01 Jan 1971

Empirical Bayesian Learning

David R. Cunningham
Missouri University of Science and Technology, drc@mst.edu

Arthur M. Breipohl

Recommended Citation

D. R. Cunningham and A. M. Breipohl, "Empirical Bayesian Learning," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC thru 1, no. 1, pp. 19 - 23, Institute of Electrical and Electronics Engineers, Jan 1971.
The definitive version is available at https://doi.org/10.1109/TSMC.1971.5408600

Empirical Bayesian Learning

DAVID R. CUNNINGHAM, MEMBER, IEEE, AND ARTHUR M. BREIPOHL, MEMBER, IEEE

Abstract—**It is shown that a certain weighted average of the prior distribution and the empirical distribution yields an estimate of the posterior distribution that is consistent with Bayes' theorem. A comparison of this approach and conventional parametric Bayesian estimation is made for some specific cases.**

Introduction

ADAPTIVE systems, pattern recognition and, indeed, all system studies where random quantities are involved, require the generation of distribution functions from sample data and experience. There are two common methods for generating such distribution functions. One is the empirical distribution function (EDF) and the other is parametric estimation based on Bayes' rule.[1] The EDF has the advantage that it will converge (with probability one) to the true distribution function as the sample size approaches infinity. However, a disadvantage of the EDF from the decision theory point of view is that no systematic method exists for determining an optimum testing plan or the maximum amount to be spent on testing.

In a Bayesian scheme, prior information can be used to develop an optimum testing plan and determine the maximum amount to be spent on testing [2], [3]. Unfortunately, if the incorrect likelihood function is chosen, there is no possibility of a Bayesian scheme converging to the true distribution function. For example, if the distribution of X is assumed to be normal, when in reality it is Laplace, and the mean is learned by Bayes' theorem, then, although the mean may be correctly learned, the distribution will not.

In this paper a method is proposed for generating a distribution function from data which combines the advantages of the EDF and Bayesian parameter estimation in that it can be used in decision theory to make testing decisions and yet will converge (with probability one) to the true distribution function. This method is referred to as Bayes' empirical distribution function (BEF). This paper defines BEF, discusses convergence, illustrates speed of convergence, and extends the concept to the multidimensional case.

BEF has two potential disadvantages: it is discontinuous and all sample data must be kept in storage. Both of these disadvantages, if significant in a specific application, can be circumvented by a method developed by Specht [11] and discussed in this paper.

Manuscript received January 12, 1970; revised July 20, 1970. This work was partially supported by the National Science Foundation.

D. R. Cunningham is with the Department of Electrical Engineering, University of Missouri, Rolla, Mo. 65401.

A. M. Breipohl was with the Department of Electrical Engineering, Oklahoma State University, Stillwater, Okla. He is now with the Department of Electrical Engineering, University of Kansas, Lawrence, Kans. 66044.

[1] Other Bayesian approaches to statistical decision problems exist [9].

Bayes' Empirical Distribution Function

Assume that the distribution function of a population is to be learned from a set of independent samples. Further assume that some prior knowledge of the distribution function is available. Using a Bayesian viewpoint, the probability distribution $F_X(x)$ of the population will be assumed to be a stochastic process with beta first-order prior density. For convenience, let

$$Q = F_X(x) \tag{1}$$

where the dependence of Q on x is understood. Thus the first-order density function of Q is

$$f_Q(q) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}, & \\ & 0 \le q \le 1,\ \alpha,\beta > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function, and α, β, and q are functions of x. Now, given q, the probability that a out of n samples would be less than or equal to x will be binomial with probability q. Thus

$$\begin{aligned} P_{A|q,n}(a \mid q) &= P\{a \text{ out of } n \text{ samples} \le x \mid q\} \\ &= \binom{n}{a} q^a(1-q)^{n-a}, \qquad a = 0,1,\cdots,n \end{aligned} \tag{3}$$

where $\binom{n}{a}$ is the binomial coefficient.

On applying Bayes' rule after sampling, the first-order density function becomes

$$f_{Q|\xi}(q) = \frac{P_{A|q,n}(a \mid q) f_Q(q)}{\displaystyle\int_{-\infty}^{\infty} P_{A|q,n}(a \mid u) f_Q(u)\, du} \tag{4}$$

where ξ is the total experience. It is well known that this posterior density is beta and

$$f_{Q|\xi}(q) = \begin{cases} \dfrac{\Gamma(n+\alpha+\beta)}{\Gamma(a+\alpha)\Gamma(n-a+\beta)} q^{a+\alpha-1}(1-q)^{n-a+\beta-1}, & \\ & 0 \le q \le 1,\ \alpha,\beta > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

A reasonable choice for the estimate of $F_X(x)$ is the expected value of Q. Therefore, the following definitions are chosen:

$$\hat{F}_X(x) \triangleq E_Q\{Q\} \tag{6}$$

$$\hat{F}_{X|\xi}(x) \triangleq E_Q\{Q \mid \xi\} \tag{7}$$

where $E_Q\{\cdot\}$ is the expected value with respect to Q. In order to be the prior distribution for X, $E_Q\{Q\}$ must be a distribution function and hence a nondecreasing function of

x. In order to assure that $\hat{F}_{X|\xi}(x)$ is also a nondecreasing function of x, it is sufficient to assume that the sum of α and β is a constant; i.e.,

$$w_0 = \alpha(x) + \beta(x) \tag{8}$$

where w_0 is not a function of x.

It can easily be shown that

$$\hat{F}_X(x) = \frac{\alpha(x)}{\alpha(x) + \beta(x)} = \frac{\alpha(x)}{w_0} \tag{9}$$

$$\hat{F}_{X|\xi}(x) = \frac{\alpha(x) + a(x)}{w_0 + n} = \frac{w_0[\alpha(x)/w_0] + n[a(x)/n]}{w_0 + n}$$

$$= \frac{w_0}{w_0 + n}\hat{F}_X(x) + \frac{n}{w_0 + n}F_n(x) \tag{10}$$

where $F_n(x)$ is the EDF. As $\hat{F}_{X|\xi}$ is the weighted average of the prior distribution function and the EDF, $\hat{F}_{X|\xi}$ can easily be shown to be a distribution function. Inspection of (10) indicates that the prior weight w_0 can be considered to be an equivalent sample size for the prior distribution. The only restrictions on w_0 and $\hat{F}_X(x)$ are

$$0 < w_0 < \infty \tag{11}$$

and

$$\hat{F}_X(x) = \text{a distribution function.} \tag{12}$$

For convenience, $\hat{F}_{X|\xi}(x)$ will be called BEF and will be denoted by

$$F_w(X) \triangleq \hat{F}_{X|\xi}(x). \tag{13}$$

The prior distribution will be noted by

$$F_{w_0}(x) \triangleq \hat{F}_X(x). \tag{14}$$

Thus

$$F_w(x) = \frac{w_0}{w_0 + n}F_{w_0}(x) + \frac{n}{w_0 + n}F_n(x). \tag{15}$$

Convergence

The following convergence theorem is of interest.

Theorem: $F_w(x)$ converges uniformly in x to $F_n(x)$ for $-\infty < x < +\infty$, and $F_w(x)$ converges uniformly in x to $F_X(x)$ with probability one for $-\infty < x < +\infty$.

The proof of the first statement follows from a simple limiting argument. The second statement follows directly from the first and the Glivenko–Cantelli theorem [5] which states $F_n(x)$ converges uniformly in x to $F_X(x)$ with probability one for $-\infty < x < +\infty$.

A convenient measure of the error associated with an estimate $F^*(x)$ of $F_X(x)$ is the integral expected square error

$$I = \int_{-\infty}^{\infty} E\{|F^*(x) - F_X(x)|^2\}\, dx.$$

It is of interest to compare the error I when $F^*(x) = F_w(x)$ and when $F^*(x)$ is obtained from a Bayesian parameter estimate. To effect the comparison some method for equating the prior weights is required.

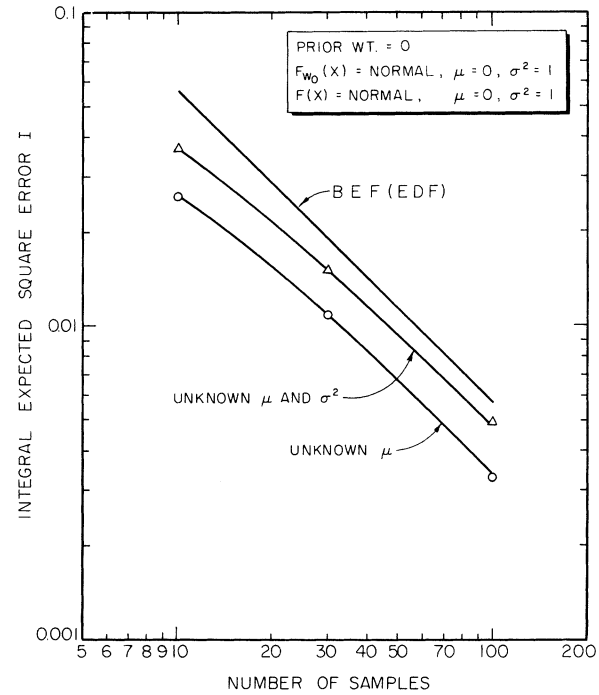

Fig. 1. Integral expected square error with zero prior weight.

Consider the case when the likelihood function is assumed to be normal with known variance σ^2 and with unknown mean μ to be estimated by Bayes' theorem. If the prior distribution of the mean is assumed normal with mean μ_0 and variance N_0^2, it can easily be shown that after a data set $\{X_1, X_2, \cdots, X_n\}$ Bayes' rule yields a normal posterior distribution of the mean with mean

$$\hat{\mu}_n = \frac{(\sigma^2/N_0^2)\mu_0 + n((1/n)\sum_{i=1}^n X_i)}{(\sigma^2/N_0^2) + n} \tag{16}$$

and variance

$$N_n^2 = \frac{\sigma^2 N_0^2}{\sigma^2 + nN_0^2}. \tag{17}$$

Using $\hat{\mu}_n$ as a point estimate for the mean of X, the EDF is normal with variance σ^2 and mean given by (16). Inspection reveals that $\hat{\mu}_n$ is a weighted average of the prior estimate of the mean and the sample mean. Thus the prior weight σ^2/N_0^2 can be considered as an equivalent prior sample size corresponding to w_0 in the BEF estimate of $F_X(x)$.

Keehn [4] has investigated Bayesian parameter estimation for the case of a multivariate normal distribution with unknown mean and covariance matrices. For the one-dimensional case his results reduce to

$$\hat{\mu}_n = \frac{w_0\mu_0 + n((1/n)\sum_{i=1}^n X_i)}{w_0 + n} \tag{18}$$

and

$$\hat{\sigma}_n^2 = \frac{v_0\phi_0 + n((1/n)\sum_{i=1}^n X_i^2 - \hat{\mu}_n^2) + w_0(\mu_0^2 - \hat{\mu}_n^2)}{v_0 + n} \tag{19}$$

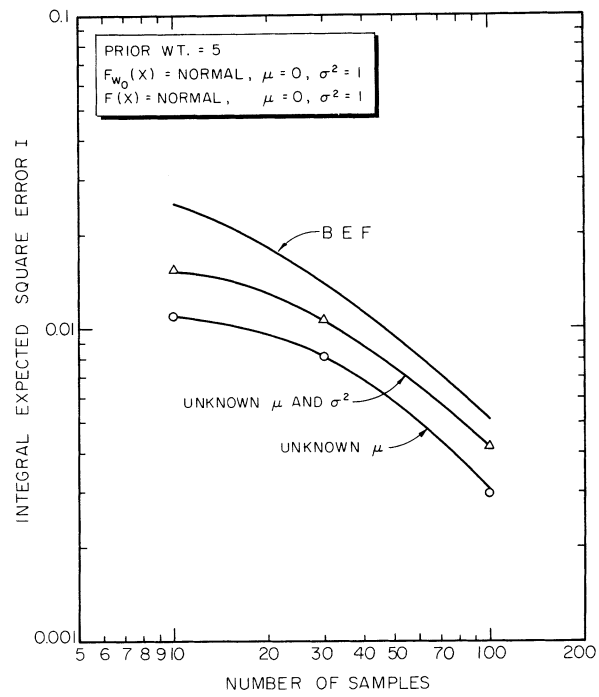

Fig. 2. Integral expected square error with correct prior weight.

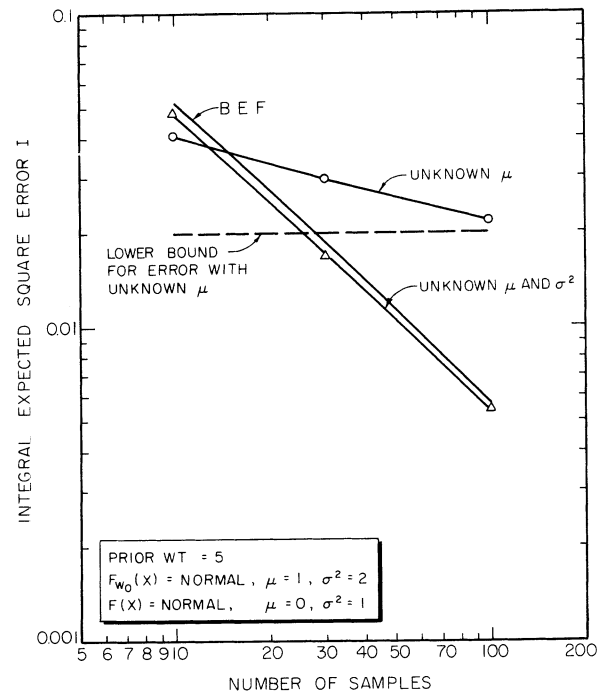

Fig. 3. Integral expected square error with incorrect prior weight.

where

$\hat{\mu}_n$	estimate of the mean
μ_0	prior estimate of the mean
w_0	prior weight of the mean
$\{X_1, X_2, \cdots, X_n\}$	data set
n	number of data samples
ϕ_0	prior estimate of the variance
v_0	prior weight of the variance
$\hat{\sigma}_n^2$	estimate of the variance.

If $v_0 = w_0$, the prior weight w_0 can be considered as an equivalent prior sample size. The estimate of the variance becomes

$$\hat{\sigma}_n^2 = \frac{w_0\phi_0 + n((1/n)\sum_{i=1}^n X_i^2 - \hat{\mu}_n^2) + w_0(\mu_0^2 - \hat{\mu}_n^2)}{w_0 + n}. \tag{20}$$

Let the true distribution function $F_X(x)$ be normal with zero mean μ and variance σ^2 equal one. The integral square error I was estimated by numerical methods for $F^*(X)$ estimated by Bayesian parameter estimation for the cases of unknown mean and unknown mean and variance and by BEF. Results for three different prior conditions are shown in Figs. 1–3.

If the correct form of the distribution (likelihood function) is chosen and the prior density of the parameter does not exclude the true value, the figures indicate the reasonable result; i.e., the conventional parametric method provides a better result than that given by BEF. Fig. 3 indicates, however, that for only a relatively small error in the form (in this example, the wrong variance) of the distribution function, the BEF estimate of the distribution may be superior (in the integral mean-square error sense) after a small number of samples. In fact, for the unknown mean case of Fig. 3, the error I will never drop below approximately 0.02. Although the mean is learned correctly in this case, the variance will always be incorrect.

Parameter Estimation

It is interesting to note that BEF can be used for Bayesian parameter estimation.[2] Consider any parameter θ such that

$$\theta = \int_a^b u(x)\, dF(x). \tag{21}$$

A reasonable estimate for θ is

$$\hat{\theta} = \int_a^b u(x)\, dF_w(x)$$
$$= \frac{w_0}{w_0 + n}\int_a^b u(x)\, dF_{w_0}(x) + \frac{n}{w_0 + n}\int_a^b u(x)\, dF_n(x). \tag{22}$$

Then the estimate of the mean given by BEF is

$$\hat{\mu} = \frac{w_0\mu_0 + n((1/n)\sum_{i=1}^n X_i)}{w_0 + n} \tag{23}$$

and the estimate of the variance is

$$\hat{\sigma}^2 = \frac{w_0\sigma_0^2 + n((1/n)\sum_{i=1}^n X_i^2 - \hat{\mu}^2) + w_0(\mu_0^2 - \hat{\mu}^2)}{w_0 + n} \tag{24}$$

[2] Another approach to empirical Bayesian parameter estimation may be found in Fu [1].

where

$\hat{\mu}$	estimate of the mean
μ_0	mean given by the prior distribution
w_0	weight of the prior distribution
$\{X_1, X_2, \cdots, X_n\}$	data set
n	number of data samples
${\sigma_0}^2$	variance given by the prior distribution.

Thus (23) and (24) are identical to (18) and (20) derived by the more conventional Bayesian approach. To reduce the data storage required for parameter estimation, (23) and (24) may, of course, be applied iteratively [12].

Multivariate Distributions

The BEF concept may easily be extended to Bayesian estimation of the joint distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ of random vectors. As the distribution of $\boldsymbol{X}$ is a real-valued function of $\boldsymbol{X}$, it is only necessary to assume that α, β, and q of (2) are functions of $\boldsymbol{X}$. It follows easily from an argument similar to the one-dimensional case that the BEF estimate $F_w(\boldsymbol{x})$ of $F_{\boldsymbol{X}}(\boldsymbol{x})$ is the weighted average

$$F_w(\boldsymbol{x}) = \frac{w_0}{w_0 + n} F_{w_0}(\boldsymbol{x}) + \frac{n}{w_0 + n} F_n(\boldsymbol{x}) \tag{25}$$

where $F_{w_0}(\boldsymbol{x})$ is the prior distribution, $F_n(\boldsymbol{x})$ is the EDF, w_0 is the weight on the prior distribution, and n is the number of samples.

Smoothing BEF

If the distribution function $F(x)$ is known to be continuous, the discontinuous nature of $F_w(x)$ may be disconcerting, if not an actual problem. Therefore, some form of smoothing may be desirable.

Parzen [8] has investigated a method for estimating the density function $f(x)$ from n independent samples $X_1, X_2, \cdots, X_n$. This estimate $g_n(x)$ of $f(x)$, where

$$F(x) = \int_{-\infty}^{x} f(u)\, du \tag{26}$$

is of the form

$$g_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x - y}{h}\right) dF_n(y) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right). \tag{27}$$

$F_n(x)$ is the familiar EDF and $K(y)$ is a weighting function satisfying certain conditions. From an engineering point of view, this is equivalent to time domain filtering, where $K(x)$ is the filter and x is analogous to time. From another point of view, it is the weighted average of n density functions[3] $(1/h)K((x - X_j)/h)$, where X_j determines the shift of K with respect to the origin and h determines the spread about X_j of K. Thus a smoothed estimate of $F(x)$ is

$$G_n(x) = \int_{-\infty}^{x} g_n(y)\, dy. \tag{28}$$

A natural application of this to BEF would be to use

$$G_w(x) = \frac{w_0}{w_0 + n} F_{w_0}(x) + \frac{n}{w_0 + n} G_n(x) \tag{29}$$

for the estimate of $F(x)$. Properties of this estimate need to be investigated. Because of the nature of $G_w(x)$, it is to be expected that $G_w(x)$ would have properties similar to $G_n(x)$. The properties of $G_n(x)$ have been investigated by several authors [6]–[8], [10], [13], [14].

Memory Reduction

A major difficulty in applying the EDF to engineering problems is that all data must be kept in storage. This difficulty carries over to BEF. Specht [10] has developed a series approximation for Parzen's method that requires a fixed storage capacity. It would appear that this approach might be used to simultaneously reduce data storage requirements and provide smoothing of BEF.

Specht's approximation chooses a weighting function $K(x)$ of the form

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{30}$$

Thus

$$g_n(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - y)^2}{2\sigma^2}\right] dF_n(y) \tag{31}$$

or

$$g_n(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{(x - X_i)^2}{2\sigma^2}\right) \tag{32}$$

where σ corresponds to h of (27). Writing

$$\exp\left[-\frac{(x - X_i)^2}{2\sigma^2}\right] = \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(\frac{xX_i}{\sigma^2}\right) \exp\left(-\frac{{X_i}^2}{2\sigma^2}\right) \tag{33}$$

and expanding $\exp(xX_i/\sigma^2)$ in a Taylor's series, (32) becomes

$$g_n(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sum_{r=0}^{\infty} C_{r,n} x^r \tag{34}$$

where

$$C_{r,n} = \frac{1}{r!\,\sigma^{2r}} \frac{1}{n} \sum_{i=1}^{n} {X_i}^r \exp\left(-\frac{{X_i}^2}{2\sigma^2}\right). \tag{35}$$

Noting that

$$C_{r,n+1} = \frac{n}{n+1} C_{r,n} + \frac{1}{r!\,\sigma^{2r}} \cdot \frac{1}{n+1} {X_{n+1}}^r \exp\left(-\frac{{X_{n+1}}^2}{2\sigma^2}\right) \tag{36}$$

[3] $K(x)$ is not necessarily a density function.

it can be seen that a recursive relation exists for $C_{r,n}$. Hence for a fixed number of terms M in the Taylor's series approximation,

$$g_n(x) \simeq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sum_{r=0}^{M} C_{r,n} x^r. \tag{37}$$

Thus the approximation requires the storage of a fixed number of terms regardless of the sample size. The memory required will, of course, depend on the precision required of the estimate. A number of properties of this approximation are investigated by Specht and should be of value in applying it to BEF.

Summary and Conclusions

Taking a specific weighted average of the prior estimate of a distribution function and the EDF as the posterior estimate of a distribution function was shown to be consistent with Bayes' theorem. This result, referred to as BEF, was compared to conventional parametric estimation. A method for extending BEF to the estimation of a distribution for finite-dimensional random vectors was outlined. The adaptation of BEF to Parzen's method was suggested for obtaining continuous estimates of a distribution function. Specht's approximation was proposed as a method for reducing data storage.

BEF offers a simple logical method for combining prior knowledge and independent sample data to estimate a distribution function. BEF can be shown to converge to the true distribution function with probability one regardless of the prior distribution. Given the true form of the distribution and a prior density for a parameter which does not exclude the true value, conventional Bayesian parametric estimation is superior to BEF. If, however, the assumed form of the distribution function is incorrect or the assumed prior excludes the true value of a parameter, BEF may yield a superior estimate after a relatively small number of samples. Thus BEF allows the use of prior information as does conventional Bayesian parametric estimation, but BEF will converge (with probability one) to the true distribution, whereas conventional Bayesian estimation may not. BEF may also be used for parametric Bayesian estimation with results very similar to the conventional Bayesian technique. For most applications it appears that BEF can replace the conventional Bayesian technique either for nonparametric estimation or parametric estimation.

References

[1] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*. New York: Academic Press, 1968.
[2] R. A. Howard, "Information value theory," *IEEE Trans. Syst. Sci. and Cybern.*, vol. SSC-2, pp. 22–26, August 1966.
[3] R. A. Howard, "Bayesian decision models for system engineering," *IEEE Trans. Syst. Sci. and Cybern.*, vol. SSC-1, pp. 36–40, November 1965.
[4] D. G. Keehn, "A note on learning for Gaussian properties," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 126–132, January 1965.
[5] M. Loeve, *Probability Theory*. New York: Van Nostrand Reinhold, 1960.
[6] V. K. Murthy, "Estimation of probability density," *Ann. Math. Statist.*, vol. 36, pp. 1029–1031, June 1965.
[7] E. A. Nadaraya, "On non-parametric estimates of density functions and regression curves," *Theory Prob. Appl.* (USSR), vol. 10, pp. 186–190, 1965.
[8] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, September 1962.
[9] H. Robbins, "The empirical approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, pp. 1–20, January 1964.
[10] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832–837, September 1956.
[11] D. E. Specht, "Series estimation of a probability density function," *Technometrics* (to be published).
[12] J. Spraggins, "A note on the iterative application of Bayes' rule," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 544–549, October 1965.
[13] G. S. Watson and M. R. Leadbetter, "On the estimation of the probability density, pt. I," *Ann. Math. Statist.*, vol. 34, pp. 480–491, June 1963.
[14] M. Woodroofe, "On the maximum deviation of the sample density," *Ann. Math. Statist.*, vol. 38, pp. 475–481, April 1967.