The Annals of Statistics
1990, Vol. 18, No. 3, 1032–1069

REMARKS ON FUNCTIONAL CANONICAL VARIATES, ALTERNATING LEAST SQUARES METHODS AND ACE[1]

BY ANDREAS BUJA

Bellcore

We discuss properties of some data-analytic methods which are intimately related to each other: alternating least squares (ALS), correspondence analysis and more recently Breiman and Friedman's ACE algorithm. The application of these methods to regression produces nonparametric estimators of nonlinear transformations, both of the response and the predictors. These procedures are among the most powerful tools for data analysis, but missing awareness of some artifacts could lead to inappropriate interpretations. We point out some anomalies as well as some curiosities in the mathematics of these methods, and we relate them to some areas in computer-aided tomography, projection pursuit regression and nonlinear devices in the theory of noise.

1. Introduction. Over many decades and in several contexts, there have emerged a host of intimately related and partly identical techniques variously known as optimal scoring, dual scaling, reciprocal averaging, simultaneous linear regression, alternating least squares, correspondence analysis, nonlinear multivariate analysis and homogeneity analysis. Prior to 1985, not much of these developments had shown up in the mainstream statistical literature, with the exception of canonical analysis of contingency tables [Kendall and Stuart (1979), 33.44–33.51; more recent is Gilula and Haberman (1988).] The emphasis of these approaches is on the analysis and quantification of categorical data as they arise typically in social sciences, but a different use for the nonlinear transformation of quantitative data has been known all along. See for instance Young, de Leeuw and Takane (1976), page 509, equation 9, for a polynomial, and de Leeuw, van Rijckevorsel and van der Wouden (1981) and van Rijckevorsel (1982) for a B-spline approach, and more recently a book edited by van Rijckevorsel and de Leeuw (1988) which covers many of these topics. Papers related in philosophy are Kruskal (1965) and Bradley, Katti and Coons (1962), who allow arbitrary monotonic transformations of a regression response. In the scaling literature, these methods are usually not introduced by way of an estimation problem based on a model involving parameters and error terms, as is done for example in the approach by Box and Cox (1964). Rather, one directly poses an optimization problem for some loss or stress

Received December 1985; revised July 1989.

[1] Work partially supported by the following grants while the author was with the faculty of the Statistics Department at the University of Washington, Seattle: NSF, MCS-83-04234 and DMS-85-04359; Office of Naval Research N00014-81-K-0095; Department of Energy DE-FG06-85ER25006.

AMS 1980 *subject classifications*. Primary 62J99; secondary 62P15.

Key words and phrases. ACE, alternating least squares, correspondence analysis, nonlinear multivariate analysis, scaling, horseshoe effect, functional canonical variates, optimal correlation, transformation of variables in regression, orthogonal polynomials, polynomial biorthogonality, series expansions of bivariate distributions, elliptic distributions.

function. This has led to the creative development of a host of algorithms and data-analytic methodology which comprise scaling versions of most of multivariate statistics: regression, analysis of variance, principal components, canonical correlations and discriminant analysis [Young (1981) and Gifi (1981).] For the purposes of the present paper, the most relevant work is that by van der Burg and de Leeuw (1983) on nonlinear canonical correlation. These exciting developments enabled researchers to analyze both quantitative and qualitative data as well as mixed data by any multivariate estimation technique. Meanwhile, statistical inference had to take a back seat, although it is being picked up for instance by the Dutch school of nonlinear multivariate analysis as the problem of "replication stability" [Gifi (1981), Section 12.4, and de Leeuw and van der Burg (1986)].

In a regression context with quantitative response and predictors, the scaling problem boils down to a search for optimal nonlinear transformations of the variables, both the response and the predictors if one likes. In a population (rather than finite-sample) formulation, the regression version of the scaling problem can be stated as that of finding arbitrary nonlinear (measurable) transformations $\theta(Y)$ and $\phi_1(X_1), \phi_2(X_2), \dots \phi_p(X_p)$ which maximize

$$\operatorname{corr}\left(\theta(Y), \sum_j \phi_j(X_j)\right).$$

The response Y and the predictors $X_1, X_2, \dots X_p$ are assumed to be random variables which take on values in arbitrary measurable spaces (!), and the transformations θ and $\phi_1, \phi_2, \dots \phi_p$ are measurable, real-valued functions defined on the respective measurable spaces such that $\theta(Y)$ and $\phi_1(X_1)$, $\phi_2(X_2), \dots \phi_p(X_p)$ are square-integrable random variables. These functions are called transformations for quantitative variables, and scalings, scorings or quantifications, for categorical variables. The most fundamental assumption is that *the response and the predictors have a joint distribution*. This applies mainly to data which may be considered as observational and representatively sampled or as results of a designed experiment where the design can be taken as an approximation to a meaningful distribution of the predictors. Perhaps even more importantly, one should notice *what is not assumed*: There is no assumption being made on the conditional distribution of the response given the predictors, such as a common error distribution for all points in predictor space or, even more specifically, normal errors. This breaks with entrenched thinking habits of statisticians who are usually preoccupied with error models and corresponding dichotomies like "smooth and rough" or "signal and noise."

The homecoming of these scaling ideas to mainstream statistics was marked by the paper by Breiman and Friedman [Breiman and Friedman (1985a); abbreviated B & F hereafter], who gave a framework in terms of populations and formulated an estimation problem for the nonlinear transformations. They propose estimation of transformations from data via an iterative algorithm which uses computationally inexpensive smoothers as building blocks.

Like much of the scaling literature, B & F's paper is centered around their algorithm which they call ACE (alternating conditional expectations). This is to some extent a special case of an alternating least squares (ALS) algorithm by van der Burg and de Leeuw (1983) except for the use of fast smoothers and some additional complications arising from the generality of the van der Burg–de Leeuw algorithm. We are not sure whether this algorithmic orientation is fortunate. Ultimately, the most successful implementations might resort to more conventional numerical linear algebra even when modern nonparametric curve estimates are used as building blocks. In spite of this minor squabble, we will use the established algorithmic acronyms and refer (redundantly and clumsily) to these methods as the ALS–ACE approach, partly to do justice to the psychometric literature which precedes B & F.

Despite its technical demands, B & F is probably easier to read than the scaling literature since it is written in a language more amenable to statisticians. It should convince statisticians that ALS–ACE is one of the most powerful and universal tools for the analysis of multivariate data due to its ability to recover frequent types of nonlinear structure and its applicability to categorical data. The publication of B & F was accompanied by a series of discussion papers by Pregibon and Vardi (1985), Buja and Kass (1985) and Fowlkes and Kettenring (1985), who all pointed out that ALS–ACE can show some unexpected behavior data analysts should be aware of. This seemingly anomalous behavior may appear upsetting at first glance, but we will show that it can be understood as a direct consequence of some simple spectral decompositions. As a method for finding transformations of regression data, ALS–ACE is more an outgrowth of canonical correlation than regression, and as such it is based on spectral theory rather than least squares principles. An understanding of the spectral behavior of ALS–ACE leads to insights into its anomalies, and in fact, their discovery by analytical means preceded their confirmation in computer simulations. The material presented in our discussion paper included the following seemingly anomalous effects:

1. ALS–ACE produces nontrivial transforms in cases which are generally considered as null situations such as joint unimodal spherical distributions. The reason is that ALS–ACE responds to any type of stochastic dependence in a distribution, and even for perfectly independent predictors and response, it will transform the data nontrivially due to sampling fluctuations which create minor deviations from independence.
2. ALS–ACE transforms can change abruptly even as the underlying distribution of the data changes very slightly. This may happen even in the most well-behaved situations where smooth densities exist and closeness is defined in the most stringent sense. The reason is that the eigenvector which belongs to the largest eigenvalue does not necessarily depend continuously on the underlying situation, a well-known fact in numerical analysis and perturbation theory.
3. Multivariate clusters may lead to approximate step functions as optimal transforms. Clusters can represent a form of dependence among variables and

ALS–ACE will pick them up since it responds to general dependence, not just the type which can be modeled by regression.

4. If data are generated according to a predictor model $\theta(Y) = \phi(X) + \varepsilon$ (X and ε independent), ALS–ACE will not necessarily find θ and ϕ. Examples can be constructed using situation 3 above with a strongly bimodal distribution for X in the model $Y = X + \varepsilon$. ALS–ACE will *not* find the identity transformations: Rather it will indicate that there are two clusters present by returning approximate step functions as optimal transforms.

5. Highly deterministic data which lead to a correlation of (or close to) one after transformation, may produce nonunique transforms; in fact, there might exist an entire infinite-dimensional space of optimal transformations.

These "anomalies" should not, however, invalidate ALS–ACE as a tool for data analysis. If properly understood, they should lose their current status of "deviant behavior." There exists a large literature which can provide insight into ALS–ACE. It falls under the topics of functional canonical variates, series expansion for bivariate distributions and maximal correlation.

Some authors in these areas are Hannan (1961), Dauxois and Pousse (1975), Chesson (1976), Naouri (1970), Barrett and Lampard (1955), Brown (1958), Lancaster (1958, 1969, 1975, 1980, 1983), Eagleson (1964, 1969), Rényi (1959) and Sarmanov (1958), among many others. Based on this literature, we present some analysis of two families of bivariate distributions which are flexible enough to exhibit features such as clustering and null situations:

(a) mixtures of product distributions (Sections 4 and 5) and
(b) some circular and elliptic probability measures (Sections 8–12).

While (a) may be new, some of the theory relating to (b) has parallels in a diversity of fields:

In engineering in the theory of noise and nonlinear instantaneous devices, McGraw and Wagner (1968) obtained results which can be read as partial answers to questions about the behavior of ALS–ACE under certain null situations. Much later and independently, Davison and Grunbaum (1981) developed similar theorems in the theory of computer-assisted tomography (CAT). Donoho and Johnstone (1986, 1988) developed an impressive and very far-reaching apparatus for projection pursuit regression (PPR) based on bivariate normal designs, which, however, can be extended in part to the distributions considered by McGraw and Wagner and Davison and Grunbaum.

Thus, the same mathematics of a peculiar family of distributions give insights into the working of methods as diverse as ALS–ACE, nonlinear devices, CAT and PPR.

The present paper is confined to populations or distributions, i.e., it does not immediately apply to finite-sample implementations of ALS–ACE. Therefore, computer simulations are necessary to confirm the qualitative validity of theoretical calculations for finite samples, where conditional expectations are estimated by curve estimates such as polynomials, splines or fast smoothers. The role of sampling fluctuations and peculiarities of the curve estimates used are probably accessible only to asymptotic theory (e.g., B & F, Appendices A.4 and A.5), hence simulations have to be carried out at any rate. All we can say offhand is that for sufficiently large sample sizes and consistent smoothers, the qualitative statements of this paper will hold to some degree of approximation. (See, however, Section 12 for examples where the connection between samples and populations breaks down.)

The limitations of a population approach are minor, however, in comparison to a deeper problem which aggravates the present author considerably more: The evidence we give for the performance of ALS–ACE on, say, null situations and clustered data, is based on examples alone—which means that our evidence is somewhat anecdotal. It is not quite clear how a theoretical feature which we derive for a particular distribution generalizes to a more general statement, such as "multivariate clusters cause ALS–ACE to generate approximate step functions." Even so, we consider the evidence hard enough to warrant reasonably general warnings of pitfalls of interpretation. We hope that future research will reveal a more complete picture and provide us with diagnostic tools for detecting potentially misleading ALS–ACE transformations.

Another limitation of this paper lies in the fact that we confine ourselves to one single predictor. This simplest case is analytically more tractable and yet ALS–ACE shows some of the behavior that distinguishes it from more conventional regression methods.

To the hurried reader we recommend having a look at Section 13, and then reading the initial paragraphs of other sections he or she might be interested in.

Abbreviations. Due to frequent citation, we will abbreviate A & S for Abramowitz and Stegun (1972), besides B & F for Breiman and Friedman (1985a).

For historical remarks and detailed references to the vast literature on the subjects of scaling, correspondence analysis and alternating least squares, see the books by Greenacre (1984), Chapter 4, Nishisato (1980), Section 1.2, and Gifi (1981). Gifi is a pseudonym for a group of authors at the Department of Data Theory, Leiden, The Netherlands. Members include de Leeuw and van der Burg who are cited on other occasions in this section.

2. Functional canonical correlation. We introduce some technical language by way of analogy with canonical correlations, of which the population ALS–ACE problem is essentially an L_2 version. Let $X_1, \dots, X_p$ and $Y_1, \dots, Y_q$ be centered random variables with finite variances. The canonical variates are defined as pairs (X, Y) which are nonzero elements of the linear spaces $\mathrm{span}(X_1, \dots, X_p)$ and $\mathrm{span}(Y_1, \dots, Y_q)$, respectively, and *stationary* elements with regard to the correlation $\mathrm{corr}[X, Y]$,

$$P_Y X = \lambda Y, \qquad P_X Y = \lambda X.$$

The operators P_X and P_Y are the orthogonal projections onto the above spaces. Orthogonality is understood with regard to the inner product given by the covariance or equivalently the product moment, since we deal with centered variables only:

$$\langle X, Y\rangle = \mathrm{Cov}[X, Y] = E[XY], \qquad \|X\|^2 = \mathrm{Var}[X] = E[X^2].$$

The population ALS–ACE problem can be viewed as the extraction of the strongest variates in a nonlinear or functional version of the canonical correlation problem. Focusing on the one-predictor situation with a single X and a single Y, we arrive at nonlinear, or better: *functional*, variates if we replace the linear spaces by

$$H(X) = \{\phi(X) | E[\phi(X)] = 0, \mathrm{Var}[\phi(X)] < \infty\},$$
$$H(Y) = \{\theta(Y) | E[\theta(X)] = 0, \mathrm{Var}[\theta(X)] < \infty\}.$$

The corresponding projection operators are just the conditional expectations given X and Y, respectively, restricted to the centered square-integrable variables. We denote them either by $E^X Z$ or $E[Z \mid X]$, and $E^Y Z$ or $E[Z|Y]$. Functional canonical variates are defined as pairs $(\phi(X), \theta(Y))$ which are again stationary with regard to the correlation $\mathrm{corr}[\phi(X), \theta(Y)]$:

$$E^Y\phi(X) = \lambda\theta(Y), \qquad E^X\theta(Y) = \lambda\phi(X).$$

This amounts to a singular value problem for the pair of operators E^Y and E^X. While the more familiar singular value problem for a matrix A amounts to finding values $\lambda \geq 0$ and unit vectors u and v satisfying $Au = \lambda v$ and $A^t v = \lambda u$, we notice that the restricted projections $E^X|H(Y)$: $H(Y) \to H(X)$ and $E^Y|H(X)$: $H(X) \to H(Y)$ are duals of each other, as is recognized from the identities

$$\langle E^Y\phi(X), \theta(Y)\rangle = \langle \phi(X), \theta(Y)\rangle = \langle \phi(X), E^X\theta(Y)\rangle.$$

The solutions which belong to the largest λ are the optimal or ALS–ACE transformations, and the value of λ is the maximal correlation. Every singular value problem is associated with two decoupled eigenproblems:

$$E^X E^Y\phi(X) = \lambda^2\phi(X), \qquad E^Y E^X\theta(X) = \lambda^2\theta(Y).$$

If none of the eigenvalues λ^2 is multiple, then any pair of eigensolutions $(\phi(X), \theta(Y))$ for the same eigenvalue λ^2 also solves the singular value problem, and one can arrange $\lambda \geq 0$ by changing the sign in one of the transforms if necessary. If multiplicity occurs, not all possible combinations of eigensolutions of the decoupled problems will satisfy the singular value conditions; this has to be reinforced by picking an eigensolution $\phi(X)$, say, and pairing it with $\theta(Y) = (1/\lambda)E^Y\phi(X)$, or vice versa. (Solutions whose stationary correlations λ are 0 cannot be sensibly matched up.)

Singular values and associated transformations do not necessarily exist unless additional restrictive assumptions are made. One way of dealing with this problem is by bypassing the existence question and recasting the framework in the language of general spectral theory using spectral measures. This approach, which does not require additional assumptions but more advanced tools from functional analysis, is carried out in Hannan (1961), Dauxois and Pousse (1975) and Chesson (1976). Unfortunately, the fact that general spectral theory does not guarantee the existence of largest eigenvalues still leaves us with a need to find more restrictive regularity conditions which single out situations in which they do exist. For a simple example where no largest eigenvalue exists, see Section 11. In those cases where one of the spaces $H(X)$ or $H(Y)$ is finite- dimensional (e.g., when the sample space is finite, or one of the variables is discrete such as in Fisher's scoring [Fisher (1970), Sec. 49.2] for contingency tables), linear algebra ensures the existence of singular value and eigendecompositions. For part of this paper (Sections 4 and 5), we will indeed consider cases which lead to finite-dimensional problems, while in most infinite-dimensional situations a common simplifying condition, compactness of the restricted projections $E^X|H(Y)$ and $E^Y|H(X)$, and hence of $E^XE^Y|H(X)$ and $E^YE^X|H(Y)$, is met. The former form a dual pair as noted above, while the latter are easily seen to be self-adjoint as mappings $H(X) \to H(X)$ and $H(Y) \to H(Y)$, respectively. The elementary spectral theorem for compact dual pairs and self-adjoint operators then grants the existence of a sequence of singular values λ_m and of associated transformations $\phi_m(X)$ and $\theta_m(Y)$ which satisfy the following conditions:

(a) each nonzero value appears with finite multiplicity at most;

(b) $\lambda_m \downarrow 0$, $\lambda_m \geq 0$;

(c) $\{\phi_m(X)\}$ and $\{\theta_m(Y)\}$ form complete orthonormal systems in $H(X)$ and $H(Y)$, respectively;

(d) $E^Y\phi_m(X) = \lambda_m\theta_m(Y)$, $E^X\theta_m(Y) = \lambda_m\phi_m(X)$.

For an elementary reference, see Naylor and Snell (1982), Sections 6.11 and 6.14, and also B & F (Section 5.3). As a corollary, we obtain that not only are the two sets of transformations orthogonal (uncorrelated) within themselves, but it also holds that

$$(*) \qquad \langle\phi_l(X), \theta_m(Y)\rangle = E[\phi_l(X)\theta_m(Y)] = \lambda_m\delta_{l,m},$$

as may be seen by a routine manipulation with the conditional expectations E^X or E^Y. Thus, the singular values are the correlations between matched pairs of transforms (which implies $|\lambda_m| \leq 1$), while all other pairs are uncorrelated. This justifies the terms *functional canonical correlations* for the singular values λ_m, and *functional canonical variates* for the corresponding transforms.

The singular value decompositions which correspond to these functional canonical correlations and variates can be written as follows:

$$E^X\theta(Y) = \sum_m \langle\theta(Y), \phi_m(X)\rangle\phi_m(X),$$

$$E^Y\phi(X) = \sum_m \langle\phi(X), \theta_m(Y)\rangle\theta_m(Y).$$

For the subsequent sections we should comment on the role of centering in the definitions of the spaces $H(X)$ and $H(Y)$. As far as the optimization of correlations is concerned, one must eliminate constants, but for the eigenproblems, we may feel free to ignore this artificial condition as long as we realize that the constants will always appear as singular transforms with singular value $\lambda = 1$ and hence should be discarded. The remaining eigensolutions will be orthogonal to the constants, i.e., centered. As we will concentrate on the singular value problem and consider the ALS–ACE optimization as the derived problem of picking the (second) largest singular value and its transforms, we may as well neglect centering in practical computations. Thus, from now on we assume that $\lambda_0 = 1$ is the first singular value with $\phi_0 = 1$ and $\theta_0 = 1$ as its eigenfunctions.

3. The singular value decomposition for bivariate distributions. While the previous section was concerned with a singular value decomposition for conditional expectation operators, the present section deals with an equivalent singular value decomposition for the underlying bivariate distribution, as described by Lancaster (1958, 1969, Chapter 6, Section 3). This decomposition will show that the two sets of eigenfunctions ϕ_m and θ_m in general capture *all* dependence between the variables X and Y. The derivation has the flavor of a null hypothesis calculation because we are to consider the Radon–Nikodym derivative

$$f(x,y) = \frac{Q_{X,Y}(dx,dy)}{Q_X(dx) \times Q_Y(dy)},$$

which may be viewed as the density of the actual distribution $Q_{X,Y}$ w.r.t. independence. This function is constant 1 ($Q_X \times Q_Y$-a.s.) iff X and Y are independent; hence, deviations from constancy can be seen as indications of dependence. For the above definition of $f(x,y)$ to make sense, we need $Q_{X,Y}(dx,dy)$ to be absolutely continuous w.r.t. $Q_X(dx) \times Q_Y(dy)$, an assumption which is not always satisfied. Cases not amenable to this approach include deterministic dependences such as $X = Y$ a.s. However, the common situation where both $Q_{X,Y}$ and $Q_X \times Q_Y$ have densities $q_{X,Y}(x,y)$ and $q_X(x)q_Y(y)$ w.r.t. a common product measure $\lambda(dx,dy)$ on $x-y$ space (usually the Lebesgue measure in the plane) is covered by

$$f(x,y) = \frac{q_{X,Y}(x,y)}{q_X(x)q_Y(y)} \quad \lambda\text{-a.e.}$$

Another useful interpretation of $f(x,y)$ is as a kernel of both conditional expectations, not w.r.t. Lebesgue measure, but the marginal distributions:

$$E^{X=x}h(Y) = E[f(x,Y)h(Y)], \qquad E^{Y=y}g(X) = E[f(X,y)g(X)].$$

Continuing in the spirit of a null hypothesis calculation, one expands $f(x,y)$ in a complete set of $Q_X \times Q_Y$ orthonormal functions. For this expansion to make sense in $L_2(Q_X \times Q_Y)$, one assumes

$$\iint f(x,y)^2 Q_X(dx) Q_Y(dy) < \infty,$$

which in terms of densities reads as

$$\iint \frac{q_{X,Y}(x,y)^2}{q_X(x)q_Y(y)} \lambda(dx,dy) < \infty,$$

i.e., assumption 5.4 of B & F. This is a common condition to assure that both conditional expectations are Hilbert–Schmidt operators, and it implies that both of them are compact. The Hilbert–Schmidt property is also equivalent to the statement that the sum of the eigenvalues of $E^X E^Y$ as well as $E^Y E^X$ is finite [Jorgens (1982), Section 6.6]. As for the expansion of $f(x,y)$, it is natural to use the functions $\psi_{l,m}(x,y) = \phi_l(x)\theta_m(y)$, $l, m = 0, 1, 2, \dots,$ since they form an orthonormal system w.r.t. the product measure $Q_X \times Q_Y$. The L_2 expansion

$$f(x,y) = \sum_{l,m=0}^{\infty} c_{l,m}\phi_l(x)\theta_m(y)$$

simplifies due to property $(*)$ of the preceding section:

$$\begin{aligned} c_{l,m} &= \iint f(x,y)\phi_l(x)\theta_m(y) Q_X(dx) Q_Y(dy) \\ &= \iint \phi_l(x)\theta_m(y) Q_{X,Y}(dx,dy) \\ &= E[\phi_l(X)\theta_m(Y)] = \lambda_m \delta_{l,m}. \end{aligned}$$

Recalling our convention $\phi_0 = 1$, $\theta_0 = 1$ and $\lambda_0 = 1$, we thus obtain:

PROPOSITION 3.1. *If the Radon-Nikodym derivative $Q_{X,Y}(dx,dy)|Q_X(dx) \times Q_Y(dy)$ exists and is square-integrable w.r.t. $Q_X \times Q_Y$, then the following expansion holds in the $L_2(Q_X \times Q_Y)$ sense*:

$$f(x,y) = 1 + \sum_{m=1}^{\infty} \lambda_m \phi_m(x)\theta_m(y).$$

In statistics, this result was introduced by Lancaster (1958, 1969), but the special case where the transformations are orthogonal polynomials (Section 7) has been considered by Barrett and Lampard (1955) earlier in problems of noise and nonlinear devices; see also Leipnik (1959). In functional analysis, this expansion has an older history which goes back at least to Schmidt's (1907) work on integral equations. The expansion may justifiably be called a singular value decomposition (svd) of the bivariate distribution $Q_{X,Y}(dx,dy)$. It is at the root of correspondence analysis of the French school [Lebart, Morineau and Warwick (1984), Section 2.4., equation 39; and Greenacre (1984)], when applied to discrete variables and estimated from data, but so-called "continuous correspondence analysis" [Naouri (1970)] is exactly the population case covered by the above proposition. For a more careful treatment of bivariate (possibly nondiagonal) expansions under weaker conditions, see Cambanis and Liu (1971).

We indicate some connections of the "density w.r.t. independence" $f(x,y)$ with some related notions which have a history in statistics. Rényi (1959) introduces what he calls "mean square contingency":

$$C(X,Y) = \left[\iint (f(x,y) - 1)^2 Q_X(dx) Q_Y(dy)\right]^{1/2}.$$

The bivariate distributions with finite mean square contingency are exactly the ones with iterated conditional expectations of the Hilbert–Schmidt type. Clearly, $C(X,Y)$ is a reasonable *measure of dependence* since it vanishes if and only if X and Y are independent, in contrast to the Pearson correlation for which the same statement holds only under additional assumptions such as normality. Since the expectation of $f(x,y)$ under independence is 1, $C(X,Y)^2$ may also be considered as the variance of $f(x,y)$ under independence. It therefore becomes

$$C(X,Y)^2 = \iint f(x,y)^2 Q_X(dx) Q_Y(dy) - 1,$$

which, written in this form, is also known as Pearson's ϕ^2 functional for the measures $Q_{X,Y}$ and $Q_X \times Q_Y$ [see Lancaster (1969), Chapter 6, Sections 1 and 3]. The mean square contingency specializes to the χ^2 functional for discrete random variables:

$$C(X,Y)^2 = \sum_{i,j} \frac{\left[q_{X,Y}(x_i,y_j) - q_X(x_i)q_Y(y_j)\right]^2}{q_X(x_i)q_Y(y_j)},$$

where x_i and y_j are the discrete values taken on by X and Y, respectively, and $\lambda(dx,dy)$ is counting measure. The singular value decomposition for $f(x,y)$ also results in a decomposition for the $C(X,Y)^2$ or ϕ^2 functional:

$$C(X,Y)^2 = \phi^2 = \sum_{m=1}^{\infty} \lambda_m^2.$$

The relevance of the above statements for ALS–ACE consists of the fact that ALS–ACE extracts the most significant "rank 1" term $\phi_1(x)\theta_1(y)$ in the expansion of the function $f(x,y) - 1$, which measures pointwise deviation from independence. ALS–ACE also extracts the maximal squared correlation as the dominant term in the decomposition of the mean square contingency, which is a global measure of dependence.

The most basic point of this section is the interpretation of population ALS–ACE as a comparison of a bivariate distribution with the null hypothesis of independence. This view prepares the ground for phenomena which would appear anomalous in the original interpretation of ALS–ACE as a regression tool.

4. Distributions of finite rank. We move the focus from generalities to some distributions which are accessible to finite-dimensional linear algebra, namely, mixtures of a finite number of pr૭duct measures:

$$Q_{X,Y} = \sum_{i=1}^{N} \alpha_i Q_X^{(i)} \times Q_Y^{(i)}, \qquad \alpha_i > 0, \sum_{i=1}^{N} \alpha_i = 1.$$

To fix notation, we introduce the associated joint, marginal and conditional densities:

$$q_{X,Y}(x,y) = \sum_{i=1}^{N} \alpha_i q_X^{(i)}(x) q_Y^{(i)}(y),$$

$$q_X(x) = \sum_{i=1}^{N} \alpha_i q_X^{(i)}(x), \qquad q_Y(y) = \sum_{i=1}^{N} \alpha_i q_Y^{(i)}(y),$$

$$q_{Y|X}(y|x) = \sum_{i=1}^{N} \alpha_i \frac{q_X^{(i)}(x)}{q_X(x)} q_Y^{(i)}(y),$$

$$q_{X|Y}(x|y) = \sum_{i=1}^{N} \alpha_i \frac{q_Y^{(i)}(y)}{q_Y(y)} q_X^{(i)}(x).$$

The "density with respect to independence" becomes:

$$f(x,y) = \sum_{i=1}^{N} \alpha_i \frac{q_X^{(i)}(x)}{q_X(x)} \frac{q_Y^{(i)}(y)}{q_Y(y)}.$$

This mixture representation resembles very much the svd of Section 3, the difference being that the component functions are not orthonormalized but have to be nonnegative. As in the previous section, we assume that $f(x,y)$ is square integrable w.r.t. $Q_X \times Q_Y$, which implies that the components $q_X^{(i)}(x)/q_X(x)$ and $q_Y^{(i)}(y)/q_Y(y)$ are elements of $L_2(Q_X)$ and $L_2(Q_Y)$, respectively. The spaces spanned by these two sets of functions or variables.

$$\operatorname{span}\left\{ \frac{q_X^{(i)}(X)}{q_X(X)} \middle| i = 1,\dots,N \right\} \quad \text{and} \quad \operatorname{span}\left\{ \frac{q_Y^{(i)}(Y)}{q_Y(Y)} \middle| i = 1,\dots,N \right\},$$

are the images of the conditional expectation operators E^X and E^Y, respectively, as may be seen from the form' of the conditional densities above. Finding the svd of the joint distribution therefore reduces to a problem of finite-dimensional linear algebra. The number of nontrivial terms in the svd is $N - 1$ rather than N, since the constants are still part of the above spaces:

$$\sum_{i=1}^{N} \alpha_i \frac{q_X^{(i)}(x)}{q_X(x)} = 1, \qquad \sum_{i=1}^{N} \alpha_i \frac{q_Y^{(i)}(y)}{q_Y(y)} = 1.$$

We may then eliminate them by considering the $N - 1$-dimensional spaces of centered variables.

PROPOSITION 4.1. *The singular value decomposition*

$$f(x,y) = 1 + \sum_{m=1}^{N-1} \lambda_m \phi_m(x)\theta_m(y)$$

of a finite mixture of product measures

$$Q_{X,Y} = \sum_{i=1}^{N} \alpha_i Q_X^{(i)} \times Q_Y^{(i)}$$

can be obtained from a canonical correlation analysis of the two spaces

$$\operatorname{span}\left\{ \frac{q_X^{(i)}(X)}{q_X(X)} - 1 \,\middle|\, i = 1,\dots, N-1 \right\},$$

$$\operatorname{span}\left\{ \frac{q_Y^{(i)}(Y)}{q_Y(Y)} - 1 \,\middle|\, i = 1,\dots, N-1 \right\}$$

according to Section 2, *using $Q_{X,Y}$ as joint distribution of the variables X and Y.*

For a proof we only have to note that these variables are indeed centered:

$$E\left(\frac{q_X^{(i)}(X)}{q_X(X)} \right) = 1 \quad \text{and} \quad E\left(\frac{q_Y^{(i)}(Y)}{q_Y(Y)} \right) = 1. \qquad \square$$

The special case $N = 2$ can be solved explicitly. The examples to be considered in the next section are examples of such "rank 2 distributions." In their svd, there is only one nontrivial term left:

$$f(x,y) = 1 + \lambda\phi(x)\theta(y),$$

and the above spaces of centered variables are one-dimensional. Therefore, any one of the two standardized elements of these spaces can serve as an eigentransform.

PROPOSITION 4.2. *The ALS–ACE transforms of a mixture of two product measures $Q_{X,Y} = (1-\alpha)Q_X^{(1)} \times Q_Y^{(1)} + \alpha Q_X^{(2)} \times Q_y^{(2)}$ are given by*

$$\phi(x) \propto \frac{q_X^{(1)}(x)}{q_X(x)} - 1 \propto \frac{q_X^{(1)}(x) - q_X^{(2)}(x)}{(1-\alpha)q_X^{(1)}(x) + \alpha q_X^{(2)}(x)},$$

$$\theta(y) \propto \frac{q_Y^{(1)}(y)}{q_Y(y)} - 1 \propto \frac{q_Y^{(1)}(y) - q_Y^{(2)}(y)}{(1-\alpha)q_Y^{(1)}(y) + \alpha q_Y^{(2)}(y)}.$$

The transform ϕ is determined by the marginal distribution of X regardless of the mixture components in the marginal distribution of Y, and vice versa for θ and Y.

REMARK. Initially, it may seem surprising that knowing the marginal distribution of X in terms of its mixture components allows one to infer the optimal transform ϕ, irrespective of the marginal distribution of Y. This fact by itself opens up the possibility of constructing "anomalous" situations, which is the program of the next section.

5. Anomalies of ALS–ACE on binary mixtures of independent sources.

Binary mixtures of product measures (which we call mixtures of "independent sources") give us a scenario for two very different situations for ALS–ACE artifacts: (1) clusters and (2) certain types of heavy tails. Using normal distributions as building blocks, we can mix standard normals with centers $-\mu$ and $+\mu$, respectively, to obtain a simple illustration for clusters. Again using normal distributions only, we arrive at an example for heavy tails by a contamination model: Mix two normals centered at the origin, a fraction $1 - \alpha$ with marginal standard deviation 1.0, and a fraction α with a marginal standard deviation $\sigma > 1$. In what follows, we denote by $\psi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ the univariate standard normal density. With Proposition 4.2 in mind, we can ignore what the mixture components of the Y distribution are and restrict our attention to the variable X alone since its transformation is independent of the Y components.

Location mixtures: *Clusters*. We consider only clusters of equal weight, i.e., $\alpha = 0.5$, and we leave the components with unit standard deviation:

$$q_X^{(1)}(x) = \psi(x - \mu), \qquad q_X^{(2)}(x) = \psi(x + \mu).$$

The optimal transform calculated according to Proposition 4.2 is

$$\phi(x) \propto \frac{\psi(x - \mu) - \psi(x + \mu)}{\psi(x - \mu) + \psi(x + \mu)} = \tanh(\mu x).$$

An example with shift $\mu = 2$ is shown in the left half of Figure 1, where both the population transform and an estimate from a sample of size 500 are plotted. The implementation of the ACE algorithm supplied by B & F was used.

Qualitatively, ϕ approaches a step function as $\mu \to \infty$, with a jump from -1 to $+1$ at $t = 0$. We could have constructed examples with nonoverlapping clusters which lead to exact step functions, e.g., a mixture of uniform distributions on the unit intervals $[-1, 0]$ and $[0, +1]$. With clusters we encounter a type of deviation from independence which is nonstandard from the point of view of regression analysis, but nevertheless, ALS–ACE as a detector of *any* type of dependence picks them up. We have seen clustering effects in real data, most recently in some market survey data at Bell Laboratories where ACE pinpointed real but visually undetectable clusters.

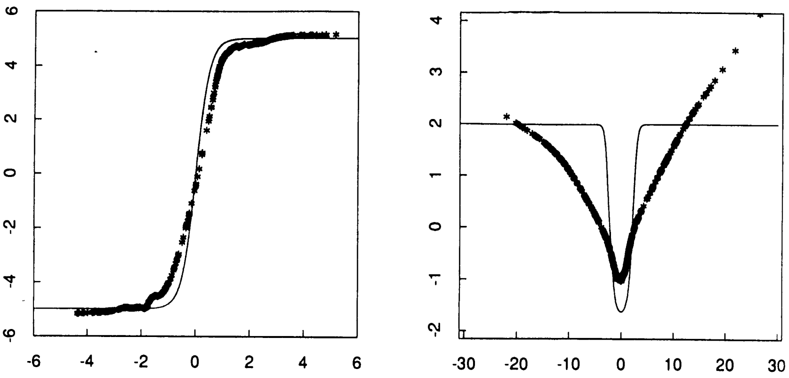

FIG. 1. *Optimal tranforms for binary mixtures of independent sources. Left: location mixture,* 50% $N(2,1)$ *and* 50% $N(-2,1)$, $n = 500$. *Right: scale mixture,* 50% $N(0,1)$ *and* 50% $N(0,100)$, $n = 500$.

Scale mixtures: Mass concentrations and heavy tails. We pick a fraction $1 - \alpha$ from a standard normal,

$$q_X^{(1)}(x) = \psi(x),$$

and another fraction α from a centered normal with standard deviation $\sigma > 1$:

$$q_X^{(2)}(x) = \psi(x/\sigma)/\sigma.$$

We obtain as optimal transforms (up to a sign change):

$$\phi(x) \propto \frac{\psi(x/\sigma)/\sigma - \psi(x)}{(1-\alpha)\psi(x) + \alpha\psi(x/\sigma)/\sigma}.$$

The right half of Figure 1 shows population transforms and estimates for a sample of 500 cases with $\alpha = 0.5$ and $\sigma = 10$. The difficulty in capturing the features of the population transform may partly be attributed to boundary effects in the underlying smoothers. This is hard to avoid as the variance–bias trade-off implicit in smoothers is tested the most near boundaries.

Major features of the transforms are on the one hand the steep valley caused by the concentration in the center, and on the other hand the asymptote which is due to the fuzzy scatter. This example is somewhat related to the effect noted by Pregibon and Vardi (1985) who show that mass concentrations in arbitrary locations can induce arbitrary values between $+1$ and -1 in the optimal correlation. Essentially, this is possible by choosing transformations which separate the mass concentration from its surroundings, very much like the valley and the asymptote do in the present example.

The scale mixture example would allow another interpretation as well: One could consider the fuzzy scatter as a heavy tail (especially if the fraction α were chosen smaller). One might then speculate that heavy tails have a flattening effect on the wings of the transforms much as the asymptotes in the

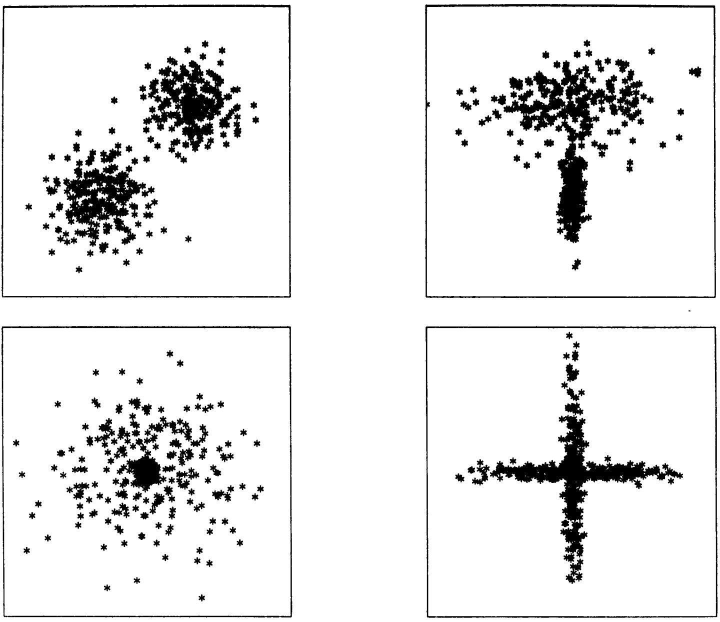

FIG. 2. *Scatterplots of binary mixtures of independent sources. Top left*: y = *location mixture*, x = *location mixture. Top right*: y = *location mixture*, x = *scale mixture. Bottom left*: y = *scale mixture*, x = *scale mixture* [y *center* $N(0,1)$ *matched with* x *center* $N(0,1)$]. *Bottom right*: y = *scale mixture*, x = *scale mixture* [y *center* $N(0,1)$ *matched with* x *tail* $N(0,100)$].

current example. We will see in Section 12 that this expectation is wrong: There we will give an example of much heavier tails with parabolas apparently being the optimal transformations. A more correct interpretation would be that the minimum at the origin and the high plateau in the wings indicate two different sources, one dominant at the origin and the other out at infinity.

By Proposition 4.2, we were able to discuss these examples using just one of the two marginals. This allows the curious possibility that the type of mixture in the X marginal may be different in the Y marginal; e.g., the two mixture components may differ in location in X while they differ in spread in Y; or, while both variables are scale mixtures, one can match up the central spike component in one variable with the flat wing component in the other. The four scatterplots of Figure 2 illustrate the shapes of samples which can be obtained by matching location and/or scales mixtures in all possible ways.

6. Symmetric bivariate distributions. We simplify the singular value decomposition (Section 3) in the case of symmetry, i.e., when the laws of (X, Y) and (Y, X) are the same. These simplifications were first exploited by Sarmanov (1958) in probabilistic terms, although it is really just a special case of what is known in analysis as Mercer's expansion of symmetric kernels.

The assumption of symmetry implies that the ranges of X and Y are the same (usually, but not necessarily $\mathbb{R}$) and that X and Y have the same marginal distribution $Q_X(dt) = Q_Y(dt) = Q(dt)$. We can then consider the conditional expectation operators $E^{X=x}h(Y) = \int P^{X=x}(dy)h(y)$ and $E^{Y=y}g(X) = \int P^{Y=y}(dx)g(x)$ as transition probabilities between the "same" spaces $H(X) = L_2(Q)$ and $H(Y) = L_2(Q)$, and, as such, they are identical: $E^X = E^Y = P$. Equivalently, we may consider P as an operator in, say, $L_2(Q_X)$ and define it as follows.

PROPOSITION 6.1. *Assuming that the distribution of* (X, Y) *is symmetric, the operator* P: $L_2(X) \to L_2(X)$, $g(X) \to P(g(X))$ *defined by* $P(g(X)) = E^X g(Y)$ *is symmetric, and all of its eigenfunctions are also eigenfunctions of* $E^X E^Y$. *The eigenvalues of the latter are the squares of the eigenvalues of* P. *If the "density w.r.t. independence"* $f(x,y)$ *is square integrable w.r.t.* $Q_X \times Q_Y$, *the svd for the distribution* $Q_{X,Y}$ *of* (X, Y) *can be written as*

$$f(x,y) = 1 + \sum_{m=1}^{\infty} \lambda_m \phi_m(x) \phi_m(y),$$

where the transforms ϕ_m *form a complete set of orthonormal eigenfunctions for* P *and the eigenvalues* λ_m *can have arbitrary signs.*

For general bivariate distributions, it makes sense to assume $\lambda_m \geq 0$ as there does not exist a more natural way to choose between ϕ_m and $-\phi_m$ and same for θ_m, but for symmetric distributions it is more convenient to adopt the following convention.

CONVENTION. *We impose* $\phi_m = \theta_m$ *and permit an arbitrary sign in* λ_m *whenever the distribution is symmetric. This convention will hold throughout this paper.*

7. Bivariate distributions with polynomial eigentransforms. We now describe a method for finding the singular value decomposition of those bivariate distributions whose eigentransforms ϕ_m and θ_m can be chosen to be polynomials of the same (!) degree. The corresponding singular value decomposition in terms of orthogonal polynomials is called a Barrett–Lampard expansion (1955) in the engineering literature. This case, which will occupy much of the remainder of this paper, is so important due to its analytical tractability that Lancaster (1975), equation 3.5, introduces a special term: *polynomial biorthogonality*. We adopt the following convention.

CONVENTION. *In polynomially biorthogonal situations, we sort the transforms* ϕ_m *and* θ_m *according to polynomial degrees rather than decreasing (absolute) singular values.*

As a consequence, the transforms ϕ_m and θ_m will both be polynomials of degree m. This convention is independent of and compatible with the one of the previous section: In symmetric *and* polynomially biorthogonal cases, both conventions can be reinforced simultaneously.

Polynomial biorthogonality is a very strong condition which implies that moments of all degrees exist. Examples with tail weight as heavy as the t distribution are therefore excluded a priori (see Section 12 for this case).

The following proposition is basic for much of the remainder of this paper. A slightly different form is referred to as Brown's criterion (1958) in the engineering literature on nonlinearities and noise. See also Csaki and Fischer (1960) in the theory of maximal correlations.

PROPOSITION 7.1. *The bivariate distribution $Q_{X,Y}$ of (X, Y) is polynomially biorthogonal if and only if the following conditions hold*:

(i) *The conditional moment $E[Y^m|X]$ is a polynomial of degree $\leq m$ in X for all m, and the same holds for $E[X^m|Y]$ in Y.*

(ii) *The powers X^m and Y^m form complete systems in the respective spaces $L_2(Q_X)$ and $L_2(Q_Y)$.*

The condition of completeness of the system of powers is generally not a problem. For probability measures with bounded support in $\mathbb{R}$, polynomials are dense in L_2 due to (a) the Weierstrass approximation theorem (on a finite interval, any continuous function can be approximated uniformly by polynomials), and (b) the fact that continuous functions are dense in L_2. For probability measures with unbounded support, we may use techniques similar to those used in moment problems based on characteristic functions.

LEMMA 7.2. *Either of the following are sufficient conditions for completeness of the system of polynomials w.r.t. a given distribution*:

(i) *The support of the distribution is bounded in $\mathbb{R}$.*

(ii) *For the even moments $M_{2n} = E[X^{2n}]$ we have* $\limsup M_{2n}^{1/2n}/n < \infty$.

As an example for (ii), for the even moments of the standard normal distribution $M_{2n} = 1 \cdot 3 \cdots \cdot (2n - 1)$, the above lim sup is 0, which proves the well-known completeness of the Hermite polynomials w.r.t. the normal distribution.

Condition (ii) must be standard. One has to show that $f = 0$ a.s. if $f \in L_2(Q)$ and $E[X^n f(X)] = 0$ for all $n = 1, 2 \dots$. This can be done by introducing the finite signed measure $d\mu(x) = f(x)\, dQ(x)$ and adapting the proof of Feller (1971), XV.4, 4.14, which applies not only to probability measures.

The next proposition provides a simple rule for calculating eigenvalues for polynomially biorthogonal distributions:

PROPOSITION 7.3. *If a bivariate distribution has the polynomial biorthogonal property, the eigenvalues λ_m^2 are the products of the leading coefficients in* *the polynomials given by the conditional moments* $E[Y^m|X]$ *and* $E[X^m|Y]$. *If the distribution is symmetric,* λ_m^2 *is the square of the leading coefficient of either polynomial.*

This is a simple consequence of the fact that $E^X E^Y$ as a mapping of $L_2(Q_X)$ onto itself is in triangular form with regard to the basis of monomials X^i, and the diagonal elements are exactly its eigenvalues. □

Here is an example of polynomial biorthogonality to illustrate the application of 7.1–7.3: Define what we may call a *triangular bivariate beta distribution* by the following density:

$$Q_{X,Y}(x,y) = \frac{a+b}{B(a,b)} x^{a-1} y^{b-1} \quad \text{for } x, y > 0 \text{ and } x + y < 1,$$

and 0 otherwise. [$B(a,b)$ is the beta function; see A & S, 6.2.] Our standard procedure, which will be repeated several times throughout this paper, consists of:

1. finding the conditional distributions and checking whether the mth conditional moment is a polynomial of degree less than or equal to m [condition (i) of Proposition 7.1];
2. finding the marginal distributions and their orthogonal polynomials (which should form a complete system along Lemma 7.2); and
3. obtaining the eigenvalues λ_m^2 via Proposition 7.3.

In this instance, the conditional distribution of Y given X for the triangular bivariate beta is $(1 - X)\beta(b, 1)$, where $\beta(b, 1)$ stands for the univariate beta distribution with parameters b and 1 (A & S, 26.1.33). With this and a dual statement for the conditional distribution of X given Y, we obtain

$$E[Y^m|X] = \frac{b}{b+m}(1-X)^m, \qquad E[X^m|Y] = \frac{a}{a+m}(1-Y)^m,$$

which are indeed polynomials. The marginal distribution of X is $\beta(a, b+1)$, which has the Jacobi polynomials with parameters $p = a + b$ and $q = a$ on the interval $(0, 1)$ as orthogonal polynomials and hence eigentransforms ϕ_m of X (A & S, 22.2.2). Similarly, we get Jacobi polynomials with $p = a + b$ and $q = b$ for θ_m. The leading coefficients in the conditional moments are $(-1)^m b/(b + m)$ and $(-1)^m a/(a + m)$, which by Proposition 7.3 give

$$\lambda_m^2 = \frac{ab}{(a+m)(b+m)}.$$

The maximal value is attained for $m = 1$, i.e., the linear transformations of X and Y. In other words, for these variables the raw correlation is the maximal correlation, and no nonlinear transformation is needed to attain it.

From a practical point of view, examples such as this one remind us that an ALS–ACE analysis resulting in linear optimal transformations does not necessarily indicate a satisfactory error structure. For example, the conditional distribution of Y given X [which is $(1 - X)\beta(b, 1)$] has inhomogeneous variance, and is also skewed except for $b = 1$ (see A & S, 26.1.33). Marginal transformations are clearly unable to rectify the situation.

A further and related conclusion is that ALS–ACE does not necessarily attempt to transform to normality, as one might initially expect, and in this light the developments in Kendall and Stuart (1979), Section 33.44, seem somewhat inappropriate in their suggested informal interpretation of the fact that for bivariate normal distributions the untransformed variables attain the maximal correlation. The unsuspecting reader might conclude that this property is peculiar to the bivariate normal, although this is not spelled out directly. In contrast, the bivariate triangular beta distribution is a first example in which the maximal correlation is attained without transformation as well.

Many other examples of polynomially biorthogonal distributions appear in the literature. Some sources are Lancaster (1969, 1975, 1983), Eagleson (1964, 1969), Griffiths (1969), McFadden (1966), Lee (1971), Rényi (1959), Sarmanov (1963) and the references therein. We will not pursue this topic in full generality but focus in the next sections on circular and elliptic distributions which are also polynomially biorthogonal.

8. Circular bivariate distributions. We apply polynomial biorthogonality to some circular distributions and derive a few facts which have implications for the performance of ALS–ACE in nonstandard cases of stochastic dependence. We also give a characterization theorem due to McGraw and Wagner in order to indicate how pervasive and/or limited the parabolic ALS–ACE transforms are in noisy data.

Nonstandard stochastic dependence and its effects on ALS–ACE. Fowlkes and Kettenring (1985) and Buja and Kass (1985) in their discussion of B & F noticed an anomaly in the behavior of ALS–ACE when applied to null situations. The transforms produced by ALS–ACE in these situations looked like parabolas. We will show here and in the next section why this is a systematic effect.

We start with the uniform distribution on the unit disk $x^2 + y^2 < 1$, which served in Buja and Kass (1985) as a simple example of an unstructured or null situation with raw correlation 0 and yet stochastic dependence. This distribution is symmetric and also polynomially biorthogonal: The conditional distribution of Y given X is uniform between $-\sqrt{1 - X^2}$ and $+\sqrt{1 - X^2}$; hence, for odd powers the conditional moments vanish, and for even powers we obtain

$$(*) \qquad E\left[Y^{2m} \middle| X\right] = (1 - X^2)^m/(2m + 1),$$

which are indeed polynomials. The eigentransforms are the orthogonal polynomials with regard to the marginal distribution of X which has density $2\pi^{-1/2}$ $(1 - x)^{1/2}$; i.e., they are the Chebyshev polynomials of the second kind (A & S,

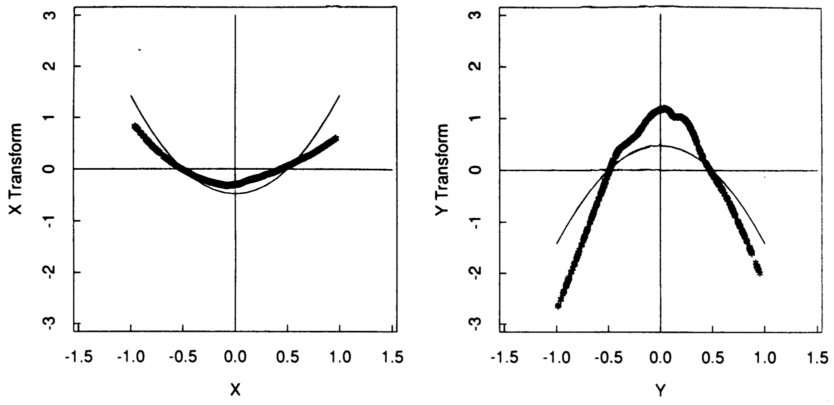

FIG. 3. *Optimal x and y transforms of the uniform distribution on the unit disk. The y variable on the right was entered first in the ACE iterations.*

22.2.5). For even order, the singular values are the coefficients of the largest power in (*) and for odd order they vanish:

$$\lambda_{2m} = (-1)^m/(2m+1), \qquad \lambda_{2m+1} = 0.$$

The absolute largest singular value is obtained for order two: $\rho_{\max} = |\lambda_2| = 1/3$, a figure that may appear large, given that we are facing what is commonly seen as a null situation.

Simple simulation experiments confirm that this qualitative behavior of optimal transformations is reflected in finite-sample applications of ALS–ACE. Figure 3 shows transformations obtained from B & F's ACE implementation on a pseudorandom sample of 500 points drawn from a uniform distribution on the unit disk. The agreement between population and sample transform is not great, but the theoretically predicted parabolic shape is recovered by the sample ACE transforms. ACE seems to favor the variable which enters the iterations first: The Y transform gets considerably more variance than its counterpart.

The effect discussed here is not just of academic value: Parabolic transformations as shown in Figure 3 do not only appear in simulations but real data as well. The author has encountered some multivariate data of defects in newborn infants and alcohol consumption of mothers where ACE produced a relatively low squared correlation of about 10% and a parabolic transform in one of the quantitative variables—a strong enough hint at a null finding.

In a sense to be made precise in Section 9, the uniform distribution on the unit disk is halfway between independence represented by the circular normal distribution and extreme dependence represented by the uniform distribution on the periphery of the unit disk (hence a degenerate measure with support on the unit circle). This latter case is almost deterministic in that the conditional distribution of Y given X puts equal mass on only two points: $-\sqrt{1-X^2}$

and $+\sqrt{1-X^2}$. Thus the conditional moments are easily computed once again, leading to polynomials in X once more:

$$E[Y^{2m}|X] = (1-X^2)^m.$$

The eigentransforms are the orthogonal polynomials with regard to the marginal distribution of X which has density $(1/\pi)(1-x^2)^{-1/2}$; i.e., they are the Chebyshev polynomials of the first kind (A & S, 22.2.4). The singular values are $\lambda_{2m} = (-1)^m$, that is, only three different values, each with infinite multiplicity:

$$\lambda_{2m+1} = 0, \qquad \lambda_{4m} = +1, \qquad \lambda_{4m+2} = -1.$$

This reflects the trivial fact that the deterministic relation $Y^2 = 1 - X^2$ entails an infinity of derived relations $Y^{2m} = (1-X^2)^m$, all being deterministic, hence adding to the multiplicities of the singular values ±1. Deterministic features in a distribution which can be represented by a relation of the form $g(Y) = f(X)$ in general lead to infinite-dimensional eigenspaces for the extremal correlations ±1 due to the possibility of transforming the relation in trivial ways such as $h(g(Y)) = h(f(X))$. This effect, too, is not merely of academic interest: Real data can come close to being deterministic when the optimal correlation is close to 1. ALS–ACE may then be ill determined, and diagnostics are required in the form of additional values $\lambda_2, \lambda_3, \dots$ in order to detect the problem.

Near multiplicity in near-deterministic bivariate distributions also provides an explanation for the behavior of ALS–ACE on the phone-call data presented by Fowlkes and Kettenring (1985). Although these data have two predictors, they behave more like a one-predictor case as one of them has very low explanatory power. See the discussion in Breiman and Friedman's rejoinder (1985b). Only in the presence of more than one strong predictor does a near-deterministic dependence not necessarily result in a proliferation of derived dependences. This is why data with high optimal correlations and only one essential predictor represent situations which are most vulnerable to ill-determined ALS–ACE transformations.

Parabolic optimal transformations: A characterization theorem by McGraw and Wagner. It is a coincidence that some very strong results relating to parabolic transformations are implicit in the engineering literature in a paper by McGraw and Wagner (1968). These authors dealt with elliptically symmetric distributions in the theory of noise and nonlinear devices where they showed in effect that any circular second-order distribution with quadratic eigentransforms belongs to one of four families of distributions. It should be noted, however, that this theorem does not yet prove that the quadratic transformations are the optimal ones. For each of the four families, one has to check optimality of the quadratic eigenvalue separately. We will perform some of these checks below in Sections 9 and 12.

THEOREM 8.1. *Any circular second-order distribution which has quadratic eigentransforms is one of the following (if suitably scaled):*

(i) *a bivariate standard normal distribution, for* $\lambda_2 = 0$;
(ii) *a degenerate uniform distribution on the unit circle, for* $\lambda_2 = -1$;
(iii) *a member of the bivariate Pearson Type II family for* $-1 < \lambda_2 < 0$:

$$q_{X,Y}(x,y) \propto (1 - x^2 - y^2)^{a-1}, \qquad a > 0,$$

for $x^2 + y^2 < 1$, *and* 0 *otherwise*;
(iv) *a member of the bivariate Pearson Type* VII *family for* $0 < \lambda_2 < 1$:

$$q_{X,Y}(x,y) \propto \frac{1}{(1 + x^2 + y^2)^{a+1}}, \qquad a > 1.$$

The Pearson Type II family is dealt with in Sections 9 and 10 and the Pearson Type VII family in Section 12. The latter is essentially a set of bivariate t distributions with continuous degrees of freedom (df $= 2a$). An important aspect of Theorem 8.1 is that moments higher than the second do not have to exist, and that Brown's criterion [Proposition 7.1(i)] of invariance of polynomial spaces under conditional expectations is assumed only for order 2. This makes it possible to include t distributions in the characterization, even though they have only finitely many existing moments.

The proof of Theorem 8.1 is based on the fact that circular symmetry links the marginal characteristic function $\zeta(t) = E[\exp iXt]$ and the bivariate characteristic function by $E[\exp i(Xs + Yt)] = \zeta(r)$, where $r = (x^2 + y^2)^{1/2}$. One can then reexpress the eigenequation $E[(Y^2 - \sigma^2)|X] = \lambda_2(X^2 - \sigma^2)$ in terms of a second-order differential equation in $\zeta(t)$ which characterizes exactly the four families of distributions listed in Theorem 8.1.

The eigenvalues λ_2 in Theorem 8.1 indicate that the tail weight determines the sign of the singular value: Lighter than normal tails (Type II) give negative values, while heavier than normal tails (Type VII) lead to positive ones. The way optimizing algorithms for sample versions of ALS–ACE work, negative optimal correlations are impossible since the sign gets pushed over to one of the transforms. Thus, in sample ALS–ACE, positive singular values are indicated if the parabolic transformations have the same orientation (Figure 5), and opposite orientation if the singular values are negative (Figure 3).

It is remarkable that the Pearson Type VII family allows λ_2 and hence maximal correlation arbitrarily close to $+1$. Strangely, one can achieve eigenvalues $\lambda_2 > 1$ by removing the second-order restriction. That this is not contradictory will be shown in Section 12.

9. Circular Pearson Type II distributions and their limiting cases.

The circular Pearson Type II distributions [Johnson and Kotz (1972), Chapter 42, Table 3] form a parametrized family which connects the circular bivariate normal (independence) with the degenerate uniform distribution on the circle (deterministic dependence), while the uniform on the disk is an intermediate case between these extremes. All distributions of this type are polynomially biorthogonal, and parabolas are the optimal transformations for all members of this family. We consider for simplicity only those members which are concentrated on the unit disk $x^2 + y^2 < 1$ with a density given by

$$q_{X,Y}(x,y) = \frac{1}{\pi B(1,a)}(1 - x^2 - y^2)_+^{a-1}, \qquad a > 0,$$

where we use $(\cdots)_+^{a-1}$ short for $(\cdots)^{a-1}$ if the value in parentheses is positive, and 0 otherwise. [$B(a,b)$ is again the beta function; see A & S, 6.2]. We will abbreviate these distributions by the symbol $\mathrm{II}_2(a)$, where the subscript 2 indicates a bivariate situation. In the engineering literature, McFadden (1966) analyzed these distributions from the point of view of series expansions. Some examples and limiting cases are the following:

1. the uniform on the unit disk, obtained for $a = 1$;
2. the degenerate uniform on the unit circle, which does not have a density but is obtained as the limit for $a \to 0$;
3. the bivariate standard normal distribution, which is the limit for $a \to \infty$ after suitable scaling, e.g., $\sqrt{2a}\,\mathrm{II}_2(a) \to$ *bivariate standard normal* in the sense of convergence of densities:

$$\frac{1}{2a\pi B(1,a)}\left(1 - \frac{x^2 + y^2}{2a}\right)_+^{a-1} \xrightarrow[a\to\infty]{} \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x^2 + y^2)\right).$$

For the marginal and conditional distributions we introduce the univariate Pearson Type II distributions, denoted by $\mathrm{II}_1(a)$:

$$q(x) = \frac{1}{B(\frac{1}{2},a)}(1 - x^2)_+^{a-1}, \qquad a > 0.$$

The moments of $\mathrm{II}_1(a)$ vanish for odd orders while for even order $m = 2k$ they are $B(k + \frac{1}{2}, a)/B(\frac{1}{2}, a)$. The systems of complete orthogonal polynomials are of the ultraspherical or Gegenbauer type (see A & S, 22.2.3; A & S use the parameter $\alpha = a + \frac{1}{2}$).

The marginal distribution of the bivariate $\mathrm{II}_2(a)$ is a univariate $\mathrm{II}_1(a + \frac{1}{2})$. However, not all univariate $\mathrm{II}_1(a)$ distributions are marginals of bivariate II_2 distributions: This is the case only for $a \geq \frac{1}{2}$, where $a = \frac{1}{2}$ stems from the uniform on the unit disk and $a > \frac{1}{2}$ from the bivariate II_2 distributions. The conditional distribution of Y given X is $\sqrt{1 - X^2}\,\mathrm{II}_1(a)$. Polynomial biorthogonality follows immediately according to Proposition 7.1: Odd order conditional moments vanish, while

$$E\left[Y^{2k}|X\right] = (1 - X^2)^k E\left[\mathrm{II}_1(a)^{2k}\right] = \frac{B(k + \frac{1}{2}, a)}{B(\frac{1}{2}, a)}(1 - X^2)^k$$

is a polynomial of order $2k$. We summarize:

PROPOSITION 9.1. *The bivariate circular* $\mathrm{II}_2(a)$ *distribution is polynomially biorthogonal. The eigentransforms are the Gegenbauer polynomials of order* a, *and the singular values are*

$$\lambda_{2k+1}^{(a)} = 0 \quad \textit{and} \quad \lambda_{2k}^{(a)} = \frac{B(k+\frac{1}{2}, a)}{B(\frac{1}{2}, a)}(-1)^k.$$

Using some standard formulas for the beta function (A & S, 6.2.2), one can rewrite the nontrivial singular values as

$$\lambda_{2k}^{(a)} = \frac{\Gamma(k+\frac{1}{2})\Gamma(a+\frac{1}{2})}{\Gamma(a+k+\frac{1}{2})\Gamma(\frac{1}{2})}(-1)^k,$$

which allows us to analyze their qualitative behavior as follows.

Proposition 9.2. *The absolute singular values* $|\lambda_{2k}^{(a)}|$ *are*

(i) *strictly monotone decreasing as functions of a for fixed* $k > 0$, *and* $|\lambda_{2k}^{(a)}| \downarrow 0$ *for* $a \to \infty$,

(ii) *strictly monotone decreasing as functions of k for fixed* $a > 0$, *and* $|\lambda_{2k}^{(a)}| \downarrow 0$ *for* $k \to \infty$.

The proof of monotonicity is probably standard, but a simple method would use logarithms and derivatives, thus deferring the problem to the Digamma function (A & S, 6.3) which is known to be monotone increasing on the positive reals. The limit at ∞ can be obtained using Stirling's formula. □

From Proposition 9.2 follow some sensible conclusions concerning the behavior of ALS–ACE:

Corollary 9.3. (i) *For* $a > 0$, *the optimal eigentransforms are parabolas.*

(ii) *For* $a \to 0$, *we obtain the singular values of the degenerate uniform on the unit circle, and the Gegenbauer polynomials specialize to the Chebyshev polynomials of the first kind.*

(iii) *For* $a \to \infty$, *we approach the singular values of the bivariate standard normal, and, if suitably scaled, the Gegenbauer polynomials converge to the Hermite polynomials.*

The first statement follows from Proposition 9.2(ii) and the second from Proposition 9.1. For the third, one uses the fact that independence is equivalent to $\lambda_m = 0$ for $m > 0$, so it follows from Proposition 9.2(i). The remarks regarding the eigenpolynomials are lifted from A & S (22.2.3, 22.2.4, 22.15.6). □

10. Elliptic bivariate distributions with polynomial eigentransforms.

This section presents theorems for polynomially biorthogonal elliptic distributions. The two main results which we reconstruct here with ALS–ACE in mind are (1) that polynomial biorthogonality carries over from circular to elliptic distributions, and (2) another characterization theorem [due to Davison and Grunbaum (1981)], which points out once more the unique role played by the bivariate Pearson Type II distributions and their limiting cases. Davison and Grunbaum obtained their result in the context of computer-aided tomography independently but much later than McGraw and Wagner (1968). We will show how to reduce the Davison–Grunbaum theorem to the one by McGraw and Wagner.

Elliptic distributions may be conveniently generated from circular distributions by suitable linear transformations. For instance, if we assume a circular distribution for (X', Y'), we may define

$$(*) \qquad X = X' \quad \text{and} \quad Y = \rho X' + \sqrt{1-\rho^2}\, Y',$$

which yields an elliptic distribution for (X, Y) with (raw) correlation ρ. For later (Section 12) we observe that ρ as a measure of ellipticity makes sense even if X and Y do not have second moments. For elliptic distributions, the identity transforms are always eigentransforms with singular value $\lambda = \rho$.

In what follows, it is important to remember that the marginal distributions of X' $(= X)$, Y' and Y are all the same because of the assumed circular symmetry in the law of (X', Y').

PROPOSITION 10.1. (i) *If the circular variables (X', Y') have conditional moments up to degree m, and if the polynomial subspaces up to degree m in X' and Y', respectively, are invariant under conditional expectations, then the same holds for all derived elliptic variables (X, Y) generated by $(*)$, regardless of the ellipticity ρ.*

(ii) *In particular, if the circular variables (X', Y') are polynomially biorthogonal, so are the elliptic variables (X, Y).*

For a proof, calculate once more the conditional moments for the variables X and Y, observing that for the circular (X', Y') the conditional expectation of odd-degree monomials vanishes:

$$\begin{aligned} E[Y^m|X] &= E\left[\left(\rho X' + \sqrt{1-\rho^2}\, Y'\right)^m \middle| X'\right] \\ &= \sum_{k=0}^{m} \binom{m}{k} \rho^{m-k} \sqrt{1-\rho^2}^{\,k} X'^{m-k} E\left[Y'^k|X'\right] \\ &= \sum_{k=0}^{[m/2]} \binom{m}{2k} \rho^{m-2k} (1-\rho^2)^k X'^{m-2k} E\left[Y'^{2k}|X'\right]. \end{aligned}$$

This is a polynomial of degree m in X' since $E[Y'^{2k}|X']$ is a polynomial of degree $2k$. □

PROPOSITION 10.2. *In an elliptic polynomially biorthogonal family of distributions, the eigenpolynomials are the same, regardless of the ellipticity ρ.*

The singular values are functions of ρ:

$$\lambda_m(\rho) = \sum_{k=0}^{[m/2]} \binom{m}{2k} \rho^{m-2k} (1-\rho^2)^k \lambda_{2k}(0).$$

Thus, we obtain the following systems of polynomials as simultaneous eigentransforms of elliptic distributions:

1. the Hermite polynomials for the standard normal (A & S, 22.215),
2. the Gegenbauer polynomials of suitable order for Pearson Type II (A & S, 22.2.3),
3. the Chebyshev polynomials of the first kind for the degenerate uniform on the unit circle and its elliptic siblings (A & S, 22.2.4).

For the Pearson Type VII family, no complete system of polynomials exists, but finitely many orthogonal eigenpolynomials exist for degrees m such that X^{2m} is integrable. In order to derive this, one could have formulated Proposition 10.2 more carefully similar to part (i) of Proposition 10.1 as follows: If polynomial subspaces up to degree m are invariant, and if moments up to degree $2m$ exist, then there exist orthogonal polynomials up to degree m which are eigentransforms for all ellipticities. One has to ask for moments up to degree $2m$ to ensure that the usual scalar product is defined for polynomials up to degree m.

PROOF OF PROPOSITION 10.2. The eigentransforms are the same for all ellipticities ρ because they are the orthogonal polynomials w.r.t. the marginal distributions of X and Y, which are the same for all ρ. The eigenvalues can be gotten from the proof of Proposition 10.1 as the leading coefficients in the polynomials $E[Y^m|X]$. To see this, use Proposition 7.3 again and recall that the leading coefficient is the $2k$th eigenvalue $\lambda_{2k}(0)$ for the circular case $\rho = 0$. □

The following is Davison and Grunbaum's (1981) characterization theorem which shows that ellipticity together with shared eigentransforms across all correlations ρ is so strong a condition that it characterizes the II_2 family among distributions on the unit disk.

THEOREM 10.3. *Make the following assumptions for a circular distribution*:

(i) *Its support is contained in the closed unit disk, but not in any smaller circular disk.*

(ii) *All the elliptic distributions derived from it have a common set of eigentransforms independent of the ellipticity* ρ, *and these eigentransforms are continuous on* $[-1, +1]$.

Then this is either a $\text{II}_2(a)$ *distribution for some* a, *or the degenerate uniform distribution on the unit circle.*

We can reduce this theorem to that of McGraw and Wagner by showing that a quadratic eigenfunction exists. To this end, we borrow an idea of Davison and Grunbaum (1981), page 90, for which we give a proof in probabilistic language.

LEMMA 10.4. *Under the assumptions of Theorem* 10.3, *we have for the singular values*:

$$\lambda_m(\rho) = \frac{\phi_m(\rho)}{\phi_m(1)}, \quad \textit{where } \phi_m(1) \neq 0.$$

PROOF. Evaluate the conditional expectation of the ϕ_m transform of $Y = \rho X' + \sqrt{1-\rho^2}\,Y'$ given X' $(= X)$ at $X' = 1$. This is one of the two points where dependence between X' and Y' is at its extreme: Given $X' = 1$, we know that $Y' = 0$. To make the argument rigorous, one would have to consider limits, whence some technical assumptions like the presence of mass arbitrarily close to the boundary of the unit disk, and continuity of the eigenfunctions. We dispense with the details and simply write:

$$E^{X'=1}\phi_m\left(\rho X' + \sqrt{1-\rho^2}\,Y'\right) = \lambda_m(\rho)\phi_m(1) = \phi_m(\rho)$$

The first equality follows from the eigenproperty of ϕ_m and the second from the dependence $Y' = 0$ at $X' = 1$. Finally, we note that $\phi_m(1) \neq 0$ since otherwise $\phi_m(\rho) = \lambda_m(\rho)\phi_m(1)$ would vanish globally, which is impossible for proper eigenfunctions. □

PROOF OF THEOREM 10.3. Since the eigenfunctions form a complete system, there must exist ϕ_m for which $\langle \phi_m(X), X^2\rangle \neq 0$. We wish to show that $\phi_m(X)$ is a parabola by showing that $\lambda_m(\rho)$ is a quadratic function of ρ. Lemma 10.4 together with Theorem 8.1 then yields the conclusion of Theorem 10.3 since among the possibilities listed in Theorem 8.1 only the Pearson Type II distributions and the degenerate uniform on the unit circle satisfy the assumption regarding support on the unit disk.

To show that the singular values are quadratic in ρ, we use

$$\langle \phi_m\left(\rho X' + \sqrt{1-\rho^2}\,Y'\right), X'^2\rangle = \lambda_m(\rho)\langle \phi_m(X'), X'^2\rangle,$$

which is a simple consequence of the eigenequation for ϕ_m. Since the distribution of $(\rho X' + \sqrt{1-\rho^2}\,Y', X')$ is symmetric, we get

$$\left\langle \phi_m(X'), \left(\rho X' + \sqrt{1-\rho^2}\,Y'\right)^2 \right\rangle = \lambda_m(\rho)\langle \phi_m(X'), X'^2\rangle.$$

Expanding the left-hand side, we can drop the cross-product term due to $E[Y'|X'] = 0$, and solve the equation for $\lambda_m(\rho)$:

$$\lambda_m(\rho) = \rho^2 + (1-\rho^2)\frac{\langle \phi_m(X'), Y'^2\rangle}{\langle \phi_m(X'), X'^2\rangle}. \qquad □$$

The invariance property of Proposition 10.1(i) is pervasive in the Donoho–Johnstone (1988) theory of projection pursuit regression (PPR) and Davison and Grunbaum's work on tomographic reconstruction (1981). The special case of the uniform distribution on the unit disk dates back to Logan and Shepp (1975) and Hamaker and Solmon (1978). The importance of the normal and Pearson Type II distributions as weight functions in tomography stems from the following fact: The least squares problem of reconstructing functions from projections can be block-diagonalized in terms of the Hermite and Gegenbauer polynomials, and thus reduced to subproblems of lower dimensionality. This derives from the simultaneous eigenproperties of these polynomials across all ellipticities.

In the theory of noise and nonlinear devices, eigenproperties of bivariate distributions have an inherent interest because they allow one to analytically investigate the covariance functions of nonlinear transforms of certain stochastic processes [see, e.g., Barrett and Lampard (1955), Nuttall (1958), Brown (1958), McGraw and Wagner (1968) Cambanis and Liu (1971) and the references given in these papers]. Nonlinear transforms of stochastic processes are interpreted as random signals which were subjected to nonlinear, instantaneous, memoryless devices.

11. Discussion of some examples of elliptic distributions. The intention in this section is to investigate the dependence of the eigentransforms and eigenvalues on the ellipticity ρ for the simplest elliptic distributions and, as one of the main points, we wish to demonstrate the phenomenon of discontinuity of ALS–ACE transforms which was described in Buja and Kass (1985). We consider (1) elliptic normal distributions, (2) uniform distributions on elliptic disks and (3) degenerate distributions on the periphery of elliptic disks. Cases (1) and (2) would probably be judged trivial for the purposes of data analysis, as linear fits seem to summarize the dependence between X and Y well and no nonlinear transformation can help in any fruitful way. Case (3), again, is different in that a nontrivial implicit equation is satisfied with probability 1, but this time we might expect it to represent a true interaction in general; i.e., the implicit equation is no longer of the form $g(x) + f(y) = 0$.

Elliptic normal distributions. The circular case amounts to independence which is always polynomially biorthogonal in a trivial sense due to $\lambda_m(0) = 0$ for $m > 0$. To determine the singular values as functions of ρ, we specialize Proposition 10.2 making use of the vanishing singular values for $\rho = 0$:

$$\lambda_m(\rho) = \rho^m,$$

which is well-known, see, e.g., Lancaster (1975), equation 4.6. The behavior of the singular values follows our expectations: As the degree increases, they decrease to 0 quite rapidly, and as the raw correlation approaches 1, the correlation of all eigenpolynomials approaches 1. In particular, the linear transforms or, equivalently, the raw data are optimal for all ρ, which is a special case of a theorem by Kolmogorov [see Lancaster (1975), Section 10.4). For $\rho \uparrow 1$, however, the suboptimal singular values cluster around 1 as well:

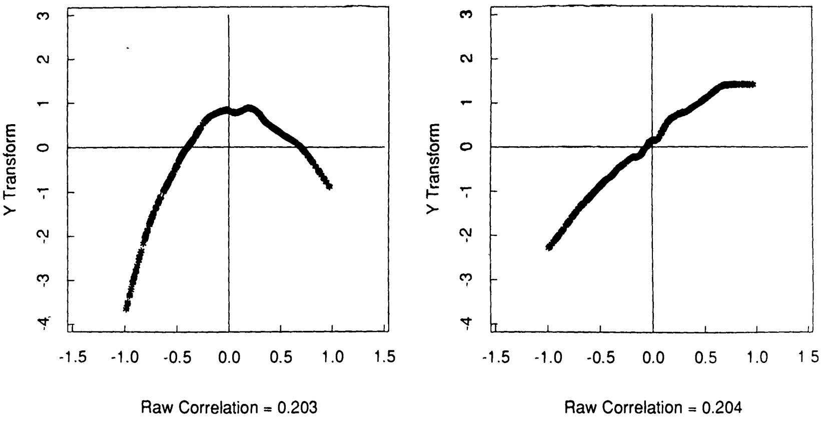

FIG. 4. *Switchover point between a linear and a quadratic optimal transform illustrated with y transforms of two nearby uniform distributions on elliptic disks. Left: approximate parabola. Right: approximate straight line.*

$\lambda_m(\rho) \uparrow 1$; i.e., we encounter badly determined eigentransformations for nearly deterministic data once again.

Uniform distributions on elliptic disks. We specialize Proposition 10.2 with $\lambda_{2m+1}(0) = 0$ and $\lambda_{2m}(0) = (-1)^m/(2m + 1)$ from Section 8. We carry this out for $m = 1, 2$ only:

$$\lambda_1(\rho) = \rho, \qquad \lambda_2(\rho) = -\tfrac{1}{3} + \tfrac{4}{3}\rho^2.$$

The optimal transform is the second-degree polynomial for ρ between 0 and $\frac{1}{4}$, and the identity transform for ρ between $\frac{1}{4}$ and 1 [Buja and Kass (1985), Section 2]. The transition from one to the other at $\rho = \frac{1}{4}$ is abrupt and represents a discontinuity in the behavior of ALS–ACE. In the low range of ρ, linear dependence is overwhelmed by a peculiar type of (generally uninteresting) circular dependence, while for larger ρ the more meaningful linear dependence takes over.

This behavior is real and reflects in finite-sample implementations. For illustration, we used a single sample from a uniform distribution on a circular disk to generate elliptic uniform samples for various ellipticities ρ. We then applied B & F's implementation of ACE. As one would expect, the locally linear smoother used by ACE favors linear transforms over quadratic ones, lowering the theoretically expected switchover point from $\frac{1}{4}$ to about 0.2035 in case of the data at hand. Figure 4 shows two Y transforms which are recognizable as approximately quadratic and linear, respectively, at ellipticities of 0.203 and 0.204, narrowing down the switchover point to an interval of size 10^{-3}.

Optimal and second suboptimal transforms similar in appearance to a straight line and a parabola, respectively, appear frequently in real quantitative data. In (possibly unordered) categorical data, however, a related effect is known to occur quite often as well: If scores of the optimal and second

second suboptimal transforms $\phi_1(X)$ and $\phi_2(X)$ are plotted against each other, one often encounters what is called in the psychometric literature a "horseshoe." If in the simplest case this can be interpreted as $\phi_2(X)$ being approximately a quadratic function of $\phi_1(X)$, this is taken as an indication that $\phi_1(X)$ recovers a good scaling or ordering of the categories of a contigency table. Schriever (1983) justified this reasoning and related the horseshoe effect to order and total positivity properties of discrete distributions. With the above example of switchover from linear to quadratic optimal transforms, however, we have an indication that the horseshoe effect can go in reverse if the optimal correlation is low enough. The oscillatory behavior of eigentransforms implied by the theory of total positivity may still appear, but the eigenvalues may not be ordered according to increasing number of oscillations.

Degenerate elliptic distributions derived from uniforms on the unit circle. The singular values are obtained by specializing Proposition 10.2 with $\lambda_{2m+1}(0) = 0$, $\lambda_{2m}(0) = (-1)^m$ from Section 8:

$$\lambda_m(\rho) = \sum_{k=0}^{[m/2]} \binom{m}{2k} \rho^{m-2k}(1-\rho^2)^k(-1)^k = \operatorname{Re}\left(\rho + \sqrt{1-\rho^2}\, i\right)^m$$
$$= \operatorname{Re}(e^{i\pi m\alpha}) = \cos(\pi m \alpha),$$

where $\cos(\pi\alpha) = \rho$ and Re denotes the real part. As a consequence, whenever $m\alpha$ is an integer, we obtain $\lambda_m = \pm 1$, and a deterministic relation of the form either $\phi_m(X) = \phi_m(Y)$ or $\phi_m(X) = -\phi_m(Y)$ holds true (i.e., an algebraic equation with separated variables). Suitable m's exist for any rational α, and in this case there are only finitely many different singular values, each with infinite multiplicity. For irrational α's, the sequence $(m\alpha) \bmod 2$ fills the half-closed interval $[0, 2)$ densely (with a uniform distribution), but it does not take on the values 0 or 1. As a consequence the spectrum

$$\{\lambda_m = \cos(\pi m \alpha) | m = 1, 2, \dots\}$$

does not attain its supremum and infimum values ± 1, and the population ALS–ACE algorithm fails to converge. The fact that the spectrum has other values besides 0 as cluster points (α irrational), or has infinite-dimensional eigenspaces for eigenvalues other than 0 (α rational), shows that the conditional expectations are not compact operators in either case.

12. Bivariate *t* distributions: Eigenvalues greater than 1. We are going to address heavy-tailed situations again, but with a scenario which differs from that of Section 5. We consider essentially the family of bivariate Pearson Type VII distributions, which contain the *t* distributions as special cases. These allow us to generate heavy-tail behavior which could not be achieved with mixtures of normal distributions. The analytical results are very peculiar: We encounter *singular values greater than* 1, and even more, they can be *arbitrarily large*. For this to happen we have to break the usual rules

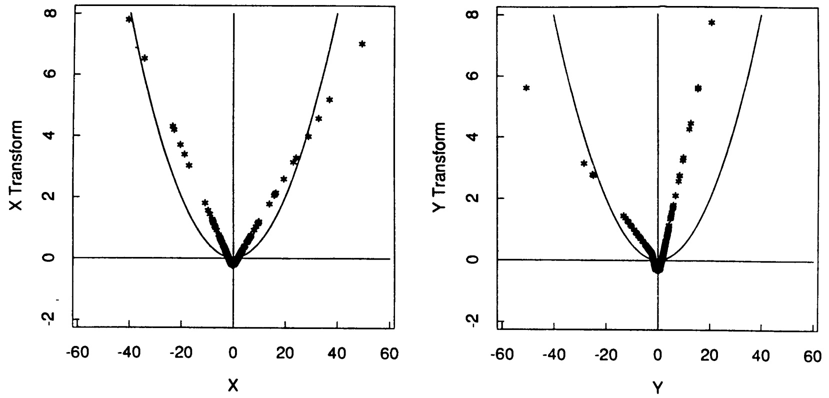

FIG. 5. *Quadratic x and y eigentransforms for a bivariate t distribution with* 1.5 *degrees of freedom.*

by abandoning marginal integrability, thus giving up both L_2 and L_1 geometry. At the root is a discrepancy between marginal and conditional integrability: Conditional expectations may still exist while marginal expectations do not.

The tail weight is the driving factor in the sense that the higher some singular values go, the heavier the tail has to be. Estimating singular values by means of correlations, as is done in finite samples, will be blind to this phenomenon. In fact, finite-sample smoothers in their simplest (non-cross-validated and nonrobustified) versions are typically shrinking linear mappings [Buja, Hastie and Tibshirani (1989)]. They are therefore unable to recover nonshrinking behavior of conditional expectations. There is no serious practical problem in this inability since samples from bivariate t distributions with low degrees of freedom generally exhibit "outliers" of such a serious nature that an analysis based on automated data transformations would be avoided on grounds of simple common sense. For illustration, we show in Figure 5 transformations obtained from a sample of (essentially) a bivariate t distribution with 1.5 degrees of freedom. The sporadic outliers determine the shape of the transforms and combine with artifacts of the locally linear smoothers to create two almost linear slopes with a sharp bend in the center. In contrast, the population eigenfunctions which come closest to these empirical transforms are parabolas once again, but the only shared qualitative feature is the presence of a valley-like shape. The reason for the presence of two opposite slopes is different in the Pearson Type II and Type VII distributions: The former have too much mass along the horizontal and vertical axes, whereas the latter have too much mass along the 45 degree lines. This is backed up below by an analysis of the "densities w.r.t. independence" (Section 3).

In order to parallel the analytics of the II_2 distributions (Section 9), we will work with a subset of the family of Pearson Type VII [Johnson and Kotz (1972), Chapter 42, Table 3; see also Chapter 37 for the multivariate t]. We consider the following bivariate densities:

$$q_{X,Y}(x,y) = \frac{a}{\pi} \frac{1}{\left(1 + x^2 + y^2\right)^{a+1}}, \qquad a > 0.$$

In analogy to the notation $\mathrm{II}_2(a)$, we write $\mathrm{VII}_2(a)$ as shorthand for a random two-vector with this density. Once again we obtain the bivariate standard normal as a limit of, e.g., $\sqrt{2a}\,\mathrm{VII}_2(a)$ as $a \to \infty$. The marginal densities are

$$q_X(t) = q_Y(t) = \frac{1}{B\left(\frac{1}{2}, a\right)} \frac{1}{\left(1 + t^2\right)^{a+1/2}}, \qquad a > 0.$$

Random variables with this law will be denoted $\mathrm{VII}_1(a)$. The $\mathrm{VII}_1(\frac{1}{2})$ distribution is just the univariate standard Cauchy, and the actual t distributions are obtained as $t_\nu = \sqrt{\nu}\,\mathrm{VII}_1(\nu/2)$, where ν is the degrees of freedom. The conditional distribution of Y given X for $\mathrm{VII}_2(a)$ is $\sqrt{1 + X^2}\,\mathrm{VII}_1(a + \frac{1}{2})$, which is reminiscent of the conditional distributions for the II_2 family.

In the following remarks, we list some curiosities in the simplest possible formulation for quadratic transforms.

Remark 12.1. The bivariate $\mathrm{VII}_2(a)$ distribution has a quadratic eigenfunction exactly for $\frac{1}{2} < a < 1$ and for $a > 1$.

This means that the Cauchy case ($a = \frac{1}{2}$) and the t distributions with 2 degrees of freedom ($a = 1$) do not have a quadratic eigenfunction, but anything in between or greater does have one.

Remark 12.2. The quadratic eigenfunctions are $\theta_2(t) = \phi_2(t) \propto t^2 - \frac{1}{2}(a-1)^{-1}$.

The "centering constant" $\frac{1}{2}(a-1)^{-1}$ is negative for $\frac{1}{2} < a < 1$. In contrast, a typical argument based on L_2 techniques and symmetry would force $\phi_2(X)$ to be orthogonal to constants, i.e., $\phi_2(X) \propto X^2 - E(X^2)$. Since $\phi_2(t) = t^2 - c$ with $c < 0$ for $\frac{1}{2} < a < 1$, the constant c cannot play the role of a marginal variance in this case.

Remark 12.3. The singular values of the quadratic eigenfunctions are $\lambda_2 = 1/(2a-1)$. Thus $\lambda_2 < 1$ for $a > 1$, but $\lambda_2 > 1$ for $\frac{1}{2} < a < 1$, and $\lambda_2 \uparrow \infty$ for $a \downarrow \frac{1}{2}$.

The value $+1$ is not taken on by λ_2. The cut point $+1$ separates the well-behaved case $\lambda_2 < 1$ (shrinking) from the pathological case $\lambda_2 > 1$ (exploding). The discontinuous behavior at $a = 1$ is explained as follows.

Remark 12.4. The two-dimensional space spanned by t^2 and 1 is invariant under conditional expectations for all $a > \frac{1}{2}$, but for $a = 1$, the action is described by $1 \to 1$ and $t^2 \to t^2 + 1$, which corresponds to a nondiagonalizable Jordan block: $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

This block has an eigenvalue $+1$ with algebraic multiplicity 2 and geometric multiplicity 1. The problem is that for $a = 1$, the eigenvalue λ_2 steps on the trivial eigenvalue $+1$ of the constant functions and ends up in degeneracy because of the nilpotent effect of the off-diagonal element.

Some background for these pathologies is given in Remark 12.5.

REMARK 12.5. The transformation X^2 is marginally square-integrable for $a > 2$, integrable for $a > 1$ and not integrable for $a \leq 1$. Yet, X^2 is conditionally integrable for all $a > \frac{1}{2}$.

The discrepancy between marginal and conditional integrability is behind the present problems: Conditional expectations may exist outside of L_1, but their familiar properties may be sacrificed. Within L_2, conditional expectations behave like symmetric shrinking mappings; in L_1, they are still shrinking, but outside any behavior is possible.

Some intuition into the properties of the VII_2 distributions may be gained by an examination of the "density w.r.t. independence" (Section 3). This density cannot be square-integrable, since otherwise L_2 theory would be applicable:

$$\iint f(x,y)^2 Q_X(dx) Q_Y(dy) \propto \iint \frac{(1+x^2)^{a+1/2}(1+y^2)^{a+1/2}}{(1+x^2+y^2)^{2a+2}}\, dx\, dy$$

$$\propto \iint \frac{(1 + r^2 + r^4 \cos^2\alpha \sin^2\alpha)^{a+1/2}}{(1+r^2)^{2a+2}} r\, dr\, d\alpha,$$

with a switch to polar coordinates ($x = r\cos\alpha$, $y = r \sin\alpha$). The integrand as a function of r is of the order $1/r$ for large values of r, hence the integral over r is infinite for all but finitely many values of α. The role of the angle α in the integrand is of particular interest: It shows that the stochastic dependence between X and Y consists of high mass concentrations along the 45 degree diagonals and mass deficiency along the axes. This makes it plausible that a quadratic transformation is able to pick up positive correlation (if defined) or at least to generate positive eigenvalues by mapping the high mass concentrations onto the 45 degree diagonal in the positive quadrant. A similar analysis for the II_2 family would show that just the opposite is true there: Mass concentrations along the coordinate axes permit the quadratic transformations to fold the mass onto a triangle in the positive quadrant with a resulting negative correlation.

In the following proposition we draw a more complete picture of the eigenpolynomials of VII_2 as far as they exist. However, the proof is too arcane to be reproduced here.

Proposition 12.6. *For the* $\mathrm{VII}_2(a)$ *distributions, polynomials in* X *and* Y *of degree* m *are marginally square integrable for* $m < a$, *marginally integrable for* $m < 2a$ *and conditionally integrable for* $m < 2a + 1$. *There exists exactly one degree which is conditionally but not marginally integrable.*

Orthogonal eigenpolynomials $\theta_m(t) = \phi_m(t)$ *exist for square-integrable degrees, i.e.,* $m < a$. *Outside of* L_2, *there generally exists a series of eigenpolynomials for all conditionally integrable degrees, i.e.,* $m < 2a + 1$. *An exception occurs when* a *is an integer, in which case no even degree eigenpolynomial of degree* $m = 2k > a$ *exists.*

The singular values are

$$\lambda_{2k} = \frac{B\left(k + \frac{1}{2}, a + \frac{1}{2} - k\right)}{B\left(\frac{1}{2}, a + \frac{1}{2}\right)} \quad \textit{and} \quad \lambda_{2k+1} = 0.$$

The behavior of the even degree singular values can be described as follows:

Only those singular values λ_{2k} *with square-integrable eigenpolynomials* (*i.e.,* $2k < a$) *can be interpreted as correlations. They form a decreasing sequence for increasing degrees.*

The singular values λ_{2k} *for marginally integrable but not square-integrable degrees* (*i.e.,* $a \le 2k \le 2a$) *are no longer correlations. They form an increasing sequence which is bounded by* $+1$.

If the conditionally but not marginally integrable degree $(2a \le m < 2a + 1)$ *is even, its singular value* λ_m *exceeds* $+1$, *unless* $m = 2a$ *when it equals* $+1$.

One could continue with an examination of the elliptic VII_2 distributions but we found it not very informative. Finally, it should be mentioned that a more complete analysis of the circular VII_2 within L_2 is possible. It involves a partly discrete and partly continuous spectral decomposition, with results which are likely to be similar to those of Wong (1964), page 270. However, we do not expect that the continuous part of the spectrum will contribute to our understanding of the empirical behavior of ALS–ACE or the theoretical problem of exploding singular values outside L_1.

13. Conclusions. In spite of Breiman and Friedman's statement that the eigenproblems for the ALS–ACE integral operators appear most intractable, we have found two rich classes of tractable situations: distributions of finite rank and polynomially biorthogonal distributions. We summarize a few practical points from these examples.

1. Linear optimal transformations do not necessarily indicate a satisfactory error structure (Section 7).
2. Null situations may result in misleading nontrivial transformations, mostly in the form of parabolas and maximal correlations up to about 0.3 (Sections 8 and 12).
3. Crossing-phenomena of eigenvalues and more general degeneracies may lead to unstable ALS–ACE transformations (Section 11).

4. In high-correlation situations we encounter indeterminacies as well, although this danger subsides if more than one strong predictor is present (Section 8).
5. Clustering is often picked up by ALS–ACE before any other structure. This may be indicated by approximate step functions as optimal transforms (Section 5).
6. There is no unique way for ALS–ACE to respond to heavy tails in null situations. Sometimes, they are indicated by a tendency to finite asymptotes in the wings (Section 5), but more often they show in steep, vaguely parabolic transformations with equal signs (i.e., positive singular values, Section 12). Light tails may be indicated by parabolas with opposite signs (i.e., negative singular values, Section 8).

On the technical side, we have the following points:

1. ALS–ACE extracts the dominant term in the singular value decomposition of a bivariate distribution (Section 3).
2. The maximal squared correlation is the dominant term in the expansion of Pearson's χ^2 or ϕ^2 functional (Section 3).
3. Correspondence analysis is ALS–ACE applied to two-way contingency tables, extracting not only the dominant but also the subdominant terms (Section 3).
4. ALS–ACE analysis on finite-rank distributions, and especially finite mixtures of independent sources, amounts to a canonical correlation analysis on certain density ratios (Section 4).
5. *Circular* polynomially biorthogonal distributions lead to *elliptic* polynomially biorthogonal distributions, for any degree of ellipticity. In terms of CAT literature this is equivalent to the following statement: If the conditional expectation operators given two orthogonal directions can be simultaneously diagonalized by polynomial eigenfunctions, then the same holds true simultaneously for *all* directions (Section 10).
6. Results by McGraw and Wagner (1968) and Davison and Grunbaum (1981) which were obtained in an engineering context and the theory of tomographic reconstruction, respectively, yield characterization theorems which explain why and where parabolic eigentransforms are likely to appear (Sections 8 and 10).

The mathematical curiosities we found concern

1. the existence/nonexistence proof of equations $f(x) = g(y)$ for ellipses (Section 11), and
2. the investigation of conditional expectations outside L_2, which lead to examples of exploding (noncontracting) conditional expectations with eigenvalues larger than 1 (Section 12).

Acknowledgments. I am indebted to the following colleagues, without whom this paper would not have been possible: J. H. Friedman for giving me the privilege of being one of the first to learn about ACE soon after its inception; R. Kass as coauthor of a discussion paper [Buja and Kass (1985)] which led to the present work; I. Johnstone for introducing me to his and D. Donoho's work on PPR as well as to the literature on CAT; S. Cambanis and J. de Leeuw for many references, especially to the engineering literature on the theory of noise and nonlinear devices; the referees and associate editors whose comments were invaluable; D. Asimov, L. Breiman, J. Chambers, D. Donoho, J. A. McDonald, T. Hastie, J. Kettenring, C. Mallows, C. Morris, A. Owen, C. Stone, W. Stuetzle and R. Tibshirani for discussions on this subject; J. Wellner for insistent encouragement which contributed greatly to the completion of this work; and once again J. de Leeuw whom the author owes a special debt of gratitude for extensive and constructive criticism which led to a major reorientation of the paper and drastic improvements in scholarship.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. E. (1972). *Handbook of Mathematical Functions*. National Bureau of Standards.

BARRETT, J. F. and LAMPARD, D. G. (1955). An expansion for some second-order probability distributions and its application to noise problems. *Trans. IRE* **IT-1** 10–15.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–243.

BRADLEY, R. A., KATTI, S. K. and COONS, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika* **27** 355–374.

BREIMAN, L. and FRIEDMAN, J. H. (1985a). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BREIMAN, L. and FRIEDMAN, J. H. (1985b). Rejoinder [to Pregibon and Vardi (1985), Buja and Kass (1985) and Fowlkes and Kettenring (1985)]. *J. Amer. Statist. Assoc.* **80** 614–619.

BROWN, J. L., JR. (1958). A criterion for the diagonal expansion of a second-order probability density in orthogonal polynomials. *IRE Trans. Inform. Theory* **IT-4** 172ff.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–510.

BUJA, A. and KASS, R. E. (1985). Some observations on ACE methodology [discussion of Breiman and Friedman (1985)]. *J. Amer. Statist. Assoc.* **80** 602–607.

CAMBANIS, S. and LIU, B. (1971). On the expansion of a bivariate distribution and its relationship to the output of a nonlinearity. *IEEE Trans. Inform. Theory* **17** 17–25.

CHESSON, P. L. (1976). The canonical decomposition of bivariate distributions. *J. Multivariate Analysis* **6** 526–537.

CSAKI, P. and FISCHER, J. (1960). On bivariate stochastic connection. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **5** 311–323.

DAUXOIS, J. and POUSSE, A. (1975). Une extension de l'analyse canonique. Quelques applications. *Ann. Inst. H. Poincaré Sect. B* (*N.S.*) **11** 355–379.

DAVISON, M. E. and Grunbaum, F. A. (1981). Tomographic reconstruction with arbitrary directions. *Comm. Pure Appl. Math.* **34** 77–119.

DE LEEUW, J., VAN RIJCKEVORSEL, J. and VAN DER WOUDEN, H. (1981). Nonlinear principal components analysis with *B*-splines. *Meth. Oper. Res.* **43** 379–393.

DE LEEUW, J. and VAN DER BURG, E. (1986). The permutational limit distribution of generalized canonical correlations. In *Data Analysis and Informatics* (E. Diday, Y. Escoufier, L. Lebart, J. P. Pagès, Y. Schektman and R. Tomassone, eds.) **4** 509–521. Elsevier, New York.

DONOHO, D. L. and JOHNSTONE, I. M. (1986). Regression approximation using projections and isotropic kernels. In *Function Estimates* (J. S. Marron, ed.). *Contemp. Math.* **59** 153–167. Amer. Math. Soc., Providence.

DONOHO, D. L. and JOHNSTONE, I. M. (1988). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106.

EAGLESON, G. K. (1964). Polynomial expansions of bivariate distributions. *Ann. Math. Statist.* **35** 1208–1215.

EAGLESON, G. K. (1969). A characterization theorem for positive definite sequences on Krawtchouk polynomials. *Austral. J. Statist.* **11** 29–38.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

FISHER, R. A. (1970). *Statistical Methods for Research Workers*, 14th ed. Hafner, New York.

FOWLKES, E. B. and KETTENRING, J. R. (1985). The ACE method of optimal transformation [discussion of Breiman and Friedman (1985)]. *J. Amer. Statist. Assoc.* **80** 607–613.

GIFI, A. (1981). Nonlinear multivariate analysis. Dept. Data Theory, Univ. Leiden, The Netherlands.

GILULA, Z. and HABERMAN, S. J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.* **83** 760–771.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic, London.

GRIFFITHS, R. C. (1969). The canonical correlation coefficients of bivariate gamma distributions. *Ann. Math. Statist.* **40** 1401–1408.

HAMAKER, C. and SOLMON, D. C. (1978). The angles between the null spaces of X-rays. *J. Math. Anal. Appl.* **62** 1–23.

HANNAN, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *J. Austral. Math. Soc.* **2** 229–242.

JOHNSON, N. L. and KOTZ, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

JORGENS, K. (1982). *Linear Integral Operators*. Pitman, Boston.

KENDALL, M. and STUART, A. (1979). *The Advanced Theory of Statistics* **2**, 4th ed. Oxford Univ. Press, New York.

KRUSKAL, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J. Roy. Statist. Soc. Ser. B* **27** 251–263.

LANCASTER, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* **29** 719–736.

LANCASTER, H. O. (1969). *The Chi-Squared Distribution*. Wiley, New York.

LANCASTER, H. O. (1975). Joint probability distributions in the Meixner classes. *J. Roy. Statist. Soc. Ser. B* **37** 434–443.

LANCASTER, H. O. (1980). Orthogonal models for contingency tables. In *Developments in Statistics* (P. R. Krishnaiah, ed.) **3**. Academic, New York.

LANCASTER, H. O. (1983). Special joint distributions of Meixner variables. *Austral. J. Statist.* **25** 298–309.

LEIPNIK, R. (1959). Integral equations, biorthonormal expansions, and noise. *J. Soc. Indust. Appl. Math.* **7** 6–30.

LEBART, L., MORINEAU, A. and WARWICK, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley, New York.

LEE, P. A. (1971). A diagonal expansion for the 2-variate Dirichlet probability density function. *SIAM J. Appl. Math.* **21** 155–165.

LOGAN, B. F. and SHEPP, L. (1975). Optimal reconstruction of a function from its projection. *Duke Math. J.* **42** 645–659.

MCFADDEN, J. A. (1966). A diagonal expansion in Gegenbauer polynomials for a class of second-order probability densities. *SIAM J. Appl. Math.* **14** 1433–1436.

MCGRAW, D. K. and WAGNER, J. G. (1968). Elliptically symmetric distributions. *IEEE Trans. Inform. Theory* **14** 110–120.

NAOURI, J. C. (1970). Analyse factorielle des correspondances continue. *Publ. Inst. Statist. Univ. Paris* **14** 1–100.

NAYLOR, A. W. and SNELL, G. R. (1982). *Linear Operator Theory in Engineering and Science.* Springer, New York.

NISHISATO, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications.* Univ. Toronto Press, Toronto.

NUTTALL, A. H. (1958). Theory and application of the separable class of random processes. Technical Report 343, Research Laboratory of Electronics, M.I.T.

PREGIBON, D. and VARDI, Y. (1985). Comment [discussion of Breiman and Friedman (1985)]. *J. Amer. Statist. Assoc.* **80** 598–601.

RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar.* **10** 441–451.

SARMANOV, O. V. (1958). The maximal correlation coefficient (symmetric case). Translated in *Selected Translations in Mathematical Statistics and Probability Theory* **4** 271–275. Amer. Math. Soc., Providence.

SARMANOV, O. V. (1963). Investigation of stationary Markov processes by the method of eigenfunction expansion. Translated in *Selected Translations in Mathematical Statistics and Probability Theory* **4** 245–269. Amer. Math. Soc., Providence.

SCHRIEVER, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. *Intern. Statist. Rev.* **51** 225–238.

SCHMIDT, E. (1907). Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. *Math. Ann.* **63** 433–476.

VAN DER BURG, E. and DE LEEUW, J. (1983). Nonlinear canonical correlation. *British J. Math. Statist. Psychol.* **36** 54–80.

VAN RIJCKEVORSEL, J. (1982). Canonical analysis with *B*-splines. In *COMPSTAT 1982* 393–398. International Association for Statistical Computing, Physica Verlag, Vienna.

VAN RIJCKEVORSEL, J. and DE LEEUW, J., eds. (1988). *Component and Correspondence Analysis.* Wiley, New York.

WONG, E. (1964). The construction of a class of stationary Markoff processes. In *Stochastic Processes in Mathematical Physics and Engineering* (R. Bellmann, ed.) 264–276. Amer. Math. Soc., Providence.

YOUNG, F. W., DE LEEUW, J. and TAKANE, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika* **41** 505–528.

YOUNG, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika* **46** 357–388.

BELLCORE
445 SOUTH STREET
BOX 1910
MORRISTOWN, NEW JERSEY 07962-1910