Some topics in statistical modelling

Peter McCullagh

Title I: Some remarks on variance components models

A variance-components model is a Gaussian model in which the mean of Y is a linear combination of given vectors x, and the covariance matrix is a linear combination of known matrices. Examples include conventional exchangeable random-effects models associated with block factors, certain models for spatial processes, spline models with generalized covariance functions, and growth-curve models for complicated non-linear trends. Since the components of Y are not statistically independent, $E(Y_i \text{ given } x)$ is linear in x and depends only on x_i, but the predicted mean $E(Y_i \text{ given } x, data)$ is not linear in x_i. Here i is an extra-sample unit, and *data* indicates the responses on the sampled units.

The theme of this lecture is that it is often effective to model the regression curve as a Gaussian process by choosing an appropriate covariance function. The predictive distribution for an extra-sample unit can then be computed, and the mean of this distribution is the Bayes estimate of the regression curve. Although the unconditional mean may be linear in the covariate, the predicted mean is not.

Title II: Spatial correlation in field trials

Fairfield Smith (1935) initiated the first systematic study of the nature of spatial correlation in field trials, The variation of yields from plots of various sizes was studied by aggregation of adjacent plots. It was found that the sample variance per unit area does not decrease in proportion to the plot area as would be expected if yields on distinct plots were independent. Fairfield Smith found that the sample variance per unit area decreases according to a power law. The index is not a universal agricultural constant: it ranges from 0.25 to 0.75 depending on the crop and on the season.

This talk describes a large-scale study of 25 uniformity trials, many of which were also studied by Fairfield Smith. The primary purpose of the study is not so much the comparison of strategies for the estimation of variety effects, as the understanding of natural or non-anthropogenic spatial variation of crop yields. The term 'field crop' is interpreted to include annual cash crops such as cereals, beans, potatoes, beets and brassicas, and also fruit crops such as oranges, lemons, peaches, apples, olives and walnuts. In each trial, yields were recorded on a semi-regular grid of rectangular plots of known size and known spacing. Geometric information is necessary in order to study deviations from isotropy and to study spatial covariances.

Our findings are as follows: (i) Agricultural processes have infinite range. (ii) The spatial component of variation is not only self-similar but also conformally invariant. (iii) Most agricultural processes are isotropic or close to isotropic. Where anisotropy is present, it is associated with the direction of drills in the field.

These conclusions point to a generalized covariance model of the form

$$\operatorname{cov}(Y(x), Y(x')) = \sigma_0^2 \delta_{x-x'} - \sigma_1^2 \log|x - x'|$$

in which the first term is white noise and the second term is the de Wijs process defined in integrated contrasts. Non-isotropic anthropogenic effects associated with rows and/or columns can be included in addition if necessary.

Title III: Likelihoods for permanent point processes

A Cox process is a Poisson point process driven by a non-negative random intensity function such as $\exp(Z(x))$ where Z is a Gaussian process. A realization of the Cox process on a bounded set

W is a finite random subset $Y \subset W$ consisting of $\#Y$ points in W. The density of Y at the point configuration y is, in essence, the probability of observing $n = \#y$ points at $(y_1, \ldots, y_n)$ or some permutation thereof. For the great majority of non-trivial Cox processes, no closed-form expression for the density exists. As a consequence, likelihood calculations are difficult.

A permanent point process is a Cox process driven by the random intensity function $|Z(\cdot)|^2$, where Z is a zero-mean complex-valued Gaussian process with covariance function K. The density at the point configuration $y \subset W$ is available in closed form, which makes the model attractive for statistical purposes. The density is proportional to the permanent of a certain Hermitian matrix of order $\#y$ derived from the covariance function K. The likelihood function is a permanent-determinant product. The Bayes estimate of the random intensity function, the conditional expected value of $|Z(\cdot)|^2$ given the data, is the Papangelou conditional intensity, which is a ratio of permanents.

Despite the closed-form expression for the density, exact likelihood computations present formidable challenges. This talk will concentrate on the process and its properties. Computational strategies will be discussed, including analytical approximations for permanents and permanent ratios.

This is joint work with Jesper Møller.