



The University of Chicago
Department of Statistics
TECHNICAL REPORT SERIES

Exchangeability and regression models

Peter McCullagh

TECHNICAL REPORT NO. 544

January 2004

5734 S. University Avenue
Chicago, IL 60637

This research was supported in part by NSF grant No. 0305009.

Exchangeability and regression models

by

Peter McCullagh

University of Chicago

January 2004

Summary

A process is said to be exchangeable if each finite-dimensional distribution is symmetric, or invariant under coordinate permutation. The definition suggests that exchangeability can have no role in statistical models for dependence, in which the units are overtly non-exchangeable on account of differences in covariate values. The theme of this paper is that this narrow view is mistaken for two reasons. First, every regression model is a set of processes in which the distributions are indexed by the finite restrictions of the covariate, and regression exchangeability is defined naturally with this in mind. Second, regression exchangeability has a number of fundamental implications connected with lack of interference (Cox, 1958a) and absence of unmeasured covariates (Greenland, Robins and Pearl 1999). This paper explores the role of exchangeability in a range of regression models, including generalized linear models, biased-sampling models (Vardi, 1985), block factors and random-effects models, models for spatial dependence, and growth-curve models. The fundamental distinction between parameter estimation and sample-space prediction is a recurring theme.

Key words: block factor; classification factor; causal model; counterfactual; crossover design; empirical Bayes; exchangeable prior process; factorial model; interaction; interference; likelihood principle; penalized likelihood; prediction; random-effects model; smoothing spline;

1 Introduction

This paper has two themes. The first is the importance of processes in the definition of a statistical model, and the second is the role of exchangeability in models for dependence, also called regression models. Our concern is with the relationship between statistical models and processes, and the role that exchangeability plays in regression models. The focus is primarily on the model as a parametric family, ignoring entirely the important role of exchangeability in the selection of a prior distribution. For factorial effects and interactions, in particular, exchangeability and partial exchangeability are appealing properties for a prior process. Experience tells us that when two factors have large main effects, their interaction is likely to be appreciable (Cox, 1984). Although it is possible to construct exchangeable processes having this property, I am not aware of an instance of their use in Bayesian models for factorial designs.

Apart from its necessity for asymptotic approximations, the main reason for emphasizing processes over distributions is that the unnatural distinction between estimation and prediction is removed. An estimated variety contrast of 50 ± 15 kg/ha is simply a prediction concerning the likely difference of yields under similar circumstances in future

This research was supported in part by NSF grant No. 0305009

seasons. Although the theory of estimation could be subsumed under a theory of prediction for statistical models, there are compelling reasons for maintaining the separation. On a purely theoretical point, estimation in the sense of inference concerning the model parameter may be governed by the likelihood principle, whereas inference in the sense of prediction is not: see section 5.1 below. In addition, apart from convenience of presentation and linguistic style, parameter estimation is the first step in naive prediction. The second step, frequently trivial and therefore ignored, is the calculation of conditional distributions or conditional expectations, as in prediction for processes in the standard probabilistic sense.

The paper has nothing to contribute to philosophical problems such as where the model comes from, nor does it have much to contribute to practical questions such as how to select a model within a class of models, how to compute the likelihood function, or how to decide whether a model is adequate for the task at hand. In addition, while the mathematical interpretation is clear, any physical interpretation of the model requires a correspondence between the mathematical objects, such as units and covariates, and the physical objects that they represent. This correspondence is usually implicit in most discussions of causality, and, although it is important, the present paper has little to say on the matter.

2 Regression models

2.1 Introduction

We begin with the presumption that every statistical model is a set of processes, one process for each parameter value $\theta \in \Theta$, aiming to explore the consequences of that condition for regression models. The reason for emphasizing processes over distributions is that a process permits inferences in the form of predictions for the response on unsampled units, including point predictions, interval predictions and distributional predictions. Without the notion of a process, the concept of further units beyond those in the sample does not exist as an integral part of the mathematical construction, which greatly limits the possibilities for prediction and inference. Much of asymptotic theory, for instance, would be impossible in the absence of a process or set of processes.

To every potential sample there corresponds an observation space, and it is the function of a process to determine a probability distribution on each of these spaces. For notational simplicity, we restrict our attention to real-valued processes in which the observation space corresponding to a sample of size n is the n -dimensional vector space of functions on the sampled units. The response value is denoted by $Y \in \mathcal{R}^n$. Other processes exist in which the observation spaces are more complicated objects such as trees or partitions, but in regression models where we have one measurement on each unit, the observation space is invariably a product set, such as \mathcal{R}^n or $\{0, 1\}^n$, of responses or functions on the sampled units.

The processes with which we are concerned here are defined on the entire set \mathcal{U} of statistical units and observed on a particular finite set called the sample. The entire set \mathcal{U} is assumed to be countably infinite, and the sample $S \subset \mathcal{U}$ is a finite subset of size n . The term sample does not imply a random sample: in a time series the sample units are usually consecutive points, and similar remarks apply to agricultural field experiments where the

sample units are usually adjacent plots in the same field. In other contexts, the sample may be stratified as a function of the covariate or classification factor.

It is worth recalling the definition of a process and how this differs from the simpler concept of a distribution on the observation space. For the purposes of this paper a process P is a function that associates with each finite sample $S \subset \mathcal{U}$ of size n a distribution P_S on the observation space \mathcal{R}^S of dimension n . Let $S \subset S'$ be a sub-sample, and let $P_{S'}$ be the distribution on $\mathcal{R}^{S'}$ determined by the process. If logical contradictions are to be avoided, $P_{S'}$ must be the marginal distribution of P_S under the operation of coordinate deletion, i.e. deletion of those units not in S' . A process is thus a collection of mutually compatible distributions of this sort, one distribution on each of the potential observation spaces.

An exchangeable process is one for which each distribution P_S is symmetric, or invariant under coordinate permutation. Sometimes the term infinitely exchangeable process is used, but the additional adjective is unnecessary when it is understood that we are dealing with a process defined on a countably infinite set. Exchangeability is a fundamental notion, and much effort has been devoted to the characterization of exchangeable processes and partially exchangeable processes (de Finetti 1974; Aldous 1981). Despite the attractions of the theory, the conventional definition is too restrictive to be of much use in applied work, where differences among units are frequently determined by a function x called a covariate.

Up to this point, we have talked of a process in terms of distributions, not in terms of a random variable. However, the Kolmogorov extension theorem guarantees the existence of a random variable Y taking values in $\mathcal{R}^{\mathcal{U}}$ such that the finite-dimensional distributions are those determined by P . As a matter of logic, however, the distributions come first and the existence of the random variable must be demonstrated, not the other way round. For the most part, the existence of the random variable poses no difficulty, and all distributional statements may be expressed in terms of random variables. However, as shown in section 4, the Kolmogorov extension does not admit the existence of counterfactuals in the literal sense. If these are deemed necessary for a theory of causal models, a fundamentally different structure is required.

To avoid misunderstandings at this point, the statistical units are the objects on which the process is defined. It is left to the reader to interpret this in a suitable operational sense depending on the application at hand. By contrast, the standard definition in the experimental design literature holds that a unit is ‘the smallest division of the experimental material such that two units may receive different treatments’ (Cox, 1958a). The latter definition implies random assignment of treatment levels to units. At a practical level, the operational definition is much more useful than the mathematical definition. While the two definitions coincide in most instances, example 3 shows that they may differ.

The purpose of the next section is to define an exchangeable regression process, or a process modulated by a covariate, and to explore briefly some of the consequences of that construction.

2.2 Regression processes

A covariate is a function on the units. It may be helpful for clarity to distinguish certain types of covariate. A quantitative covariate is a function $x: \mathcal{U} \rightarrow \mathcal{R}$ or $x: \mathcal{U} \rightarrow \mathcal{R}^p$ taking values in a finite-dimensional vector space. This statement does not exclude instances in

which x is a bounded function or a binary function. A qualitative covariate or factor is a function $x: \mathcal{U} \rightarrow \Omega$ taking values in a set Ω called the set of levels or labels. These labels may have no additional structure, in which case the term nominal scale is used, or they may be linearly ordered or partially ordered or they may constitute a tree or a product set. The exploitation of such structure is a key component in successful model construction, but that is not the thrust of the present paper. For the moment at least, a covariate is a function $x: \mathcal{U} \rightarrow \Omega$ taking values in an arbitrary set Ω . Ordinarily, of course, the values of x are available only on the finite sampled subset $S \subset \mathcal{U}$, but we must not forget that the aim of inference is ultimately to make statements about the likely values of the response on unsampled units whose x -value is specified. If statistical models have any value, we must be in a position to make predictions about the response distribution on such units, possibly even on units whose covariate value does not occur in the sample.

At this point the reader might want to draw a distinction between estimation and prediction, but this distinction is more apparent than real. If a variety contrast is estimated as 50 ± 15 kg/ha, the prediction is that the mean yield for other units under similar conditions will be 35–65 kg/ha higher for one variety than the other. Without the concept of a process to link one statistical unit with another, it is hard to see how inferences or predictions of this sort are possible. Nonetheless, Besag (2002, p. 1271) makes it clear that this point of view is not universally accepted.

Let $x: \mathcal{U} \rightarrow \Omega$ be a given function on the units. Recall that a real-valued process is a function P that associates with each finite subset $S \subset \mathcal{U}$ a distribution P_S on \mathcal{R}^S , and that these distributions are mutually compatible with respect to sub-sampling of units. A regression process having the following property for every integer n is called regression exchangeable or exchangeable modulo x .

(RE) *Two finite samples $S = \{i_1, \dots, i_n\}$ and $S' = \{j_1, \dots, j_n\}$ of equal size, ordered such that $x(i_r) = x(j_r)$ for each r , determine the same distribution $P_S = P_{S'}$ on \mathcal{R}^n .*

Exchangeability modulo x is the condition that if x takes the same value on two samples, the distributions are also the same. The vast majority of models that occur in practical work have this property, without which little progress would be possible. The property is a consequence of the definition of a statistical model as a functor on a certain category in the sense of McCullagh (2001) or Brøns (2001). For any function g defined on Ω , exchangeability modulo $g(x)$ implies exchangeability modulo x , and $g(x) \equiv 0$ reduces to the standard definition of exchangeability. Exchangeability modulo x is not to be confused with partial exchangeability as defined by Aldous (1981) for random rectangular matrices.

The following are consequences of regression exchangeability.

(C1) Two units having the same value of x also have the same response distribution: $x(i) = x(j)$ implies $P_{\{i\}} = P_{\{j\}}$ or $Y_i \sim Y_j$.

(C2) Lack of interference: The distribution of Y_i depends only on the value of x on that specific unit, not on the values of x on other units.

The first consequence of exchangeability, that differences between distributions are determined by differences between covariate values, is related, at least loosely, to the assumption of ‘no unmeasured confounders’ (Greenland, Robins and Pearl, 1999). This is a critical assumption in so-called causal models. I am not aware of anything approaching a mathematical definition of causality in the literature, so it is hard to say precisely what the connection is. Neutral terminology is used here to avoid any suggestion of causality.

We say that any difference between distributions is associated with a difference between covariate values in the ordinary mathematical sense without implying or suggesting a causal interpretation.

Lack of interference (Cox, 1958a, p. 19) is also a key assumption in experimental design and in clinical trials. It is known, however, that biological interference can and does occur. Such effects are usually short-range, such as root interference or fertilizer diffusion, so typical field trials incorporate guard strips to minimize the effect. This sort of interference can be accommodated within the present framework by including in x the necessary information about nearby plots. Interference connected with carry-over effects in a crossover trial can be accommodated by defining the statistical units as subjects rather than subjects at specific time points (example 3 below).

2.3 Interaction

Suppose the model is such that responses on different units are independent, and that $x = (v_0, v)$ with v_0 a binary variable indicating treatment level, and v a covariate or other classification factor. The response distributions at the two treatment levels are $P_{0,v}$ and $P_{1,v}$. It is conventional in such circumstances to define ‘the treatment effect’ by a function or functional of the two distributions

$$\text{Treatment effect} = T(P_{1,v}) - T(P_{0,v}) = \tau(v)$$

such that $T(P_{1,v}) = T(P_{0,v})$ implies $P_{1,v} = P_{0,v}$ for all model distributions. For example, T might be the difference between the two means (Cox, 1958a), the difference between the two log-transformed means (Box and Cox, 1964), the difference between the two medians, the log ratio of the two means, the log odds ratio (Cox, 1958b), the log hazard ratio (Cox, 1972), or the log variance ratio. In principle, T is chosen for ease of expression in summarizing conclusions and in making predictions concerning differences to be expected in future. Ideally, T is chosen so that, under the model in its simplest form, the treatment effect is constant over the values of v , in which case we say that there is no interaction between treatment and other covariates. In practice, preference is given to scalar functions because these lead to simpler summaries, but the definition does not require this.

If $P_{1,v} = P_{0,v}$ the model distributions do not depend on the treatment level, and the treatment effect $\tau(v)$ is zero. Conversely, a zero treatment effect in the model implies equality of distributions. If $\tau(v)$ is identically zero, we say that there is no treatment effect. The process is then exchangeable modulo v , i.e. ignoring treatment. This is a stronger condition than regression exchangeability with treatment included as a component of x .

If the treatment effect is constant and independent of v , we say that there is no interaction, and the single numerical value suffices to summarize the difference between distributions. Although the process is not now exchangeable in the sense of the preceding paragraph, the adjustment for treatment is usually of a very simple form, so much of the simplicity of exchangeability remains. If the treatment effect is not constant, we say that there is interaction. By definition, non-zero interaction implies a non-constant treatment effect, so a zero treatment effect in the presence of non-zero interaction is a logical contradiction.

It is possible to define an average treatment effect $\text{ave}(\tau(v))$, averaged with respect to a given distribution on v , and some authors refer to such an average as the ‘main effect

of treatment'. Such averages may be useful in limited circumstances as a summary of the treatment effect in a specific heterogeneous population. However, if the interaction is appreciable, and in particular if the sign of the effect varies across sub-groups, we would usually want to know the value in each of the sub-groups. A zero value of the average treatment effect does not imply exchangeability in the sense discussed above, so a zero average rarely corresponds to a hypothesis of mathematical interest. Nelder (1977) and Cox (1984) argue that statistical models having a zero average main effect in the presence of interaction are seldom of scientific interest. McCullagh (2000) reaches a similar conclusion using an argument based on algebraic representation theory in which selection of factor levels is a permissible operation.

3 Examples of exchangeable regression models

Virtually all exchangeable regression models that occur in practice have independent components, in which case it is sufficient to specify the marginal distributions for each unit. The first four examples are of that type, but the fifth example shows that the component variables in an exchangeable regression process need not be independent or conditionally independent.

Example 1. Classical regression models. In the classical multiple regression model, the covariate x is a function on the units taking values in a finite-dimensional vector space $\mathcal{V} \cong \mathcal{R}^p$, which we call the covariate space. Each point $\theta = (\beta, \sigma)$ in the parameter space consists of a linear functional $\beta \in \mathcal{V}'$ plus a real number σ , and the parameter space consists of all such pairs. If the value of the linear functional β at $v \in \mathcal{V}$ is denoted by $v^T \beta$, the value on unit i is $x_i^T \beta$. In the classical linear regression model the response distribution for unit i is normal with mean equal to $x_i^T \beta$ and variance σ^2 . The model may be modified in a number of minor ways, for example by restricting σ to be non-negative to ensure identifiability.

From the point of view of exchangeability, generalized linear models or heavy-tailed versions of the above model are not different in any fundamental way. For example, the linear logistic model in which $\eta_i = x_i^T \beta$ and $Y_i \sim \text{Ber}(1/(1 + \exp(-\eta_i)))$ is an exchangeable regression model in which the parameter space consists of linear functionals on \mathcal{V} .

Example 2. Treatment and classification factors. A treatment or classification factor is a function x on the units taking values in a set, usually a finite set, called the set of levels. It is conventional in applied work to draw a strong distinction between a treatment factor and a classification factor (Cox, 1984). The practical distinction is an important one, namely that the level of a treatment factor may, in principle at least, be determined by the experimenter, whereas the level of a classification factor is an immutable property of the unit. Age, sex and ethnic origin are examples of classification factors: medication and dose are examples of treatment factors. I am not aware of any mathematical construction corresponding to this distinction, so the single definition covers both. A block factor as defined in section 5 is an entirely different sort of mathematical object in which the concept of a set of levels is missing.

Let Ω be the set of treatment levels, and let $\tau: \Omega \rightarrow \mathcal{R}$ be a function on the levels. In conventional statistical parlance, τ is called the vector or list of (treatment) effects, and

differences such as

$$\tau(M) - \tau(F), \quad \text{or} \quad \tau(\text{Kerr's pinks}) - \tau(\text{King Edward})$$

are called contrasts. A critical point to note is that the parameter space \mathcal{R}^Ω has a preferred basis determined by the factor levels, and a preferred basis is essential for the construction of an exchangeable prior process for the effects. Without a preferred basis, no similar construction exists for a general linear functional β in a regression model.

In the standard linear model with independent components, the distribution of the response on unit i is Gaussian with mean $\tau(x(i))$ and variance σ^2 . The parameter space is the set of all pairs (τ, σ) in which τ is a function on the levels and σ is a real number. Once again, the extension to generalized linear models presents no conceptual difficulty.

Example 3. Crossover design. In a two-period crossover design, one observation is made on each subject under different experimental conditions at two times sufficiently separated that carry-over effects can safely be neglected. If we regard the subjects as the statistical units, which we are at liberty to do, the design determines the observation space \mathcal{R}^2 for each unit. The observation space corresponding to a set of n units is $(\mathcal{R}^2)^n$. Let x be the treatment regime, so that (x_{i1}, x_{i2}) is the ordered pair of treatment levels given to subject i . In the conventional statistical model the response distribution for each unit is bivariate Gaussian with covariance matrix $\sigma^2 I_2$. The mean vector is

$$E \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_i + \tau_{x_{i1}} \\ \alpha_i + \tau_{x_{i2}} + \delta \end{pmatrix}$$

in which α is a function on the subjects, and δ is a common temporal trend. The parameter space consists of all functions α on the n subjects, all functions τ on the treatment levels, plus the two scalars (δ, σ) , so the effective dimension is $n + 3$ for a design with n subjects and two treatment levels. Since the parameter space for this model depends on the design, the initial declaration in section 2 that every model is a set of processes, is not entirely correct. The more general definition given by McCullagh (2002) is necessary, but the implications are ignored here.

An alternative Gaussian model, with units and observation spaces defined in the same manner, has the form

$$E \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \tau_{x_{i1}} \\ \tau_{x_{i2}} + \delta \end{pmatrix}, \quad \text{cov} \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

with a fixed parameter space independent of the design. The temporal effect δ is indistinguishable from a carryover effect that is independent of the initial treatment. If there is reason to suspect a non-constant carry-over effect, the model may be extended by writing

$$E \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \tau_{x_{i1}} \\ \tau_{x_{i2}} + \gamma_{x_{i1}} + \delta \end{pmatrix}, \quad \text{cov} \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

If there are two treatment levels and all four combinations occur in the design, the difference $\gamma_1 - \gamma_0$ is estimable.

Example 4. *Biased sampling.* We consider a biased-sampling model in which observations on distinct units are independent and, for notational convenience, real valued. The covariate w associates with the i th unit a bias function w_i such that $w_i(x) \geq 0$ for each real number x . The parameter space is either the set of probability distributions or the set of non-negative measures on the Borel sets in \mathcal{R} such that each integral $\int w_i(x) dF(x)$ is finite (Vardi 1985; Kong, McCullagh, Meng, Nicolae and Tan 2002). For each F in the parameter space, the response distribution on unit i is the weighted distribution such that $dF_i(x) \propto w_i(x) dF(x)$. Thus, to each point in the parameter space the model associates a process with independent but non-identically distributed components. Two units having the same bias function also have the same distribution, so the process is regression-exchangeable.

The simplest example is one in which $w_i(x) = 1$ identically for each unit, in which case the maximum-likelihood estimator \hat{F} is the empirical distribution at the observations. Size-biased sampling corresponds to $w(x) = |x|$ or some power of $|x|$. In general, the maximum-likelihood estimator \hat{F} is a distribution supported at the observation points, but with unequal atoms at these points.

Example 5. *Prediction and smoothing models.* Consider the modification of the simple linear regression model in which unit i has covariate value x_i , and, in a conventional but easily misunderstood notation,

$$Y_i = \beta_0 + \beta_1 x_i + \eta(x_i) + \epsilon_i. \quad (1)$$

The coefficients (β_0, β_1) are parameters to be estimated, ϵ is a process with i.i.d. $N(0, \sigma^2)$ components, and η is a zero-mean stationary process on the real line, independent of ϵ , with covariance function

$$\text{cov}(\eta(x), \eta(x')) = \sigma_\eta^2 K(x - x').$$

If η is a Gaussian process, the response distribution for any finite collection of n units may be expressed in the equivalent distributional form

$$Y \sim N(X\beta, \sigma^2 I_n + \sigma_\eta^2 V) \quad (2)$$

where $V_{ij} = K(x_i - x_j)$ are the components of a positive semi-definite matrix. This model with four-dimensional parameter space is regression-exchangeable with non-independent components.

The linear combination $\eta(x_i) + \epsilon_i$ in (1) is a convenient way of describing the distribution of the process as a sum of more elementary processes. The treacherous aspect of the notation lies in the possibility that η (or ϵ) might be mistaken for a parameter to be estimated from the data, which is not the intention. The alternative parametric model with independent components and parameter space consisting of all smooth functions η is very different from (2) and is not considered here.

The simplest way to proceed for estimation and prediction is first to estimate the parameters $(\sigma^2, \sigma_\eta^2, \beta_0, \beta_1)$ by maximum-likelihood, or by some closely related procedure such as REML for the variance components followed by weighted least squares for the regression parameters. With prediction in mind, Wahba (1984) recommends generalized cross-validation over REML on the grounds that it is more robust against departures from

the stochastic model. Efron (2001) considers a range of estimators and seems to prefer the REML estimator despite evidence of bias. Suppose that this has been done, and that we aim to predict the response value $Y(i^*)$ for a new unit i^* in the same process whose covariate value is $x^* = x(i^*)$. Proceeding as if the parameter values were given, the conditional expected value of $Y(i^*)$ given the values on the sampled units is computed by the formula

$$\hat{Y}_{i^*} = E(Y(i^*) | Y) = \beta_0 + \beta_1 x^* + k^* \Sigma^{-1} (Y - \mu) \quad (3)$$

where μ, Σ are the estimated mean and covariance matrix for the sampled units, and $k_i^* = \sigma_\eta^2 K(x^*, x_i)$ is the vector of covariances. The conditional distribution is Gaussian with mean (3) and constant variance independent of Y . Interpolation features prominently in the geostatistical literature where linear prediction is called Kriging (Stein, 1999)

If η is Brownian motion with generalized covariance function $-|x - x'|$ on contrasts, the prediction function (3) is continuous and piecewise linear: if $K(x, x') = |x - x'|^3$, the prediction function is a cubic spline (Wahba 1990; Green and Silverman 1994). Of course, K is not necessarily a simple covariance function of this type: it could be in the Matérn class or it could be a convex combination of simple covariance functions. The cubic and linear splines illustrated in Fig. 1 are obtained by fitting model (2) to simulated data (Wahba 1990, p. 45), in which $\eta(x)$ is the smooth function shown as the dashed line.

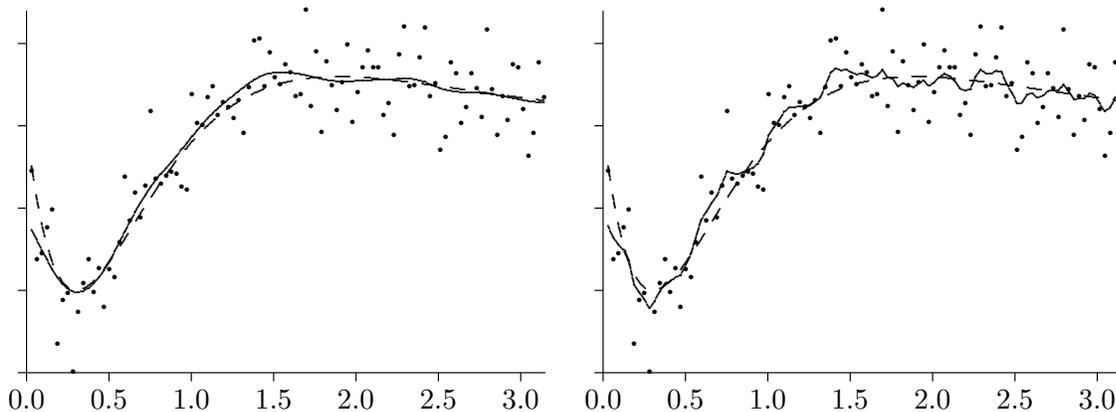


Fig. 1. Cubic and linear spline prediction graphs (solid line) fitted by REML. The ideal predictor $\eta(x)$ (dashed line) is taken from Wahba (1990, p. 45).

In statistical work, the adjective ‘Bayesian’ usually refers to the operation of converting a probability $p(A | B)$ into a probability of the form $p(B | A)$ by supplying additional information and using Bayes’s theorem. The transformation from the joint distribution of (Y^*, Y) as determined by the process (2), to the conditional distribution $Y^* | Y$, does not involve prior information or Bayes’s theorem. Nonetheless, it is possible to cast the argument leading to (3) in a Bayesian mould, so the majority of authors use the term Bayesian or empirical Bayes in this context (Wahba 1990; Efron 2001). A formal Bayesian analysis begins with a prior distribution π on the parameters $(\beta, \sigma^2, \sigma_\eta^2)$, and uses the likelihood function to obtain the posterior distribution. The process is such that the predictive distribution for a new unit has mean (3) and constant variance depending on the parameters. The Bayesian predictive distribution is simply the posterior average, or mixture, of these

distributions. In this context, the adjective ‘Bayesian’ refers to the conversion from prior and likelihood to posterior, not to (3).

Since η is not a parameter to be estimated, the introduction of extra-likelihood criteria such as penalty functions or kernel density estimators to force smoothness is, in principle at least, unnecessary. In practice, if the family of covariance functions for η includes a smoothness parameter such as the index in the Matérn class, the likelihood function seldom discriminates strongly. Two covariance functions such as $-|x - x'|$ and $|x - x'|^3$ achieving approximately the same likelihood value, usually produce prediction graphs that are pointwise similar. For the data in Fig. 1, the REML log likelihood values are 42.2 for the model in which $K(x, x') = -|x - x'|$, 42.1 for the ‘quadratic’ version $|x - x'|^2 \log |x - x'|$, and 41.3 for the cubic. Visually the prediction graphs are very different, so aesthetic considerations may determine the choice for graphical presentation.

To illustrate one crucial difference between an estimator and a predictor, it is sufficient to note that the prediction function (3) is not a projection on the observation space in the sense that the least-squares fit is a projection on the observation space. However, it is a projection on a different space, the combined observation-prediction space. Let S_0 be the n sampled units, let S_1 be the m unsampled units for which prediction is required, and let $S = S_0 \cup S_1$ be the combined set. The covariance matrix Σ for the joint distribution in \mathcal{R}^S may be written in partitioned form with components $\Sigma_{00}, \Sigma_{01}, \Sigma_{11}$, and the model matrix may be similarly partitioned, X_0 for the sampled units and X_1 for the unsampled units. The prediction function is linear in the observed value Y_0 and may be written as the sum of two linear transformations as follows.

$$\begin{pmatrix} \hat{Y}_0 \\ \hat{Y}_1 \end{pmatrix} = \begin{pmatrix} P_0 & 0 \\ X_1(X_0^T \Sigma_{00}^{-1} X_0)^{-1} X_0^T \Sigma_{00}^{-1} & 0 \end{pmatrix} \begin{pmatrix} Y_0 \\ \star \end{pmatrix} + \begin{pmatrix} Q_0 & 0 \\ \Sigma_{10} \Sigma_{00}^{-1} Q_0 & 0 \end{pmatrix} \begin{pmatrix} Y_0 \\ \star \end{pmatrix} \quad (4)$$

where $P_0 = X_0(X_0^T \Sigma_{00}^{-1} X_0)^{-1} X_0^T \Sigma_{00}^{-1}$ and $Q_0 = I - P_0$. Evidently $\hat{Y}_0 = Y_0$ as it ought. The first transformation is the least-squares projection $P: \mathcal{R}^S \rightarrow \mathcal{R}^S$ onto $\mathcal{X} \subset \mathcal{R}^S$ of dimension p . The second transformation is a projection $T: \mathcal{R}^S \rightarrow \mathcal{R}^S$, self-adjoint with respect to the inner product Σ^{-1} , and thus an orthogonal projection. Its kernel consists of all vectors of the form (x, \star) , i.e. all vectors in $\mathcal{X} + \mathcal{R}^{S_1}$ of dimension $m + p$, and the image is the orthogonal complement. Direct calculation shows that $PT = TP = 0$ so the sum $P + T$ is also a projection.

Most computational systems take $X_1 = X_0$, so the predictions are for new units having the same covariate values as the sampled units. The first component in (4) is ignored, and the prediction graphs in Fig. 1 show the conditional mean \hat{Y}_1 as a function of x , with Σ estimated in the conventional way by marginal maximum likelihood based on the residuals.

The preceding argument assumes that K is a proper covariance function, so Σ is positive definite, which is not the case for the models illustrated in Fig. 1. However, the results apply to generalized covariance functions under suitable conditions permitting pointwise evaluation provided that the subspace \mathcal{X} is such that K is positive semi-definite on the contrasts in \mathcal{X}^0 (Wahba, 1990).

Example 6. Functional response model. Consider a growth-curve model in which the stature or weight of each of n subjects is measured at a number of time points over the relevant period. To keep the model simple, the covariate for subject i is the schedule of

measurement times alone. By exchangeability, the response distributions for all subjects having the same schedule are equal.

If we denote by $Y_i(t)$ the measured height of subject i at time t , the simplest sort of additive growth model may be written in the form

$$Y_i(t) = \alpha_i + m(t) + \eta_i(t) + \epsilon_i(t)$$

in which α is an i.i.d. process on the subjects, m is a smooth random function of time with mean μ , η is a zero-mean process continuous in time and independent for distinct subjects, and ϵ is white noise. All four processes are assumed to be independent and Gaussian. The distributions are such that Y is Gaussian with mean $E(Y_i(t)) = \mu(t)$ and covariance matrix/function

$$\text{cov}(Y_i(t), Y_j(t')) = \sigma_\alpha^2 \delta_{ij} + \sigma_m^2 K_m(t, t') + \sigma_\eta^2 K_\eta(t, t') \delta_{ij} + \sigma_\epsilon^2 \delta_{ij} \delta_{t-t'}.$$

If the functions μ , K_m and K_η are given or determined up to a small set of parameters to be estimated, all parameters can be estimated by maximum likelihood or by marginal maximum likelihood. The fitted growth curve, or the predicted growth curve for a new subject from the same process, can then be obtained by computing the conditional expectation of $Y_{i^*}(t)$ given the data. Note that if σ_m^2 is positive, $Y_{i^*}(t)$ is not independent of the values on other subjects, so the predicted value is not the fitted mean $\mu(t)$.

The model shown above is intended to illustrate the general technique of additive decomposition into simpler processes followed by prediction for unobserved subjects. It is ultimately an empirical matter to decide whether such decompositions are useful in practice, but real data are likely to exhibit departures of various sorts. For example, the major difference between subjects may be a temporal translation, as in the alignment of time origins connected with the onset of puberty. Further, growth measurements are invariably positive and seldom decreasing over the interesting range. In addition, individual and temporal effects may be multiplicative, so the decomposition may be more suitable for log transformed process. Finally, there may be covariates, and if a covariate is time-dependent, i.e. a function of t , the response distribution at time t could depend on the covariate history.

4 Causality and counterfactuals

4.1 Notation

It is conventional in both the applied and the theoretical literature to write the linear regression model in the form

$$E(Y_i | x) = x_i^T \beta, \quad \text{var}(Y_i | x) = \sigma^2,$$

at least when it is understood that the components are independent. The notation suggests that $x^T \beta$ is the conditional mean of the random variable Y and σ^2 is the conditional variance, as if x were a random variable defined on the same probability space as Y . Despite the notation and the associated description, that is not what is meant because x is an ordinary function on the units, not a random variable.

The correct statement runs as follows. First, x is a function on the units taking values in \mathcal{R}^p . Second, the values taken by x on a finite set of n units may be listed as a matrix X of order $n \times p$ with rows indexed by sampled units. Third, to each parameter point (β, σ) , the model associates a distribution on \mathcal{R}^n by the formula $N(X\beta, \sigma^2 I_n)$, or by a similar formula for generalized linear models. In this way, the model determines a set of real-valued processes, one process for each parameter point. Each process is indexed by the finite restrictions of x , with values in the observation space \mathcal{R}^n . No conditional distributions are involved at any point in this construction.

Even though no conditional distributions are implied, the conventional notation and the accompanying description are seldom seriously misleading, so it would be pedantic to demand that they be corrected. However, there are exceptions or potential exceptions.

The linear regression model associates with each parameter point $\theta = (\beta, \sigma)$ a univariate process: it does not associate a bivariate process with a pair of parameter values. As a consequence, it is perfectly sensible to compare the probability $P_{x,\theta}(E)$ with the probability $P_{x,\theta'}(E)$ for any event $E \subset \mathcal{R}^n$. But it makes no sense to compare the random variable Y in the process determined by θ with the random variable in the process determined by θ' . A question such as ‘How much larger would Y_i have been had β_1 been 4.3 rather than 3.4?’ is meaningless within the present construction because the two processes are not defined on the same probability space. The alternative representation of a linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \sigma \epsilon_i$$

is potentially misleading on this point because it suggests the answer $(4.3 - 3.4)x_i$.

4.2 Counterfactuals

The linear regression model determines a process Y indexed by the finite restrictions of x : if $x' \neq x$ is another function on the same units, a similar construction determines a second process Y' indexed by the finite restrictions of x' . No bivariate process is associated with the pair (x, x') . As a consequence, it is perfectly sensible to compare the probability $P_{x,\theta}(E)$ with the probability $P_{x',\theta}(E)$ for any event $E \subset \mathcal{R}^n$. But it makes no sense to compare the random variable Y in the process determined by x with the random variable in the process determined by x' because these are not defined on the same probability space. Questions such as ‘If I had taken aspirin would my headache be gone?’ or ‘How much longer would longer would Cynthia Crabb have survived had she been given a high dose of chemo rather than a low dose?’ are meaningless within the present construction. Counterfactual comparisons of this literal sort, requiring a subject (unit) to be in two disjoint sets, are beyond the scope of regression models as defined here. For a more philosophical discussion, the reader is referred to Dawid (2000) who reaches a similar conclusion.

If the preceding question is not interpreted as a literal counterfactual, it is possible to make progress by interpreting the question in a distributional sense as follows. Among the infinite set of subjects on which the process is defined, a subset exists having the same covariate values as Cynthia Crabb except that they have the high dose of chemotherapy. Conceptually, there is no difficulty in supposing that there is an infinite number of such subjects, identical in all covariate respects to Cynthia Crabb except for the dose of chemotherapy. By exchangeability, all such subjects have the same survival distribution.

Provided that we are willing to interpret the question as a comparison of the actual survival time of Cynthia Crabb with the distribution of survival times for patients in this subset, the mathematical difficulty of finding Cynthia Crabb in two disjoint sets is avoided. The question may now be answered, and the answer is a distribution that may in principle be estimated given sufficient data.

It is clear from the discussion of Dawid (2000) that most statisticians are unwilling to forego counterfactual statements. The reason for this seems to be a deep-seated human need to assign credit or blame, to associate causes with observed effects, this sentence being an instance of the phenomenon it describes. My impression is that most practical workers interpret counterfactual matters such as unit-treatment additivity in this distributional sense, sometimes explicitly so (Cox, 1958a, 2000).

Another way of trying to make sense of the counterfactual question is first to define an exchangeable bivariate process on the units, thereby ensuring that all random variables exist in the mathematical sense. The first component is interpreted as the survival time at low dose, the second component is interpreted as the survival time at high dose, and the difference $Y_{i2} - Y_{i1}$ or ratio Y_{i2}/Y_{i1} is the desired counterfactual difference in survival times. The process determines a bivariate survival distribution, and the covariate serves to reduce the bivariate process to a univariate process by indicating which of the two components is observed. The bivariate process is such that the one-dimensional marginal distribution for the observables is the same as above. However, unless the observation space is bivariate, the counterfactual variable does not take values in the observation space so the issue of counterfactual prediction cannot arise. Moreover, the design provides no evidence that the mathematical counterfactual has a physical counterpart.

Model constructions involving latent or unobserved random processes are used frequently and successfully in statistical models. The net result in this case is a univariate marginal process defined on the finite restrictions of the covariate. The introduction of the latent component is technically unnecessary, but it is sometimes helpful as a pedagogical device. Provided that inferences are restricted to estimation and prediction for this marginal process, the technique is uncontroversial. Counterfactual predictions for the latent bivariate survival process are beyond the index set on which the marginal process is defined, and thus cannot be derived from the marginal model alone. Nonetheless, with heroic assumptions, such as independence of the two survival components that are unverifiable in the marginal process, counterfactual predictions for the bivariate process may be technically possible. However, the absence of a physical counterpart to the mathematical counterfactual makes it hard to understand what such a statement might mean in practice or how it might be checked.

4.3 *Exchangeability and causality*

Why did Cynthia Crabb receive the high dose when other patients with the same covariates received the low dose? If treatment assignment is done by objective randomization, exchangeability is maintained and the comparison is fair, at least in an averaged sense. If it is done on doctor's advice, either the doctor was inconsistent in dispensing her advice, or she had additional information, or the different patients had different doctors. By assumption, the doctor information is not included as an effect in the model, so the distributions are exchangeable even if the doctors are different. The propriety of this model assumption is certainly debatable: a different model could have included a block effect for doctors or

hospitals. The issue is more complicated if the protocol allows patients to select their own dose level because that information is arguably more naturally included as a component of the response. At any rate, the model distribution may depend on the protocol, and without further assumptions, there is little reason to suppose that the response distribution under the randomization protocol is the same as the distribution under the elective protocol.

A central theme in much of causal inference is the attempt to deduce or to predict in a probabilistic sense what would have occurred had the design protocol been different than it actually was. Since the theory of prediction for processes does not extend beyond the index set on which the process is defined, this sort of prediction requires an explicit broader and perhaps fundamentally different foundation. One can envisage a compound doubly-randomized design in which the first arm is a conventional randomized experiment, the response on each individual being 5-year survival. In the elective arm, patients are permitted to select the drug or dose, so the response is bivariate. This sort of process, in which the observation space itself depends on the covariate, is certainly unconventional, but it is not fundamentally different from the definition given in section 2. The definition of regression exchangeability is unchanged, but would usually be considered in a modified form in which the conditional distribution of survival times given the chosen treatment is the same as the distribution of survival times for the assigned treatment in the randomized arm. In other words, the survival distribution depends on treatment and other baseline covariates, but not on whether the treatment is randomly assigned or freely selected. With this modified concept of exchangeability in the extended process, it is possible to extrapolate from the randomized experiment, by making predictions for the outcomes under a different protocol.

It is invariably the case in matters of causal assessment that closer inspection reveals additional factors or an intermediate sequence of events that could affect the interpretation of treatment contrasts had they been included in the model, i.e. if a different model had been used. A good example can be found in the paper by Versluis, Schmitz, von der Heydt and Lohse (2000) on the sound-causing mechanism used by the snapping shrimp *Alpheus heterochaelis*. Since the shrimp tend to congregate in large numbers, the combined sound is appreciable and can interfere with naval sonar. The loud click is caused by the extremely rapid closure of the large snapper claw, in the sense that a sound is heard every time the claw snaps shut and no sound is heard otherwise. It had been assumed that the sound was caused by mechanical contact between hard claw surfaces, and the preceding statement had been universally interpreted in that way. However, closer inspection reveals a previously unsuspected mechanism, such that the claw is not the source of the sound. During the rapid claw closure a high-velocity water jet is emitted with a speed that exceeds cavitation conditions, and the sound coincides with the collapse of the cavitation bubble not with the closure of the claw. In light of this information, what are we to say of causal effects? The initial statement that the rapid closure of the claw causes the sound, although defensible in many ways, suggests a causal mechanism that is now known to be false. Whether or not the statement is legally correct, it is scientifically misleading and must be avoided in circumstances where it could lead to false conclusions or inappropriate decisions. The complementary statement that the closure of the claws does not cause the sound, although equally defensible, is no better.

Rarely, if ever, does there exist a most proximate cause for any observed phenomenon, so the emphasis must ultimately be on mechanisms. Confusion results when the words

‘cause’ or ‘causal’ are used with one mechanism, or no specific mechanism, in mind, and interpreted in the context of a different mechanism. For clinical trials where biochemical pathways are complicated and unlikely to be understood in sufficient detail, the word mechanism is best replaced by protocol. The natural resolution is to refuse to discuss causality except in the context of a specific mechanism or protocol, which might, but need not, involve manipulation or intervention. Thus, the closure of the claw causes the sound through an indirect mechanism involving cavitation. This statement does not exclude the possibility that cavitation itself is a complex process with several stages.

Unless the protocol is well defined, an unqualified statement concerning the causal effect of a drug or other therapy is best avoided. Thus, following a randomized trial in which a drug is found to increase the 5-year survival rate, the recommendation that it be approved for general use is based on a model assumption, the prediction that a similar difference will be observed on average between those who elect to use the drug and those who elect not to use it. Equality here is a model assumption, a consequence of regression exchangeability in the modified sense discussed above. As with all model assumptions, this one may prove to be incorrect in specific applications. Unlike counterfactuals, the assumption can be checked in several ways, by direct comparison in a compound doubly-randomized experiment, by comparisons within specific sub-groups or by comparing trial results with subsequent performance. In the absence of exchangeability, there is no mathematical reason to suppose that the 5-year survival rate among those who elect to use the drug should be similar to the rate observed in the randomized experiment. It is not difficult to envisage genetic mechanisms such that those who elect not to use the drug have the longer 5-year survival, but all such mechanisms imply non-exchangeability or the existence of potentially identifiable sub-groups.

5 Exchangeable block models

5.1 Block factor

The distinction between a block factor and a treatment factor, closely related to the distinction between fixed and random effects, is a source of confusion and anxiety for students and experienced statisticians alike. As a practical matter, the distinction between a block factor and a treatment or classification factor is not a rigid one. The key distinguishing feature is the anonymous or ephemeral nature of the levels of a block factor. Cox (1984) uses the term non-specific, while Tukey (1974) prefers the more colourful phrase ‘named and faceless values’ to make a similar distinction.

Even if it is more rigid and less nuanced, a similar distinction can be made in the mathematics. A block factor B is defined as an equivalence relation on the units, a symmetric binary function $B: \mathcal{U} \times \mathcal{U} \rightarrow \{0, 1\}$ that is reflexive and transitive. Equivalently, but more concretely, B is a partition of the units into disjoint non-empty subsets called blocks such that $B(i, j) = 1$ if units i, j are in the same block and zero otherwise. The number of blocks may be finite or infinite. For the observed set of n units, B is a symmetric positive semi-definite binary matrix whose rank is the number of blocks in the sample.

A treatment factor $x: \mathcal{U} \rightarrow \Omega$ may be converted into a block factor by the elementary device of ignoring factor labels, a forgetful transformation defined by

$$B(i, j) = \begin{cases} 1 & \text{if } x(i) = x(j) \\ 0 & \text{otherwise.} \end{cases}$$

If X is the incidence matrix for the treatment factor on the sampled units, each column of X is an indicator function for the units having that factor level, and $B = XX^T$ is the associated block factor. It is not possible to convert a block factor into a treatment factor because the label information, the names of the factor levels, is not contained in the block factor.

One can define a process on the finite restrictions of the equivalence relation B in much the same way that a regression process is defined on the finite restrictions of a function or covariate. Whether it is indexed by the finite restrictions of x or B , a process is called block-exchangeable if the following condition is satisfied for each n . Two samples $\{i_1, \dots, i_n\}$ and $\{j_1, \dots, j_n\}$, ordered in such a way that $B(i_r, i_s) = B(j_r, j_s)$ for each r, s , determine the same distribution on \mathcal{R}^n . The implications of block-exchangeability are different from regression exchangeability because the label information is not available. Block exchangeability automatically implies that all one-dimensional marginal distributions are the same, and there are only two distinct two-dimensional marginal distributions depending on whether $B(i, j)$ is true or false.

For a sample of size n , the image or range of X is the same subspace $\mathcal{X} \subset \mathcal{R}^n$ as the range of B in \mathcal{R}^n , the set of functions that are constant on each block. Consider the following four linear Gaussian specifications:

- (i) $Y \sim N(X\beta, \sigma^2 I_n)$,
- (ii) $Y \sim N(B\gamma, \sigma^2 I_n)$,
- (iii) $Y \sim N(\mu \in \mathcal{X}, \sigma^2 I_n)$,
- (iv) $Y \sim N(\mu \in \mathbf{1}, \sigma^2 I_n + \sigma_b^2 B)$.

The non-identifiability of $\gamma \in \mathcal{R}^n$ can be resolved if necessary by restriction to \mathcal{X} . In the sense that they determine precisely the same set of distributions on the observation space, the first three forms are equivalent up to re-parameterization. Even so, the models are very different in crucial respects.

In (i), β is a function on the treatment levels, so inference for specific levels or specific contrasts is immediate. In (ii) or (iii), one can transform from γ to $\beta = X^T\gamma$ only if the information about treatment levels is available, i.e. coded in the indexing system of the process. As a purely mathematical question, the first three versions are equivalent only if the process is indexed by the finite restrictions of the treatment factor x . Expressions (ii)-(iv) may be regarded as processes indexed by the finite restrictions of the block factor B , in which case the information concerning treatment labels is unavailable for inferential purposes such as estimation. As processes indexed on B , the second and third forms are equivalent, but they are not block-exchangeable and they are not equivalent to the first. As models indexed by B , one can talk of consistent estimation in (iv) where the parameter space is a fixed set, but the concept of consistent estimation is meaningless for the full parameter in (ii) or (iii), where the parameter space depends on the sample or design.

The fourth form, the standard random effects model with i.i.d. block effects, is different from the other three: it is block-exchangeable. The expression may be regarded as defining a process on the finite restrictions of x , or a process on the finite restrictions of the block factor B . In that sense (iv) is ambiguous, as are (ii) and (iii). In practice, it would usually be assumed that the block names are available for use if necessary, as for example in animal breeding experiments (Robinson, 1991). As a process indexed by x , it

is possible to make inferences or predictions about contrasts among specific factor levels. The conditional distribution of the response on a new unit with $x(\cdot) = 1$ is Gaussian with mean and variance

$$\frac{\sigma^2\mu + n_1\sigma_b^2\bar{y}_1}{\sigma^2 + n_1\sigma_b^2}, \quad \sigma^2\left(1 + \frac{\sigma_b^2}{n\sigma_b^2 + \sigma^2}\right), \quad (5)$$

where n_1 is the number of units in the sample for which $x(\cdot) = 1$ and \bar{y}_1 is the average response. In practice, the parameter values must first be estimated from the available data. As a process defined on B , inference for specific factor levels or specific contrasts is impossible. Nonetheless, since each new unit i^* comes with block information in the form of the extended equivalence relation B , it is possible to make predictive statements about new units such that $B(i^*, 4) = 1$, i.e. new units that are in the same block as unit 4 in the sample. The formula for the conditional distribution is much the same as that given above, so the mathematical predictions have a similar form except that the block does not have a name. It is also possible, on the basis of the model, to make predictive statements about new units that are not in the same block as any of the sample units. Whether such predictions are reliable is a matter entirely dependent on specific details of the application.

It is worth remarking at this point that, for a balanced design, the sufficient statistic for model (iv) is the sample mean plus the between- and within-block mean squares. Even for an unbalanced design, an individual block mean such as \bar{y}_1 is not a function of the sufficient statistic. Accordingly, two observation points $y \neq y'$ producing the same value of the sufficient statistic will ordinarily give rise to different predictions in (3) or (5). In other words, the conclusions are not a function of the sufficient statistic. One of the subtleties of the likelihood principle as stated, for example, by Cox and Hinkley (1970, p. 39) or Berger and Wolpert (1988, p. 19) is the clause ‘conclusions about θ ’, implying that it is concerned solely with parameter estimation. Since (3) and (5) are statements concerning events in the observation space, not estimates of model parameters or statements about θ , there can be no violation of the likelihood principle.

5.2 Example: homologous factors

We consider in this section a further, slightly more complicated, example of an exchangeable block model in which the covariate $x = (x_1, x_2)$ is a pair of homologous factors taking values in the set $\Omega = \{1, \dots, n\}$ (McCullagh, 2000). If there is only a single replicate, the observation Y is a square matrix of order n with rows and columns indexed by the same set of levels. More generally, the design is said to be balanced with r replicates if, for each cell (i, j) there are exactly r units u for which $x(u) = (i, j)$. For notational simplicity, we sometimes assume $r = 1$, but in fact the design need not be balanced and the assumption can lead to ambiguities regarding the model.

The following models are block-exchangeable, but the processes are more naturally indexed by the finite restrictions of the classification factor x rather than by the block factor:

$$\begin{aligned} Y_{ij} &= \mu + \eta_i + \eta_j + \epsilon_{ij} \\ Y_{ij} &= \mu + \eta_i - \eta_j + \epsilon_{ij} \\ Y_{ij} &= \eta_i - \eta_j + \epsilon'_{ij} \end{aligned}$$

In these expressions η/σ_η and ϵ/σ_ϵ are independent i.i.d. standard Gaussian processes, so the parameter space for the first two consists of the three components $(\mu, \sigma_\eta^2, \sigma_\epsilon^2)$. In the

third model, $\epsilon'_{ij} = -\epsilon'_{ji}$, so the observation matrix Y is skew-symmetric.

These expressions suggest that Y is a process indexed by ordered pairs of integers, and in this respect the notation is misleading. The ‘correct’ version of the first model is

$$Y(u) = \mu + \eta_{x_1(u)} + \eta_{x_2(u)} + \eta'_{x(u)} + \epsilon(u) \quad (5)$$

making it clear that Y and ϵ are processes indexed by the units, and there may be several units such that $x(u) = (i, j)$. In plant breeding experiments, the units such that $x(u) = (i, i)$ are called self-crosses; in round-robin tournaments, self-competition is usually meaningless, so there are no units such that $x(u) = (i, i)$. In the absence of replication, the interaction process η' and the residual process ϵ are not separately identifiable: only the sum of the two variances is estimable. However, absence of replication in the design does not imply absence of interaction in the model. To put it another way, two distinct models may give rise to the same set of distributions for a particular design. Aliasing of interactions in a fractional factorial design is a well-known example of the phenomenon.

If there is a single replicate, the three models may be written in the equivalent distributional form as follows:

$$Y \sim N(\mu\mathbf{1}, \sigma_\eta^2 K + \sigma_\epsilon^2 I_{n^2})$$

$$Y \sim N(\mu\mathbf{1}, \sigma_\eta^2 K' + \sigma_\epsilon^2 I_{n^2})$$

$$Y \sim N(0, \sigma_\eta^2 K' + \sigma_\epsilon^2 I'_{n^2}).$$

The matrices K, K', I' are symmetric of order $n^2 \times n^2$ given by

$$K_{ij,kl} = \delta_{ik} + \delta_{jl} + \delta_{il} + \delta_{jk},$$

$$K'_{ij,kl} = \delta_{ik} + \delta_{jl} - \delta_{il} - \delta_{jk},$$

$$I'_{ij,kl} = \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}.$$

Note that δ_{ik} is the block factor for rows, δ_{jl} is the block factor for columns, and the remaining terms δ_{il}, δ_{jk} are meaningless unless the two factors have the same set of levels. Each of the three model distributions is invariant under permutation, the same permutation being applied to rows as to columns. Accordingly, the models depend on the rows and columns as block factors, not as classification factors.

In the standard Bradley-Terry model for ranking competitors in a tournament, the component observations are independent, and the competitor effect $\{\eta_i\}$ is a parameter vector to be estimated (Agresti 1996, p. 436). Such models are closed under permutation of factor levels and under restriction of levels, but they are not invariant, and thus not block-exchangeable. By contrast, all three models shown above are block-exchangeable, and competitor effects do not occur in the parameter space. To predict the outcome of a match between competitors i, j , we first estimate the variance components by maximum likelihood. In the second stage, the conditional distribution $Y(u^*) | Y$ for a new unit such that $x(u^*) = (i, j)$ is computed by the standard formulae for conditional distributions, and this is the basis on which predictions are made. This exercise is straightforward provided that the variance components required for prediction in (5) are identifiable at the design. An allowance for errors of estimation along the lines of Barndorff-Nielsen and Cox (1996) is also possible.

6 Concluding remarks

Exchangeability is a consequence of requiring distributions to be well-behaved under sub-sampling and finite permutation, the two operations that generate the category of injective maps on finite sets. As such, exchangeability has implications in diverse areas well beyond the range of topics considered in this paper. From this vantage, one can talk of exchangeable random trees, or exchangeable tree-valued processes as a natural extension of real-valued processes. The most obvious statistical applications are connected with prior processes for factorial effects in Bayesian models. The following are some additional suggestions for future work.

- (i) Construct an exchangeable prior process for treatment effects that permits multiple comparisons in a formal Bayesian framework.
- (ii) Construct a partially exchangeable prior process in the sense of Aldous (1981) for the main effects and interactions in a factorial model that has the following property. Two factors that have large main effects are likely to exhibit appreciable interaction. Explore the consequences of using such a prior in a Bayesian model.
- (iii) What is the most natural extension of example 5 to non-linear or non-Gaussian models? One possibility is to use penalized likelihood or h -likelihood (Lee and Nelder, 1996). Is there a version that avoids extra-likelihood criteria.
- (iv) Modify the prediction model in example 5 for density prediction.
- (v) Modify the model in section 5.2 so that it can be used for binary responses.
- (vi) What is the natural notion of exchangeability for trees or tree-indexed processes? What are the implications of using such a process for the coefficients in a wavelet decomposition?
- (vii) For a replicate experiment on the same process, the test statistic $T(Y^*)$ may or may not exceed the value $T(Y)$ observed in the original experiment: the two statistics are exchangeable and thus have the same distribution. The exceedance probability or p -value is a prediction on the duplicate sample space $\text{pr}(T(Y^*) \geq T(Y) | Y)$, and as such is not subject to the likelihood principle. The subsequent inference, that a small p -value is evidence against the model or null hypothesis, if interpreted as evidence in favour of specific parameter points in a larger parameter space, is an inference potentially in violation of the likelihood principle. Bearing in mind the distinction between estimation and prediction, clarify the nature of the LP violation (Berger and Wolpert, 1988).

References

- Agresti, A. (1996) *Categorical Data Analysis*. J. Wiley & Sons, New York.
- Aldous, D. (1981) Representations for partially exchangeable arrays of random variables. *J. Mult. Analysis* **11**, 581–598.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1996) Prediction and asymptotics. *Bernoulli* **2**, 319–340.
- Berger, J. and Wolpert, R.L. (1988) *The Likelihood Principle*. IMS Lecture Notes **6.**,
- Besag, J. (2002) Discussion of McCullagh (2002) *Ann. Statist.* **30**, 1267–1277.
- Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc. B* **26**, 211–252.
- Brøns, H. (2002) Discussion of McCullagh (2002) *Ann. Statist.* **30**, 1279–1283.

- Cox, D.R. (1958a) *Planning of Experiments*. J. Wiley & Sons, New York.
- Cox, D.R.(1958b) The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc. B* **20**, 215–242.
- Cox, D.R. (1984) Interaction (with discussion). *Int. Statist. Rev.***52**, 1–31.
- Cox, D.R. (2000) Comment on Dawid (2000). *J. Am. Statist. Assoc.* **95**, 424–425.
- Cox, D.R. and Snell, E.J. (1981) *Applied Statistics*. Chapman and Hall, London.
- Cox, D.R. and Hinkley, D.V. (1970) *Theoretical Statistics*. Chapman and Hall, London.
- Dawid, A.P. (2000) Causal inference without counterfactuals (with discussion). *J. Am. Statist. Assoc.* **95**, 407–448.
- Efron, B. (2001) Selection criteria for scatterplot smoothers. *Ann. Statist.* **29**, 470–454.
- deFinetti, B. (1974) *Theory of Probability, vols 1, 2*. J. Wiley & Sons, New York.
- Green, P. and Silverman, B. (1994) *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.
- Greenland, S., Robins, J.M. and Pearl, J. (1999), Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.
- Kong,, A. McCullagh, P., Meng, X-L., Nicolae, D. and Tan, Z. (2003) A theory of statistical models for Monte Carlo integration (with discussion). *J. R. Statist. Soc. B* **65**, 585–618.
- Lee, J. and Nelder, J.A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc. B* **58**, 619–678.
- McCullagh, P. (2000) Invariance and factorial models (with discussion) (2000) *J. R. Statist. Soc. B* **62**, 209–256.
- McCullagh, P. (2002) What is a statistical model? (with discussion). *Ann. Statist.* **30**, 1225–1310.
- Nelder, J.A. (1977) A reformulation of linear models (with discussion). *J. R. Statist. Soc. A* **140**, 48–77.
- Robinson, G.K. (1991) That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* **6**, 15–50.
- Stein, M.L. (1999) *Interpolation of Spatial data*. Springer-Verlag, New York.
- Tukey, J.W. (1974) Named and faceless values: An initial exploration in memory of Prasanta C. Mahalanobis. *Sankhya*, Series A **36**, 125–176.
- Vardi, Y. (1985) Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.
- Versluis, M., Schmitz, B., von der Heydt, A. and Lohse, D. (2000) How snapping shrimp snap: through cavitating bubbles. *Science* **289**, 2114–2117.
- Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–1402.
- Wahba, G. (1990) *Spline Models for Observational Data*. SIAM Philadelphia.