981). However, the WLE differs from the MLE. We compare them in Figure 2.

In the light of the comparisons made in the figures, we find the WLE superior to the MLE for two normal samples and the mean-square-error criterion. However, the WLE also seems superior to the MLE in that it imposes the condition $|\mu_2 - \mu_1| \leq c$ at "level 2" rather than "level 1" of the estimation process. Hence it seems likely to be much more robust than the MLE against violations in that assumption of a mean difference bounded by c.

A much more extensive treatment of this problem for $n = 2$ can be found in van Eeden & Zidek (2000). In particular, they show that with the Akaike optimum weights proposed in Section 4, the optimal estimator is

$$\hat{\mu}_1 = y_1 + \beta(\Delta) z, \tag{14}$$

where $\beta(\Delta) = \sigma_1^2(\sigma_Z^2 + \Delta^2)^{-1}$, $\Delta = \mu_2 - \mu_1$ and $\sigma_Z^2 = \sigma_1^2 + \sigma_2^2 = \text{var}(Y_2 - Y_1)$. However, Δ is unknown. Observe that if, in the optimal estimator, we set

a) $\Delta^2 = \infty$, then we obtain the classical estimator of μ_1;

b) $\Delta^2 = c^2$, then we obtain the estimator derived above;

c) $\Delta^2 = \min(c^2, z^2)$, then we obtain the natural naive plug-in estimator;

d) $\Delta^2 = 0$, then we obtain the classical unbiased estimator for the case of two samples with $\mu_2 = \mu_1$.

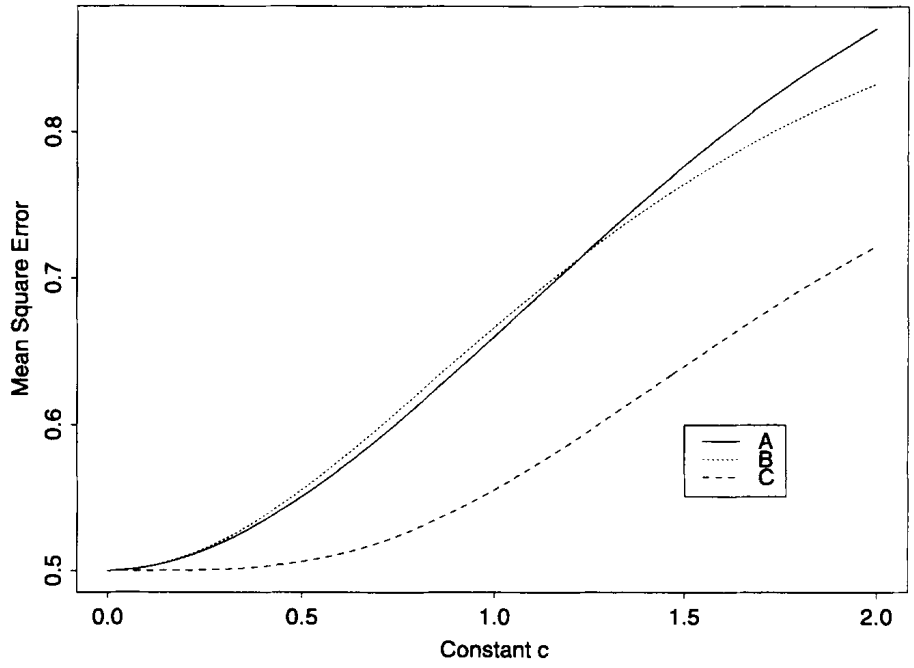

FIGURE 1: A comparison of the maximum WLE and MLE when $\sigma_1 = \sigma_2 = 1$. Curve A represents the maximum MSE of MLE when $\mu_1 = \mu_2$, B the maximum MSE of WLE over whole parameter space and C, the MSE of WLE on the line $\mu_1 = \mu_2$ (where it is constant).

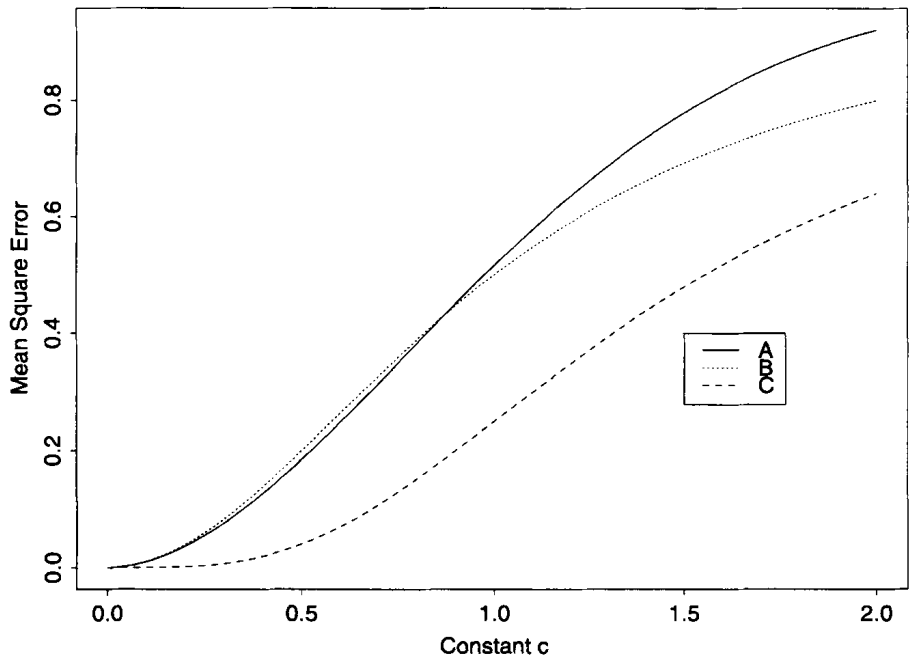

FIGURE 2: A comparison of the maximum MSEs of WLE and MLE when $\sigma_2 = 0$. Curve A represents the maximum MSE of MLE when $\mu_1 = \mu_2$, B the maximum MSE of WLE over the whole parameter space and C the maximum MSE of WLE when $\mu_1 = \mu_2$ (where it is constant).

These are all relevance weighted estimators corresponding to positive weights that would be expected to do well when $\Delta^2 \to 0$, i.e., the populations have nearly identical means. They show that each of the last three estimators improves on the classical estimator in having a uniformly smaller mean square error over the range $\Delta^2 \leq c^2$.

Other ways of restricting the parameter space arise naturally and provide the basis for relevance weighting. For example, van Eeden & Zidek (2001) consider the well-studied problem where $\mu_2 - \mu_1 \geq 0$. In that case, one anticipates that $Y_2 - Y_1 < 0$ will point to the possibility that in fact, $\mu_2 - \mu_1 = 0$, at least aproximately. Hence, in that case, one might treat Y_1 and Y_2 as replicates and potentially realize substantial gains in terms of MSE of estimation. To realize these gains, one would naturally estimate Δ by $\max(0, Y_2 - Y_1)$ in $\beta(\Delta)$ to get an adaptively weighted likelihood estimator.

In fact, van Eeden & Zidek (2001) consider this and a number of other adaptively weighted likelihood estimators that successfully trade bias for variance in more subtle ways. For example, they study the "Pitman" estimator

$$\hat{\mu}_{P1} = Y_1 - \sqrt{\frac{\sigma_1^4}{\sigma_1^2 + \sigma_2^2}} \frac{\phi\left(\frac{Z}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)}{Z\Phi\left(\frac{W}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)} Z,$$

where ϕ and Φ denote, respectively, the density function and the distribution function of the standard normal distribution.

The Pitman estimator was proposed and studied by Cohen & Sackrowitz (1970). (However, note that the formula above for $\hat{\mu}_{P1}$ differs from theirs when the variances are unequal. owing to an erroneous assumption they make in that case.) In this formula, $\beta(\Delta)$ has a complicated negative estimator. The successful bias-variance trade-off is in fact realized away from the point where $\Delta = 0$, the point where relevance weighting will maximize the gains achievable from combining all the sample information.

Before concluding this section, we observe that the Stein effect cannot obtain in the case $n = 2$, as is well known. However, the estimator obtained by taking $\Delta = z$ in equation (14) corresponds to the classical James–Stein estimator, albeit without being minimax in the case when Δ is unrestricted. In the next section, we derive that celebrated estimator.

5.2. Simultaneous estimation.

Suppose now that the μ_i are to be simultaneously estimated under the assumption of equi-relevance of all populations $j \neq i$ for the estimation of μ_i for all i. By this, in equation (4), we mean that $c_{ij} = c$ for all $i \neq j$ and some constant $c \geq 0$.

The assumption of equi-relevance suggests taking $\lambda_{ij} = \lambda$ or 1 according as $j \neq i$ or $j = i$. We readily find the WLE for μ_i to be

$$\hat{\mu}_{i\alpha} = (1 - \alpha)y_i + \alpha\bar{y}, \quad i = 1, \ldots, n,$$

where $\alpha = n\lambda/\{1 + (n-1)\lambda\}$.

Guided by equation (11), we compute

$$\mathrm{E}\left[\int f_i(y \mid \theta_i) \ln\{f_i(y \mid \hat{\theta}_{i\lambda_i})\}\, d\mu(y) \;\middle|\; \theta\right].$$

Here $\hat{\theta}_{i\lambda_i}$ becomes $\hat{\mu}_{i\alpha}$ while

$$\ln\{f_i(y \mid \hat{\theta}_{i\lambda_i})\} = -\tfrac{1}{2}(y - \hat{\mu}_{i\alpha})^2 + K,$$

where K represents a constant whose exact value is of no relevance to the argument. To find the equivalent of $\lambda(\theta)$ in equation (11), say $\alpha(\mu)$ with $\mu = (\mu_1, \ldots, \mu_n)$, we need to compute

$$\sum_{i=1}^{n} \mathrm{E}\left\{\left(Y_i^f - \hat{\mu}_{i\alpha}\right)^2 \mid \mu\right\}$$

and minimize the resulting function of α. Here we have used Y_i^f to represent a future independent copy of Y_i. Equivalently, we need to find the α that minimizes

$$\sum_{i=1}^{n} \mathrm{E}\left\{\left(\mu_i - \hat{\mu}_{i\alpha}\right)^2 \mid \mu\right\}.$$

A straightforward calculation gives $\alpha(\mu) = 1/(1+\tau^2)$, where

$$\tau^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mu_i - \bar{\mu})^2.$$

When $\mu_i \equiv \bar{\mu}$, $\alpha(\mu) = 1$ making the WLE $\hat{\mu}_{i\alpha} = \overline{Y}$, as one would wish. At the other extreme, $\tau \to \infty$ makes $\hat{\mu}_{i\alpha} \to y_i$, again in keeping with our intuition.

To estimate the unknown $\alpha(\mu)$ and get say $\hat{\alpha}$, we note that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

is an unbiased estimator of $1+\tau^2$ suggesting the possibility $\hat{\alpha} = K/s^2$ for some constant K. Although we cannot have $\hat{\alpha} = 1$ when $\mu_i \equiv \bar{\mu}$ we can have $\mathrm{E}(\hat{\alpha} \mid \mu) = 1$ under those circumstances by choosing $K = (n-3)/(n-1)$, leading us to finally select

$$\alpha = (n-3) / \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

With $\hat{\alpha}$ chosen as above, we get for the WLE

$$\begin{aligned} \hat{\mu}_{i\hat{\alpha}} &= y_i + \hat{\alpha}(\bar{y} - y_i), \\ &= \bar{y} + \left(1 - \frac{n-3}{\sum_{i'=1}^{n} (y_{i'} - \bar{y})^2}\right)(y_i - \bar{y}), \quad i = 1, \ldots, n, \end{aligned}$$

a well-known variant of the James–Stein estimator.

5.3. Higher-way layouts.

We may generalize the results of the previous subsection to the context of experimental design. There we can find simultaneous estimators of cell means for higher-way layouts encountered in the analysis of balanced experiments with two or more factors $A, B, \ldots$, each at several levels.

Suppose, for definiteness, that we consider the case of just two factors A and B. Denote the cell mean at levels i and j, respectively, for these factors by μ_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Reparameterize these means in the usual way in terms of main effects α_i^A, α_j^B and interactions α_{ij}^{AB} for all i and j with the usual side conditions for balanced designs. Then $\mu_{ij} = \mu + \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB}$, where μ denotes the overall mean of the cell means.

Suppose that independent measurements $X_{ij} \sim \mathsf{N}(\mu_{ij}, 1)$ are made, one for each cell, and that simultaneous estimates of the main effects and interactions are required. Let $\hat{\alpha}_i^A = \overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}$ and so on denote the least squares estimators of the effects of interest.

We assume a high degree of exchangeability among the cell means. To be precise, assume with respect to μ_{ij} that the $\mu_{ij'}$ with $j' \neq j$ are exchangeable, the degree of exchangeability being the same for all i, for any row i and column j. A similar assumption is made for columns. Finally, all the cell means for $i' \neq i$ and $j' \neq j$ are assumed to be exchangeable.

These exchangeability assumptions lead to a WL and, in turn, to WL estimators of the cell means that may be represented as

$$\hat{\mu}_{ij} = \hat{\mu} + \gamma^A \hat{\alpha}_i^A + \gamma^B \hat{\alpha}_j^B + \gamma^{AB} \hat{\alpha}_{ij}^{AB},$$

where the weights γ^r must be specified for $r = A, B, AB$.

To use the Akaike criterion as we did in the previous sections, we first find the weights that minimize

$$\mathrm{E}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\mu}_{ij} - \mu_{ij})^2\right\}.$$

We readily find, for example,

$$\gamma_{\mathrm{opt}}^A = 1 - \frac{I-1}{J\|\alpha^A\|^2 + (I-1)},$$

where α^A denotes the vector of main effects for treatment A. Analogous expressions are found for γ_{opt}^B and $\gamma_{\mathrm{opt}}^{AB}$.

Observe that when $\|\alpha^A\|^2 = 0$, then $\gamma_{\mathrm{opt}}^A = 0$ as one would expect on intuitive grounds; the main effects would not be used in estimating the cell mean when those effects are zero. However, the unknown γ_{opt}^A must be estimated and its estimator cannot have that intuitively appealing property. Instead, we embrace that property in the requirement that the estimator be an unbiased estimator of 0 when $\|\alpha^A\|^2 = 0$ by choosing $\hat{\gamma}^A = 1 - (I-3)\|\hat{\alpha}^A\|^{-2}/J$. For analogous reasons, we adopt

$$\hat{\gamma}^B = 1 - \frac{(J-3)}{I\|\hat{\alpha}^B\|^2} \quad \text{and} \quad \gamma_{\mathrm{opt}}^{AB} = 1 - \frac{(I-1)(J-1)-2}{\|\hat{\alpha}^{AB}\|^2}.$$

We thereby obtain the simultaneous estimators

$$\hat{\mu}_{ij} = \hat{\mu} + \hat{\gamma}^A \hat{\alpha}_i^A + \hat{\gamma}^B \hat{\alpha}_j^B + \hat{\gamma}^{AB} \hat{\alpha}_{ij}^{AB}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Sun (1996) proves that these estimators dominate the maximum likelihood estimators X_{ij} of the μ_{ij} in that they have uniformly smaller expected sum of mean square errors.

We can easily extend these estimators to higher-way layouts. In the three-way layout for example, the estimated optimal weight attached to the least squares estimator of the three-way interaction between factors A, B and C is

$$\hat{\gamma}_{ABC} = 1 - \frac{(I-1)(J-1)(K-1)-2}{\|\hat{\alpha}^{ABC}\|^2}.$$

We do not know if these resulting simultaneous estimators of the cell-means dominate their maximum likelihood counterparts as do those for the corresponding means in the two-way layout above.

6. GENERALIZED SHEWHART CHARTS

Suppose a product has gone through preliminary design and redesign phases as well as a final product capability analysis. At this stage, the product's quality characteristic X has a normal distribution with specified mean μ_0 and variance σ_0^2, that is, $X \sim \mathsf{N}(\mu_0, \sigma_0^2)$. In production, the

quality inspection program requires a sample of independent Y measurements to be taken at each inspection time-point t. Let X_t represent that sample's average. On the basis of $X_1, \ldots, X_t$, the hypothesis that the process has remained in control, $\mathcal{H}_0 : \mu_1 = \ldots = \mu_t = \mu_0$, must be tested. Thus, inferential interest concerns population t.

To define a suitable version of the WL for use here, we impose the additional condition that $f_t = f_t(\cdot \mid \theta_t)$, $\theta_t \in \Theta_t$, Θ_t being an open subset of Euclidean space and f_t known for all t, so that only the θ_t are unknown. Define the WL for population t by

$$\theta \to \prod_{j=1}^{t} \prod_{\ell=1}^{n_j} f_t^{\lambda_j/n_j}(y_{j\ell} \mid \theta_t),$$

λ_j/n_j being 0 when $n_j = 0$. The WLE is defined as any estimator yielding a noninfinite maximum of the WL. Armed with the WLE, we can develop a likelihood-ratio-test (LRT) in certain situations like that confronted here.

Below we specialize the WL for use in the Gaussian context of our analysis. However, the generality of the WL suggests that the method could lead to generalizations of other control charts such as those used for attribute or count data.

In the Gaussian case, various nonsequential methods for carrying out the test of $\mathcal{H}_0$ are available (see Montgomery 1991), and each of these leads to a test statistic and associated chart with μ_0 as the centerline and control limits (CLs) above and below that centerline. The standard test based on just X_t, gives Shewhart's $3 - \sigma$ control limits for example, where σ refers to σ_0. Two other popular methods derive from the moving average with span ℓ, and the exponentially weighted moving average which puts geometrically declining weight on successively older X_i values. We will show that each of these control charts are special cases of the generalized control chart derived here from the WL.

6.1. Generalized control charts.

Since the production process must initially be in control, the alternative to $\mathcal{H}_0$ would be that a change has occurred at some time prior to t, that is, $\mathcal{H}_1 : \mu_1 = \ldots = \mu_\tau = \mu_0$ and $|\mu_t - \mu_0| > 0$, $\tau \leq t - 1$. We can easily derive that test. To simplify our presentation, we consider a one-sided alternative, $\mathcal{H}_1 : \mu_1 = \ldots = \mu_\tau = \mu_0$ and $\mu_t > \mu_0$, $\tau \leq t - 1$. However, the same optimal choice of the weights obtains in the case of the two-sided alternative.

Since in this case $n_j \equiv 1$, we can readily compute the generalized weighted likelihood ratio (GWLR)

$$\Lambda = \frac{L(\mu_0)}{L(\mu)},$$

where

$$L(\mu) \propto \exp\left\{ -\sum_{j=1}^{t} \lambda_j \frac{(x_j - \mu)^2}{2\sigma_0^2} \right\},$$

ignoring irrelevant positive constants of proportionality, and $\mu = \sum_{j=1}^{t} \lambda_j x_j / \sum_{j=1}^{t} \lambda_j$. Rejection if $\Lambda < \Lambda_0$ can be translated as

$$\frac{\sum_{j=1}^{t} \lambda_j (x_j - \mu_0)}{\sqrt{\sum_{j=1}^{t} \lambda_j^2}} > C_0, \tag{15}$$

where C_0 must be chosen to get the required size α when $\mathcal{H}_0$ holds. Thus $C_0 = z_\alpha \sigma_0$, where $\mathrm{P}(Z > z_\alpha) = \alpha$ defines z_α.

We find the power function of the proposed test to be

$$\mathrm{P}\left\{ \frac{\sum_{j=1}^{t} \lambda_j (x_j - \mu_0)}{\sqrt{\sum_{j=1}^{t} \lambda_j^2}} > C_0 \,\middle|\, \mathcal{H}_1 \right\} = \mathrm{P}\left\{ Z > C_0 + \frac{\sum_{j=1}^{t} \lambda_j (\mu_0 - \mu_j)}{\sqrt{\sum_{j=1}^{t} \lambda_j^2}} \right\},$$

where $Z \sim \mathsf{N}(0, \sigma_0^2)$. Observe that

$$\sum_{j=1}^{t} \lambda_j(\mu_0 - \mu_j) = \left\{ \sum_{j=1}^{t} \lambda_j(\mu_t - \mu_0) - \sum_{j=1}^{t} \lambda_j(\mu_t - \mu_j) \right\}.$$

To optimize our choice of the weights and the power of our test, we adopt a minimax approach. First, we specify a worst-case scenario, a bound $1 \leq \tau_0 < \tau$ and constants M_j^*, $j = \tau_0+1, \ldots, t$, such that $|\mu_t - \mu_j| \leq M_j^*$, $j = \tau_0+1, \ldots, t$ ($M_t^* = 0$). These constraints allow us to express a variety of extreme departures from the centerline when the process goes out of control. We can allow for abrupt changes when loss of control will be due to a sudden shift. We can also accommodate the gradual changes one might associate with "wear out." Under any such pattern of change, we may choose the λ_j to maximize the minimum power achievable under the worst-case scenario determined by the associated constraints. We also assume an indifference zone, specifically that $\mu_t - \mu_0 > M_t$ for a specified M_t.

The power function attains that minimum over the region specified in the worst-case scenario at

$$\inf_{|\mu_t - \mu_j| \leq M_j^*} \frac{\sum_{j=1}^{t} \lambda_j(\mu_t - \mu_0) - \sum_{j=1}^{t} \lambda_j(\mu_t - \mu_j)}{\sqrt{\sum_{j=1}^{t} \lambda_j^2}} = \frac{\sum_{j=1}^{t} \lambda_j(M_t - M_j^*)}{\sqrt{\sum_{j=1}^{t} \lambda_j^2}}, \tag{16}$$

where for simplicity we have put $M_j^* = M_t$ when necessary, so that we can formally extend the summation in the last expression to all j from just $j = \tau_0 + 1, \ldots, t$.

Observe that the infimum in equation (16) is homogeneous in the λ_j, which may be multiplied by any constant without altering it. To identify a particular point at which that function attains its maximum, we need to add a condition on the weights, $\lambda_1 + \cdots + \lambda_t = 1$ being a natural choice. Observe that we can take $\lambda_j = 0$ when $M_t - M_j^* \leq 0$ to concentrate the remaining positive λ_j on the j for which $M_t - M_j^* > 0$. With these observations, we may relabel the λ_j and without loss of generality assume $M_t - M_j^* > 0$ for all j. Now t plays the role of $t - \tau_0$ less the number of j for which $M_t - M_j^* \leq 0$.

We find the maximum of the function in equation (16) to be attained at

$$\lambda_j = \frac{M_t - M_j^*}{t(M_t - \overline{M}^*)},$$

where $\overline{M}^* = \sum_{j=1}^{t} M_j^*/t$. The corresponding optimal test rejects $\mathcal{H}_0$ if

$$\frac{\sum_{j=1}^{t}(M_t - M_j^*)(x_j - \mu_0)}{\sqrt{\sum_{j=1}^{t}(M_t - M_j^*)^2}} > C_0.$$

One special case of interest reflects the abrupt change scenario with $t = \tau_0 + 1$, so that $M_t - M_j^* = 0$ for all $j < t$. Then we obtain the classic Shewhart $3 - \sigma$ chart. Another notable special case obtains when $M_j^* = 0$ for $j = \tau_0 + 1, \ldots, t$. Now we get the moving average chart. Finally, we get the exponentially weighted moving average chart by letting

$$\lambda_{t-j} = \left\{1 - (1-\gamma)^{(t-\tau_0)}\right\}\gamma(1-\gamma)^j, \quad j = 0, \ldots, t - \tau_0 - 1,$$

and

$$\lambda_{t-j} = 0, \quad j = t - \tau_0, \ldots, t - 1,$$

that is, $M_{t-j}^* = \left\{1 - (1-\gamma)^j\right\}M_t$, $j = 0, \ldots, t - \tau_0 - 1$. The corresponding optimal test statistic is

$$\frac{\sqrt{1-(1-\gamma)^2}\sum_{j=\tau_0+1}^{t}(1-\gamma)^{t-j}(x_j - \mu_0)}{\sqrt{1-(1-\gamma)^{2(t-\tau_0)}}}.$$

To obtain a chart not in the current tool box for on-line quality management, suppose that if the process goes out of control, it is thought that the mean will drift away from μ_0 gradually, changing linearly over successive production runs. This heuristic view suggests the choice $M_j^* = M_t(t-j)/(t-\tau_0)$ for $j = \tau_0+1, \ldots, t$, while as before, $M_j^* = M_t$ for $j = 1, \ldots, \tau_0$. Our theory then gives us an alternative to the moving average chart, that above in equation (15) with

$$\lambda_j = \frac{2(j-\tau_0)}{(t-\tau_0)(t-\tau_0+1)}. \tag{17}$$

The optimal test statistic becomes

$$\frac{\sqrt{6}}{\sqrt{(t-\tau_0)(t-\tau_0+1)\{2(t-\tau_0)+1\}}} \sum_{j=\tau_0+1}^{t} (j-\tau_0)(x_j-\mu_0).$$

For two-sided testing, equation (15) becomes

$$\frac{\sum_{j=1}^t \lambda_j(x_j-\mu_0)}{\sqrt{\sum_{j=1}^t \lambda_j^2}} > C_1 \quad \text{or} \quad < -C_1,$$

where $C_1 = z_{\alpha/2}\sigma_0$. Its power function is

$$\mathrm{P}\left\{\frac{\sum_{j=1}^t \lambda_j(x_j-\mu_0)}{\sqrt{\sum_{j=1}^t \lambda_j^2}} > C_1 \text{ or } < -C_1 \middle| H_1\right\} = \mathrm{P}\left\{Z > C_1 - \frac{\sum_{j=1}^t \lambda_j(\mu_0-\mu_j)}{\sqrt{\sum_{j=1}^t \lambda_j^2}}\right\} + \mathrm{P}\left\{Z < -C_1 - \frac{\sum_{j=1}^t \lambda_j(\mu_0-\mu_j)}{\sqrt{\sum_{j=1}^t \lambda_j^2}}\right\}.$$

The infimum corresponding to equation (16) turns out to be identical to it. Thus the optimal weights in equation (15) are also optimal in the case of a two-sided alternative, although now the critical value for rejection will have z_α replaced by $z_{\alpha/2}$.

6.2. Discussion.

At the practical level, much would need to be done before the charts of the generalized Shewhart type could be recommended for use in practice. In particular, its operating characteristics would need to be worked out. At the theoretical level, more needs to be done to extend the work described here to non-Gaussian situations.

We have left further development to future work. Our emphasis here is on the new methodology itself. We have shown that it formalizes and unifies control charting within the ambit of likelihood theory. Furthermore, it extends that theory, providing a way of customizing charts to accommodate context-specific knowledge of wear-out or failure mechanisms in the production process. As the example leading to equation (17) demonstrates, the construction of such charts is straightforward, at least at the conceptual level. Thus we feel the method may be of potential interest even at this early stage of development.

7. APPLICATION: PREDICTING HOCKEY SCORES

In this section, we give an illustrative application of the WL. The example expands on that given by Hu & Zidek (2001). To describe that application, let $Y_{ij} \sim \text{Poisson}(\zeta_{ij})$ be independently distributed random variables representing the number of goals team i scores in any one game played against team j in the National Hockey League's (NHL's) 1996–97 season. (In fact, we found a small negative correlation between Y_{ij} and Y_{ji} but for simplicity have ignored this as well as the home advantage.) Our specific goal will be the prediction of (Y_{12}, Y_{21}), where "1"

and "2" denote respectively the Vancouver Canucks and the Calgary Flames, a pair of teams that met five times during that season.

Given estimates $\hat{\zeta}_{ij}$, we have the predictive probability of $Y_{ij} = y_{ij}$ for any y_{ij} given by $\exp(-\hat{\zeta}_{ij})(\hat{\zeta}_{ij})^{y_{ij}}/y_{ij}!$ In turn, these distributions may be used to find the probabilities of a winning, losing and drawing the game. (Professor David A. Sprott pointed out in personal communication that these probabilities are actually point estimators of the true probabilities. In particular, they ignore the uncertainty in the population mean estimates. That would make confidence intervals desirable but we have not yet investigated this possibility. It should be noted that while Professor Sprott's comment is very reasonable, the outcome of further analysis is by no means a foregone conclusion. In particular, Professor Richard Smith has shown in an unpublished 1997 manuscript on his web page entitled "Predictive inference, rare events and hierarchical models," that "The usual Bayesian method may well be inferior to a crude maximum likelihood 'plug-in' approach.")

The use of the weighted likelihood seems especially appealing near the beginning of the season since little "direct" information about the relative strengths of teams 1 and 2 will be available. In fact, the maximum likelihood estimator (the average number of goals scored in previous matches between these two teams) would not even be available at the time of the first game. In contrast, the maximum weighted likelihood estimator would be able to bring in the information from games that each of these teams played against other teams in the league. That is, the WLE would use the information in the relevant sample in addition to the information in the direct sample.

To create the weights, we assume the scores for Vancouver in games played against teams other than Calgary to be exchangeable. The analogous assumption is made about the Calgary scores, with the weights implied by this assumption, we find the maximum weighted likelihood estimate of ζ_{ij} to be

$$\hat{\zeta}_{ij}^{\mathrm{WLE}} = \bar{y}_{ij} + \alpha_{ij}(\bar{y}_{i(j)} - \bar{y}_{ij}), \quad i, j = 1, 2, \tag{18}$$

where $\bar{y}_{ij}$ denotes the average number of goals for team i in the k_{ij} previous games played against team j, while $\bar{y}_{i(j)}$ represents the corresponding average number of goals for team i in the $k_{i(j)}$ games played against teams other than j.

Let us adopt the approximate Akaike criterion for selecting the weight α. We then find an optimal weight that may be estimated by

$$\hat{\alpha}_{ij} = \frac{\bar{y}_{ij}k_{ij}^{-1}}{\bar{y}_{ij}k_{ij}^{-1} + \bar{y}_{i(j)}k_{i(j)}^{-1} + (\bar{y}_{i(j)} - \bar{y}_{ij})^2} \tag{19}$$

for $i, j = 1, 2$. Note that in equation (19) the ratios $\bar{y}_{i(j)}k_{i(j)}^{-1}$ and $\bar{y}_{ij}k_{ij}^{-1}$ are estimates of the lack-of-precision (i.e., variance) of the respective sample average estimates for the goals of team i in games which do and do not involve team j, respectively. At the same time, the term $(\bar{y}_{i(j)} - \bar{y}_{ij})^2$ in equation (19) represents the estimated squared bias stemming from the use of the relevant but not directly relevant games involving teams other than j.

Notice that both (i) a lack of precision in the biased estimate of the population average number of goals for i when playing j $(i, j = 1, 2)$ and (ii) a large bias can result in a diminished value for the estimated optimal weight, in agreement with intuition. Notice also that the value of that estimated weight will change as the season wears on and both k_{ij} and $k_{i(j)}$ increase in size. In fact, as k_{ij} increases, the estimate will tend to zero so that we would tend to rely less and less on the less directly relevant data.

We now compare the WLE in equation (18) with estimated α_{ij} to the maximum likelihood estimator as point predictors of the goals in a future game. In Table 1, we show those predictions for each of the last four games when Vancouver met Calgary during the season.

TABLE 1: A game-by-game comparison of the maximum weighted likelihood and maximum likelihood predictors of Vancouver and Calgary goals for the last four games of the 1996–97 season.

Vancouver Score	Calgary Score	WLE Predictor Vancouver	WLE Predictor Calgary	MLE Predictor Vancouver	MLE Predictor Calgary
4	3	3.2	1.4	3.0	1.0
0	3	3.2	2.4	3.5	2.0
5	2	2.8	2.6	2.3	2.3
3	3	3.1	2.6	3.0	2.3
MSE		2.1		3.3	

Notice that as a predictor of the numbers of goals for the two teams, the WLE has a smaller overall mean square error. Not surprisingly, the WLE predictors tend to be more stable over these games. In the case of Game 2, for example, the MLE has to rely on the results of just a single game, Game 1, whereas the WLE relies on all the previous games played by Vancouver and Calgary.

TABLE 2: The estimated game-by-game probabilities (in percentage) of a win, loss or draw for the Vancouver Canucks against the Calgary Flames derived from the maximum weighted likelihood and maximum likelihood estimators for the last four games of the 1996–97 season.

Game Number	WLE Win	WLE Tie	WLE Lose	MLE Win	MLE Tie	MLE Lose
2	72	14	14	77	13	10
3	55	16	29	66	15	20
4	43	18	39	41	19	41
5	50	17	33	51	17	32

We can also use the estimators above to construct predictive distributions for the number of goals to be scored by each of the teams and, in turn, to estimate the probabilities of a win, loss or draw for Vancouver. Those estimates appear in Table 2. Observe how the predictive probabilities from the MLE and the WLE converge as the number of Vancouver–Calgary games increases; we see at most small differences in the last two of the four games above. On the other hand, we see substantial differences in the first two games.

8. RELATED WORK

The idea of "weighting" seems to go back a long way in the history of statistics, at least as far as the important paper of Stone (1977). Even before this idea was linked to the likelihood, it was used by Cleveland (1979) in another important development, locally weighted polynomial regression. Weerahandi & Zidek (1988, 1992) arrive at a Bayesian version of this approach when they tackle the thorny problem of using prior knowledge in estimating a smooth function (an infinite-dimensional parameter).

The idea of using local weighting in conjunction with the likelihood is suggested by Brillinger in his discussion of the paper by Stone (1977). That fact does not seem to have been generally recognized, however. Instead, commonly the so-called "locally weighted likelihood" is attributed to Tibshirani & Hastie (1987). These authors trim out likelihood components corresponding to points outside a window located at any given point in regressor space while assigning unit

weights to components corresponding to points inside that window. In other words, they choose 0–1 weights and obtain a special case of the likelihood suggested by Brillinger (see Stone 1977).

The idea of Tibshirani and Hastie has been substantially developed within the domain of nonparametric regression, summaries of that work being found in Fan & Gijbels (1996). In particular, Staniswalis (1989) refines the idea of Tibshirani and Hastie in that context. The idea has been also adapted for use with estimating equations (see Fan, Heckman & Wand 1995). As well, Brillinger (1989, 1990, 1992, 1994) contributed to the development of that theory in a series of applications. The notion of the weighted likelihood moved outside the framework of nonparametric regression seemingly with the paper of Field & Smith (1994). Strictly speaking, their method is a weighted estimating equation rather than a weighted likelihood approach, since their weights $w(y_i, \theta)$ depend not only on the observed y, but also on the unknown parameter θ to be estimated. So the weights would also need to be differentiated in finding the likelihood equations. Nevertheless the name has stuck (Dupuis & Morgenthaler 2002). Smith and Field actually get their estimating equation by correcting that obtained using these weights in conjunction with the derivative of the log-likelihood to give Fisher consistency. To achieve a robust estimator of θ, the weights down-weight likelihood components corresponding to exceptional y_i. However, there need not be an x here, as there were in the case of nonparametric regression.

Eguchi & Copas (1998) present another version of the WL that the authors say unifies versions presented by Copas (1995), Hjort & Jones (1996) and Loader (1996). Their weights, unlike those of Field & Smith (1994), take a rather special form, $K\{h^{-1}(y_i - t)\}$, so that the y_i close to t can be more heavily weighted, although the resulting estimating equations are corrected to be unbiased. The likelihood of Field & Smith (1994) is more general than that of Eguchi and Copas since the weights are allowed to depend on θ as well.

However, we are not aware of any formalization of the weighted likelihood prior to Hu (1994). In fact, Hu (1994) and Efron (1996) seem to have been the first to explicitly use the term "relevance" in the sense of Hu (1997), Hu & Zidek (1993a, 1993b, 1997, 2001), and Hu, Rosenberger & Zidek (2000).

When observations are selected sequentially, by some sequential design, the probability distribution of the responses may be time heterogeneous. Altman & Royston (1988) describe several examples of time heterogeneity in clinical trials. In these cases, the potential for time trends could bias results from traditional likelihood analyses. It is desired to use weighted likelihood methodology to obtain an estimator which takes the time trend into account. Hu & Rosenberger (2000) consider an application to a neurophysiology experiment. In that paper, the authors find that the weighted likelihood method significantly reduced both the bias and the mean square error.

9. CONCLUDING REMARKS

The theory of the weighted likelihood sketched in this paper unifies a diverse group of applications in which it has been used at least implicitly. We have seen how James–Stein estimation and nonparametric regression can be embraced within it. We have shown further how the theory provides a unified approach to quality control charting, providing hitherto unknown charts and yielding the well-known charts as special cases. Nonparametric smoothing for parameters other than the location can be done with the help of the WL. The theory extends the classical theory of Fisher and Wald and in particular, has an asymptotic theory resembling that of Wald.

The relevance of auxiliary data to inferences for a specified population can vary depending on the population characteristic under study. For example, if the population variances were thought to be identical, one would use unit weights for finding the appropriate WLE of any one population variance. At the same time, variable weights might be used in finding the WLE of their means if these were thought to be merely similar. Thus for each coordinate of a vector-valued parameter, the appropriate WLE might be found from a different set of weights. The Akaike entropy criterion could, however, still be used as above to find the optimal sample-based weights.

It is natural to consider the relationship and comparative advantages of our approach to the empirical Bayes approach. We cannot give a definitive answer at this time. We would observe, however, that the WL-based approach can ask much less in the way of prior specification than its Bayesian counterpart while at the same time allowing available prior knowledge to be incorporated. At the same time, our approach lacks the philosophical underpinnings and coherence of the Bayesian approach.

We find it curious but inexplicable that in certain cases like those of Section 5.2, our method gives a WLE identical to its classical empirical Bayesian counterpart. Another example can be found for the problem of simultaneously estimating the means of independent Poisson populations. To obtain that result, assume $X_i \sim \text{Poisson}(\mu_i)$, $i = 1, \ldots, n$, are independently distributed. Then under the sort of exchangeability assumptions made in the last section for normal means, we obtain the simultaneous WLEs given by $\hat{\mu}_i = \overline{X} + \beta(X_i - \overline{X})$, where β represents the weight. That weight may be optimized using the Akaike criterion. Using the sample moments to estimate the mean and variance in the result, we obtain

$$\hat{\beta} = (1 - \overline{X} S_X^{-2})_+, \tag{20}$$

where S_X^2 denotes the usual unbiased variance for the sample $X_1, \ldots, X_n$. It turns out that this estimator is an empirical Bayes estimator (Berger 1985, p. 297).

ACKNOWLEDGEMENTS

We are grateful to Dr. Nancy E. Heckman for her valuable comments on the work leading up to this paper. We are also indebted to Professor Howell Tong for his encouragement and stimulating comments. The work described in this paper was partially completed while the second author was on leave in the Department of Mathematics at the National University of Singapore and later in the Institute of Mathematics and Statistics at the University of Kent in Canterbury. Both institutions generously provided the necessary facilities and support. Finally, we owe a big debt of gratitude to the anonymous Associate Editor and the referees whose comments greatly improved our exposition.

REFERENCES

H. Akaike (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory: Held at Tsahkadsor, Armenia, September 2–8, 1971* (B. N. Petrov & F. Csáki, eds.), Akadémiai Kiadó, Budapest, pp. 276–281.

H. Akaike (1977). On entropy maximization principle. In *Applications of Statistics: Proceedings of the Symposium held at Wright State University, Dayton, Ohio, 14–18 June 1976* (P. R. Krishnaiah, ed.), North-Holland, Amsterdam, pp. 27–41.

H. Akaike (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30, 9–14.

H. Akaike (1982). On the fallacy of the likelihood principle. *Statistics & Probability Letters*, 1, 75–78.

H. Akaike (1983). Information measures and model selection. *Bulletin de l'Institut International de Statistique*, 50 (1), 277–290.

H. Akaike (1985). Prediction and entropy. In *A Celebration of Statistics: The ISI Centenary Volume* (A. C. Atkinson & S. E. Fienberg, eds.), Springer-Verlag, New York, pp. 1–24.

D. G. Altman & J. P. Royston (1988). The hidden effects of time. *Statistics in Medicine*, 7, 629–637.

J. O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition. Springer-Verlag, New York.

P. J. Bickel (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9, 1301–1309.

D. R. Brillinger (1989). Mapping aggregate birth data. In *Analysis of Data in Time: Proceedings of the 1989 International Symposium* (A. C. Singh & P. Whitridge, eds.), Statistics Canada Symposium Series no. 6, Statistics Canada, Ottawa, pp. 77–83.

D. R. Brillinger (1990). Spatial-temporal modelling of spatially aggregate birth data. *Survey Methodology*, 16, 255–269.

D. R. Brillinger (1992). Locally weighted analysis of spatially aggregate birth data: uncertainty estimation and display. In *Spatial Issues in Statistics: Symposium 91* (M. March & C. Weiss, eds.), Statistics Canada Symposium Series no. 8, Statistics Canada, Ottawa, pp. 71–79.

D. R. Brillinger (1994). Examples of scientific problems and data analyses in demography, neurophysiology, and seismology. *Journal of Computational and Graphical Statistics*, 3, 1–22.

G. Casella & W. E. Strawderman (1981). Estimating a bounded normal mean. *The Annals of Statistics*, 9, 870–878.

A. Charras & C. van Eeden (1991). Bayes and admissibility properties of estimators in truncated parameter spaces. *The Canadian Journal of Statistics*, 19, 121–134.

W. S. Cleveland (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

A. Cohen & H. B. Sackrowitz (1970). Estimation of the last mean of a monotone sequence. *The Annals of Mathematical Statistics*, 41, 2021–2034.

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

D. J. Dupuis & S. Morgenthaler (2002). Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics*, 30, 17–36.

A. W. F. Edwards (1984). *Likelihood*, First Paperback Edition. Cambridge University Press.

B. Efron (1996). Empirical Bayes methods for combining likelihoods (with discussion). *Journal of the American Statistical Association*, 91, 538–565.

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

P. L. Eubank (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

J. Fan & I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

J. Fan, N. E. Heckman & W. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90, 141–150.

C. A. Field & B. Smith (1994). Robust estimation—a weighted maximum likelihood approach. *International Statistical Review*, 62, 405–424.

W. Härdle (1990). *Applied Nonparametric Regression.* Cambridge University Press.

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

F. Hu (1994). *Relevance Weighted Smoothing and a New Bootstrap Method.* Unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, Vancouver.

F. Hu (1997). Asymptotic properties of relevance weighted likelihood estimations. *The Canadian Journal of Statistics*, 25, 45–60.

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

F. Hu, W. F. Rosenberger & J. V. Zidek (2000). The relevance weighted likelihood for dependent data. *Metrika*, 51, 223–243.

F. Hu & J. V. Zidek (1993a). *A Relevance Weighted Quantile Estimator.* Technical Report no. 134, Department of Statistics, The University of British Columbia, Vancouver.

F. Hu & J. V. Zidek (1993b). Relevant samples and their information. Mimeo.

F. Hu & J. V. Zidek (1997). Smoothing with the relevance weighted likelihood. Mimeo.

F. Hu & J. V. Zidek (2001). The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference* (S. E. Ahmed & N. M. Reid, eds.), Springer-Verlag, New York, pp. 211–234.

W. James & C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Contributions to the Theory of Statistics* (J. Neyman, ed.), University of California Press, pp. 361–379.

C. R. Loader (1996). Change point estimation using nonparametric regression. *The Annals of Statistics*, 24, 1667–1678.

D. C. Montgomery (1991). *Introduction to Statistical Quality Control*, Second Edition. Wiley, New York.

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

C. Stein (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Contributions to the Theory of Statistics* (J. Neyman, ed.), University of California Press, pp. 197–206.

S. M. Stigler (1990). The 1988 Neyman Memorial Lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5, 147–155.

C. Stone (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595–620.

L. Sun (1996). Shrinkage estimation in the two-way multivariate normal model. *The Annals of Statistics*, 24, 825–840.

R. J. Tibshirani & T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559–567.

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

C. van Eeden & J. V. Zidek (2000). Combining the data from two normal populations to estimate the mean of one when their means difference is bounded. Available from http://hajek.stat.ubc.ca/~jim/pubs

C. van Eeden & J. V. Zidek (2002). Combining sample information in estimating ordered normal means. *Sankhyā Series A*, 64, 588–610.

M. P. Wand & M. C. Jones (1995). *Kernel Smoothing*. Chapman & Hall, London.

X. Wang (2001). *Maximum Relevance Weighted Likelihood Estimation*. Unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, Vancouver.

X. Wang, C. van Eeden & J. V. Zidek (2001). Asymptotic properties of the maximum weighted likelihood estimator. *Journal of Statistical Planning and Inference*, in press.

X. Wang & J. V. Zidek (2002). Derivation of mixture distributions and the weighted likelihood function. Mimeo.

S. Weerahandi & J. V. Zidek (1988). Bayesian nonparametric smoothers. *The Canadian Journal of Statistics*, 16, 61–74.

J. V. Zidek & S. Weerahandi (1992). Bayesian predictive inference for samples from smooth processes. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, April 15–20, 1991* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds.), Oxford University Press, pp. 547–563.

Received 18 January 2002
Accepted 6 September 2002

Feifang HU: fh6e@virginia.edu
Department of Statistics, University of Virginia
Charlottesville, VA 22904, USA

James V. ZIDEK: jim@stat.ubc.ca
Department of Statistics, The University of British Columbia
333-6356 Agricultural Road, Vancouver
British Columbia, Canada V6T 1Z2