# 4
# Exchangeability and regression models

**Peter McCullagh**

## 4.1 Introduction

Sir David Cox's statistical career and his lifelong interest in the theory and application of stochastic processes began with problems in the wool industry. The problem of drafting a strand of wool yarn to near-uniform width is not an auspicious starting point, but an impressive array of temporal and spectral methods from stationary time series were brought to bear on the problem in Cox (1949). His ability to extract the fundamental from the mundane became evident in his discovery or construction of the eponymous Cox process in the counting of neps in a sample of wool yarn Cox (1955$a$). Subsequent applications include hydrology and long-range dependence (Davison and Cox 1989, Cox 1991), models for rainfall (Cox and Isham 1988, Rodriguez-Iturbe *et al.* 1987, 1988), and models for the spread of infectious diseases (Anderson *et al.* 1989).

At some point in the late 1950s, the emphasis shifted to statistical models for dependence, the way in which a response variable depends on known explanatory variables or factors. Highlights include two books on the planning and analysis of experiments, seminal papers on binary regression, the Box–Cox transformation, and an oddly titled paper on survival analysis. This brief summary is a gross simplification of Sir David's work, but it suits my purpose by way of introduction because the chief goal of this chapter is to explore the relation between exchangeability, a concept from stochastic processes, and regression models in which the observed process is modulated by a covariate.

A stochastic process is a collection of random variables, $Y_1, Y_2, \ldots$, usually an infinite set, though not necessarily an ordered sequence. What this means is that $\mathcal{U}$ is an index set on which $Y$ is defined, and for each finite subset $S = \{u_1, \ldots, u_n\}$ of elements in $\mathcal{U}$, the value $Y(S) = \big(Y(u_1), \ldots, Y(u_n)\big)$ of the process on $S$ has distribution $P_S$ on $\mathbb{R}^S$. This chapter is a little unconventional in that it emphasizes probability distributions rather than random variables. A real-valued process is thus a consistent assignment of probability distributions to observation spaces such that the distribution $P_n$ on $\mathbb{R}^n$ is the marginal distribution of $P_{n+1}$ on $\mathbb{R}^{n+1}$ under deletion of the relevant co-ordinate. A notation such as $\mathbb{R}^n$ that puts undue emphasis on the incidental, the dimension of the observation space, is not entirely satisfactory. Two samples of equal size need not have the same distribution, so we write $\mathbb{R}^S$ rather than $\mathbb{R}^n$ for the set of real-valued functions on the sampled units, and $P_S$ for the distribution.

90                     *Exchangeability and regression models*

A process is said to be exchangeable if each finite-dimensional distribution is symmetric, or invariant under co-ordinate permutation. The definition suggests that exchangeability can have no role in statistical models for dependence, in which the distributions are overtly non-exchangeable on account of differences in covariate values. I argue 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 that in mind. Second, regression exchangeability has a number of fundamental implications connected with lack of interference (Cox 1958$a$) and absence of unmeasured covariates (Greenland *et al.* 1999). This chapter 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; see Examples 5 and 6 below.

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 $50 \pm 15$ kg/ha is simply a prediction concerning the likely difference of yields under similar circumstances in future 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 4.5.1 below. Second, 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. Finally, parameter estimation is equivariant under non-linear transformation: the evidence from the data in favour of $g(\theta) \in S$ is the same as the evidence for $\theta \in g^{-1}S$. Pointwise prediction, in the sense of the conditional mean of the response distribution on a new unit, is not equivariant under non-linear response transformation: the mean of $g(Y)$ is not $g(\mathrm{E}Y)$.

We do not aim to contribute to philosophical matters such as where the model comes from, nor to answer 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, covariates and observations, and the physical objects that they represent. This correspondence is usually implicit in most discussions of models. Despite its importance, the present chapter has little to say on the matter.

## 4.2 Regression models

### 4.2.1 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 each potential sample, survey or experiment there corresponds an observation space, and it is the function of a process to associate a probability distribution with 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 $Y \in \mathbb{R}^n$. Other processes exist in which an observation is a more complicated object such as a tree or partition (Kingman 1978), but in a regression model where we have one measurement on each unit, the observation space is invariably a product set, such as $\mathbb{R}^n$ or $\{0,1\}^n$, of responses or functions on the sampled units.

The processes with which we are concerned here are defined on the set $\mathcal{U}$ of statistical units and observed on a finite subset called the sample. The entire set or population $\mathcal{U}$ 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. A process $P$ is a function that associates with each finite sample $S \subset \mathcal{U}$ of size $n$ a distribution $P_{\mathrm{S}}$ on the observation space $\mathbb{R}^{\mathrm{S}}$ of dimension $n$. Let $S \subset S'$ be a subsample, and let $P_{\mathrm{S}'}$ be the distribution on $\mathbb{R}^{\mathrm{S}'}$ determined by the process. If logical contradictions are to be avoided, $P_{\mathrm{S}}$ must be the marginal distribution of $P_{\mathrm{S}'}$ under the operation of co-ordinate deletion, that is, 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 $P_{\mathrm{S}}$ is symmetric, or invariant under co-ordinate 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, Kingman 1978). Despite the attractions of the theory, the conventional definition of exchangeability 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 or sequence of random variables. However, the Kolmogorov extension theorem guarantees the existence of a random variable $Y$ taking values in $\mathbb{R}^{\mathcal{U}}$ 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. The process $Y$ is a function on the units, usually real-valued but possibly vector-valued, so there is one value for each unit.

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 1958$a$). The latter definition implies random or deliberate 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.

### 4.2.2   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 $x\colon \mathcal{U} \mapsto \mathbb{R}$ or $x\colon \mathcal{U} \mapsto \mathbb{R}^p$ 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 $x\colon \mathcal{U} \mapsto \Omega$ taking values in a set $\Omega$ 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 work. For the moment at least, a covariate is a function $x\colon \mathcal{U} \mapsto \Omega$ taking values in an arbitrary set $\Omega$. 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 $50 \pm 15$ 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, a prediction about the difference of infinite averages. 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. My impression is that the prediction step is so obvious and natural that its mathematical foundation is taken for granted.

Let $x \colon \mathcal{U} \mapsto \Omega$ 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 $P_\mathrm{S}$ on $\mathbb{R}^\mathrm{S}$, and that these distributions are mutually compatible with respect to subsampling of units. A process having the following property for every integer $n$ is called regression exchangeable or exchangeable modulo $x$.

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

Exchangeability modulo $x$ is the condition that if $x$ takes the same value on two samples, the distributions are also the same. Any distinction between units, such as name or identification number that is not included as a component of $x$, has no effect on the response distribution. The majority of models that occur in practical work have this property, but Example 3 below shows that there are exceptions. The property is a consequence of the definition of a statistical model as a functor on a certain category, the injective maps, in the sense of McCullagh (2002) or Brøns (2002), provided that the parameter space is a fixed set independent of the design.

Exchangeability in the conventional sense implies that two samples of equal size have the same response distribution regardless of the covariate values, so exchangeability implies regression exchangeability. For any function $g$ defined on $\Omega$, 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 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 *et al.* 1999). This is a key assumption in the literature on causality. At this stage, no structure has been imposed on the set $\mathcal{U}$, and no structure has been ruled out. In most discussions of causality the units have a temporal structure, so the observation on each unit is a time sequence, possibly very brief, and the notion of 'the same unit at a later time' is well defined (Lindley 2002, Singpurwalla 2002, Pearl 2002). The view taken here is that causality is not a property of a statistical model but of its interpretation, which is very much context dependent. For example, Brownian motion as a statistical model has a causal interpretation

in terms of thermal molecular collisions, and a modified causal interpretation may be relevant to stock-market applications. In agricultural field work where the plots are in linear order, Brownian motion may be used as a model for one component of the random variation, with no suggestion of a causal interpretation. This is not fundamentally different from the use of time-series models and methods for non-temporal applications (Cox 1949). Thus, where the word causal is used, we talk of a causal interpretation rather than a causal model. We say that any difference between distributions is associated with a difference between covariate values in the ordinary mathematical sense without implying a causal interpretation.

The second consequence of regression exchangeability, that the distribution of $Y_i$ depends only on the value of $x$ on unit $i$, and not on the values on other units, is a key assumption in experimental design and in clinical trials called lack of interference (Cox 1958$a$, p. 19). 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 discard 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; see Example 3.

### 4.2.3   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 baseline 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 1958$a$), 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 1958$c$), 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 $P_{1,v} = P_{0,v}$ for each $v$, the model distributions do not depend on the treatment level, and the treatment effect $\tau(v)$ is identically 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 the baseline covariates $v$, 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 subgroups, we would usually want to know the value in each of the subgroups. 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$a$) 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.

## 4.3 Examples of exchangeable regression models

The majority of 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 $\mathcal{V}$, which we call the covariate space. Each point $\theta = (\beta, \sigma)$ in the parameter space consists of a linear functional $\beta \in \mathcal{V}'$, where $\mathcal{V}'$ is the space dual to $\mathcal{V}$, plus a real number $\sigma$, and the parameter space consists of all such pairs. If the value of the linear functional $\beta$ at $v \in \mathcal{V}$ is denoted by $v^{\mathrm{T}}\beta$, the value on unit $i$ is $x_i^{\mathrm{T}}\beta$. In the classical linear regression model the response distribution for unit $i$ is normal with mean equal to $x_i^{\mathrm{T}}\beta$ and variance $\sigma^2$. The model may be modified in a number of minor ways, for example by restricting $\sigma$ to be non-negative to ensure identifiability.

96 *Exchangeability and regression models*

From the point of view of regression 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 $\eta_i = x_i^{\mathrm{T}}\beta$ and $Y_i$ is Bernoulli with parameter $1/\{1 + \exp(-\eta_i)\}$ 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$a$). 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 4.5 is an entirely different sort of mathematical object from which the concept of a set of levels is missing.

Let $\Omega$ be the set of treatment levels, and let $\tau\colon \Omega \mapsto \mathbb{R}$ be a function on the levels. In conventional statistical parlance, $\tau$ 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. We note in passing that the parameter space $\mathbb{R}^{\Omega}$ 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 should this be required. Without a preferred basis, no similar construction exists for a general linear functional $\beta$ 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 $\sigma^2$. The parameter space is the set of all pairs $(\tau, \sigma)$ in which $\tau$ is a function on the levels and $\sigma$ is a real number. For each parameter point, condition (RE) is satisfied by the process. 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 $\mathbb{R}^2$ for each unit. The observation space corresponding to a set of $n$ units is $(\mathbb{R}^2)^n$. 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

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

in which $\alpha$ is a function on the subjects, and $\delta$ is a common temporal trend. The parameter space consists of all functions $\alpha$ on the $n$ subjects, all functions $\tau$ on the treatment levels, plus the two scalars $(\delta, \sigma)$, so the effective dimension is $n+3$ for a design with $n$ subjects and two treatment levels.

This model is not regression exchangeable because two units $i, j$ having the same treatment regime $(x_{i1}, x_{i2}) = (x_{j1}, x_{j2})$ do not have the same response distribution for all parameter values: the difference between the two means is $\alpha_i - \alpha_j$. This model is a little unusual in that the parameter space is not a fixed set independent of the design: the dimension depends on the sample size. Nonetheless, it is a statistical model in the sense of McCullagh (2002).

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

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

with a fixed parameter space independent of the design. Two samples having the same covariate values also have the same distribution, so condition (RE) is satisfied. The temporal effect $\delta$ is indistinguishable from a carry-over 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

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

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 $\mathbb{R}$ such that each integral $\int w_i(x)\,\mathrm{d}F(x)$ is finite (Vardi 1985, Kong *et al.* 2003). For each $F$ in the parameter space, the response distribution on unit $i$ is the weighted distribution such that $\mathrm{d}F_i(x) \propto w_i(x)\,\mathrm{d}F(x)$. 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 $|x|$, a phenomenon that arises in a wide range of applications from the wool industry (Cox 1962, §5.4) to stereology and auditing (Cox and Snell 1979). 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,

$$Y_i = \beta_0 + \beta_1 x_i + \eta(x_i) + \epsilon_i. \tag{4.1}$$

The coefficients $(\beta_0, \beta_1)$ are parameters to be estimated, $\epsilon$ is a process with independent $N(0, \sigma^2)$ components, and $\eta$ is a zero-mean stationary process on the real line, independent of $\epsilon$, with covariance function

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

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

$$Y \sim N(X\beta, \sigma^2 I_n + \sigma_\eta^2 V), \tag{4.2}$$

where $V_{ij} = K(x_i, x_j)$ are the components of a positive semidefinite matrix. For each value of $(\beta, \sigma^2, \sigma_\eta^2)$, two samples having the same covariate values determine the same distribution on the observation space, so this model is regression-exchangeable with non-independent components.

The linear combination $\eta(x_i) + \epsilon_i$ in eqn (4.1) is a convenient way of describing the distribution of the process as a sum of more elementxchary processes. The treacherous aspect of the notation lies in the possibility that $\eta$ or $\epsilon$ 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 $\eta$, is also regression exchangeable, but very different from eqn (4.2).

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 estimation, or by some closely related procedure such as residual maximum likelihood estimation—or REML—for the variance components followed by weighted least squares estimation of the regression parameters. With prediction in mind, Wahba (1985) recommends use of generalized crossvalidation over residual maximum likelihood estimation 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 residual maximum likelihood 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

$$\hat{Y}_{i^*} = \mathrm{E}\{Y(i^*) \,|\, Y\} = \beta_0 + \beta_1 x^* + k^* \Sigma^{-1}(Y - \mu), \qquad (4.3)$$

where $\mu, \Sigma$ 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 (4.3) and constant variance independent of $Y$. Interpolation features prominently in the geostatistical literature, where linear prediction is called kriging (Stein 1999).

If $\eta$ is Brownian motion with generalized covariance function $-|x - x'|$ on contrasts, the prediction function (4.3) 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 (Matérn 1986) or it could be a convex combination of simple covariance functions. The cubic and linear splines illustrated in Figure 4.1 are obtained by fitting model (4.2) to simulated data (Wahba 1990, p. 45), in which $\eta(x)$ is the smooth function shown as the dashed line.

In statistical work, the adjective 'Bayesian' usually refers to the operation of converting a probability $\mathrm{pr}(A \,|\, B)$ into a probability of the form $\mathrm{pr}(B \,|\, A)$ by supplying additional information and using Bayes's theorem. The transformation from the joint distribution of $(Y^*, Y)$ as determined by the process (4.2), 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 eqn (4.3) 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 $\pi$ on the parameters $(\beta, \sigma^2, \sigma_\eta^2)$, and uses the likelihood function to obtain the posterior distribution. The process is such that
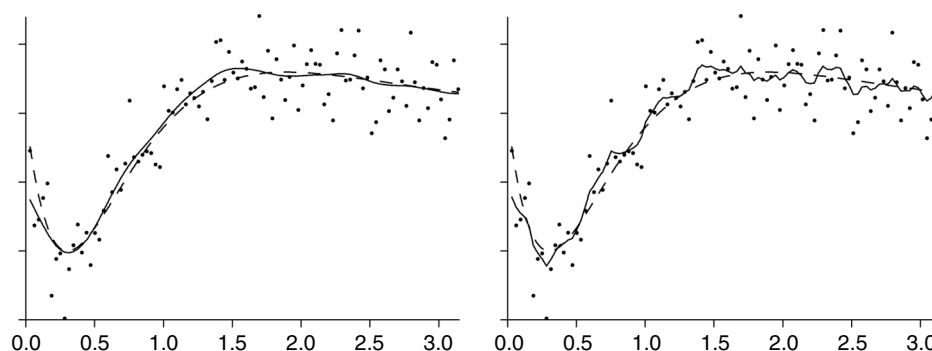


FIG. 4.1. Spline prediction graphs (solid lines) fitted by residual maximum likelihood, cubic on the left, linear on the right. The ideal predictor $\eta(x)$ (dashed line) is taken from Wahba (1990, p. 45).

the predictive distribution for a new unit has mean (4.3) 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 eqn (4.3).

Since $\eta$ 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 $\eta$ 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 Figure 4.1, the maximum log residual 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 (4.3) 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 $\Sigma$ for the joint distribution in $\mathbb{R}^S$ 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,

$$\begin{pmatrix} \hat{Y}_0 \\ \hat{Y}_1 \end{pmatrix} = \begin{pmatrix} P_0 & 0 \\ X_1(X_0^{\mathrm{T}}\Sigma_{00}^{-1}X_0)^{-1}X_0^{\mathrm{T}}\Sigma_{00}^{-1} & 0 \end{pmatrix} \begin{pmatrix} Y_0 \\ \star \end{pmatrix} + \begin{pmatrix} Q_0 & 0 \\ \Sigma_{10}\Sigma_{00}^{-1}Q_0 & 0 \end{pmatrix} \begin{pmatrix} Y_0 \\ \star \end{pmatrix}, \quad (4.4)$$

where $P_0 = X_0(X_0^{\mathrm{T}}\Sigma_{00}^{-1}X_0)^{-1}X_0^{\mathrm{T}}\Sigma_{00}^{-1}$, $Q_0 = I - P_0$, and $\star$ is the unobserved value. Evidently $\hat{Y}_0 = Y_0$ as it ought. The first transformation is the least-squares projection $P \colon \mathbb{R}^S \mapsto \mathbb{R}^S$ onto $\mathcal{X} \subset \mathbb{R}^S$ of dimension $p$. The second transformation is a projection $T \colon \mathbb{R}^S \mapsto \mathbb{R}^S$, self-adjoint with respect to the inner product $\Sigma^{-1}$, and thus an orthogonal projection. Its kernel consists of all vectors of the form $(x, \star)$, i.e. all vectors in $\mathcal{X} + \mathbb{R}^{S_1}$ 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 eqn (4.4) is ignored, and the prediction graphs in Figure 4.1 show the conditional mean $\hat{Y}_1$ as a function of $x$, with $\Sigma$ 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 $\Sigma$ is positive definite, which is not the case for the models illustrated in Figure 4.1. 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 semidefinite 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. 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 $\alpha$ is an exchangeable process on the subjects, $m$ is a smooth random function of time with mean $\mu$, $\eta$ is a zero-mean process continuous in time and independent for distinct subjects, and $\epsilon$ is white noise. All four processes are assumed to be independent and Gaussian. The distributions are such that $Y$ is Gaussian with mean $\mathrm{E}\{Y_i(t)\} = \mu(t)$ and covariance matrix/function

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

If the functions $\mu$, $K_m$ and $K_\eta$ 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 $\sigma_m^2$ 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 satisfies condition (RE), so it is regression exchangeable even though the components are not independent. It 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.

## 4.4   Causality and counterfactuals

### 4.4.1   Notation

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

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

when it is understood that the components are independent. The notation suggests that $x^{\mathrm{T}}\beta$ is the conditional mean of the random variable $Y$ and $\sigma^2$ 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 $\mathbb{R}^p$, and 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. Second, to each parameter point $(\beta, \sigma)$, the model associates a distribution on $\mathbb{R}^n$ 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 $\mathbb{R}^n$. No conditional distributions are involved at any point in this construction. At the same time, the possibility that the regression model has been derived from a bivariate process by conditioning on one component is not excluded.

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 $E \subset \mathbb{R}^n$, as in a likelihood ratio. But it makes no sense to compare the random variable $Y$ in the process determined by $\theta$ with the random variable in the process determined by $\theta'$. A question such as 'How much larger would $Y_i$ have been had $\beta_1$ been 4.3 rather than 3.4?' is meaningless within the present construction because the two processes need not be 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 $(4.3 - 3.4)x_i$.

### 4.4.2   Exchangeability and counterfactuals

A counterfactual question is best illustrated by examples such as 'If I had taken aspirin would my headache be gone?' or 'How much longer would Cynthia Crabb

have survived had she been given a high dose of chemotherapy rather than a low dose?' The presumption here is that $\mathcal{U}$ consists of subjects or patients, that $x$ is a function representing treatment and other baseline variables, and that $Y$ is the outcome. Formally, $i = Cynthia\ Crabb$ is the patient name, $x(i) = low$ is the treatment component of the covariate, $P_{\{i\}}$ is the survival distribution, and $Y \colon (\Omega, \mathcal{F}, P) \mapsto \mathbb{R}^{\mathcal{U}}$ is a random variable whose $i$th component $Y_{\mathrm{i}}(\omega)$ is the outcome for Cynthia Crabb. Since the constructed process is a real-valued function on the units, there is only one survival time for each subject. Within this framework, it is impossible to address a question that requires Cynthia Crabb to have two survival times. For a more philosophical discussion, the reader is referred to Dawid (2000) who reaches a similar conclusion, or Pearl (2000) who reaches a different conclusion.

If the preceding question is not interpreted as a literal counterfactual, it is possible to make progress by using regression exchangeability and 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 regression 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 1958*a*, 2000). However, Pearl (2000) argues to the contrary, that counterfactual statements are testable and thus not metaphysical.

The directness and immediacy of counterfactual statements are appealing by way of parameter interpretation and model explanation. Another way of trying to make sense of the notion is to invoke a latent variable, a bivariate process with two survival times for each subject, so 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 $Y_{\mathrm{i}2} - Y_{\mathrm{i}1}$ or ratio $Y_{\mathrm{i}2}/Y_{\mathrm{i}1}$ is the desired counterfactual difference in survival times. The treatment value serves to reduce the bivariate process to a univariate process by indicating which of the two components is observed. The net result is a univariate process whose distributions determine the

104 *Exchangeability and regression models*

exchangeable regression model described above. From the present point of view the same process is obtained by an indirect route, so nothing has changed. Unless the observation space for some subset of the units is bivariate, the counterfactual variable is unobservable, so counterfactual prediction cannot arise.

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 to establish a mechanistic interpretation (Cox 1992, §5.1). 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 observation space 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.

### 4.4.3  Exchangeability and causality

Why did Cynthia Crabb receive the low dose when other patients with the same non-treatment covariates received the high dose? A statistical model does not address questions of this sort. However, regression exchangeability is a model assumption implying that, in the absence of a treatment effect, the responses for all patients in Cynthia Crabb's baseline covariate class are exchangeable. As a consequence, all patients whose non-treatment covariate values are the same as those of Cynthia Crabb have the same response distribution. Any departure from exchangeability that is associated with treatment assignment may then be interpreted as evidence against the null model of no treatment effect. In applications where it is a feasible option, objective randomization is perhaps the most effective way to ensure that this model assumption is satisfied.

Like all model assumptions, regression exchangeability may prove unsatisfactory in specific applications. If the treatment assignment is done on doctor's advice on the basis of information not available in $x$, the exchangeability condition may well be violated. Likewise, if the protocol allows patients to select their own dose level, the choice may be based on factors not included in $x$, in which case there is little assurance that the patients are exchangeable modulo $x$.

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 from what 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 five-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 4.2. 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 *et al.* (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, in which 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, is obviously correct in most reasonable senses. It satisfies the conditions of reproducibility, consistency with subject-matter knowledge, and predictability by well-established theory, as demanded by various authors such as Bradford Hill (1937), Granger (1988), and Cox (1992). However, its very consistency with subject-matter knowledge invites an interpretation that is now known to be false. Whether or not the statement is legally correct, it is scientifically misleading, and is best avoided in applications where it might lead to false conclusions. For example, the observation that a shrimp can stun its prey without contact, simply by clicking its claw, might lead to the false conclusion that snails are sensitive to sonar. The complementary statement that the closure of the claw does not cause the sound, although equally defensible, is certainly not better.

Rarely, if ever, does there exist a most proximate cause for any observed phenomenon, so the emphasis must ultimately be on processes and mechanisms (Cox 1992). 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 avoid the term 'causal' 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 five-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 subgroups or by comparing trial results with subsequent performance. In the absence of exchangeability, there is no mathematical reason to suppose that the five-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 five-year survival, but all such mechanisms imply non-exchangeability or the existence of potentially identifiable subgroups.

## 4.5 Exchangeable block models

### 4.5.1 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 is not a rigid one. The key distinguishing feature is the anonymous or ephemeral nature of the levels of a block factor. Cox (1984$a$) 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 $B : \mathcal{U} \times \mathcal{U} \mapsto \{0, 1\}$ 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 semidefinite binary matrix whose rank is the number of blocks in the sample.

A treatment or classification factor $x\colon \mathcal{U} \mapsto \Omega$ is a list of levels, one for each unit. It 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 $B = XX^{\mathrm{T}}$ 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.

Since the information in the block factor $B$ is less than the information in $x$, exchangeability modulo $B$ is a stronger condition than exchangeability modulo $x$. A process is called block-exchangeable if the following condition is satisfied for each $n$. Two samples $\{i_1, \ldots, i_n\}$ and $\{j_1, \ldots, j_n\}$, 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 $\mathbb{R}^n$. Block exchangeability implies that the label information has no effect on distributions. 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 (one or zero). More generally, the $n$-dimensional distribution is invariant under those coordinate permutations that preserve the block structure, i.e. permutations $\pi$ such that $B(i,j) = B(\pi_i, \pi_j)$ for all $i, j$,

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

(i)  $Y \sim N(X\beta, \sigma^2 I_n), \qquad \beta \in \mathbb{R}^\Omega, \sigma > 0,$
(ii)  $Y \sim N(B\gamma, \sigma^2 I_n), \qquad \gamma \in \mathcal{X}, \sigma > 0,$
(iii)  $Y \sim N(\mu, \sigma^2 I_n + \sigma_b^2 B), \qquad \mu \in \mathbf{1}, \sigma > 0, \sigma_b \geq 0,$

in which $\mathbf{1} \subset \mathbb{R}^n$ is the one-dimensional subspace of constant functions. In the sense that they determine precisely the same set of distributions on the observation space, the first two forms are equivalent up to reparameterization. Even so, the models are very different in crucial respects.

By definition in (i), $\beta \in \mathbb{R}^\Omega$ is a function on the treatment levels, so inference for specific levels or specific contrasts is immediate. In (ii), one can transform from $\gamma$ to $\beta = X^{\mathrm{T}}\gamma$ only if the block labels are available. Block labels are not used in (ii), so the formulation does not imply that this information is available. Nonetheless, in most applications with which I am familiar, the function $x\colon \mathcal{U} \mapsto \Omega$ would be available for use, in which case the two formulations are equivalent. Both are regression exchangeable but neither formulation is block exchangeable.

108                    *Exchangeability and regression models*

The third form, the standard random-effects model with independent and identically distributed block effects, is different from the others: 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 (iii) is ambiguous, as is (ii). 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). Given that $x$ is available, it is possible to make inferences or predictions about contrasts among specific factor levels using model (iii). The conditional distribution of the response on a new unit with $x(\cdot) = 1$ is Gaussian with mean and variance

$$\frac{\sigma^2 \mu + n_1 \sigma_b^2 \bar{y}_1}{\sigma^2 + n_1 \sigma_b^2}, \qquad \sigma^2 \left( 1 + \frac{\sigma_b^2}{n \sigma_b^2 + \sigma^2} \right), \tag{4.5}$$

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. If the function $x$ is unavailable, the blocks are unlabelled so 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.

An alternative version of (iii) may be considered, in which the inverse covariance matrix, or precision matrix, is expressed as a linear combination of the same two matrices, $I_n$ and $B$. If the blocks are of equal size, the inverse of $\sigma_0^2 I_n + \sigma_1^2 B$ is in fact a linear combination of the same two matrices, in which case the two expressions determine the same set of distributions on $\mathbb{R}^n$, and thus the same likelihood function after reparameterization. However, the second formulation does not determine a process because the marginal $(n-1)$-dimensional distribution after deleting one component is not expressible in a similar form, with a precision matrix that is a linear combination of $I_{n-1}$ and the restriction of $B$. The absence of a process makes prediction difficult, if not impossible.

It is worth remarking at this point that, for a balanced design, the sufficient statistic for model (iii) 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 eqn (4.3) or eqn (4.5). 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 (1974, p. 39)

or Berger and Wolpert (1988, p. 19) is the clause 'conclusions about $\theta$', implying that it is concerned solely with parameter estimation. Since eqns (4.3) and (4.5) are statements concerning events in the observation space, not estimates of model parameters or statements about $\theta$, there can be no violation of the likelihood principle. On the other hand, a statement about $\theta$ is a statement about an event in the tail $\sigma$-field of the process, so it is not clear that there is a clear-cut distinction between prediction and parametric inference.

### 4.5.2    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 $\Omega = \{1, \ldots, n\}$ (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 of balance can lead to ambiguities in notation.

The following models are block exchangeable:

$$Y_{ij} = \mu + \eta_i + \eta_j + \epsilon_{ij},$$
$$Y_{ij} = \mu + \eta_i - \eta_j + \epsilon_{ij},$$
$$Y_{ij} = \eta_i - \eta_j + \epsilon'_{ij}.$$

In these expressions $\eta/\sigma_\eta$ and $\epsilon/\sigma_\epsilon$ are independent 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

$$Y(u) = \mu + \eta_{x_1(u)} + \eta_{x_2(u)} + \eta'_{x(u)} + \epsilon(u), \tag{4.6}$$

making it clear that $Y$ and $\epsilon$ 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 $\eta'$ and the residual process $\epsilon$ 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:

$$Y \sim N\big(\mu\mathbf{1},\ \sigma_\eta^2 K + \sigma_\epsilon^2 I_{n^2}\big),$$
$$Y \sim N\big(\mu\mathbf{1},\ \sigma_\eta^2 K' + \sigma_\epsilon^2 I_{n^2}\big),$$
$$Y \sim N\big(0,\ \sigma_\eta^2 K' + \sigma_\epsilon^2 I'_{n^2}\big).$$

The matrices $K, K', I'$ are symmetric of order $n^2 \times n^2$ and are 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 $\delta_{ik}$ is the block factor for rows, $\delta_{jl}$ is the block factor for columns, and the remaining terms $\delta_{il}, \delta_{jk}$ 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 2002, 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 of $Y(u^*)$ given $Y$ 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 eqn (4.6) are identifiable at the design. An allowance for errors of estimation along the lines of Barndorff-Nielsen and Cox (1996) is also possible.

## 4.6   Concluding remarks

Kolmogorov's definition of a process in terms of compatible finite-dimensional distributions is a consequence of requiring probability distributions to be well behaved under subsampling of units. Exchangeability is a different sort of criterion based on egalitarianism, the assumption that distributions are unaffected by permutation of units. Regression exchangeability is also based on egalitarianism, the assumption that two sets of units having the same covariate values also have the same response distribution. The range of examples illustrated in Section 4.3 shows that the assumption is almost, but not quite, universal in parametric statistical models.

Regression exchangeability is a fairly natural assumption in many circumstances, but the possibility of failure is not to be dismissed. Failure means that there exist two sets of subjects having the same covariate values that have different distributions, presumably due to differences not included in the covariate. If the differences are due to an unmeasured variable, and if treatment assignment is determined in part by such a variable, the apparent treatment effect is a combination of two effects, one due to the treatment and the other due to the unrecorded variable. Randomization may be used as a device to guard against potential biases of this sort by ensuring that the exchangeability assumption is satisfied, at least in the unconditional sense.

As a function on the units, a covariate serves to distinguish one unit from another, and the notion in an exchangeable regression model is that differences between distributions must be explained by differences between covariate values. However, a covariate is not the only sort of mathematical object that can introduce inhomogeneities or distributional differences. The genetic relationship between subjects in a clinical trial is a function on pairs of subjects. It is not a covariate, nor is it an equivalence relation, but it may affect the distribution of pairs as described in Section 4.5. Two pairs having the same covariate value may have different joint distributions if their genetic relationships are different. The relevant notion of exchangeability in this context is that two sets of units having the same covariate values and the same relationships also have the same joint distribution.

Exchangeability is a primitive but fundamental concept with implications in a wide range of applications. Even in spatial applications, if we define the relationship between pairs of units to be their spatial separation, the definition in Section 4.5 is satisfied by all stationary isotropic processes. The concept is not especially helpful or useful in the practical sense because it does not help much in model construction or model selection. Nonetheless, there are exceptions, potential areas of application in which notions of exchangeability may provide useful insights. The following are four examples.

In connection with factorial decomposition and analysis of variance, Cox (1984*a*, §5.5) has observed that two factors having large main effects are more likely to exhibit interaction than two factors whose main effects are small. To mimic this phenomenon in a Bayesian model, it is necessary to construct a partially exchangeable prior process in the sense of Aldous (1981) that exhibits the desired property. Does exchangeability allow this? If so, describe such a process and illustrate its use in factorial models.

Given a regression-exchangeable process, one can duplicate an experiment in the mathematical sense by considering a new set of units having the same covariate values as the given set. 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 excedent probability or *p*-value is a prediction on the combined sample space $\mathrm{pr}\{T(Y^*) \geq T(Y) \,|\, Y\}$, 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 likelihood-principle violation (Berger and Wolpert 1988).

A mixture of processes on the same observation spaces is a process, and a mixture of exchangeable processes is an exchangeable process. An improper mixture of processes is not a process in the Kolmogorov sense. The fact that improper mixtures are used routinely in Bayesian work raises questions connected with definitions. What sort of process-like object is obtained by this non-probabilistic operation? Is it feasible to extend the definition of a process in such a way that the extended class is closed under improper mixtures? For example, the symmetric density functions

$$f_\mathrm{n}(x_1,\ldots,x_\mathrm{n}) = n^{-1/2}\Gamma\{(n-\nu)/2\}\pi^{-n/2}\left\{\sum_{i=1}^{n}(x_\mathrm{i}-\bar{x}_\mathrm{n})^2\right\}^{-(n-\nu)/2}$$

are Kolmogorov-compatible in the sense that the integral of $f_{n+1}$ with respect to $x_{n+1}$ gives $f_\mathrm{n}$. For $n \geq 2$, the ratio $f_{n+1}/f_\mathrm{n}$ is a transition density, in fact Student's $t$ on $n-\nu$ degrees of freedom centred at $\bar{x}_\mathrm{n}$. In symbols, for $n \geq 2$,

$$X_{n+1} = \bar{x}_\mathrm{n} + \left\{\frac{(n^2-1)s_\mathrm{n}^2}{n(n-\nu)}\right\}^{1/2}\epsilon_\mathrm{n},$$

in which $s_\mathrm{n}^2$ is the sample variance of the first $n$ components, and the components of $\epsilon$ are independent with distributions $\epsilon_\mathrm{n} \sim t_{\mathrm{n}-\nu}$. However, $f_\mathrm{n}$ is not integrable on $\mathbb{R}^n$, so these functions do not determine a process in the Kolmogorov sense, and certainly not an exchangeable process. Nonetheless, the transition densities permit prediction, either for one value or averages such as $\bar{X}_\infty = \bar{x}_\mathrm{n} + s_\mathrm{n}\epsilon_\mathrm{n}/[(n-1)/\{n(n-\nu)\}]^{1/2}$.

In the preceding example, the transition density $p_\mathrm{n}(x;t) = f_{n+1}(x,t)/f_\mathrm{n}(x)$ is a function that associates with each point $x = (x_1,\ldots,x_\mathrm{n})$ a probability density on $\mathbb{R}$. In other words, a transition density is a density estimator in the conventional sense. By necessity, the joint two-step transition density $p_\mathrm{n}^2(x;t_1,t_2) = f_{n+2}(x,t_1,t_2)/f_\mathrm{n}(x)$ is a product of one-step transitions

$$p_\mathrm{n}^2(x;t_1,t_2) = p_\mathrm{n}(x;t_1)\,p_{n+1}\{(x,t_1),t_2\}.$$

For an exchangeable process, this density is symmetric under the interchange $t_1 \leftrightarrow t_2$. Both marginal distributions of $p_\mathrm{n}^2$ are equal to $p_\mathrm{n}$, so two-step-ahead prediction is the same as one-step prediction, as is to be expected in an exchangeable process. Ignoring matters of computation, a sequence of one-step predictors determines a two-step predictor, so a one-step density estimator determines a two-step density estimator. For commercial-grade kernel-type density estimators, it appears that these estimators are not the same, which prompts a

number of questions. Is the difference between the two estimators an indication that density estimation and prediction are not equivalent activities? If so, what is the statistical interpretation of the difference? Is the difference appreciable or a matter for concern? If it were feasible to compute both estimators, which one would be preferred, and for what purpose?

**Acknowledgement**