

SOME ASPECTS OF NON-STANDARD MULTIVARIATE ANALYSIS

By

H. S. Battey

Department of Mathematics, Imperial College London, SW7 2AZ, UK

and

D. R. Cox

Nuffield College, University of Oxford OX1 1NF, UK

SUMMARY

A broad review is given of some areas of multivariate analysis that are not frequently emphasized. We start with situations in which underlying distributions are far from multivariate normal form, so that standard methods of multivariate analysis based on covariances are likely to be unsatisfactory. We emphasize the important distinction between internal and external analyses associated with multiple outcomes. A second theme relates to multiple outcomes generated by time or spatial series in which long-range dependence operates. The implications are summarized.

Some key words: binary outcomes; non-normal multivariate dependencies; scaling.

1 Introduction

In many contexts it is appropriate in the various phases of an investigation to consider data as arranged in separate study units on each of which observations are available, possibly of several different kinds. Much of the extensive older literature concentrates implicitly largely on essentially linear relations, for example involving one or more multivariate Gaussian distributions and thus often characterized by correlation or covariance matrices or by functions of those. The issues involved may be divided broadly into internal and external ones, for example into studies of correlation contrasted with studies of regression.

The useful broad distinction, to be interpreted flexibly, between internal and external analyses is that in the former, but not the latter, component variables are in some sense all on an equal footing, it being required to find suitable summary variables that in some sense encapsulate most of the variation present, whereas in external formulations there are two or indeed several or many types of variable with interest focused on appropriate conditional dependences taken in appropriate sequence. Such chains of dependency aim at a deeper understanding of the data-generating process often implicitly or explicitly evolving in time.

By multivariate analysis, multiple outcomes are usually understood and we focus on that here. For methods of analysis largely emanating from the normal distribution and extending to distributions with ellipsoidal contours of equal density, we defer to excellent accounts such as [3] or [28]. Instead we focus on less standard aspects. We start by outlining some possibilities where any underlying distributions are far from multivariate normal form. In particular linear combinations of components have the potential to be unsatisfactory. Our emphasis here is on binary component variables. A second class of non-standard multivariate analysis problems concerns time or spatial series. In

some special contexts, for example when data defined over ever extending spatial or temporal ranges are involved, effective variances may decay more slowly than inversely as sample size, as found initially in agricultural trials [16], and turbulence ([22; 18]) and in hydrological studies of the River Nile [20]. The properties and implications of processes with such long-range dependence are reviewed in Section 4.

2 Multivariate binary outcomes

While the more common situation is univariate, involving a single binary outcome and several or perhaps many potential explanatory variables, here we discuss models and approaches to analysis for multivariate binary data, i.e. data with two or more binary outcomes. There may be other observable variables thought to influence these outcomes jointly, necessitating formulations analogous or similar to the logistic model for univariate binary responses; this is discussed in Section 2.3. We start with the simpler situation in which only outcome variables are observed. The primary goal is then to study the association between these outcomes rather than their dependence on others. This requires a model of their joint distribution, a compromise between achieving an adequate order of approximation to the unknown data generating mechanism and allowing useful interpretable conclusions.

The relative theoretical simplicity of multivariate Gaussian dependencies stems from the distribution being completely described by its mean and covariance matrix. A key result [35] is that the (i, j) th entry of the concentration (inverse covariance) matrix is zero if and only if variables i and j are independent conditional on all the other variables. It is natural to seek an analogue of this result for multivariate binary random variables.

One old approach advocated by Karl Pearson treats the binary variables as dichotomizations of unobserved and perhaps hypothetical continuous random variables. For example, with Z_i and Z_j obtained by dichotomization at zero of two normal random variables, each of mean zero, unit variance and correlation ρ_{ij} , the covariance of Z_i and Z_j is, by a formula of Sheppard [31],

$$\text{cov}(Z_i, Z_j) = \frac{2 \sin^{-1}(\rho_{ij})}{\pi}, \quad (1)$$

by the properties of the standard bivariate normal distribution function. We do not comment further on this approach, preferring instead, except when the latent variables have a tangible interpretation, to treat the binary random variables themselves as primary, leading in the simplest case of two binary outcomes to the following geometric representation.

2.1 Some geometry of models for two binary outcomes

The case of two binary outcomes can be depicted geometrically, similarly to the geometric representation of pure 2×2 contingency tables given by Fienberg and Gilbert [17]. For a detailed history of the analysis of 2×2 tables and an authoritative rebuttal of various sources of historical contention, see Yates [38]. Note also the strong conceptual distinction between the situation in which both variables are treated as outcomes and that in which only one is treated as an outcome, the other potentially explanatory.

For a pair of binary variables, Y_1, Y_2 , each coded as $\{0, 1\}$, the probabilities of the four outcomes are

$$\begin{array}{cc|c} \pi_{00} & \pi_{01} & \pi_{0.} \\ \pi_{10} & \pi_{11} & \pi_{1.} \\ \hline \pi_{.0} & \pi_{.1} & 1 \end{array}$$

where $\pi_{j.} = \pi_{j0} + \pi_{j1}$ are the marginal probabilities for variable 1, $\pi_{.k} = \pi_{0k} + \pi_{1k}$ are those for variable 2 and $\pi_{.0} + \pi_{.1} = \pi_{0.} + \pi_{1.} = 1$ by construction.

A convex basis for the space of all bivariate distributions of binary outcomes, written in the above tabular form, is

$$B_1 = \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 0 \\ \hline \end{array}, \quad B_2 = \begin{array}{|c|c|} \hline 0 & 1 \\ \hline 0 & 0 \\ \hline \end{array}, \quad B_3 = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array}, \quad B_4 = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & 1 \\ \hline \end{array},$$

so that any tabulized joint distribution t may be written as

$$t = \lambda_1 B_1 + \lambda_2 B_2 + \lambda_3 B_3 + \lambda_4 B_4,$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, barycentric coordinates for t , belong to the unit simplex in $\mathbb{R}^4$. The resulting tetrahedron, $\mathcal{S}_3$ say, is a three-dimensional object embedded in $\mathbb{R}^4$ whose four vertices correspond to $\lambda = (1, 0, 0, 0)$, $\lambda = (0, 1, 0, 0)$, $\lambda = (0, 0, 1, 0)$ and $\lambda = (0, 0, 0, 1)$.

The two binary variables associated with t are independent if and only if the probability of the joint event is equal to the product of the marginal probabilities, i.e. $\pi_{jk} = (\pi_{j0} + \pi_{j1})(\pi_{0k} + \pi_{1k})$ for all four combinations of j and k . This holds if and only if the cross-product ratio

$$\theta = (\pi_{00}\pi_{11})/(\pi_{01}\pi_{10}),$$

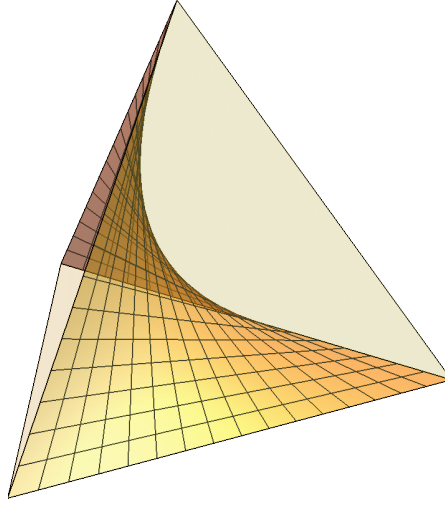
is equal to one (e.g. [10], p.393), an observation that forms the basis for testing independence in the pure 2×2 table. This constraint can be parameterized by two variation-independent parameters r and s , both constrained to the unit interval. Indeed $\theta = 1$ if and only if

$$t = \begin{array}{|c|c|} \hline rs & r(1-s) \\ \hline (1-r)s & (1-r)(1-s) \\ \hline \end{array}, \quad \text{or} \quad t = \begin{array}{|c|c|} \hline rs & (1-r)s \\ \hline r(1-s) & (1-r)(1-s) \\ \hline \end{array}.$$

Thus, for vectors $v_1, \dots, v_4$ of unit length, representing vertices of the unit three-simplex, the parametric surface

$$rsv_1 + r(1-s)v_2 + (1-r)sv_3 + (1-r)(1-s)v_4, \quad (r, s) \in [0, 1]^2, \quad (2)$$

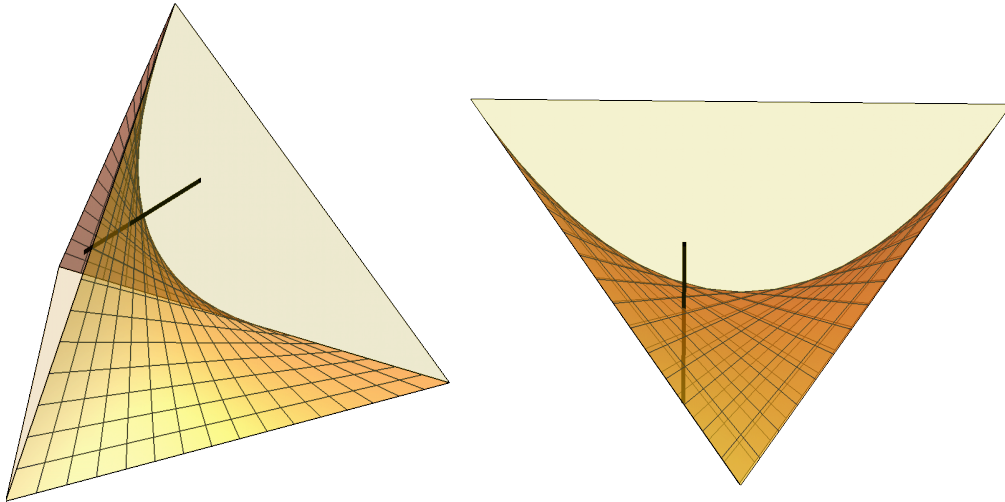
characterizes the set of all bivariate binary distributions with independent marginals. This was noted for pure 2×2 contingency tables by Fienberg and Gilbert [17]. Fig. 1 plots equation (2) together with $\mathcal{S}_3$, the convex hull of $\{v_1, \dots, v_4\}$. Note that the choice of coordinates for $v_1, \dots, v_4$ is immaterial except in the visual appearance of Fig. 1, a more regular triangular pyramid being preferred.



108

109 **Fig. 1.** The convex hull of the four vertices represents the space of all bivariate dis-
 110 tributions for binary outcomes. Those distributions with independent marginals are
 111 represented by the surface (2).

112 If the marginal distributions are considered known, this restricts the space of joint
 113 distributions to a one dimensional subspace of $\mathcal{S}_3$ (Fig. 2). Due to non-negativity of
 114 the probabilities, the most extreme distributions with marginals $\pi_{0\cdot}, \pi_{1\cdot}, \pi_{\cdot 0}, \pi_{\cdot 1}$ are de-
 115 termined by specifying, for each j and k in turn, $\pi_{jk} = \min\{\pi_{j\cdot}, \pi_{\cdot k}\}$. All other prob-
 116 abilities are subsequently determined by the marginal totals. Of the four distributions
 117 constructed in this way, two are the same. Thus there are two “most extreme” distribu-
 118 tions and all others with the same marginals are convex combinations of these, including
 119 as a special case, the one in which the two binary outcomes are independent. This is one
 120 way of interpreting why one degree of freedom is used in independence tests for 2×2
 121 tables rather than three.



122

123 **Fig. 2.** The straight line represents the one-dimensional subspace of $\mathcal{S}_3$ with marginal
 124 totals $\pi_{0\cdot} = 0.4, \pi_{1\cdot} = 0.6, \pi_{\cdot 0} = 0.2, \pi_{\cdot 1} = 0.8$. The surface of independence from Fig. 1
 125 is also depicted from two different vantage points.

2.2 Models for multivariate binary outcomes

Similarly, the most extensive model for p binary outcomes $Y_1, \dots, Y_p$ is a multinomial distribution, the $2^p - 1$ independent parameters corresponding to 2^p possible values of $Y_1, \dots, Y_p$. The case of independence is a sub-model of this, as in the simpler illustration of Section 2.1.

Logistic models are intermediate between the fully general multinomial representation and the independence model. Write $Z_i = 2Y_i - 1$ so that the binary variables are now coded as $\{-1, 1\}$. The log probabilities of the 2^p distinct observations can be modelled as

$$\log \text{pr}(Z_1 = z_1, \dots, Z_p = z_p) = \alpha_1 z_1 + \dots + \alpha_p z_p + \alpha_{12} z_1 z_2 + \dots + \alpha_{(p-1)p} z_{p-1} z_p + \dots - \Lambda,$$

where Λ is a normalizing constant ensuring the probabilities sum to one. Under this model, with $p = 2$,

$$\log \frac{\text{pr}(Z_1 = 1, Z_2 = 1) \text{pr}(Z_1 = -1, Z_2 = -1)}{\text{pr}(Z_1 = 1, Z_2 = -1) \text{pr}(Z_1 = -1, Z_2 = 1)} = \log \theta = 4\alpha_{12},$$

so that the independence model corresponds to $\alpha_{12} = 0$. For three variables, the constraint $\alpha_{123} = 0$ and $\alpha_{23} = 0$ gives a model in which Z_2 and Z_3 are independent given $Z_1 = z_1$, as can be seen from the calculation

$$\text{pr}(Z_2 = z_2, Z_3 = z_3 \mid Z_1 = z_1) = \frac{\exp\{\alpha_2 z_2 + \alpha_{12} z_1 z_2\} \exp\{\alpha_3 z_3 + \alpha_{13} z_1 z_3\}}{\sum_{z_2} \exp\{\alpha_2 z_2 + \alpha_{12} z_1 z_2\} \sum_{z_3} \exp\{\alpha_3 z_3 + \alpha_{13} z_1 z_3\}}$$

after multiplying and dividing by $e^{2\alpha_1 z_1}$.

More generally, if $\alpha_1, \dots, \alpha_p$ are the only non-zero parameters, the independence model is recovered, while if the all interactions up to order p are allowed, the resulting saturated model is in effect a reparameterization of the multinomial model. Useful sub-models are obtained by setting some parameters to zero or imposing other, ideally interpretable, constraints on the parameter space. For example, by setting all odd-order interaction parameters to zero, a subclass of the palindromic Bernoulli distributions is obtained [26]. The simplest models for dependence allow two-way but not higher-order interactions, leading to the so-called quadratic exponential binary distribution ([9; 12; 13]).

Without further restrictions there are $p(p+1)/2$ unknown parameters in the quadratic exponential binary distribution. Thus it is essentially equivalent in complexity to the unconstrained multivariate normal model and it is natural to seek further plausible restrictions, as in the multivariate normal situation. Equal correlation in the Gaussian setting is analogous to $\alpha_{ij} = \alpha$ for all i and j , while analogues of the class of multivariate totally positive distributions of order 2 [21], more commonly known as MTP_2 distributions, are obtained when $\alpha_{ij} \geq 0$ for all i and j . Cox and Wermuth [12] show that although the unconstrained quadratic exponential binary distribution does not retain its exact form under marginalization and conditioning, such closure can be approximately achieved under this MTP_2 restriction. This implies Simpson's paradox is not possible, to the order of approximation considered, under the $\alpha_{ij} \geq 0$ constraint. See also [23] for a recent discussion in a more general context.

2.3 External analysis of multivariate binary outcomes

We briefly extend this discussion of multivariate binary data to accommodate potential explanatory features. This is part of a common and in a sense richer situation that may be called external analysis, in general terms concerned with how one or more variables of interest depend on other specified variables, and what process or sequence of processes can be regarded as generating observed patterns of outcomes. In the simplest version of this there are just two types of variable, response and explanatory respectively. Often more realistically the variables can be divided into blocks, or arranged in time or space, relations within a block, being taken suitably conditionally, in particular given observations on “previous” blocks, and taken marginally over “subsequent” blocks. This encourages a multi-phased synthesis of the data-generating process and in suitable circumstances may point towards a deeper interpretation, sometimes even reasonably called causal.

In the notation of Section 2.2, suppose it is of interest to study how the joint distribution of $Y_1, \dots, Y_p$, or equivalently $Z_1, \dots, Z_p$, depends on a vector of covariates x . Although log-linear models [25] have frequently been advocated for this task, these make no explicit distinction between responses and covariates and consequently have a number of shortcomings [9]. McCullagh [27] proposed the multivariate logistic model, later developed by Glonek and McCullagh [19]. This is closed under marginalization and provides a natural interpretation of the first-order parameters as the logistic transforms of marginal probabilities, i.e. p univariate logistic regression models [8] are obtained when all interaction terms are zero. This is a sub-model of the class of marginal log-linear models considered by Evans and Richardson [15].

McCullagh [27] provides the following example in which questions of this type emerge. The data concerns two respiratory ailments of working coalminers who were smokers and apparently free of pneumoconiosis. Each individual was classified according to their age (9 ranges) and whether or not they suffered from breathlessness or wheeze, and the counts reported as effectively a $9 \times 2 \times 2$ contingency table. The aim of the study was to uncover how breathlessness and wheeze and their interaction are related to age. One tangential criticism at this point is that passage of time is not in itself causal, only processes that occur over time, so that age is not to be considered the underlying cause of these ailments, although it may be a useful predictor. McCullagh [27] raised further substantial reservations over the use of these data to reveal generalizable epidemiological conclusions. Nevertheless, the example provides a useful illustration. It was found, as expected, that the relevant parameter estimates from fitting the multivariate logistic model were almost identical to those obtained by fitting separate logistic models to each of the response variables. They were also slightly more efficient. The joint model further allowed estimation of the effect of the covariate on the interaction between these two binary response variables.

2.4 Detection of higher-order dependencies

Here is a conceptually simple but not trivial illustration posed in a form to illustrate difficulties with commonly used multivariate approaches based on the normal distribution, such as factor analysis and principal components analysis. Observations are available on a large number of independent individuals recording the values of a large number of binary variables, all having the same marginal probability $1/2$ and pairwise uncorrelated. In fact a possibly large unknown subset of variables are all simultaneously equal with

208 appreciable probability, even with probability one. The formal problem is to detect that
 209 subset as efficiently as possible.

210 Pairwise study as a generalization of standard multivariate techniques would not
 211 detect this. There are various generalizations, for example the presence of several such
 212 blocks. The challenge is partly to uncover such structures effectively and then to attach
 213 reasonable measures of uncertainty to what is found in order to make a more secure
 214 subject-matter interpretation. Division of the data into supposedly independent sub-
 215 sections each subject to the proposed analysis may formally compromise efficiency of
 216 estimation but may be the best route to cautious interpretation.

217 Suppose now that the dependence of a particular binary variable on the others is
 218 of interest and that its value is totally random unless a subset of size s of the binary
 219 explanatory variables are all at a particular combination of their levels, in which case
 220 the interest variable is, for example, at its upper level with high probability. This is
 221 again a situation that may be difficult to detect. Even if it were known which variables
 222 were jointly explanatory, there are 2^s possible combinations of levels that might be
 223 responsible for the positive outcome.

224 3 Other multivariate non-normal distributions

225 The previous ideas have natural extensions, especially to modelling multivariate Poisson
 226 variables. The geometric aspects of Section 2.1 apply directly to pure 2×2 contingency
 227 tables after standardization by the sample size with these standardized counts then inter-
 228 preted as estimates of probability. Similarly for higher-order tables. The considerations
 229 involved in that discussion are also closely related to so-called compositional data [1],
 230 which entails the data, rather than the model parameters, belonging to the unit simplex
 231 in $\mathbb{R}^d$. There has been a resurgence of interest in the statistical analysis of compositional
 232 data due to its importance in microbiomics (e.g. [24; 32; 33; 7]), an application that leads
 233 to the additional difficulty that observations are high-dimensional. When covariates are
 234 of compositional form, there are difficulties with the interpretation of standard linear
 235 regression coefficients as the notion of hypothetically altering one variable while others
 236 are treated as fixed is not physically possible; it is impossible to change one proportion
 237 without altering at least one other. Logarithmic transformation resolves this difficulty,
 238 although the regression coefficients are then required to sum to zero. See [1].

239 Olkin [29] outlined the construction of some other multivariate non-normal distribu-
 240 tions, some of which stem from the multivariate normal via the Wishart distribution. He
 241 briefly commented on physical explanations that could conceivably produce multivariate
 242 exponential random variables.

243 Multivariate outcomes of various types routinely emerge in matched pair and other
 244 matched strata experiments. We focus on pairwise comparisons for concreteness. Thus,
 245 pairs of individuals are matched based on their intrinsic features and a treatment is ran-
 246 domized to one unit per pair, leaving the other as the untreated control. The objective is
 247 to eliminate sources of variation other than the treatment. Given n pairs, let $(T_i, U_i)_{i=1}^n$
 248 denote the observations on an outcome of interest for the treated individual and the un-
 249 treated control. Provided there are no interactive effects between the treatment and the
 250 intrinsic characteristics of the pairs, an assumption that may need separate checks, it is
 251 reasonable to represent the distributions of T_i and U_i as being equal up to the presence
 252 of a treatment parameter, ψ say, and otherwise known modulo a pair-specific nuisance

parameter, λ_i say. Inclusion of a nuisance parameter for each pair avoids having to specify in detail those aspects of the probabilistic model that are not of subject-matter interest. There are limitations to this approach. Notably any effect-modification due to intrinsic characteristics would not be detectable.

The point we wish to make is that if the pair-specific nuisance parameters $(\lambda_i)_{i=1}^n$ are treated as unknown constants, then T_i and U_i are independent, while if these nuisance parameters are treated as random, from a known or unknown distribution parameterized by θ , this induces dependence within pairs leading to a multivariate formulation. The methods of analysis for these two situations are very different. For the former, typically preliminary manoeuvres are applied to eliminate the nuisance parameters before applying maximum likelihood to the transformed random variables to estimate ψ . On the other hand, with $(\lambda_i)_{i=1}^n$ treated as random, joint estimation of θ and ψ by some form of maximum likelihood is reasonable, although there could be serious difficulties if the assumed distribution for $(\lambda_i)_{i=1}^n$ is an inadequate representation. See [5] for a detailed discussion of matched exponential pairs. Other special distributional forms raise similar issues.

4 Outcomes with long-range dependence

We turn now to a completely different situation generating multiple outcomes: that of processes evolving over time or space. We focus in particular on the special issue of long-range dependence and its connection with second-order statistical analysis of time series. This account closely follows the early sections of [11].

Let $\{X_t\}_{t \in \mathbb{Z}}$ be a second-order stationary process with autocovariance function $\gamma_X(h) = \text{cov}(X_t, X_{t+h})$, variance $\tau_X = \gamma_X(0)$, autocorrelation function $\rho_X(h) = \gamma_X(h)/\tau_X$ and spectral density function

$$g_X(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_X(h) e^{-ih\omega}, \quad -\pi < \omega < \pi.$$

Since, by Herglotz's theorem (e.g. [6], p.117),

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\omega} g_X(\omega) d\omega,$$

the autocovariance function and the spectral density function provide equivalent second-order characterizations. A third characterization, in terms of sequences of averaged processes, reveals additional insights.

A new second-order stationary process $\{X_t^{(m)}\}_{t \in \mathbb{Z}}$ can be formed by averaging the original series in nonoverlapping blocks of size $m \in \{1, 2, \dots\}$, that is

$$X_t^{(m)} = (X_{tm-m+1} + \dots + X_{tm})/m. \quad (3)$$

Let $\gamma_m(h)$ and $v_m = \gamma_m(0)$ be the autocovariance function and variance of $\{X_t^{(m)}\}_{t \in \mathbb{Z}}$, and let $\rho_m(h) = \gamma_m(h)/v_m$ be the autocorrelation function. On considering $\{v_m\}_{m \in \mathbb{N}}$ as a sequence, with $v_1 = \tau_X$, we have [39]

$$v_m = \frac{\tau_X}{m} + \frac{2}{m^2} \sum_{h=1}^m (m-h) \gamma_X(h) = \frac{\tau_X}{m} + \frac{2}{m^2} \sum_{s=1}^{m-1} \sum_{h=1}^s \gamma_X(h). \quad (4)$$

Thus, for $h \in \{2, 3, \dots\}$

$$\gamma_X(h-1) = \Delta^2 \left\{ \frac{h^2}{2} \left(v_h - \frac{\tau_X}{h} \right) \right\}$$

where Δ^2 denotes the second-difference operator. This is seen most clearly by expanding the double sum in (4) and noting that, for a sequence $\{Z_n\}_{n \in \mathbb{Z}}$, $\Delta^2(Z_n) = \Delta\{\Delta(Z_n)\} = Z_n - 2Z_{n-1} + Z_{n-2}$. Thus we have established the equivalence of $g_X(\omega)$, $\{\gamma_X(h)\}_{h \in \mathbb{Z}}$ and $\{v_m\}_{m \in \mathbb{N}}$ in characterizing the second-order properties of a stationary sequence.

The qualitative distinctions between short-range and long-range dependent processes are captured by the following essentially equivalent conditions. For short-range dependent processes:

- (SR.1) $\sum_{h=0}^{\infty} \gamma_X(h)$ is convergent;
- (SR.2) $g_X(0)$ is finite;
- (SR.3) v_m is asymptotically of the form τ_X/m for large m ;
- (SR.4) $\{X_t^{(m)}\}_{t \in \mathbb{Z}}$ tends to second-order pure noise as $m \rightarrow \infty$.

By contrast, for long-range dependent processes:

- (LR.1) $\sum_{h=0}^{\infty} \gamma_X(h)$ is divergent;
- (LR.2) $g_X(\omega)$ is singular near $\omega = 0$;
- (LR.3) mv_m diverges as $m \rightarrow \infty$.

We shall restrict attention to a rich but more manageable subclass of long-range dependent processes of index $\beta \in (0, 1)$, characterized by:

- (LR- β .1) $\gamma_X(h) \sim \gamma' h^{-\beta}$ ($h \rightarrow \infty$);
- (LR- β .2) $g_X(\omega) \sim g' \omega^{-(1-\beta)}$ ($\omega \rightarrow 0$);
- (LR- β .3) $v_m \sim v' m^{-\beta}$ ($m \rightarrow \infty$);
- (LR- β .4) $\rho_m(h) \rightarrow (1/2)\delta^2(h^{2-\beta})$ as $m \rightarrow \infty$
 $\sim (1/2)(2-\beta)(1-\beta)h^{-\beta}$ as $h \rightarrow \infty$,

where γ' , g' and v' are functionally related and $f(x) \sim g(x)$, ($x \rightarrow x_0$) means $f(x) = g(x)(1 + o(1))$ as $x \rightarrow x_0$.

Property (LR- β .1) specifies that the tails of the autocovariance function decay as a slow power law. A major statistical implication of Property (LR- β .3) is that the variance of a sample mean decays more slowly than the inverse sample size. This arises due to the non-degenerate correlation structure of the averaged process $\{X_t^{(m)}\}_{t \in \mathbb{Z}}$ as $m \rightarrow \infty$.

A detailed treatment of long-range dependence and self-similarity is given in [30], where the conditions above are discussed in Sections 2.2, 2.3 and 2.8.

For spatial processes the results are qualitatively similar ([36; 37]), although there is inherent multidimensionality. A further feature, often ignored with little effect due to the scales involved, are the topological features of the area over which the spatial series is observed. For instance, in a study of crop yields, the space over which the process is observed can reasonably be treated as two-dimensional Euclidean, while if the process of interest is the spatial arrangement of lipids or proteins on the cellular membranes of bacteria or other microorganisms (e.g. [34]), the geometry of the surface usually cannot be ignored.

Cox [11] called Property LR- β .4 asymptotic second-order self-similarity because of its resemblance to exact second order self-similarity. Specifically, a process $\{X_t\}_{t \in \mathbb{Z}}$ for which

$$\rho_X(h) = (1/2)\Delta^2(h^{2-\beta}) \quad (5)$$

has $\rho_m(h) = \rho_X(h)$ and $v_m = v_1 m^{-\beta}$ and is exactly second-order self-similar (cf LR- β .4).
In other words, sums of consecutive values such as

$$X_1 + \cdots + X_m, X_{m+1} + \cdots + X_{2m}, \cdots$$

have the same autocorrelation function as the original process.

Recently Cox [14] modelled the accretion of big data as evolving implicitly in time or space, with new sources of variability entering in a time-dependent Poisson process and operating permanently thereafter. The work illustrates, in idealized form, the considerable potential for producing estimates of misleadingly high apparent precision, owing to strong implicit assumptions of at most weak dependence. This is perhaps most problematic in external analyses, which are less central to the present paper with its focus on problems of a multivariate nature. There are a number of issues involved both in assessing the real errors of estimation present and indeed in considering whether sources of variability that are negligible with modest amounts of data must be assessed critically when large amounts of data are involved. Indeed systematic errors that are relatively unimportant with modest amounts of data may become critical as the amount of data gets large.

Acknowledgements. The invitation to contribute to this notable anniversary is much appreciated.

References

- [1] Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc. B*, 44, 139–177.
- [2] Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, 71, 323–330.
- [3] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.
- [4] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.
- [5] Battey, H. S. and Cox, D. R. (2020). High-dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3, 119–148.
- [6] Brockwell, P. J. and Davis, R. A. (2006). *Time Series: Theory and Methods*. Springer, New York.
- [7] Cao, Y., Zhang, A. and Li, H. (2020). Multi-sample estimation of bacterial composition matrix in metagenomics data. *Biometrika*, 107, 75–92.
- [8] Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc. B*, 20, 215–242.

- [9] Cox, D. R. (1972). The analysis of multivariate binary data. *J. R. Statist. Soc. C*, 21, 113–120.
- [10] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
- [11] Cox, D. R. (1984). Long-range dependence: a review. In *Proceedings 50th Anniversary Conference, Iowa State Statistical Laboratory*.
- [12] Cox, D. R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika*, 81, 403–408.
- [13] Cox, D. R. and Wermuth, N. (2002). On some models for multivariate binary variables parallel in complexity with the multivariate Gaussian distribution. *Biometrika*, 89, 462–469.
- [14] Cox, D. R. (2015). Big data and precision. *Biometrika*, 102, 712–716.
- [15] Evans, R. J. and Richardson, T. S. (2013). Marginal log-linear parameters for graphical Markov models. *J. R. Stat. Soc. Ser. B*, 75, 743–768.
- [16] Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28, 1–23.
- [17] Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *J. Amer. Statist. Assoc.*, 65, 694–701.
- [18] Frisch, U. (1996). *Turbulence and the legacy of A. N. Kolmogorov*. Cambridge University Press.
- [19] Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *J. R. Statist. Soc. B*, 57, 533–546.
- [20] Hurst, H. E. (1950). Long term storage capacity of reservoirs. *Proceedings of the American Society of Civil Engineers*, 116, 770–799.
- [21] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities I. Multivariate totally positive distributions, *J. Multivar. Anal.*, 10, 467–498.
- [22] Kolmogorov, A. N. (1941; 1991). The local structure of turbulence in incompressible viscous fluid for very large Reynolds’ numbers. *Proc. R. Soc. Lond. A*, 434, 9–13. First published in Russian in *Dokl. Akad. Nauk SSSR*, 30.
- [23] Lauritzen, S. L., Uhler, C. and Zwiernik, P. (2021). Total positivity in exponential families with application to binary variables. *Ann. Statist.*, to appear.
- [24] Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Ann. Rev. Statist. Applic.*, 2, 73–94.
- [25] Mantel, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. *Biometrics*, 22, 83–95.

- 398 [26] Marchetti, G. M. and Wermuth, N. (2016). Palindromic Bernoulli distributions.
399 *Electron. J. Statist.* 10, 2435–60.
- 400 [27] McCullagh, P. (1989). Models for discrete multivariate responses. *Bulletin of the*
401 *International Statistical Institute*, 53, 407–418.
- 402 [28] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley.
- 403 [29] Olkin, I. (1994). Multivariate non-normal distributions and models of dependency.
404 *Multivariate Analysis and its Applications*. Vol. 24, Institute of Mathematical Statis-
405 tics, Monograph Series, 37–53.
- 406 [30] Pipiras, V. and Taqqu, M. S. (2017). *Long-range dependence and self-similarity*.
407 Cambridge University Press, Cambridge.
- 408 [31] Sheppard, W. F. (1898). On the geometric treatment of the ‘normal curve’
409 of statistics with particular reference to correlation and the theory of errors.
410 *Proc. R. Soc. Lond.*, 62, 170–173.
- 411 [32] Shi, P., Zhang, A. and Li, H. (2016). Regression analysis for microbiome composi-
412 tional data. *Ann. Appl. Statist.*, 10, 1019–1040.
- 413 [33] Shi, P., Zhou, Y. and Zhang, A. (2021). High-dimensional log-error-in-variable re-
414 gression with applications to microbial compositional data analysis. *Biometrika*, to
415 appear.
- 416 [34] Ward, S., Cohen, E. A. K., and Adams, N. (2021). Testing for complete spatial
417 randomness on three-dimensional bounded convex shapes. *Spat. Statist.*, 41, 100489.
- 418 [35] Wermuth, N. (1976). Analogies between multiplicative models in contingency tables
419 and covariance selection. *Biometrics*, 32, 95–108.
- 420 [36] Whittle, P. (1956) Variation of yield variance with plot size. *Biometrika*, 43, 337–
421 343.
- 422 [37] Whittle, P. (1962). Topographic correlation, power-law covariance functions and
423 diffusion. *Biometrika*, 49, 305–314.
- 424 [38] Yates, F. (1984). Tests of significance of 2×2 contingency tables (with discussion).
425 *J. Roy. Statist. Soc. A*, 147, 426–463.
- 426 [39] Yule, G. U. (1945). On a method of studying time series based on their internal
427 correlation. *J. Roy. Statist. Soc.*, 108, 208–225.