Discussion Paper No. 15

SEQUENTIAL LINEAR MINIMUM VARIANCE ESTIMATION OF PROBABILITIES: PROBABILITIES THAT MAY BE TIME OR SPACE DEPENDENT

by

Brian Schaefer

September 25, 1972

This research was supported through the Urban Systems Engineering Center, Northwestern University, Evanston, Illinois.

1. INTRODUCTION

The estimation of probabilities has been a central problem to mathematical statistics and applied probability theory since the time of Laplace, Bayes, and Bernoulli. Among the most popular methods in use for estimating probabilities are maximum likelihood, method of moments, unbiased minimum variance (non-Bayesian), and Bayesian methods. The method described in this paper is an application of a result in stochastic linear systems theory, whose generality permits the estimation of probabilities that may be varying in space or time, perhaps according to a stochastic process. Samples (of possibly different size,) taken at different points in time or space, are combined to obtain a linear minimum variance estimate. The method is Bayesian,but because of the structure imposed, only the first two moments of the a priori distribution need be assumed, and computed for the a posteriori distribution. This fact and the sequential nature of the method make it ideal for computer implimentation. Also due to a bounding theorem which we shall quote below, it is possible to assume the least favorable a priori variance and obtain a bound on the a posteriori variance.

The paper does not consider the cost of sampling and is most applicable therefore in cases where this cost is either negligible or the data is provided regularly by government statistics,census data, stock market quotations, etc. The emphasis is on applications where a probability may not only be expected to change from one observation time to the next, but where there is some relationship between the probabilities at different observation times. Some typical applications in which this is true are:

(i) Medical labs are often interested in estimating the proportion of abnormal blood cells in a blood sample, on a series of days in which the condition of a patient may be expected to slowly improve

or deteriorate.

(ii) Retail stores, interested in estimating the effects of their advertising and customer satisfaction policies may wish to estimate the time dependent proportions of new and repeat customers.

(iii) An analyst may try to estimate the probability of waiting longer than a fixed amount of time in a non-stationary queue.

(iv) Geneticists are often interested in the evolving proportion of a particular genotype, and health workers may wish to estimate the proportion of the population that has contacted, is immune to, or carries a particular disease.

We will also assume that the observations are obtained from simple random sampling. For finite population the method will thus be applicable if sampling occurs with replacement or if the population is so large that the finite multiplier is insignificant.

II. TIME DEPENDENT PROBABILITIES

Suppose (Ω,A,p) is a probability space, and $\{X_t, t\in T\}$ is an n-dimensional stochastic process such that for every $t\in T$, X_t is A-measureable. T is an index set and may be either countable or an interval of the real line R. The process X_t induces a farily of probability spaces (R^n,B^n,P_t) where B^n is the σ-algebra of n-dimensional Borel sets. Suppose we are given a set of events $\{B_t^i: B_t^i\in B^n;\ t\in T;\ i=1,..,k\}$, and a set of observation times $\{t': t'\in T'\subset T\}$, where T' is a countable index set. The problem may then be stated as: given an arbitrary sequence of observations $\{X_{t'}: s_1\leq t'\leq s_2,\ t'\in T'\}$ of the process X_t, estimate the probabilities $P_t(B_t^i)$, $t\in T$, $i=1,..,k$. Without loss of generality we can let $s_1=0$. If $t > s_2$ the problem is usually known as prediction [1],[2], if $t = s_2$ as filtering, and if $t < s_2$ as smoothing. In this paper we will estimate $P(B_t) = \{P_t(B_t^1), P_t(B_t^2),\ldots,P_t(B_t^k)\}$ sequentially, filtering at each observation and predicting between observations.

In the standard estimation of probabilities problem it is assumed that $X_t = X_s$, and $B_t = B_s$, $\forall\ r,s \in T$; there is thus no need for the index set T. In this case there are many estimation methods available, some of which are mentioned in the introduction. We will allow a more general relationship between the random variables X_t and X_s, and the events B_t and B_s. We will assume that the process $\{P(B_t):\ t \in T\}$ is either deterministic, a Markov process, a wide sense Markov process, or a martingale that may be modelled as

$$(1)\quad P(B_t) = F(s,t)P(B_s) + u(s,t) + w(s,t)\quad s,\ t \in T,\ s < t$$

where $F(s,t)$, $u(s,t)$ and $u(s,t)$ are such that $0 \leq P(B_t) \leq 1\ \forall\ 0 \leq P(B_s) \leq 1$; 0 and 1 are to be interpreted as the 0 and 1 K vectors. $F(s,t)$ and $u(s,t)$ are known functions of s and t, and $w(s,t)$ is a stochastic process with

$$(2)\quad E[w(s,t)] = 0$$

$$(3)\quad E[w(s,t')w(t',t)'] = 0 \qquad s < t' < t$$
(kxk 0 matrix)

$$(4)\quad E[w(s,t)w(s,t)'] = Q(s,t)$$

$$(5)\quad E[w(s,t)P'(B_s)] = 0$$

Trivial examples of a model such as (1) would be

$$P(B_t) = P(B_s)$$

or $P(B_t) = .5\ P(B_s) + .2 + w$, where w is uniformly distributed on $[-.1,.1]$.

If T is countable it may be more convenient to represent (1) as a stochastic difference equation

$$(6)\quad P(B_{t_i}) = f(t_i)P(B_{t_{i-1}}) + u(t_i) + w(t_i),$$

or if T is continuous, as an Ito stochastic differential equation

$$(7)\quad dP(B_t) = f(t)P(B_t) + u(t) + dz(t)$$

where z(t) is a Brownian process. (1) may be viewed as a solution of either (6) or (7).

If the process $P(B_t)$ must be modelled by a nonlinear equation

$$(8)\quad P(B_t) = F(P(B_s),s,t) + u(s,t) + w(s,t)$$

then the results of the method given below for estimating $P(B_t)$ are in a practical sense very useful, although the calculated a posteriori covariance of the estimate is only an approximation [see Nahi [3] for a discussion of a linearization approximation].

The criterion we will use to evaluate the estimate $\hat{P}_\tau(B_t)$ of $P(B_t)$ is the trace of the unconditional a posteriori covariance matrix

$$(9) \quad E[(P(B_t) - \hat{P}_\tau(B_t))'(P(B_t) - \hat{P}_\tau(B_t))]$$

and we will constrain $\hat{P}_\tau(B_t)$ to be a linear function of a sufficient statistic of X_s, $s\in T'$, $0 \leq s \leq \tau$. The main results of the theory necessary for obtaining the optimal $\hat{P}_\tau(B_t)$ are summarized in the next section.

III. Main Results of Sequential Linear Minimum Variance Estimation

Suppose Z_t is a stochastic process observable only through the related process Y_t. Z_t may be a continuous or discrete time process with $t\in T$, where T is an interval or countable index set. Y_t is a discrete time process with $t\in T'\subset T$. Let H_z be the Hilbert space of random variables spanned by the process Z_t, $t\in T$ [see Parzen [4] or Cramer and Leadbetter [5]]. The inner product on H_z is defined by

$$(z_1,z_2) = E[z_1' z_2] \qquad \forall\, z_1 z_2 \in H_z$$

which induces the norm

$$\|z\|^2 = E[z'z] \qquad \forall\, z \in H_z$$

Let H_{Y_t} be Hilbert space spanned by the process Y_s, $0 \leq s \leq t$, $s\in T'$. Given the observations $y(s)$, $0 \leq s \leq t$, $s\in T'$ of the process Y_t we would like to find the linear function of $y(s)$, $0 \leq s \leq t$, $s\in T'$ that is the minimum variance estimator of Z_t. The main result of optimal linear estimation theory is due to the well known projection theorem in Hilbert spaces [see for example [6]].

Projection Theorem

There exists a unique $\hat{z}_\tau(t) \in H_{Y_\tau}$ such that

(10) $E[(z_t - \hat{z}_\tau(t))'(z_t - \hat{z}_\tau(t))]$

is minimized. A necessary and sufficient condition for $\hat{z}_\tau(t)$ is

(11) $E[(z_t - \hat{z}_\tau(t))'y] = 0 \qquad \forall\ y \in H_{Y_\tau}$

If the stochastic process z_t is a Markov process or martingale described by

(12) $z_t = F(s,t)z_s + u(s,t) + w(s,t)$

where $w(s,t)$ is a stochastic process with properties given in (2)-(4) and $E[z_s w'(s,t)] = 0$, and if the observation process Y_t is linearly related to z_t by

(13) $Y_t = C(t)z_t + v(t)$

where

(14) $E[v(t)] = 0$

(15) $E[v(t)z_s'] = 0 \qquad \forall\ t \in T$

(16) $E[v(t)v'(s)] = 0 \qquad \forall\ t \neq s \in T$

(17) $E[v(t)v'(t)] = R(t)$

then the Projection Theorem implies the following theorem, obtained in the recursive form by Kalman [1].

Sequential Optimal Estimation Theorem

Suppose that z_t and Y_t are stochastic processes defined by (12) and (13) and that at time s, (11) is satisfied for all $y \in H_{Y_T}$ by $\hat{z}_\tau(s)$ with

(18) $E[(z_s - \hat{z}_\tau(s))(z_s - \hat{z}_\tau(s))'] = S_{\tau,s}$

then for $t > s$

(19) $\hat{z}_\tau(t) = F(s,t)\hat{z}_\tau(s) + u(s,t)$

(20) $\hat{z}_t(t) = \hat{z}_\tau(t) + H(\tau,t)[y(t) - C(t)\hat{z}_\tau(t)]$

where

$$(21)\quad K(s,t) = S_{\tau,t}C'(t)[C(t)S_{\tau,t}C'(t) + R(t)]^{-1}$$

$$(22)\quad S_{\tau,t} = F(s,t)S_{\tau,s}F'(s,t) + Q(s,t)$$

$$(23)\quad S_{t,t} = (I - K(t,t)C(t))S_{\tau,t}$$

Often, in applications, one is not sure of the covariances $S_{\tau,s}$, $Q(s,t)$ and $R(t)$, and it would be useful to know if one were pessimistic in his assumptions about these covariances, would the computed value of the a posteriori covariance $S_{t,t}$ also be pessimistic. To make this concept precise it is necessary to introduce the following notation.

Using equations (22) and (23) it is possible to calculate $S_{\tau,t}$ recursively, however due to possible erroneous assumptions about the covariances $Q(s,t)$ or $R(t)$ it may be that this calculated covariance is not equal to the actual covariance. Nahi and Schaefer [9] discuss a method of detecting and correcting this difference. We will let

$S^{a}_{\tau,t}$ ≡ actual covariance

$S^{c}_{\tau,t}$ ≡ computed covariance

$Q^{a}(s,t)$ ≡ actual covariance of $w(s,t)$

$Q^{c}(s,t)$ ≡ assumed covariance of $w(s,t)$ used in calculations

$R^{a}(t)$ ≡ actual covariance of $v(t)$

$R^{c}(t)$ ≡ assumed covariance of $v(t)$ used in calculations

The matrix inequality $A \geq B$ implies that $A-B$ is positive semi-definite. The following bounding theorem was obtained partially by Nishimura [7] and extended by Heffes [8].

Bounding Theorem

If $S^{c}_{\tau,s} \geq S^{a}_{\tau,s}$, $Q^{c}(s,t) \geq Q^{a}(s,t)$ and $R^{c}(t) \geq R^{a}(t)$ then $S^{c}_{\tau,t} \geq S^{a}_{\tau,t}$

and $S^{c}_{t,t} \geq S^{a}_{t,t}$.

In the following section we will apply the results of the above theorems to the estimation of probabilities problem.

IV. Estimation of Probabilities

We will assume that probabilities to be estimated $P(B_t) = (P(B^1_t), P(B^2_t),\ldots,P(B^k_t))$ may be modelled as a stochastic process as given by equation (1) with

$$(24)\quad E[P(B_o)] = \hat{P}_o(B_o)$$

$$(25)\quad E[(P(B_o) - \hat{P}_o(B_o))(P(B_o) - \hat{P}_o(B_o))'] = S_{o,o}$$

In the formulation of the observation process we will assume simple random sampling. Suppose that at each time $t\in T'$, we have $n(t)$ independent observations $(x_1(t),\ldots,x_{n(t)}(t))$ of the random variable X_t. A sufficient statistic for estimating the probability $P(B^i_t)$ is the total number of observations that are contained within B^i_t, which we shall denote by $y^i(t)$. If we let

$$y(t) = (y^1(t),y^2(t),\ldots,y^k(t)) \text{ and } C(t) = n(t) ,$$

then the observation process may be modelled as

$$(26)\quad y(t) = C(t)\, P(B_t) + v(t) .$$

If we can verify that equations (14)-(16) are true for this observation process, then the model consisting of equations (1) and (26) is in a form in which the results of section III can be applied.

Since $y^i(t)$ is the number of times the random variable X_t falls within B^i_t, $y^i(t)$ is a binomial random variable with

$$(27)\quad E[y^i(t) \mid P(B^i_t)] = n(t)P(B^i_t)$$

and

$$(28)\quad Var[y^i_t(t) \mid P(B^i_t)] = n(t)P(B^i_t)[1-P(B^i_t)]$$

thus from (26)

$$(29)\quad E[v(t)] = E[E[v(t) \mid P(B_t)]]$$
$$= E[E_t y(t) - C(t)P(B_t) \mid P(B_t)]]$$
$$= 0$$

Thus (14) is verified. (15) may be verified by noting that

$$(30)\quad E[v(t)P'(B_t)] = E[E[v(t)P'(B_t) \mid P(B_t)]]$$
$$= E[E[v(t) \mid P(B_t)]\, P'(B_t)]$$
$$= 0$$

If we assume that the $v(s,t)$ process is uncorrelated with Y_s, then from (5) and (26)

$$(31)\quad E[v(s,t)v'(s)] = 0$$

Thus from (1),(30) and (31) it follows that

$$(32)\quad E[v(t)P'(B_s)] = 0 \qquad \forall\ s,\ t \in T\ .$$

If $P(B_t)$ and $P(B_s)$ are known, then $y(t)$ and $y(s)$ are independent, hence

$$(33)\quad E[v(t)v'(s)] = E[E[v(t)v'(s) \mid P(B_t)P(B_s)]]$$
$$= E[E[v(t) \mid P(B_t)]E[v(s) \mid P(B_s)]]$$
$$= 0$$

and therefore (16) is verified. To compute $R(t)$ given in equation (17) we note that

$$(34)\quad E[v(t)v'(t)] = E[E[v(t)v'(t) \mid P(B_t)]]$$
$$= \text{diag}\ E[n(t)P(B_t)[1-P(B_t)]]$$
$$= n(t)\ \text{diag}\ [E[P(B_t)] - E[P^2(B_t)]]$$
$$= R(t)$$

where diag $E[P(B_t)]$ signifies that the terms $E[P(B_t^i)]$ are on the diagonal of the matrix and all off diagonal elements are zero. To compute $R(t)$ recursively we obtain

$$(35)\quad E[P(B_t)] = F(s,t)E[P(B_s)] + u(t,s)$$

with $\qquad E[P(B_o)] = \hat{P}_o(0)$

and

$$(36)\quad E[P^2(B_t)] = Var[P(B_t)] + E^2[P(B_t)]$$

$$(37)\quad Var[P(B_t)] = F(s,t)Var[P(B_s)]F'(s,t) + Q(s,t)$$

with $$Var[P(B_o)] = S_{o,o}$$

Summarizing, we can say that if the process $P(B_t)$ is modelled by equation (26), then the $P_\tau(B_t)$ that minimizes (9) may be computed sequentially using (19)-(23) with the obvious replacement of $\hat{z}_\tau(t)$ by $\hat{P}_\tau(B_t)$.

In the Bounding Theorem of section III, it was required that $S^c_{\tau,s} \geq S^a_{\tau,s}$, $Q^c(s,t) \geq Q^a(s,t)$, $R^c(t) \geq R^a(t)$. Often in practice it is difficult to ensure that assumed covariances have this property; however, for the estimation of probabilities this property is guaranteed if we let

$$(38)\quad S^c_{o,o} = Q^c(s,t) = \text{diag}(\tfrac{1}{4})$$

$$(39)\quad R^c(t) = \text{diag}\left(\frac{1}{4n(t)}\right)$$

The next section will give several examples of the procedure described in this and the previous two sections.

V. Examples

(i) Simple Probability

In this example we will show that in the conventional case of assumed constant probability, with maximum a priori covariance, the above estimation reduces to the sample mean.

Suppose we wish to estimate the probability of an event B given a sequence of observations $\{y_t: t=0,1,2,\ldots.\}$ where $y_t = 1$ or 0 depending on whether or not the event B occurred. We assume that

$$B_t = B_{t-1} \qquad i=0,1,2,\ldots.$$

$$P(B_t) = P(B_{t-1})$$

$$y(t) = P(B_t) + v(t)$$

$$E[P(B_o)] = \hat{P}_o(B_o)$$

$$E[(P(B_o) - \hat{P}_o(B_o))^2] = S_{o,o}$$

Thus from (19)-(23) the linear minimum variance estimate of $P(B_t)$ may be obtained sequentially from

$$\hat{P}_t(B_t) = \hat{P}_{t-1}(B_{t-1}) + H(t)[y(t) - \hat{P}_{t-1}(B_{t-1})]$$

$$H(t) = \frac{S_{t-1,t-1}}{S_{t-1,t-1} + \hat{P}_o(B_o) - \hat{P}_o^2(B_o) - S_{o,o}}$$

$$S_{t,t} = (1-H(t))\, S_{t-1,t-1}$$

For this simple case it is apparent that the sequential procedure may be reduced to a closed form. Substituting repeatedly for $S_{t-1,t-1}$ yields

$$S_{t,t} = \frac{(\hat{P}_o(B_o) - \hat{P}_o^2(B_o) - S_{o,o})S_{o,o}}{\hat{P}_o(B_o) - \hat{P}_o^2(B_o) + (t-1)\, S_{o,o}}$$

and

$$H(t) = \frac{S_{o,o}}{\hat{P}_o(B_o) - \hat{P}_o^2(B_o) + (t-1)\, S_{o,o}}$$

Therefore

$$\hat{P}_t(B_t) = \frac{1}{1+t\,\dfrac{S_{o,o}}{\hat{P}_o(B_o) - \hat{P}^2(B_o) - S_{o,o}}}\,\hat{P}_o(B_o) + \frac{1}{(t-1) + \dfrac{\hat{P}_o(B_o) - \hat{P}^2(B_o)}{S_{o,o}}}\sum_{j=1}^{t} y(j)$$

Thus as t increases, the influence of $\hat{P}_o(B_o)$ decreases and the estimate approaches the maximum likelihood estimate and the sample mean. Also if we choose the maximum a priori covariance which occurs if

$$Pr(P_o(B_o) = 0) = Pr(P_o(B_o) = 1) = \tfrac{1}{2}$$

so that $S_{o,o} = \hat{P}_o(B_o) - \hat{P}_o^2(B_o)$

then $\hat{P}_t(B_t) = \frac{1}{t}\sum_{j=1}^{t} y(j)$

(ii) Multinomial

The extension of multinomial probabilities is an obvious extension of the above example. The model may be put in the canonical form

$$P(B_t^1) = P(B^1)$$
$$P(B_t^2) = P(B^2)$$
$$\vdots$$
$$\vdots$$
$$P(B_t^k) = P(B^1) + P(B^2) + \ldots P(B_t^{k-1})$$

(iii) Population Processes

Suppose we wish to estimate the probability of incurring a disabling disease or accident, $P(D_t)$, the probability of a birth with a congenital disability, $P(B_t)$, and the total number of disabled persons in the population, $n(t)$, in time period t.

Because of a lack of more explicit information we assume that $P(D_t)$ and $P(B_t)$ are martingales

$$E[P(D_t) \mid P(D_{t-1})] = P(D_{t-1})$$

and

$$E[P(B_t) \mid P(B_{t-1})] = P(D_{t-1})$$

with

$$\mathrm{Var}[P(D_t) \mid P(D_{t-1})] = Q_D(t)$$
$$\mathrm{Var}[P(B_t) \mid P(B_{t-1})] = Q_B(t)$$

We also assume that from census data and birth and death statistics we are given

$N(t)$ = total population in time period t.

$\mathcal{N}(t)$ = number of births in time period t.

$P(DD_t)$ = probability of dying in time period t.

Then the model describing the evolution of the disabled population may be postulated to be

$$\pi(t+1) = \pi(t) + P(D_t)\pi(t) + P(B_t)\chi_i(t) - P(DB)_t\pi(t)$$

$$P(D_t) = P(D_{t-1}) + w^1(t)$$

$$P(B_t) = P(B_{t-1}) + w^2(t)$$

where $E[(w^1(t))^2] = Q_1(t)$

$E[(w^2(t))^2] = Q_2(t)$

$E[w^3(t)] = E[w^4(t)] = 0$

A sample of size n(t) is taken in each time period t and the number of disabled people $y_1(t)$, the number who were disabled due to disease or accident during time period t, $y_2(t)$, and the number of disabled births, $y_3(t)$, are noted.

Then

$$y_1(t) = \frac{n(t)}{N(t)}\pi(t) + v_1(t)$$

$$y_2(t) = n(t)\,P(D_t) + v_2(t)$$

$$y_3(t) = n(t)P(B_t) + v_3(t)$$

The covariance of $v_i(t)$ are obtained from (35)-(37). Equations (19)-(23) may thus be used to obtain $\hat{\pi}_t(t)$, $\hat{P}_t(D_t)$, and $\hat{P}_t(B_t)$.

Conclusions

A sequential method of obtaining linear minimum variance estimators of time varying probabilities has been formulated. Several simple examples are described, and a method of obtaining an upper bound on the mean square error when the a priori covariances are unknown is discussed.

References

[1] R. Kalman, "A New Approach to Linear Filtering and Prediction", Trans. ASME, J. Basic Eng., Series D, Vol. 82, pp. 35-44, March 1960.

[2] N. Wiener, The Interpolation, Extrapolation and Smoothing of Stationary Time Series, John Wiley & Sons, Inc., New York, N. Y., 1949.

[3] N. Nahi, Estimation Theory and Applications, John Wiley & Sons, Inc., New York, N. Y., 1969.

[4] E. Parzen, "An Approach to Time Series Analysis", Ann. Mach. Stat., Vol. 32, pp. 951-989, 1961.

[5] H. Cramer, M. R. Leadbetter, Stationary and Related Stochastic Processes, John Wiley & Sons, Inc., New York, N. Y., 1967.

[6] G. F. Simmons, Topology and Modern Analysis, McGraw Hill, 1963.

[7] T. Nishimura, "On the A Priori Information in Sequential Estimation Problems", IEEE Trans. on Automatic Control, Vol. AC-11, pp. 197-204, April 1966.

[8] H. Heffes, "The Effect of Erroneous Models on the Kalman Filter Response", IEEE Trans. on Automatic Control, Vol. AC-13, pp. 699-702, Dec. 1968.

[9] N. Nahi and B. Schaefer, "Decision-Directed Adaptive Recursive Estimators: Divergence Prevention", IEEE Trans. on Automatic Control, Vol. AC-17, pp. 61-68, Feb. 1972.