The Annals of Mathematical Statistics
1969, Vol. 40, No. 4, 1477–1482

SMOOTHING BY CHEATING

By James M. Dickey

State University of New York at Buffalo

1. Summary. This self-contained note extends and generalizes smoothing procedures proposed by Whittle (1957), (1958) and Dickey (1968). The use of linear filters, chosen through analysis of the data to be smoothed, is advocated. Hence, smoothing is nonlinear. The setting can be viewed as a multivariate extension of empirical Bayes settings in which, usually, there are strong independence assumptions.

2. General theory. Imagine that we have a collection of numbers, v_x, indexed by x, each an error-plagued measure of its unknown ideal dual number u_x. Suppose that the unknowns u_x are a realization of a probability structure on the vector $\mathbf{u}$ of u_x's. Our "measuring instrument" is described by a conditional probability structure on the vector $\mathbf{v}$ of v_x's given the vector $\mathbf{u}$. In short, $\mathbf{u}$ and $\mathbf{v}$ have a joint probability structure.

As the reader will discover, for the beginning of this discussion the indices x of u_x and v_x need have no correspondence. Summations below may have interpretations as integrals.

We desire to estimate $\mathbf{u}$ from $\mathbf{v}$, $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\mathbf{v})$; and a penalty will be imposed by a squared Euclidean norm, $1(\hat{\mathbf{u}}, \mathbf{u}) = (\hat{\mathbf{u}} - \mathbf{u}, \hat{\mathbf{u}} - \mathbf{u}) = \sum_x \sum_x \lambda_{x,y}(\hat{u}_x - u_x)(\hat{u}_y - u_y)$ the matrix of $\lambda_{x,y}$'s positive definite. Hence, given the square integrability (expectability) of $\mathbf{u}$, $\mathbf{v}$, and $\hat{\mathbf{u}}$, we seek to maximize the expected utility, or equivalently, minimize $E(\hat{\mathbf{u}} - \mathbf{u}, \hat{\mathbf{u}} - \mathbf{u})$, by choosing the function $\hat{\mathbf{u}}(\mathbf{v})$ as the conditional expectation of $\mathbf{u}$ given $\mathbf{v}$, $\hat{\mathbf{u}} = E(\mathbf{u} \mid \mathbf{v})$, since for $\hat{\mathbf{u}}^* \neq \hat{\mathbf{u}}$, $E(\hat{\mathbf{u}}^* - \mathbf{u}, \hat{\mathbf{u}}^* - \mathbf{u}) = E(\hat{\mathbf{u}}^* - \hat{\mathbf{u}}, \hat{\mathbf{u}}^* - \hat{\mathbf{u}}) + E(\hat{\mathbf{u}} - \mathbf{u}, \hat{\mathbf{u}} - \mathbf{u})$, the cross-product term vanishing. The optimal estimate $\hat{\mathbf{u}}$ is thus seen to be independent of the choice of (positive-definite) penalty inner product $1(\hat{\mathbf{u}}, \mathbf{u})$.

The conditional expectation $\hat{u}_x = E(u_x \mid \mathbf{v})$ can be viewed as a Hilbert-space projection of u_x on the subspace of square-integrable random variables measurable with respect to $\mathbf{v}$ (linearly generated by polynomials in the v_x's). This Hilbert-space inner product is the usual second-order moment,

$$(1) \qquad (u_x, u_y)_H = Eu_x \cdot u_y .$$

Assume for now that $\mathbf{u}$ and $\mathbf{v}$ are jointly normally distributed. Then, as is well known, the conditional expectation $\hat{u}_x = E(u_x \mid \mathbf{v})$ takes the linear form,

$$(2) \qquad \hat{u}_x = \sum_y w_x(y) \cdot (v_y - Ev_y) + Eu_x ,$$

where

$$(3) \qquad \sum_z \operatorname{Cov}(v_y, v_z) \cdot w_x(z) = \operatorname{Cov}(u_x, v_y), \qquad \text{all } y.$$

Received 19 February 1968; revised 4 December 1968.

The optimal estimate $\hat{u}_x$ can now be viewed as the Hilbert-space projection of u_x on the subspace spanned (linearly) by the v_x's and a nonzero constant.

Whether or not $\mathbf{u}$ and $\mathbf{v}$ are normal, the optimal linear function of the v_x's and a constant, the so called "wide-sense conditional expectation," is given by (2) and (3), still independently of the choice of penalty inner product. To prove this, note first that if (2) and (3) hold for one coordinatization of $\mathbf{u}$ and $\hat{\mathbf{u}}$, then they hold for all, and that any two penalty inner products are related by a change of coordinates. The sum-of-squared-errors penalty is easily seen to lead to (3), by minimizing the expectation of an individual summand $(\hat{u}_x - u_x)^2$, assuming at first, zero means for u_x and $\mathbf{v}$.

Note that the best linear estimate of the form,

$$\hat{u}_x^* = \sum_y w_x^*(y) \cdot v_y , \tag{2*}$$

satisfies

$$\sum_z (Ev_yv_z)w_x^*(z) = Eu_xv_y , \tag{3*}$$

and that $\hat{u}_x^*$ and $\hat{u}_x$ (as given by (2) and (3)) coincide if a linear combination of the v_y's is a nonzero constant.

Much applied mathematics has centered on equations (2) and (3), including the theory of linear filtering and prediction for stationary time series. Practical problems concerning the solution of equation (3) (or the analogous integral equation) given the low-order moment structure, or, equivalently, the spectral structure, have been treated at length.

Attention is here directed to problems of inference about the needed low-order moments. Subcases are considered of the case of "signal plus noise,"

$$E(v_x \mid \mathbf{u}) = u_x , \tag{4}$$

where we have now established a correspondence between the indices x of u_x and v_x.

Equation (4) implies

$$Eu_x = Ev_x , \tag{5}$$

so that the expectations of the v_x's are the only first-order moments required. To infer Ev_x from the data $\mathbf{v}$, we assume first-order stationarity of the v_x's (and hence of the u_x's). The two needed first-order moments can then be estimated by the empirical mean of v_x, formed by averaging over x,

$$\hat{E}u_x = \hat{E}v_x \equiv \sum_x v_x / \sum_x 1. \tag{6}$$

(The notation $\hat{E}$, as used here, should not be confused with Doob's (1953) wide-sense conditional expectation operator.)

Note that when $E(\mathbf{v} \mid \mathbf{u}) = \mathbf{u}$, the right-hand sides of equations (3) and (3*) become Cov (u_x, u_y) and Eu_xu_y, respectively, which can be related to the overall second-order moments of $\mathbf{v}$ by the familiar conditional variance formula,

$$\text{Cov } (v_x, v_y) = \text{Cov } (u_x, u_y) + E \text{ Cov } (v_x, v_y \mid \mathbf{u}), \tag{7}$$

or

$$(7^*) \qquad Ev_x \cdot v_y = Eu_x \cdot u_y + E \operatorname{Cov}(v_x, v_y \mid \mathbf{u}).$$

Hence, if the expected second-order structure of the "measuring instrument," $E \operatorname{Cov}(v_x, v_y \mid \mathbf{u})$, is known, the needed second-order moments follow from an estimate of the overall moments of $\mathbf{v}$,

$$(8) \qquad \hat{\operatorname{Cov}}(v_x, v_y) = \sum_z (v_{x+z} - \hat{E}v_x)(v_{y+z} - \hat{E}v_x)/D,$$

or

$$(8^*) \qquad \hat{E}v_x \cdot v_y = \sum_z v_{x+z}v_{y+z}/D,$$

where the denominator $D = \sum_x 1$, or $(\sum_x 1) \pm 1$, or a favorite other divisor. (In smoothing problems of P. Whittle, taken up below, the conditional second-order moment structures, $\operatorname{Cov}(v_x, v_y \mid \mathbf{u})$, are known quadratic functions of the unknowns $\mathbf{u}$.)

The use of the serial estimate (8) requires second-order stationarity of the v_x's, which, in turn, requires a shift structure $x \to$ "$x + z$." No ergodicity, or approximate ergodicity, of v_x is here assumed. For one, ergodic subjective probability structures hardly ever arise in practice (Leonard J. Savage, personal communication). In addition, the moments actually estimated by the empirical averages are expectations conditional on the shift-invariant event realized. The estimate $\hat{u}_x$ (2) based on these conditional moments minimizes the conditional, and hence the unconditional expected loss among all functions of the shift-invariant events and, linearly, of $\mathbf{v}$ and a nonzero constant. (Pointed out by Richard A. Olshen, personal communication.)

3. Subcases of "signal plus noise." (a) *Joint normality of* $\mathbf{u}$ *and* $\mathbf{v}$. Bayes' theorem has been appealed to in contexts where the probability structure of $\mathbf{u}$ and the conditional probability structure of $\mathbf{v}$ given $\mathbf{u}$ are at hand. If these are multivariate normal distributions, as is well known and easily derived, $\mathbf{u}$ given $\mathbf{v}$ is conditionally normal with mean,

$$(9) \qquad \hat{\mathbf{u}} = E(\mathbf{u} \mid \mathbf{v}) = [\operatorname{Var}(\mathbf{v} \mid \mathbf{u})][\operatorname{Var} \mathbf{v}]^{-1}E\mathbf{u} + [\operatorname{Var} \mathbf{u}][\operatorname{Var} \mathbf{v}]^{-1}\mathbf{v},$$

and variance matrix,

$$(10) \qquad \operatorname{Var}(\mathbf{u} \mid \mathbf{v}) = [\operatorname{Var} \mathbf{u}][\operatorname{Var} \mathbf{v}]^{-1}[\operatorname{Var}(\mathbf{v} \mid \mathbf{u})],$$

where

$$(11) \qquad \operatorname{Var} \mathbf{v} = \operatorname{Var} \mathbf{u} + \operatorname{Var}(\mathbf{v} \mid \mathbf{u}).$$

Equation (11) is the matrix form of the homoscedastic case of equation (7); it can be used to determine a third matrix from any pair, of which at least one must be known in advance.

One might be interested in whether equations (9)–(11) apply in any way to the general nonnormal case when $\hat{\mathbf{u}} = E(\mathbf{u} \mid \mathbf{v})$ and $\hat{\mathbf{v}} = E(\mathbf{v} \mid \mathbf{u}) = \mathbf{u}$ are inter-

preted as *wide-sense* conditional expectations. Indeed, these equations with Var $(\mathbf{u} \mid \mathbf{v})$ and Var $(\mathbf{v} \mid \mathbf{u})$ denoting $E(\mathbf{u} - \hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})'$ and $E(\mathbf{v} - \hat{\mathbf{v}})(\mathbf{v} - \hat{\mathbf{v}})'$, respectively, hold as geometric theorems in Hilbert space, since they then refer only to the marginal first and second-order moments of the probability structure.

The estimate $\hat{\mathbf{u}}$ given by (9) minimizes the sum of two quadratic forms in $\mathbf{u}$,

$$(12)\quad (\mathbf{u} - E\mathbf{u})'[\text{Var } \mathbf{u}]^{-1}(\mathbf{u} - E\mathbf{u}) + (\mathbf{v} - \mathbf{u})'[\text{Var } (\mathbf{v} \mid \mathbf{u})]^{-1}(\mathbf{v} - \mathbf{u})$$
$$= [\mathbf{u} - E(\mathbf{u} \mid \mathbf{v})]'[\text{Var } (\mathbf{u} \mid \mathbf{v})]^{-1}[\mathbf{u} - E(\mathbf{u} \mid \mathbf{v})] + C,$$

where C is constant in $\mathbf{u}$. Bellman and Kalaba and Lockett (1965) treat the numerical problem of minimizing (12) when the prior (marginal) variance matrix of $\mathbf{u}$ exhibits high correlations for pairs of unknowns, u_x and u_y, when x and y are "close."

In some applications, such as optical filtering, the measurements v_x and the unknowns u_x are positive, and one might try to minimize (12) under this constraint. It may be easier to transform v_x and u_x, say to their logarithms, then minimize (12), and then transform back. If one is serious about the quadratic loss function, instead of transforming back, one should recall the formula for the mean of a lognormal variate, exp (u_x) given $\mathbf{v}$,

$$(13)\qquad E(\exp(u_x) \mid \mathbf{v}) = \exp[E(u_x \mid \mathbf{v}) + \tfrac{1}{2} \text{Var } (u_x \mid \mathbf{v})].$$

The formula for a general mixed moment follows trivially from (13).

Whittle (1957), (1958) proposed use of the smoothed estimates $\hat{u}_x^*$ given by (2*) and (3*) in the following contexts.

(b) *The v_x's independent and exponentially distributed with means u_x.* The classical periodogram ordinates are asymptotically so distributed.

(c) *The v_x's independent and Poisson distributed with means u_x.* This applies, for one, to probability estimation from the cell counts $\mathbf{v}$ of a random-size sample; the probability estimate is then $\hat{u}_x / \sum \hat{u}_y$.

(d) *The (Nv_x)'s multinomially distributed with cell probabilities u_x and total cell count N.* This applies, of course, to probability estimation from the cell frequencies $\mathbf{v}$ of a sample of fixed size N. Note that in this context $\hat{u}_x$ (equation (2)) coincides with $\hat{u}_x^*$ (equation (2*)) since $\sum v_x \equiv 1$.

In these three contexts (b)–(d), the variance structure of the "measuring instrument" is a quadratic function Q of the "true values" $\mathbf{u}$,

$$\text{Cov } (v_x, v_y \mid \mathbf{u}) \doteq Q(\mathbf{u})$$
$$(14b)\qquad = \delta_{x,y} u_x^2$$
$$(14c)\qquad = \delta_{x,y} u_x$$
$$(14d)\qquad = N^{-1}[\delta_{x,y} u_x - u_x u_y],$$

where $\delta_{x,y} = 1$ or 0 according to whether or not $x = y$. Hence, the expected "instrumental" variance structure, needed for use of (7) and (8), is a linear function $F(F^*)$ of the first two moments of the unobservable $\mathbf{u}$.

$$E \operatorname{Cov}(v_x, v_y \mid \mathbf{u}) = F(\operatorname{Var} \mathbf{u}, E\mathbf{u})$$

$$= \delta_{x,y}[\operatorname{Cov}(u_x, u_y) + (Eu_x)^2] \tag{15b}$$

$$= \delta_{x,y} Eu_x \tag{15c}$$

$$= N^{-1}[(\delta_{x,y} - Eu_y)Eu_x - \operatorname{Cov}(u_x, u_y)], \tag{15d}$$

or

$$E \operatorname{Cov}(v_x, v_y \mid \mathbf{u}) = F^*(E\mathbf{u}\mathbf{u}', E\mathbf{u})$$

$$= \delta_{x,y} Eu_x^2 \tag{15b*}$$

$$= \delta_{x,y} Eu_x \tag{15c*}$$

$$= N^{-1}[\delta_{x,y} Eu_x - Eu_x u_y]. \tag{15d*}$$

Equations (6), (7), (8) and (15) yield the estimates $G(G^*)$,

$$\hat{\operatorname{Cov}}(u_x, u_y) = G(\hat{\operatorname{Var}} \mathbf{v}, \hat{E}\mathbf{v})$$

$$= [\hat{\operatorname{Cov}}(v_x, v_y) - \delta_{x,y}(\hat{E}v_x)^2]/(1 + \delta_{x,y}) \tag{16b}$$

$$= \hat{\operatorname{Cov}}(v_x, v_y) - \delta_{x,y}\hat{E}v_x \tag{16c}$$

$$= [\hat{\operatorname{Cov}}(v_x, v_y) + N^{-1}(\hat{E}v_y - \delta_{x,y})\hat{E}v_x]/(1 - N^{-1}) \tag{16d}$$

or

$$\hat{E}u_x u_y = G^*(\hat{E}\mathbf{v}\mathbf{v}', \hat{E}\mathbf{v})$$

$$= \hat{E}v_x v_y/(1 + \delta_{x,y}) \tag{16b*}$$

$$= \hat{E}v_x v_y - \delta_{x,y}\hat{E}v_x \tag{16c*}$$

$$= [\hat{E}v_x v_y - N^{-1}\delta_{x,y}\hat{E}v_x]/(1 - N^{-1}), \tag{16d*}$$

which together with (6) and (8) supply the necessary moments for use of (2) and (3) in determining $\hat{\mathbf{u}}$.

According to its nonnormal interpretation, equation (10) supplies an estimate of the moments needed for a linear calculation of the overall expected loss for optimal linear $\hat{\mathbf{u}}(\mathbf{v})$ as given by (2), (3).

The probability-estimation context (d) received more detailed treatment in Dickey (1968), which paper should have made reference to famous work by I. J. Good (1963, 1965, 1966). Good considered the global smoothing problem and has proposed the use of prior distributions based on peeking at the data.

Acknowledgment. I am grateful to Leonard J. Savage, Richard A. Olshen, and Frederick Mosteller for their help.

REFERENCES

BELLMAN, RICHARD, KALABA, ROBERT and LOCKETT, JOANN (1965a). Dynamic programming and ill-conditioned linear systems. *J. Math. Anal. Appl.* **10** 206–215.

BELLMAN, RICHARD, KALABA, ROBERT and LOCKETT, JOANN (1965b). Dynamic programming and ill-conditioned linear systems—II. *J. Math. Anal. Appl.* **12** 393–400.

DICKEY, JAMES M. (1968). Smoothed estimates for multinomial cell probabilities. *Ann. Math. Statist.* **39** 561–566.

DICKEY, JAMES M. (1969). Estimation of disease probabilities conditioned on symptom variables. To appear in *Math. Biosciences.*

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

GOOD, I. J. (1963). Maximum entropy for hypothesis formulation especially for multi-dimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

GOOD, I. J. (1965). *The Estimation of Probabilities.* M.I.T. Press, Cambridge, Mass.

GOOD, I. J. (1966). How to estimate probabilites. *J. Inst. Maths. Applics.* **2** 364–383.

WHITTLE, P. (1957). Curve and periodogram smoothing. *J. Roy. Statist. Soc. Ser. B* **19** 38–47.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20** 334–343.