The Annals of Statistics
1983, Vol. 11, No. 3, 1004–1007

ORTHOGONAL SERIES METHODS FOR BOTH QUALITATIVE AND QUANTITATIVE DATA

BY PETER HALL

Australian National University

We introduce and describe orthogonal series methods for estimating the density of qualitative, quantitative or mixed data. The techniques are completely nonparametric in character, and so may be used in situations where parametric models are difficult to construct. Just this situation arises in the context of mixed—both qualitative and quantitative—data, where there are few parametric models.

1. Introduction. An advantage of orthogonal series methods of density estimation is that they are readily extended from the univariate case to multivariate distributions, provided the dimension is not too high. Our main aim in the present paper is to show that this advantage extends from the purely continuous case, where it is by now well known, to the case of *mixed* multivariate data, in which the components have both discrete and continuous distributions. We also demonstrate that orthogonal series methods may be used to smooth unbounded, discrete data. This requires us to construct a complete sequence of orthonormal functions on an ordered, countably infinite set, and is undertaken in Section 2. We handle the case of mixed data in Section 3.

Orthogonal series methods for the estimation of a continuous density were introduced by Whittle (1958), Čencov (1962), Schwartz (1967) and Kronmal and Tarter (1968). Ott and Kronmal (1976) suggested using Walsh series to estimate the density of a multivariate binary distribution. We tackle the problem differently from Ott and Kronmal, in that we consider general orthogonal sequences which are constructed by taking the product of sequences for univariate data. This yields the example of Walsh functions on a binary set as a special case. Our technique is completely nonparametric in character, and so may be used in circumstances where parametric models are difficult to construct. Just this situation arises in the context of mixed data.

2. Unbounded, discrete data. We shall assume that the range of the data is the set $\{0, 1, 2, \cdots\}$, and base our orthogonal functions on the Poisson distribution. Expand the function $G(x, t) = (1 + t)^x e^{-\lambda t}$ as a power series in t, $G(x, t) = \sum_{i=0}^{\infty} \psi_i(x) t^i/i!$ The following properties of the polynomials ψ_i may be derived:

$$\psi_i(x) = \sum_{j=0}^{i} \binom{i}{j} (-\lambda)^{i-j} x(x-1) \cdots (x-j+1), \quad i \geq 0;$$

$\psi_{i+1}(x) = x\psi_i(x-1) - \lambda\psi_i(x)$, $i \geq 0$ (this formula is easily used to generate the functions ψ_i numerically); and if the variable Z has the Poisson distribution with parameter λ, $E\{\psi_i(Z)\psi_j(Z)\} = \delta_{ij} i!\lambda^i$, where δ_{ij} is the Kronecker delta. Therefore if we define

$$\phi_i(x) = \psi_i(x)(i!\lambda^i)^{-1/2}\{\lambda^x e^{-\lambda}/\Gamma(x+1)\}^{1/2}, \quad i \geq 0 \quad \text{and} \quad x \geq 0,$$

then $\sum_{i=0}^{\infty} \phi_i(n)\phi_j(n) = \delta_{ij}$, so that the functions ϕ_i are orthonormal. The functions ψ_i are the Charlier polynomials (see Abramowitz and Stegun, 1965, page 788), and the first six are listed in Table 1.

Received May 1982; revised December 1982.

AMS 1980 *subject classifications.* Primary, 62G05; secondary, 62E20

Key words and phrases. Classification, cross-validation, density estimator, discrete data, discrimination, near neighbour, orthogonal series, qualitative data, quantitative data, smoothing.

TABLE 1
Polynomials orthogonal in Poisson (λ) *distribution*

$$\begin{aligned}
\psi_0(x) &= 1; \quad \psi_1(x) = x - \lambda; \quad \psi_2(x) = x^2 - (1 + 2\lambda)x + \lambda^2;\\
\psi_3(x) &= x^3 - 3(1 + \lambda)x^2 + (2 + 3\lambda + 3\lambda^2)x - \lambda^3;\\
\psi_4(x) &= x^4 - 2(3 + 2\lambda)x^3 + (11 + 12\lambda + 6\lambda^2)x^2 - 2(3 + 4\lambda + 3\lambda^2 + 2\lambda^3)x + \lambda^4;\\
\psi_5(x) &= x^5 - 5(2 + \lambda)x^4 + 5(7 + 6\lambda + 2\lambda^2)x^3\\
&\quad - 5(10 + 11\lambda + 6\lambda^2 + 2\lambda^3)x^2 + (24 + 30\lambda + 20\lambda^2 + 10\lambda^3 + 5\lambda^4)x - \lambda^5.
\end{aligned}$$

Let p be a density on $\{0, 1, 2, \cdots\}$. The generalized Fourier series for p is $p(x) = \sum_{r=0}^{\infty} c_r\phi_r(x)$, where $c_r = \sum_{x=0}^{\infty} p(x)\phi_r(x)$, and an unbiased estimator of c_r based on the random sample $X_1, \cdots, X_n$, is given by $\hat{c}_r = n^{-1}\sum_{j=1}^{n}\phi_r(X_j)$. A "smoothed" estimator of p can be constructed by weighting the estimated Fourier coefficients: $\hat{p}(x) = \sum_{r=0}^{\infty} w_r\hat{c}_r\phi_r(x)$ for weights w_r. The problem of selecting the weights is examined in a very general context in the next section.

The parameter λ plays a similar role to the kernel type in classical nonparametric density estimation. Note that λ is not a smoothing parameter, and that changing the value of λ has very little effect on the results. For example, $\sum_{r=0}^{\infty} \operatorname{var}\{\phi_r(X_1)\}$ does not depend on λ.

3. General orthogonal series estimators. We shall show how to construct orthogonal series estimators of mixed multivariate densities, based on univariate components. The univariate, continuous case is well known; see Kronmal and Tarter (1968). We have just examined the case of a univariate density on an unbounded discrete set. If the set is univariate and bounded, the orthogonal functions may be taken proportional to *orthogonal contrasts*. For example, vectors $(\phi_i(0), \cdots, \phi_i(4))$ defining functions orthogonal on $(0, 1, \cdots, 4)$, may be taken equal to $5^{-1/2}(1, 1, 1, 1, 1)$, $2^{-1}(1, 1, 0, -1, -1)$, $2^{-1}(1, -1, 0, 1, -1)$, $2^{-1}(1, -1, 0, -1, 1)$ and $(20)^{-1/2}(1, 1, -4, 1, 1)$.

More generally, suppose data take values on the d-variate sample space $S = S_1 \times \cdots \times S_d$. If $\{\phi_i^{(j)}, i \geq 0\}$ is an orthonormal basis for the space of functions on S_j, and if we define

$$\phi_r(x) = \prod_{j=1}^{d} \phi_{r_j}^{(j)}(x_j) \text{ for each } x = (x_j) \in S \text{ and vector } r = (r_j),$$

then $\{\phi_r\}$ is an orthonormal basis for the space of functions on S. It is a little cumbersome to write down the orthogonality relations in complete generality. For example, if $S = \{0, 1, \cdots, m\} \times \{0, 1, 2, \cdots\} \times (a, b)$, the relations take the form

$$\sum_{x_1=0}^{m}\sum_{x_2=0}^{\infty}\int_a^b dx_3\phi_r(x_1, x_2, x_3)\phi_s(x_1, x_2, x_3) = \delta_{rs}.$$

Let us agree to write this as

$$\int_S \phi_r(x)\phi_s(x)\,dx = \delta_{rs}. \tag{3.1}$$

A density p on S admits the expansion $p(x) = \sum_r c_r\phi_r(x)$, where in the notation suggested by (3.1), $c_r = \int \phi_r(x)p(x)\,dx$. A "smoothed" orthogonal series estimator of p based on a random sample $X_1, \cdots, X_n$, is given by $\hat{p}(x) = \sum_r w_r\hat{c}_r\phi_r(x)$, where $\hat{c}_r = n^{-1}\sum_{j=1}^{n}\phi_r(X_j)$ and the w_r are weights. Generalized weighted mean "integrated" square error (MISE), with integration in the sense of (3.1), may be defined by

$$\int E\{\hat{p}(x) - p(x)\}^2\,dx = n^{-1}\Sigma_r w_r^2 \operatorname{var}\{\phi_r(X_1)\} + \Sigma_r(w_r - 1)^2c_r^2. \tag{3.2}$$

For most choices of orthonormal sequences, the function ϕ_0 will be a constant, and then the condition $\sum_x \hat{p}(x) = 1$ may be imposed by simply insisting that $w_0 = 1$.

If each subspace S_j is discrete, then the estimator defined by taking $w_r = 1$ for each r is well defined, and equals the cell proportion estimator $\hat{p}_0(x)$. However, if one or more of the subspaces is continuous, and each $w_r = 1$, then the variance component of MISE diverges. Even in the purely discrete case, we could choose the weights so as to minimise MISE. The solution to this problem is given by

$$w_r = c_r^2/[c_r^2 + n^{-1}\text{var}\{\phi_r(X_1)\}],$$

for each r. (Compare Watson and Leadbetter, 1963.) Unbiased estimators of the numerator and denominator are given by $(n\hat{c}_r^2 - \hat{b}_r^2)/(n-1)$ and $\hat{c}_r^2$, respectively, where $\hat{b}_r^2 = n^{-1}\sum_{i=1}^{n}\phi_r^2(X_i)$. Therefore we might consider the estimator

$$\hat{p}_1(x) = \sum_{r\in Q_1}(n\hat{c}_r^2 - \hat{b}_r^2)\{(n-1)\hat{c}_r\}^{-1}\phi_r(x), \tag{3.3}$$

where Q_1 denotes the set of indices r with $n\hat{c}_r^2 \geq \hat{b}_r^2$. We shall prove in Appendix I that if each S_j is bounded and discrete, and each c_r is nonzero, then $n^{1/2}\{\hat{p}_1(x) - \hat{p}_0(x)\} \to 0$ in probability as $n \to \infty$. Therefore $\hat{p}_1$ satisfies the same central limit theorem as $\hat{p}_0$. This result is an analogue of part (b) of the Theorem of Wang and van Ryzin (1981). The case where $c_r = 0$ for some r is discussed in Appendix I.

An alternative way of selecting the weights is to take each w_r equal to 0 or 1. Arguing as in Section 4.1 of Ott and Kronmal (1976), we are led to the estimator $\hat{p}_2(x) = \sum_{r\in Q_2}\hat{c}_r\phi_r(x)$, where Q_2 equals the set of indices r with $(n+1)\hat{c}_r^2 \geq 2\hat{b}_r^2$. Again, if each S_j is bounded and discrete, and each c_r is nonzero, $n^{1/2}\{\hat{p}_2(x) - \hat{p}_0(x)\} \to 0$ in probability. Interestingly, this estimator is almost identical to that constructed by least-squares cross-validation, which is defined by $\hat{p}_3(x) = \sum_{r\in Q_3}\hat{c}_r\phi_r(x)$, where Q_3 equals the set of indices r with $n\hat{c}_r^2 \geq 2(1 - 1/2n)\hat{b}_r^2$. See Appendix II.

Suppose the sample space can be written as $S = S^{(1)} \times S^{(2)}$, where $S^{(1)}$ is purely continuous and $S^{(2)}$ purely discrete. Let $r = (r^{(1)}, r^{(2)})$ denote the analogous decomposition of r. An estimator of the marginal density on $S^{(1)}$ would usually be constructed using weights $u_{r^{(1)}}$ which equal 1 if $r^{(1)} \leq m$ (for some vector m), 0 otherwise. The marginal density on $S^{(2)}$ can be constructed separately with a very different weighting scheme $v_{r^{(2)}}$ such as that given in (3.3), and the two combined into a product series estimator,

$$\hat{p}_4(x) = \sum_{r=(r^{(1)},r^{(2)})}u_{r^{(1)}}u_{r^{(2)}}\hat{c}_r\phi_r(x).$$

Alternatively, for $x = (x^{(1)}, x^{(2)}) \in S^{(1)} \times S^{(2)}$ we could define

$$\hat{p}_5(x) = \sum_{z\in S^{(2)}}\hat{w}(x^{(2)}, z)\hat{p}_0(x^{(1)}, z),$$

where $\hat{p}_0$ is an unsmoothed estimator and $\hat{w}(\cdot,\cdot)$ is a system of weights designed for smoothing the cell proportion estimator on $S^{(2)}$. There are many possible choices for $\hat{w}$. For example, if $S^{(2)} = \{0, 1\}^k$, one could use weights associated with a kernel estimator, as in Aitchison and Aitken (1976). An alternative would be to construct the weights for a near neighbour estimator of order ℓ, for any ℓ in the range $0 \leq \ell \leq k$, using methods of Hills (1967) or Hall (1981). In this regard, Dr. H. J. Trampisch has kindly pointed out that the equations for w_{00} and $\tilde{w}_{00}$ at the top of page 574 of Hall (1981), which represent an alternative method of computing the vector w_{00} at the foot of page 573, hold only in the case $\ell = k$, since that condition is required for $h = P1_{\ell+1}$. In the case $\ell < k$, w_{00} can be estimated using the maximum likelihood estimate of P.

APPENDICES

(I) *Limiting distributions of* $\hat{p}_1, \hat{p}_2$. If $c_r \neq 0$ then $(n\hat{c}_r^2 - \hat{b}_r^2)\{(n-1)\hat{c}_r\}^{-1} = \hat{c}_r + O_p(n^{-1})$ and $P(r \in Q_1) \to 1$, while if $c_r = 0$, $n^{1/2}(n\hat{c}_r^2 - \hat{b}_r^2)\{(n-1)\hat{c}_r\}^{-1}I(r \in Q_1) \to b_r(Z - Z^{-1})I(|Z| > 1)$ in distribution, where Z is $N(0, 1)$ and $b_r^2 = E\hat{b}_r^2$. This results in a normal $N\{0, p(r) - p^2(r)\}$ limit for $n^{1/2}\{\hat{p}_1(r) - p(r)\}$ if each c_r is nonzero, and a nonnormal limit if some c_r vanishes and the corresponding b_r is nonzero. In the same way, $\hat{p}_2$ can have a nonnormal limit if some c_r vanishes, and similar results may be proved for Ott and

Kronmal's estimators. However, these results are of a pathological nature, since insisting that $c_r = 0$ for some r places a linear constraint on the class of probability densities, and so is unlikely from a Bayesian viewpoint.

(II) *Cross-validation.* Let $\hat{p}(x) = \Sigma_r w_r \hat{c}_r \phi_r(x)$ be an estimator in which each $w_r = 0$ or 1, let $\hat{p}_{ni}$ denote the version of $\hat{p}$ calculated for the $(n - 1)$-sample with X_i deleted, and let δ_i be the density concentrated at X_i. Following Bowman (1982) we choose the weights to minimise $n^{-1} \sum_i \int \{\hat{p}_{ni}(x) - \delta_i(x)\}^2 \, dx$, which results in the rule: set $w_r = 1$ iff $n^2 \hat{c}_r^2 \geq (2n - 1)\hat{b}_r^2$.

Acknowledgments. The comments of two referees have contributed greatly to the conciseness of this paper.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1965). *Handbook of Mathematical Functions.* Dover, New York.

AITCHISON, J. and AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420.

BOWMAN, A. W. (1982). A comparative study of some kernel-based nonparametric density estimators. Manchester-Sheffield Research Report No. 84/AWB/1.

ČENCOV, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559–1562.

HALL, P. (1981). Optimal near neighbour estimator for use in discriminant analysis. *Biometrika* **68** 572–575.

HILLS, M. (1967). Discrimination and allocation with discrete data. *Appl. Statist.* **16** 237–250.

KRONMAL, R. A. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

OTT, J. and KRONMAL, R. A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *J. Amer. Statist. Assoc.* **71** 391–399.

SCHWARTZ, S. C. (1967). Estimation of probability density by an orthogonal series. *Ann. Math. Statist.* **38** 1261–1265.

WANG, M.-C. and VAN RYZIN, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika* **68** 301–310.

WATSON, G. S. and LEADBETTER, M. R. (1963). On the estimation of the probability density, II. *Ann. Math. Statist.* **34** 480–491.

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

DEPARTMENT OF STATISTICS
THE FACULTIES
THE AUSTRALIAN NATIONAL UNIVERSITY
P.O. BOX 4
CANBERRA ACT 2600
AUSTRALIA