NAVAL HEALTH RESEARCH CENTER

COMPUTER ASSISTED IMPROVEMENT OF MEAN SQUARED ERROR IN STATISTICAL ESTIMATION

J. E. Angus

Report No. 91-17

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92186 – 5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Mathematics and Computers in Simulation 35 (1993) 1–13
North-Holland

MATCOM 924

Computer-assisted improvement of mean squared error in statistical estimation

John E. Angus

Department of Mathematics, The Claremont Graduate School, Claremont, CA, United States

Abstract

Angus, J.E., Computer-assisted improvement of mean squared error in statistical estimation, Mathematics and Computers in Simulation 35 (1993) 1–13.

A computer-assisted method for improving the mean squared error (MSE) in estimation for parametric models is presented. Assuming existence of nontrivial sufficient statistics, the method involves generation of Monte Carlo samples from the conditional distribution of the observables, given the sufficient statistic(s). The method is illustrated in connection with a simple back-propagation neural network model for estimating a logistic regression function, and a specific numerical example related to logistic regression is presented.

1. Introduction and basic method

Suppose that X is a random d-vector having density belonging to the dominated family $\{f(\cdot;\theta), \theta \in \Theta\}$ with dominating σ-finite measure μ. Suppose for the moment that θ is real-valued, and that it is desired to estimate θ using squared error loss. Typically, many estimators are available, and sometimes an estimator that is optimum in some sense (e.g., a uniform minimum variance unbiased estimator) can be shown to exist in theory. Often, the statistician is forced to use a suboptimum estimator. For example, this can happen if the optimum estimator is analytically intractable, or if economic considerations dictate that pre-existing algorithms, created without regard to optimality, must be used without modification. Examples of the former abound, while the latter situation exists, for example, in the case where a nonlinear regression function is estimated using a back-propagation neural network. White [13] describes such neural networks, and shows that the back-propagation algorithm leads to estimators of the network weights that are less efficient (i.e., have greater asymptotic variances) than ordinary nonlinear least squares estimators. See also [7,8,12] for descriptions of the back-propagation algorithm, which is a version of stochastic gradient descent. Angus [1] discusses connections between back-propagation neural networks and statistical nonlinear least squares. White's [13] landmark paper connecting neural networks with concepts in asymptotic statistical estimation presents a method for "correcting" the back-propagation algorithm to make its asymptotic efficiency equivalent to that of nonlinear least squares. The correction is analogous to the method of scoring in efficient likelihood estimation as presented in [11], for

Correspondence to: Prof. J.E. Angus, Department of Mathematics, The Claremont Graduate School, Blaisdell House, 143 E. Tenth Street, Claremont, CA 91711-3988, United States.

0378-4754/93/$06.00

example, and amounts to taking one nonlinear least squares Newton–Raphson step from the back-propagation solution.

A statistic $T = T(X)$ is by definition sufficient for the parameter θ if the conditional distribution of X given T does not depend on θ. The usual method for finding sufficient statistics is the application of the factorization criterion, which states that under fairly general conditions the statistic T is sufficient for θ if the density function of X factors as $f(x; \theta) = h(x)g(T(x); \theta)$, where the function h does not depend on θ, and the function g depends on x only through $T(x)$. When a sufficient statistic T is available, and $\hat{\theta} = \hat{\theta}(X)$ is an estimator of θ having finite second moment, then the Rao–Blackwell theorem implies that the new estimator defined by $\delta = \delta(T) = E(\hat{\theta} \mid T)$, the conditional expectation of $\hat{\theta}$ given T, has smaller mean squared error than $\hat{\theta}$. That is,

$$\mathrm{MSE}(\delta) = E(\delta - \theta)^2 \leq \mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

See [6] for further details on sufficiency, the factorization criterion and the Rao–Blackwell theorem.

It happens often that the improved estimator δ cannot be used. For example, the conditional expectation $E(\hat{\theta} \mid T)$ may be difficult to compute analytically and/or economic constraints may dictate that only the algorithms for computing $\hat{\theta}$ can be used. For example, a "canned" algorithm package that cannot be modified may be the only resource available. In the case of a neural network an advantage of the back-propagation learning algorithm, a relatively (statistically) inefficient parameter estimation algorithm, is that it can be implemented in hardware or firmware by essentially running the network architecture in reverse. However, the user of such a network would not be free to modify the estimation algorithm to achieve greater efficiency.

If it is relatively easy to generate (i.e., simulate) independent and identically distributed observations from the conditional distribution of X given T, then the following algorithm can be used to approximate $\delta(T)$ after taking the observation of X. Here, θ can now be a vector-valued parameter, and the abbreviation "iid" stands for "independent and identically distributed".

(i) If $X = x$ is observed, calculate the observed value t of T by $t = T(x)$.
(ii) Generate $X_1^*, \ldots, X_n^*$, iid according to the conditional distribution of X given $T = t$.
(iii) Compute (1)

$$\delta^* = \delta^*(X_1^*, \ldots, X_n^*, t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(X_i^*)$$

as the approximation to $\delta(t)$.

By Kolmogorov's strong law of large numbers [10], δ^* converges with probability 1 (with respect to the probability space on which the X_i^*'s are defined) to the mean of the conditional distribution of $\hat{\theta}(X)$ given $T = t$, i.e., to $\delta(t) = E(\hat{\theta}(X) \mid T = t)$. Also, it will be shown that δ^* achieves a reduction in mean squared error over the original estimator $\hat{\theta}$, although not as great a reduction as that achieved by δ. Moreover, an advantage of algorithm (1) is that the algorithms already existing for computation of $\hat{\theta}(X)$ can be reused with the simulation data without modification. That is, no (substantial) new algorithms are needed. For example, in the

case of the back-propagation neural network, the initial observation X (called an "examplar") is a random vector containing an observation of a set of inputs along with the corresponding output. To apply algorithm (1), n simulated exemplars $X_1^*, \ldots, X_n^*$ would be generated independently and identically distributed from the conditional distribution of X given a sufficient statistic T. The network would then be retrained n times (once for each simulated examplar), and the resulting weights from each of the n training sessions would be averaged to form the new estimated weights having improved mean squared error. Of course, for this approach to be successful, the form of the distribution of the exemplar vector X must be known and amenable to extraction of a nontrivial sufficient statistic.

By enlarging the original probability space, it may be assumed that all the above random variables are defined on the same probability space. Suppose for the moment that θ is real-valued. If S is a random variable with $E(S^2) < \infty$, the conditional variance of S given T is defined by $\mathrm{Var}(S \mid T) = E((S - E(S \mid T))^2 \mid T)$. The relative merits of the three estimators $\hat{\theta}$, δ and δ^* in terms of mean squared errors are summarized as follows (see the next section for a derivation of these):

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\delta) + E\big(\mathrm{Var}(\hat{\theta} \mid T)\big), \tag{2a}$$

$$\mathrm{MSE}(\delta) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\delta), \tag{2b}$$

$$\mathrm{MSE}(\delta^*) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\delta) + \frac{1}{n} E\big(\mathrm{Var}(\hat{\theta} \mid T)\big), \tag{2c}$$

where $\mathrm{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$. Thus, in theory, the computer-assisted estimator δ^* can be constructed to have mean squared error arbitrarily close to that of the improved estimator δ by simply increasing n. In particular, if $\hat{\theta}$ is unbiased for θ (i.e., $E(\hat{\theta}) = \theta$ for all $\theta \in \Theta$) and δ is the uniform minimum variance unbiased estimator of θ, then δ^* is also unbiased, the mean squared errors in (2) become variances, and δ^* can be made to have variance arbitrarily close to the optimum.

2. Mathematical background and notation

Following are mathematical and statistical notations and concepts that are used in the remainder of the paper. It is assumed that the reader is familiar with the rudiments of measure-theoretic probability, further details of which may be found in [4,10,11].

Vectors will be taken to be column vectors, and a superscript "t" will denote matrix transpose. $\mathbf{R}^d$ is the d-dimensional Euclidean space with the usual norm $\|\cdot\|$; $\mathfrak{R}^d$ represents the class of Borel subsets of $\mathbf{R}^d$.

Random d-vectors X are measurable $\mathbf{R}^d$-valued functions defined on a common probability space with sample space Ω, σ-field of events $\mathfrak{I}$, probability measure P and expectation operator E. If $d = 1$, X is referred to as a random variable. $N_d(m, \Lambda)$ will signify (depending on context) either the d-dimensional normal or Gaussian distribution with mean vector m and variance-covariance matrix Λ, or a random vector having this distribution. If X is a random d-vector, $\mathrm{Var}(X)$ denotes its variance covariance matrix $E((X - EX)(X - EX)^t)$.

A sequence of random d-vectors $\{X_n; n \geq 1\}$ converges in distribution to the random

d-vector X and $n \to \infty$, written $X_n \Rightarrow X$ as $n \to \infty$, if $P\{X_n \in A\} \to P\{X \in A\}$ as $n \to \infty$ for all Borel sets $A \in \mathfrak{R}^d$ with boundary ∂A satisfying $P\{X \in \partial A\} = 0$. The term "almost surely" (a.s.) is synonymous with "on a set of probability 1". Thus, for example, $X_n \to^{a.s.} X$ as $n \to \infty$ means that for all ω in a set having probability 1, $X_n(\omega) \to X(\omega)$ as $n \to \infty$. The indicator function of a set A, denoted by I_A, satisfies $I_A(\omega) = 1$ if $\omega \in A$, and is 0 otherwise.

If X and Y are random variables with $E|Y| < \infty$, then (one version of) the conditional expectation of Y given X, denoted by $E(Y|X)$, is a measurable function of X, $g(X)$, having the property that $\int_A g(X)\,dP = \int_A Y\,dP$ for any event A in the σ-field generated by X. With this definition of $E(Y|X)$, the convention is made that $E(Y|X=x) \equiv g(x)$ for all x in the range of X. If X and Y have joint density f_{XY} with respect to Lebesgue measure, and X has marginal density f_X with respect to Lebesgue measure, then $g(x) \equiv E(Y|X=x)$ may be computed using the formula

$$E(Y|X=x) = \int_{-\infty}^{\infty} y \frac{f_{XY}(x, y)}{f_X(x)}\,dy.$$

Conditional expectations have the following properties:

$E(E(Y\|X)) = E(Y)$,	if $E\|Y\| < \infty$,
$E(h(X)Y\|X) = h(X)E(Y\|X)$ a.s.,	if h is a measurable function of X, $E\|h(X)Y)\| < \infty$, $E\|Y\| < \infty$,
$E(Y_1 + Y_2\|X) = E(Y_1\|X) + E(Y_2\|X)$ a.s.,	if $E\|Y_1\| < \infty$, $E\|Y_2\| < \infty$.

From these properties, the relations (2) in Section 1 can be verified assuming that $E(\hat{\theta}^2) < \infty$. First, $E(\delta) = E(E(\hat{\theta}|T)) = E(\hat{\theta})$, and $E(\delta^*|T) = \delta(T)$, almost surely. Hence,

$$\mathrm{MSE}(\delta) = E(\delta - \theta)^2 = E(\delta - E\delta + E\delta - \theta)^2 = \mathrm{Var}(\delta) + \mathrm{bias}^2(\hat{\theta}),$$

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - \delta + \delta - \theta)^2 \\ &= \mathrm{MSE}(\delta) + E\Big(E\big((\hat{\theta} - \delta)^2 | T\big)\Big) + 2E\big((\delta - \theta)E(\hat{\theta} - \delta | T)\big) \\ &= \mathrm{MSE}(\delta) + E\big(\mathrm{Var}(\hat{\theta}|T)\big),\end{aligned}$$

$$\begin{aligned}\mathrm{MSE}(\delta^*) &= E(\delta^* - \theta)^2 = E(\delta^* - \delta + \delta - \theta)^2 \\ &= \mathrm{MSE}(\delta) + E\Big(E\big((\delta^* - \delta)^2 | T\big)\Big) + 2E((\delta - \theta)E(\delta^* - \delta | T)) \\ &= \mathrm{MSE}(\delta) + \frac{1}{n}E\big(\mathrm{Var}(\hat{\theta}|T)\big).\end{aligned}$$

3. A central limit theorem

In applying the algorithm (1), it is of interest to know whether the conditional distribution of $\sqrt{n}(\delta^* - \delta)$ given T approximates, in some sense, the unconditional distribution of $\sqrt{n}(\delta - \theta)$

when a random sample $X_1, \ldots, X_n$ is available. The former is the distribution that would result from repeated independent applications of algorithm (1) using the same fixed value of t on each application, and it depends on the initial estimator $\hat{\theta}$ as well as the choice of sufficient statistic T. This conditional distribution is a type of "bootstrap" distribution (see [5]), because it involves Monte Carlo resampling from a conditional distribution that depends on the outcome of the original sample. This question of approximation is difficult to answer in general, but can at least be addressed asymptotically for important special cases.

To be definite, assume that $X_1, X_2, \ldots, X_n$ are random d-vectors that constitute a random sample from a d-parameter regular exponential family with density $f(x; \eta) = \exp(\eta' x - c(\eta))$ with respect to a σ-finite measure μ on $(\mathbb{R}^d, \Re^d)$. Here, $\eta \in \mathcal{T} \subset \mathbb{R}^d$ where $\mathcal{T}$ contains an open rectangle in $\mathbb{R}^d$, and the function c is twice differentiable and satisfies

$$E(X_1) = \frac{\partial c(\eta)}{\partial \eta}, \qquad \operatorname{Var}(X_1) = E(X_1 X_1') - (EX_1)(EX_1)' = \frac{\partial^2 c(\eta)}{\partial \eta \, \partial \eta'}.$$

Let $\theta = \partial c(\eta)/\partial \eta = E(X_1)$ be the parameter of interest. It can be shown that $\operatorname{Var}(X_1)$ depends on η only through θ, and that the Fisher information matrix $I(\theta)$ for θ satisfies $I^{-1}(\theta) = \operatorname{Var}(X_1)$ (see [9, p. 127], for example). Suppose that it is of interest to estimate $\theta = E(X_1)$ efficiently and in unbiased fashion. For this problem, it is easy to show (see [9, Example 5.3, p. 438, and Section 6.5]) that the most efficient unbiased estimator of θ is given by $\delta(X_1, \ldots, X_n) = (1/n)\sum_{i=1}^n X_i = S_n/n$. Suppose that the initial estimator of θ is taken to be $\hat{\theta} = X_1$. It follows from the multivariate central limit theorem that $\sqrt{n}(\delta - \theta) \Rightarrow N_d(0, I^{-1}(\theta))$ as $n \to \infty$. If algorithm (1) is applied in this context with $T = S_n$, it would be desirable to have, with probability 1, the conditional distribution of $\sqrt{n}(\delta^* - \delta)$ given T also converge to $N_d(0, I^{-1}(\theta))$ in some sense. The following theorem, which applies to more general situations, helps address this issue. The result of this theorem is very similar to the fundamental Bootstrap Central Limit Theorem (see [2], for example). Before stating and proving Theorem 2, the following lemma is needed. Its proof is essentially the same as that of the result of [11, p. 147, Problem 4.7].

Lemma 1. *For each $n \geq 1$, let $X_{n1}, X_{n2}, \ldots, X_{nn}$ be independent $\mathbb{R}^d$-valued random vectors with $EX_{ni} = 0$, $EX_{nk}X_{ni}' = \Lambda_{ni}$, and let μ_{ni} be the probability distribution of X_{ni}. Suppose that as $n \to \infty$,*

(i) $$\frac{1}{n}\sum_{i=1}^{n} \Lambda_{ni} \to \Lambda;$$

(ii) *for every $\epsilon > 0$,*

$$\frac{1}{n}\sum_{i=1}^{n} \int_{(x: \|x\| > \epsilon\sqrt{n})} \|x\|^2 \mu_{ni}(dx) \to 0, \quad \text{as } n \to \infty.$$

Then $(X_{n1} + \cdots + X_{nn})/\sqrt{n} \Rightarrow N_d(0, \Lambda)$ as $n \to \infty$.

Theorem 2 can now be stated and proved.

Theorem 2. *Assume that $X_1, X_2, \ldots$ are iid random d-vectors defined on a fixed probability space $(\Omega, \mathfrak{I}, P)$, each having covariance matrix Λ. Let $S_n = X_1 + \cdots + X_n$, and for each n, let $\mu_n(S_n, \cdot)$ be a regular conditional probability distribution of X_1 given S_n. Conditioned on S_n, let $X_{n1}^*, \ldots, X_{nn}^*$ be iid according to $\mu_n(S_n, \cdot)$. Then, with probability 1, as $n \to \infty$,*

$$P\left\{\frac{X_1^* + \cdots + X_n^* - S_n}{\sqrt{n}} \in A \mid S_n\right\} \to P\{N_d(0, \Lambda) \in A\},$$

for every Borel set $A \in \mathfrak{R}^d$ with boundary ∂A satisfying $P\{N_d(0, \Lambda) \in \partial A\} = 0$.

Proof. Clearly, $E(X_1^* \mid S_n) = E(X_1 \mid S_n) = S_n/n$ almost surely. The idea of the proof is to apply Lemma 1 to the sequence $\{X_{ni}^* - S_n/n;\ 1 \leq i \leq n\}$, conditioned on S_n, and find a set of probability 1 where the conditions of the lemma apply. Now let

$$Y_{ni} = X_{ni}^* - E(X_1 \mid S_n) \overset{\text{a.s.}}{=} X_{ni}^* - \frac{S_n}{n}.$$

Then using obvious symmetries,

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_{ni}Y_{ni}^t \mid S_n) \overset{\text{a.s.}}{=} E(X_1X_1^t \mid S_n) - \frac{S_n}{n}\left(\frac{S_n}{n}\right)^t.$$

By the strong law of large numbers, $S_n/n \to^{\text{a.s.}} EX_1$ as $n \to \infty$. Let σ_n be the σ-field generated by $\{S_n, S_{n+1}, \ldots\}$. Then $E(X_1X_1^t \mid S_n) = E(X_1X_1^t \mid \sigma_n)$. But σ_n decreases to $\bigcap_{n>1}\sigma_n$, the tail σ-field of the sequence $\{S_n, n > 1\}$, and it follows [4, p. 228] that

$$E(X_1X_1^t \mid S_n) \overset{\text{a.s.}}{\to} E\left(X_1X_1^t \mid \bigcap_{n>1}\sigma_n\right),$$

as $n \to \infty$. But $\bigcap_{n>1}\sigma_n$ is contained in the σ-field of permutable events [10, pp. 373, 374] and hence contains only events of probability 0 or 1 by the Hewitt–Savage 0-1 law, so that the latter conditional expectation is constant almost surely [10, p. 374]. But since

$$E\left(E\left(X_1X_1^t \mid \bigcap_{n>1}\sigma_n\right)\right) = E(X_1X_1^t),$$

that constant must be $E(X_1X_1^t)$. It follows that

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_{ni}Y_{ni}^t \mid S_n) \overset{\text{a.s.}}{\to} \Lambda,$$

as $n \to \infty$. This establishes (i) in Lemma 1.

Note that it is sufficient (and necessary) to verify condition (ii) of Lemma 1 only for rational $\epsilon > 0$. Again, employing obvious symmetries, and letting $C > 0$ be a constant,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E\left(\|Y_i\|^2 I\left\{\|Y_i\| > \epsilon\sqrt{n}\right\} \mid S_n\right) \\
&\quad = E\left(\|X_1 - E(X_1 \mid S_n)\|^2 I\left\{\|X_1 - E(X_1 \mid S_n)\| > \epsilon\sqrt{n}\right\} \mid S_n\right) \\
&\quad \leqslant \left(C + \left\|\frac{S_n}{n}\right\|\left(\left\|\frac{S_n}{n}\right\| + 2\sqrt{C}\right)\right) P\left\{\|X_1 - E(X_1 \mid S_n)\| > \epsilon\sqrt{n} \mid S_n\right\} \\
&\qquad + \left\|\frac{S_n}{n}\right\|^2 P\left\{\|X_1\|^2 > C \mid S_n\right\} + E\left(\|X_1\|^2 I\left\{\|X_1\|^2 > C\right\} \mid S_n\right) \\
&\qquad + 2\left\|\frac{S_n}{n}\right\| E\left(\|X_1\| I\left\{\|X_1\|^2 > C\right\} \mid S_n\right),
\end{aligned}$$

with all inequalities being understood as "almost sure" inequalities. By Chebyshev's inequality,

$$P\left\{\|X_1 - E(X_1 \mid S_n)\| > \epsilon\sqrt{n} \mid S_n\right\} \leqslant (\sqrt{n}\epsilon)^{-1}\left(E(\|X_1\| \mid S_n) + \left\|\frac{S_n}{n}\right\|\right),$$

which converges almost surely to zero by using arguments similar to those used in verifying (i) of Lemma 1, and the strong law of large numbers. Using Chebyshev's inequality again as well as the same arguments just mentioned, it follows that almost surely

$$\begin{aligned}
&\left\|\frac{S_n}{n}\right\|^2 P\left\{\|X_1\|^2 > C \mid S_n\right\} + E\left(\|X_1\|^2 I\left\{\|X_1\|^2 > C\right\} \mid S_n\right) \\
&\qquad + 2\left\|\frac{S_n}{n}\right\| E\left(\|X_1\| I\left\{\|X_1\|^2 > C\right\} \mid S_n\right) \\
&\quad \leqslant C^{-1}\left\|\frac{S_n}{n}\right\|^2 E\left(\|X_1\|^2 \mid S_n\right) + E\left(\|X_1\|^2 I\left\{\|X_1\|^2 > C\right\} \mid S_n\right) \\
&\qquad + 2\left\|\frac{S_n}{n}\right\| E\left(\|X_1\| I\left\{\|X_1\|^2 > C\right\} \mid S_n\right),
\end{aligned}$$

which converges almost surely as $n \to \infty$ to

$$\begin{aligned}
&C^{-1}\|EX_1\|^2 E\left(\|X_1\|^2\right) + E\left(\|X_1\|^2 I\left\{\|X_1\|^2 > C\right\}\right) \\
&\qquad + 2\|EX_1\| E\left(\|X_1\| I\left\{\|X_1\|^2 > C\right\}\right).
\end{aligned}$$

Since C was arbitrary, letting $C \to \infty$, and applying the Dominated Convergence Theorem [4] to the last two terms, the entire expression tends to 0 and it follows that condition (ii) of Lemma 1 is satisfied almost surely for all rational $\epsilon > 0$. The number of exceptional ω sets where the aforementioned calculations and inequalities fail, as well as those where the μ_n, $n \geqslant 1$, fail to

be probability measures, is at most countable, and each has probability 0. Hence, the conclusion of the theorem follows. □

It follows from Theorem 2 that if algorithm (1) is applied in the context of the discussion at the beginning of this section, then along almost all sample sequences $X_1, X_2, \ldots$, the conditional distribution of $\sqrt{n}(\delta^* - \delta)$ given S_n converges in distribution to $N_d(0, I^{-1}(\theta))$ as $n \to \infty$, which is the same limiting distribution as that of $\sqrt{n}(\delta - \theta)$.

4. Application to a back-propagation neural network implementation of logistic regression

To illustrate the method of applying algorithm (1), consider the logistic regression model. In this model, a dichotomous random variable Y has the conditional probability mass function $P\{Y = 1 \mid X\} = 1 - P\{Y = 0 \mid X\} = F(X'\theta)$ where F is the cumulative distribution function of the logistic distribution, namely $F(x) = e^x/(1 + e^x)$. Here, X is a random p-vector of explanatory variables, and θ is a p-vector of unknown parameters. This model arises in bioassay, medical diagnosis, linear discriminant analysis and many other statistical contexts. Suppose that random samples $Y_{i,1}, \ldots, Y_{i,m_i}$ from the conditional distributions of $Y \mid X_i$, $i = 1, \ldots, K$, are available and it is desired to estimate θ. Several statistical techniques are available for estimating θ, including maximum likelihood, and minimum logit chi-square (see [3] and Section 5). In fact, Berkson [3] studies the problem of improving the mean squared error of the minimum logit chi-square estimator of θ through the use of the Rao–Blackwell theorem. The initial estimation problem, however, also fits naturally into a simple two-layer back-propagation neural network with logistic sigmoid response function (see Fig. 1).

K exemplars consisting of the pairs $(X_1, \hat{p}_1), \ldots, (X_K, \hat{p}_K)$, would be presented to the network of Fig. 1, where $\hat{p}_i = (1/m_i)\sum_{j=1}^{m_i} Y_{i,j}$. The network, using the back-propagation algorithm, would then "learn" the connection weights (the θ_i's) that provide a minimum of the quantity $\sum_{i=1}^{K}(\hat{p}_i - F(X_i'\theta))^2$. (A modified value of $\hat{p}_i$ that lies strictly between 0 and 1 may be necessary to avoid numerical instabilities, see Section 5.) The mean squared errors of the estimators of the θ_i's thus obtained can be improved using the method of Section 1 as follows. It is assumed that the analyses are all conditional on the values of the X_i's. That is, the X_i's are treated as constants.

It follows directly from the factorization criterion for sufficiency that $T = T(Y) = \sum_{i=1}^{K}\sum_{j=1}^{m_i} X_i Y_{i,j}$ is sufficient for θ, where $Y = (Y_{1,1}, \ldots, Y_{1,m_1}, Y_{2,1}, \ldots, Y_{2,m_2}, \ldots, Y_{K,1}, \ldots, Y_{K,m_K})'$. Denoting by y the observed value of Y, it is easy to show that the conditional distribution of Y,

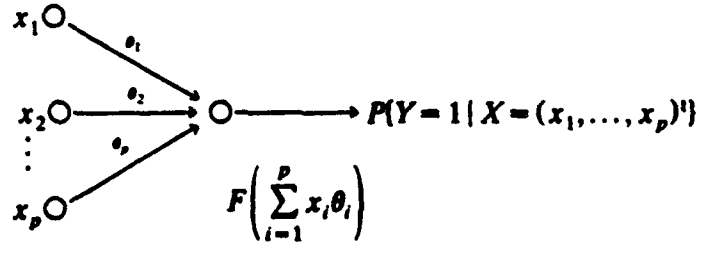

Fig. 1. Simple two-layer back-propagation neural network for logistic regression.

given $T = T(y)$, is uniform over the region defined by $D_y = \{z: T(z) = T(y)\}$. To implement the method of Section 1, independent and identically distributed observations $Y^*(1), \ldots, Y^*(n)$, each uniformly distributed over D_y, are generated via a simulation, and the back-propagation network is retrained on each of n "new" exemplars

$$(X_1, \hat{p}_1^*(v)), \ldots, (X_K, \hat{p}_K^*(v)), \quad \hat{p}_i^*(v) = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{i,j}^*(v), \; v = 1, \ldots, n,$$

yielding the set $\hat{\theta}^*(1), \ldots, \hat{\theta}^*(n)$ of back-propagation estimates of θ. The improved MSE estimator of θ is then $(1/n)\sum_{v=1}^{n} \hat{\theta}^*(v)$.

5. Numerical example of a computer-assisted improvement in mean squared error

The following numerical example is intended to illustrate the many different concepts presented in this paper in connection with algorithm (1). Assume the logistic regression model of the previous section with $p = 1$ (one independent variable), and

$$P\{Y = 1 \mid X\} = 1 - P\{Y = 0 \mid X\} = \frac{e^{\theta X}}{1 + e^{\theta X}} \equiv F(\theta X).$$

In order to obtain closed-form solutions so that comparisons can be made and for ease of exposition, assume that instead of minimizing $\sum_{i=1}^{K} (\hat{p}_i - F(\theta X_i))^2$, θ is estimated by minimizing Berkson's [3] logit χ^2 defined by

$$\chi_L^2(\theta) = \sum_{i=1}^{K} m_i \hat{p}_i (1 - \hat{p}_i)(\theta X_i - \text{logit}(\hat{p}_i))^2,$$

where $\text{logit}(p) = \ln(p/(1-p))$, $p \in (0, 1)$. To avoid singularities, $\hat{p}_i$ will be taken to be the modified estimator

$$\hat{p}_i = \frac{\sum_{j=1}^{m_i} Y_{i,j} + \tau_1}{m_i + \tau_1 + \tau_2},$$

for fixed $\tau_1, \tau_2 > 0$, which corresponds to the Bayes estimator of $p_i = F(\theta X_i)$ using squared error loss and a beta(τ_1, τ_2) prior distribution. It is easily shown by differentiation that $\chi_L^2(\theta)$ has a unique minimum at the value of θ given by

$$\hat{\theta} = \frac{\sum_{i=1}^{K} X_i m_i \hat{p}_i (1 - \hat{p}_i) \, \text{logit}(\hat{p}_i)}{\sum_{i=1}^{K} X_i^2 m_i \hat{p}_i (1 - \hat{p}_i)}.$$

Now, suppose that $K = 2$, $m_1 = m_2 = 5$, $\tau_1 = \tau_2 = 0.01$, $X_1 = 1$ and $X_2 = 2$. To simplify notation, define $Y_1 = \sum_{i=1}^{5} Y_{1,i}$ and $Y_2 = \sum_{i=1}^{5} Y_{2,i}$, so that $\hat{p}_1 = (Y_1 + 0.01)/5.02$, $\hat{p}_2 = (Y_2 +$

0.01)/5.02, the sufficient statistic for θ becomes $T = T(Y_1, Y_2) = Y_1 + 2Y_2$ and the minimum logit χ^2 estimator of θ is

$$\hat{\theta} = \hat{\theta}(Y_1, Y_2) = \frac{\hat{p}_1(1-\hat{p}_1)\,\text{logit}(\hat{p}_1) + 2\hat{p}_2(1-\hat{p}_2)\,\text{logit}(\hat{p}_2)}{\hat{p}_1(1-\hat{p}_1) + 4\hat{p}_2(1-\hat{p}_2)}.$$

Note that Y_1 and Y_2 are independent random variables, with $Y_i \sim$ binomial(5, $F(\theta i)$), $i = 1, 2$. In practice, θ is unknown, but suppose that its true value is $\theta = 1$. In this example, the expected value of $\hat{\theta}$ is easily computed by

$$E(\hat{\theta}) = \sum_{i,j=0}^{5} \hat{\theta}(i, j) \frac{5!5!}{i!(5-i)!\,j!(5-j!)} F(\theta)^i (1-F(\theta))^{5-i} F(2\theta)^j (1-F(2\theta))^{5-j},$$

and the mean squared error of $\hat{\theta}$ is

$$\text{MSE}(\hat{\theta}) = \sum_{i,j=0}^{5} \left(\hat{\theta}(i, j) - \theta\right)^2 \frac{5!5!}{i!(5-i)!\,j!(5-j!)} \times F(\theta)^i (1-F(\theta))^{5-i} F(2\theta)^j (1-F(2\theta))^{5-j},$$

from which

$$E(\hat{\theta}) = 0.81296 \quad \text{and} \quad \text{MSE}(\hat{\theta}) = 1.15163, \qquad \text{for } \theta = 1.$$

Thus, the estimator of θ is biased. Denoting $S_t = \{(y_1, y_2): y_1 + 2y_2 = t, 0 \leq y_1, y_2 \leq 5\}$, the conditional distribution of (Y_1, Y_2) given $T = Y_1 + 2Y_2 = t$, is

$$P\{(Y_1, Y_2) = (y_1, y_2) \mid T = t\} = \binom{5}{y_1}\binom{5}{y_2}\left[\sum_{(y_1,y_2)\in S_t} \binom{5}{y_1}\binom{5}{y_2}\right]^{-1},$$

for $(y_1, y_2) \in S_t$, $t \in \{0, 1, \ldots, 15\}$, and is 0 elsewhere. (Note that this distribution is not "uniform", since in this example, the observables were reduced initially by sufficiency to Y_1 and Y_2, whereas the variables $Y_{i,j}$ in Section 3 were not.) The improved estimator of θ is $\delta(T) = E(\hat{\theta} \mid T)$, which, in this example, can be computed fairly easily for any given $t \in \{0, 1, \ldots, 15\}$ by the formula

$$\delta(t) = E(\hat{\theta} \mid T = t) = \sum_{(y_1,y_2)\in S_t} \hat{\theta}(y_1, y_2)\binom{5}{y_1}\binom{5}{y_2}\left[\sum_{(y_1,y_2)\in S_t} \binom{5}{y_1}\binom{5}{y_2}\right]^{-1}.$$

The unconditional mean squared error of δ is

$$\text{MSE}(\delta) = \sum_{t=0}^{15} \left(E(\hat{\theta} \mid T = t) - \theta\right)^2 P\{T = t\},$$

where $P\{T = t\}$ is computed from the joint unconditional binomial distribution of (Y_1, Y_2). Hence, from (2),

$$E\left(\text{Var}(\hat{\theta} \mid T)\right) = \text{MSE}(\hat{\theta}) - \text{MSE}(\delta).$$

Table 1
Summary of computations

t	$S_t = \{(y_1, y_2): y_1 + 2y_2 = t, 0 \le y_1, y_2 \le 5\}$	$P(T = t)$	$\delta(t) = E(\hat{\theta} \mid T = t)$
0	(0, 0)	$3 \cdot 10^{-8}$	−3.60000
1	(1, 0)	$4.6 \cdot 10^{-1}$	−1.09429
2	(0, 1), (2, 0)	$3.8 \cdot 10^{-6}$	−0.23644
3	(1, 1), (3, 0)	$2.4 \cdot 10^{-5}$	−0.45605
4	(0, 2), (2, 1), (4, 0)	0.00012	−0.21953
5	(1, 2), (3, 1), (5, 0)	0.00051	−0.26307
6	(0, 3), (2, 2), (4, 1)	0.00184	−0.03796
7	(1, 3), (3, 2), (5, 1)	0.00576	−0.06173
8	(0, 4), (2, 3), (4, 2)	0.01564	0.06173
9	(1, 4), (3, 3), (5, 2)	0.03704	0.03796
10	(0, 5), (2, 4), (4, 3)	0.07533	0.26307
11	(1, 5), (3, 4), (5, 3)	0.13179	0.21953
12	(2, 5), (4, 4)	0.19290	0.45605
13	(3, 5), (5, 4)	0.22472	0.23644
14	(4, 5)	0.20362	1.09429
15	(5, 5)	0.11070	3.60000

Note: The values of $\delta(t) = E(\hat{\theta} \mid T = t)$ are independent of the true value of θ, while the values of $P(T = t)$ are computed assuming $\theta = 1$.

When $\theta = 1$, these computations yield

$$\mathrm{MSE}(\delta) = 1.11699 \quad \text{and} \quad E\big(\mathrm{Var}(\hat{\theta} \mid T)\big) = 0.03464.$$

Table 1 lists the possible values of T, along with the sets S_t, values of $P(T = t)$ (assuming $\theta = 1$) and the value of the estimator $\delta(T)$ at $T = t$.

In practice, the data will be collected, yielding an observed value for T, say $T = t = 11$, but typically it will happen that the computation of $\delta(t)$ is intractable. Algorithm (1) would then allow one to approximate the estimate $\delta(t)$ as follows. Take $n = 1000$, for example.

(i) Generate $(Y_1^*(i), Y_2^*(i))$, $i = 1, \ldots, 1000$, independent and identically distributed from the distribution

$$p(y_1, y_2) = \binom{5}{y_1}\binom{5}{y_2}\Bigg[\sum_{(y_1, y_2) \in S_{11}} \binom{5}{y_1}\binom{5}{y_2}\Bigg]^{-1}, \quad (y_1, y_2) \in S_{11}.$$

(ii) Compute

$$\delta^* = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\theta}(Y_1^*(i), Y_2^*(i)).$$

Carrying out this simulation algorithm yielded a value of $\delta^* = 0.2215$, which is very close to the exact value of $\delta(11) = 0.21953$.

By formula (2), for this example, this procedure yields an estimator δ^* with unconditional mean squared error

$$\mathrm{MSE}(\delta^*) = \mathrm{MSE}(\delta) + 0.001\ E\big(\mathrm{Var}(\hat{\theta}|T)\big) = 1.11699 + (0.001)(0.03464) = 1.11702.$$

Hence, while the computer-assisted estimator of θ, namely δ^*, has the same unconditional mean as the original estimator $\hat{\theta}$ (i.e., $E(\delta^*) = E(\hat{\theta}) = 0.81296$), its mean squared error about the true value of $\theta = 1$ has been improved ($\mathrm{MSE}(\delta^*) = 1.11702$ compared to $\mathrm{MSE}(\hat{\theta}) = 1.14163$).

Clearly, by the definition of the estimate $\delta(11)$, this algorithm yields an approximation to $\delta(11)$. Of course, in this example, $\delta(11)$ is easily computed and the algorithm is not necessary. In other situations, however, the generation of random samples from the conditional distribution of the observable given the sufficient statistic will be easy compared to the direct computation of $\delta(t)$.

5. Summary and conclusions

A computer-assisted Monte Carlo method for improving the mean squared error in parametric estimation has been presented. The method assumes that it is possible to derive a nontrivial sufficient statistic for the unknown parameter(s), and that is is relatively straightforward to simulate the conditional distribution of the observables given the sufficient statistic. In addition, a central limit theorem, similar to the central limit theorem for the bootstrap mean, has been proven, which demonstrates that for an important class of problems the Monte Carlo distribution of the computer-assisted estimator (asymptotically) approximates the sampling distribution of the "ideal" estimator for the problem (i.e., the estimator that the computer-assisted estimator is aiming to approximate). The application of the method to the improved estimation of the coefficients of a logistic regression function that has been implemented on a simple back-propagation neural network has been illustrated. Finally, a numerical example related to the logistic regression function has demonstrated the mean squared error improvement as well as the concepts behind and implementation of the algorithm.

The idea of computer-assisted statistical analysis is not new. An idea similar in spirit to that considered in this paper is the technique of bootstrapping [5], which has revolutionized the practice of applied statistics. In addition, a careful consideration of the theory behind the algorithm presented in this paper shows that the algorithm is just a simple Monte Carlo technique for evaluating an integral, that is, the (possibly vector-valued) integral representing the first moment(s) of the conditional distribution of the basic estimator given the sufficient statistic. Therefore, the method could be improved further by implementing improved Monte Carlo integration techniques, an area that is currently being accelerated by the work of Wozniakowski [14] and others. In this direction, more information would be needed concerning the conditional distributions than simply an ability to generate random samples from them. Also, the current improved integration techniques would address only integrals involving distributions having densities with respect to Lebesgue measure, whereas the simple algorithm (1) works in general. Whether the additional improvement in mean squared error (if any)

achievable by improved Monte Carlo integration is worth the cost of the added computational complexity for certain estimation problems is an area for further research.

Acknowledgements

A substantial portion of this research was supported by an American Society for Engineering Education summer faculty research fellowship at the Medical Information Systems and Operations Research Department, Naval Health Research Center, San Diego, CA. The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of Defense, nor the U.S. Government. I would also like to thank an anonymous referee whose comments on an earlier draft greatly improved the presentation.

References

[1] J.E. Angus, On the connection between neural network learning and multivariate nonlinear least squares estimation, *Internat. J. Neural Networks* 1 (1989) 42–47.
[2] J.E. Angus, A note on the central limit theorem for the bootstrap mean, *Comm. Statist.* A18 (1989) 1979–1982.
[3] J. Berkson, Maximum likelihood and minimum χ^2 estimates of the logistic function, *J. Amer. Statist. Assoc.* 50 (1955) 130–162.
[4] R. Durrett, *Probability: Theory and Applications* (Brooks/Cole, Belmont, CA, 1990).
[5] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (SIAM, Philadelphia, PA, 1982).
[6] T. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach* (Wiley, New York, 1967).
[7] R. Hecht-Nielsen, *Neurocomputing* (Addison-Wesley, Reading, MA, 1991).
[8] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[9] E.L. Lehmann, *Theory of Point Estimation* (Wiley, New York, 1983).
[10] M. Loeve, *Probability Theory I* (Springer, New York, 1977).
[11] C.R. Rao, *Linear Statistical Inference and its Applications* (Wiley, New York, 2nd ed., 1977).
[12] D.E. Rummelhart, J.L. McClelland and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1* (MIT Press, Cambridge, MA, 1986).
[13] H. White, Some asymptotic results for learning in single hidden-layer feedforward network models, *J. Amer. Statist. Assoc.* 84 (1989) 1003–1013.
[14] H. Wozniakowski, Average case complexity of multivariate integration, *Bull. Amer. Math. Soc.* 24 (1991) 185–194.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	July 1991	Final

4. TITLE AND SUBTITLE
Computer-assisted Improvement of Mean Squared Error in Statistical Estimation

5. FUNDING NUMBERS
Program Element:
Work Unit Number:
ASEE Summer Faculty Navy Research Program

6. AUTHOR(S)
John E. Angus

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

8. PERFORMING ORGANIZATION
Report No. 91-17

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 2
Bethesda, MD 20889-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES First printed as technical report, "Computer Assisted Improvement of the Estimatin Mean Squared Error with Application to Back Propagation Neural Networks." Published in: Mathematics and Computers in Simulation, 1993, 35, 1-13.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A computer-assisted method for improving the mean squared error (MSE) in estimation for parametric models is presented. Assuming existence of nontrivial sufficient statistics, the method involves generation of Monte Carlo samples from the condition distribution of the observables, given the sufficient statistic(s). The method is illustrated in connection with a simple back-propagation neural net work model for estimating a logistic regression function, and a specific numerical example related to logistic regression is presented.

14. SUBJECT TERMS
Parametric estimation; Neural networks; Central limit theorem; Exponential families; Monte Carle Simulation; Nonlinear model; Computer Intensive methods

15. NUMBER OF PAGES 13

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited