Marginal likelihood for Gaussian models

by

Peter McCullagh *
Department of Statistics
University of Chicago

1. Introduction

Variance-component models are considered in which the model for the mean response $E(Y) = X\beta$ is linear, the error distribution is Gaussian, and the family of covariance matrices $\Sigma = \sigma^2 V(\gamma)$ is closed under scalar multiplication. The marginal likelihood based on the residuals is derived and shown to coincide with the usual REML likelihood (Patterson and Thompson 1971, 1974; Harville 1978). We take this construction a step further by deriving the marginal distribution of the scaled residual vector, showing that the resulting likelihood coincides with Diggle's (1988) REML profile likelihood. The latter function is a particularly convenient tool for studying variance-component models because it is a genuine likelihood function that depends only on the variance-component ratios γ. When these ratios are determined or estimated, the regression parameters may be computed by weighted least squares in the usual way. This paper is concerned with the estimation of variance-component ratios in a variety of linear Gaussian models.

Threee areas of application are considered, the first being the classical application with additive random block effects. A plot of the marginal log likelihood shows the variance-component ratios that are compatible with the observed data. This can be followed by a simple maximum-likelihood analysis or a fully-fledged Bayesian analysis for the treatment effects.

The second area studied in greater detail involves spatial variation in field trials. For a variety of reasons, we concentrate on a very special two-parameter family of Gaussian processes, closely related to the de Wijs process, and closed under convolution. Each process in the family is stationary, isotropic, self-similar and conformally invariant. One slight complication here is that the processes are defined only on spatial contrasts. The means and variances of plot yields are not defined, but means and variances of plot contrasts are well defined. The fact that yields are invariably positive is thus not an objection to the Gaussian assumption.

* Support for this research was provided in part by NSF grant DMS-0071726. The data used and analyzed in this paper are available electronically from the author's web site `www.stat.uchicago.edu/~pmcc/reml`

The third area of application is a new multivariate Gaussian clustering model, which aims to partition the units into homogeneous clusters. The marginal likelihood is especially appropriate in this case because it depends on the data only through the configuration, the maximal invariant under the general affine group. Thus each affine transformation of the data necessarily produces the same likelihood function and determines the same clustering, or posterior distribution on clusterings.

2. Marginal likelihood

2.1 *The scaled Gaussian distribution*

Let $\mathcal{E}^n$ denote n-dimensional Euclidean space, and let $Y \sim N_n(0, \Sigma)$ be a zero-mean Gaussian random variable taking values in $\mathcal{E}^n$. We consider in this paper various parametric statistical models $\Sigma \in \Theta$ for the covariance matrix, in which the parameter set Θ is closed under positive scalar multiplication. Some familiar examples are as follows.

(i) In the standard normal-theory linear model, $\Theta = \{\sigma^2 V \mid \sigma^2 > 0\}$ for some known positive definite matrix V, usually the identity.

(ii) In the standard variance-components model, the covariance matrix is expressible as a non-negative linear combination $\Sigma = \sigma_1^2 V_1 + \cdots + \sigma_k^2 V_k$ of known symmetric matrices $V_1, \ldots, V_k$. Here V_1 is positive definite, $\sigma_1^2 > 0$, the remaining matrices and coefficients being semi-definite.

(iii) All standard linear Gaussian models, either autoregressive or moving average.

(iv) All zero-mean spatial Gaussian models such as the stationary isotropic processes with covariance functions

$$\operatorname{cov}\bigl(Y(x), Y(x')\bigr) = \sigma_0^2 \delta_{x-x'} + \sigma_1^2 \exp(-\lambda r)$$
$$\operatorname{cov}\bigl(Y(x), Y(x')\bigr) = \sigma_0^2 \delta_{x,x'} + \sigma_1^2 \lambda r\, K_1(\lambda r)$$

Here $\sigma_0^2, \sigma_1^2, \lambda$ are arbitrary positive numbers, $\sigma^2 \delta_{x-x'}$ is the white noise variance or nugget effect, $r = |x - x'|$ is the spatial separation, K_1 is the Bessel function, and $rK_1(r)$ is the covariance function of Whittle's (1954) stationary Gaussian process in the plane.

Other examples will emerge later in the paper.

For each $y \in \mathcal{E}^n$, the length or norm of y is well defined, so we may transform to the scaled vector $\check{y} = y/|y|$ provided that $y \neq 0$. It is evident from the geometry that the distribution of $\check{Y}$ under Σ is the same as the distribution under $\lambda\Sigma$ for each $\lambda > 0$. In fact, the distribution of $\check{Y}$ can be derived directly by integration as follows. First transform to spherical polar coordinates $y \mapsto (r, \check{y})$ where $r = |y|$ and $\check{y}$ is a point on the unit sphere. The Jacobian is $dy = r^{n-1}\, dr\, d\mu(\check{y})$ where μ is Lebesgue measure on the unit sphere. The

Gaussian density in $\mathcal{E}^n$ transforms by

$$|\Sigma|^{-1/2} \exp(-y'\Sigma^{-1}y/2)\, dy = |\Sigma|^{-1/2} \exp(-r^2 \check{y}'\Sigma^{-1}\check{y}/2)\, r^{n-1}\, dr\, d\mu(\check{y})$$

Now put $u = r^2$, transform and integrate over u, giving the density of the scaled, or projected, Gaussian distribution in the form $\check{f}(y;\Sigma)\, d\mu(y)$ where

$$\check{f}(y;\Sigma) = \tfrac{1}{2}\Gamma(\tfrac{1}{2}n)\, \pi^{-n/2}\, |\Sigma|^{-1/2} (y'\Sigma^{-1}y)^{-n/2}. \tag{1}$$

Note that $\check{f}(y;\lambda\Sigma) = \check{f}(y;\Sigma)$, as claimed. The marginal log likelihood function for Σ based on the projected observation $y/|y|$ is thus

$$\begin{aligned} \check{l}(\Sigma; y) &= -\tfrac{1}{2}\log|\Sigma| - \tfrac{n}{2}\log(y'\Sigma^{-1}y) \\ &= -\tfrac{1}{2}\log|\Sigma| - \tfrac{n}{2}\log(y'\Sigma^{-1}y) + \text{const} \end{aligned} \tag{2}$$

where the additive constant $n \log|y|$ can be ignored.

In example (ii) above, it is natural to write

$$\Sigma = \sigma^2(\gamma_1 V_1 + \cdots + \gamma_k V_k)$$

re-parameterizing in such a way that the variance-component ratios satisfy either $\gamma_1 = 1$ or the more symmetrical condition $\sum \gamma_j = 1$. Ordinarily, these components are non-negative, so γ may be regarded as a point in the probability simplex. The marginal log likelihood (2) is then a function of γ alone, so we write $\check{l}(\gamma; y)$ in place of $\check{l}(\Sigma; y)$. The elimination of σ from the likelihood function by marginalization is most useful when the number of variance components is two or three, in which case marginal likelihood plots are extremely convenient and informative.

If $n = 2$, the projected variable $\check{y} = y/|y|$ is a point on the unit circle, so we may write $\check{y} = (\cos\phi, \sin\phi)$ with $d\mu(\check{y}) = d\phi$. For simplicity, it is helpful to take $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, in which case the quadratic form reduces to

$$\check{y}'\Sigma^{-1}\check{y} = \tfrac{1}{2}\big(\sigma_1^2 + \sigma_2^2 - (\sigma_1^2 - \sigma_2^2)\cos(2\phi)\big) \,/\, (\sigma_1^2\sigma_2^2).$$

The projected density is on the unit circle is

$$\frac{(\sigma_1\sigma_2)\, d\phi}{\pi\big(\sigma_1^2 + \sigma_2^2 - (\sigma_1^2 - \sigma_2^2)\cos(2\phi)\big)}.$$

which depends only on the ratio σ_1/σ_2. For general Σ, the projected distribution is bipolar, and symmetric about both principal axes of Σ. The real-valued random variable $\tan\phi = \tan\arg(\check{Y})$ has the Cauchy distribution.

2.2 *Regression models and residual likelihood*

The Gaussian model described above has zero mean, which severely limits its usefulness in applied work. Suppose therefore that $Y \sim N_n(\mu, \Sigma)$ in which the mean vector lies in the subspace $\mathcal{X}$ of dimension p in $\mathcal{E}^n$. It is conventional to write $\mu = X\beta$ in matrix notation, in which the columns of X are a basis for the subspace $\mathcal{X}$. The assumptions made regarding Σ are those described above, so that the extended parameter space is the Cartesian product $(\mu, \Sigma) \in \mathcal{X} \times \Theta$. That is to say, the regression parameters and the dispersion parameters are assumed to be variation independent. Note that $\mathcal{X}$ is a vector space, but Θ is a subset of the positive definite matrices, and thus cannot be a vector space.

By the term 'residual' or 'residual vector' we understand any linear transformation $A\colon \mathcal{E}^n \to \mathcal{V}$ such that $\ker A = \mathcal{X}$, together with the random point $AY \in \mathcal{V}$. Although not essential to the definition, it is extremely convenient in subsequent work to choose the transformation so that $A\mathcal{E}^n = \mathcal{V}$, so that, as a matrix, A has full rank. The primary condition says that $A\mathcal{X} = 0$, or $AX = 0$ in matrix notation, and the supplementary condition that $Ax = 0$ implies $x \in \mathcal{X}$. All linear transformations having the same kernel are regarded as equivalent residuals. Evidently, the distribution of the residual is the same for all values of $\mu \in \mathcal{X}$, so the mean may be ignored in all distributional calculations that follow.

To derive the distribution of the residual, we integrate over the kernel as follows. Let $\mathcal{W}$ be any subspace of $\mathcal{E}^n$ complementary to $\mathcal{X}$, so that each point $y \in \mathcal{E}^n$ can be written in the form $y = y_1 + y_2$ with $y_1 \in \mathcal{X}$ and $y_2 = Qy \in \mathcal{W}$. Here Q is the projection onto $\mathcal{W}$ along $\mathcal{X}$, and QY is a residual according to the definition. There is no suggestion that $\mathcal{X}$ and $\mathcal{W}$ are orthogonal subspaces.

The quadratic form in the exponent of the normal density can be written in the form

$$\begin{aligned} y'\Sigma^{-1}y &= (P_\Sigma y)'\Sigma^{-1}P_\Sigma y + (Q_\Sigma y)'\Sigma^{-1}Q_\Sigma y \\ &= (P_\Sigma y)'\Sigma^{-1}P_\Sigma y + (Q_\Sigma y_2)'\Sigma^{-1}Q_\Sigma y_2 \end{aligned}$$

where $P_\Sigma = X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$, $Q_\Sigma = I - P_\Sigma$, and $y_2 \in \mathcal{W}$. Each point $P_\Sigma y = Xb$ in $\mathcal{X}$ is a linear combination of the basis elements, so the quadratic form reduces to

$$y'\Sigma^{-1}y = b'X'\Sigma^{-1}Xb + (Q_\Sigma y_2)'\Sigma^{-1}Q_\Sigma y_2.$$

The Jacobian of the transformation $b \mapsto Xb$ is $|X'X|^{1/2}$. Thus, integration over $\mathcal{X}$ gives the marginal distribution of the residual QY at $y_2 \in \mathcal{W}$ in the form

$$\begin{aligned} &(2\pi)^{-n/2}|\Sigma|^{-1/2}|X'X|^{1/2}\exp\bigl(-(Q_\Sigma y_2)'\Sigma^{-1}Q_\Sigma y_2/2\bigr) \times \int_{\mathcal{R}^p} \exp(-b'X'\Sigma^{-1}Xb/2)\,db \\ &= (2\pi)^{-n/2}|\Sigma|^{-1/2}|X'X|^{1/2}\exp\bigl(-y'(\Sigma^{-1}Q_\Sigma)y/2\bigr) \times (2\pi)^{p/2}|X'\Sigma^{-1}X|^{-1/2}. \qquad (3) \end{aligned}$$

The marginal log likelihood based on the residual is thus

$$l_m(\Sigma; y) = -\tfrac{1}{2} y'\Sigma^{-1} Q_\Sigma y - \tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\log|X'\Sigma^{-1}X|, \tag{4}$$

which is also called the residual log likelihood or the REML log likelihood or the restricted log likelihood (Patterson and Thompson 1971, 1974; Harville 1977, 1978)

It is useful to note that the residual likelihood, obtained by sample-space integration over $y \in \mathcal{X}$, is not the same as the profile likelihood obtained by parameter-space maximization over $\mu \in \mathcal{X}$ for fixed Σ. The profile log likelihood for Σ is the sum of the first two of the three terms in (4). Diggle (1988) remarks that the reasons for choosing the residual likelihood over the profile likelihood are largely intuitive. I interpret this remark to mean that the REML likelihood function is clearly the correct one to use for variance-component estimation, but that convincing arguments are hard to find. To my mind, the two facts that support the use of the REML likelihood are as follows.

(i) Both functions depend on the same quadratic function of the residuals, so there can be no difference between the profile and REML likelihoods in information content.

(ii) One is a likelihood function and the other is not.

For example the derivative of (4) has zero expectation, and the higher-order Bartlett relations are satisfied. The profile log likelihood does not have these properties, which accounts for much of the bias in maximum-likelihood estimation of variance components. Note that if the mean-value component of the model is a manifold contained in $\mathcal{X}$, the derivation of the marginal likelihood by integration over $\mathcal{X}$ is unaffected. However, the profile likelihood is then not equal to the first two terms in (4). The profile likelihood is not quadratic in y, so the argument (i) above fails.

Following the line of argument in the preceding section, the marginalization argument may be carried a step further by calculating the marginal distribution of the scaled residual and the associated likelihood function. The argument used in section 2.1 yields the log likelihood

$$\check{l}(\Sigma; y) = -\tfrac{n-p}{2}\log(y'\Sigma^{-1}Q_\Sigma y) - \tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\log|X'\Sigma^{-1}X|, \tag{5}$$

which is constant on scalar multiples of Σ.

Provided that the parameter space is closed under positive scalar multiplication, expression (5) may also be derived as a profile log likelihood, by maximizing the residual log likelihood (4) over the scale parameter. Following Diggle (1988), we write the covariance matrix in the form $\sigma^2 V(\gamma)$ and maximize over σ^2 for fixed positive definite $V \equiv V(\gamma)$. The re-parameterized log likelihood (4) is

$$l_m(\sigma^2 V; y) = -\tfrac{1}{2} y'V^{-1}Q_\Sigma y/\sigma^2 - \tfrac{1}{2}\log|V| - n\log\sigma - \tfrac{1}{2}\log|X'V^{-1}X| + p\log\sigma.$$

Differentiation with respect to σ gives

$$\begin{aligned} \partial\check{l}/\partial\sigma &= y'V^{-1}Q_V y/\sigma^3 - (n-p)/\sigma \\ (n-p)\hat{\sigma}^2 &= y'V^{-1}Q_V y \end{aligned}$$

where $Q_V = Q_\Sigma$. The marginal likelihood estimator of σ^2 for fixed V is the weighted mean squared residual, the familiar unbiased estimate. Substitution into (4) yields the marginal profile likelihood in the form (5) with Σ replaced by $V(\gamma)$. We prefer to think of both (4) and (5) as marginal likelihoods based on certain functions of the data, rather than profile likelihoods. Accordingly, we compromise in our terminology by referring to (5) as Diggle's profile log likelihood. Note that the derivation of the marginal log likelihoods (4) and (5) by sample-space integration requires only that the set of mean-values be a subset of $\mathcal{X}$, and does not require Θ to be closed under scalar multiplication.

2.3 *Multivariate response*

For later work, it is worthwhile extending the standard residual likelihood to multivariate regression models in which the response Y_i on unit i is a q-variate normal vector with covariance matrix Δ regarded as an unknown parameter. That is to say, Y is a matrix of order $n\times q$, and $Y \sim N(X\beta, \Sigma\otimes\Delta)$, where the coefficient matrix β is of order $p\times q$. Given Σ, the weighted least squares estimate of β is the familiar expression $\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$, independent of the value of Δ. The covariance matrix of the least-squares estimate is $(X'\Sigma^{-1}X)^{-1}\otimes\Delta$.

The expression $E(Y) = X\beta$ with unrestricted coeficient matrix $\beta \in \mathcal{R}^{pq}$ means that the model space is the direct sum $\mathcal{X}^{\oplus q}$ of q copies of $\mathcal{X}$, one copy for each of the response variables. The residual is any linear transformation $y \mapsto Ay$ having $\mathcal{X}^{\oplus q}$ as kernel, and the residual or REML log likelihood is

$$\begin{aligned} l_m(\Sigma,\Delta;Qy) &= -\tfrac{1}{2}\operatorname{tr}(y'\Sigma^{-1}Q_\Sigma y\Delta^{-1}) - \tfrac{1}{2}\log|\Sigma\otimes\Delta| - \tfrac{1}{2}\log|X'\Sigma^{-1}X\otimes\Delta| \\ &= -\tfrac{1}{2}\operatorname{tr}(y'\Sigma^{-1}Q_\Sigma y\Delta^{-1}) - \tfrac{q}{2}\big(\log|\Sigma| - \log|X'\Sigma^{-1}X|\big) - \tfrac{n-p}{2}\log|\Delta|. \end{aligned}$$

which has a form similar to the univariate REML log likelihood. The marginal log likelihood based on the scaled residual matrix is

$$\check{l}(\Sigma;Qy) = -\tfrac{n-p}{2}\log|y'\Sigma^{-1}Qy| - \tfrac{q}{2}\log|\Sigma| - \tfrac{q}{2}\log|X'\Sigma^{-1}X|, \tag{6}$$

which does not depend on the matrix Δ. Note that, for each non-singular $q\times q$ matrix L,

$$\check{l}(\Sigma;QyL) = \check{l}(\Sigma;Qy) - (n-p)\log|L|,$$

so the marginal log likelihood (6) is unaffected by such transformation. The similarity with (4) and (5) is evident.

2.4 *Score statistic*

Let V_0 be a positive definite symmetric matrix of order n, and let V_1 be any other semi-definite symmetric matrix. We use the marginal log likelihood (5) for the scaled residuals to compute the derivative of l at V_0 in the direction V_1. That is to say, we consider the parametric family of covariance matrices proportional to $V_0+\gamma V_1$ for a range of small values of γ, and compute the derivative of l with respect to γ at $\gamma=0$. This derivative, called the score statistic, is often easier to compute than the full log likelihood, and serves as a test of the hypothesis that the covariance matrix of Y is $\sigma^2 V_0$.

The calculations are straightforward but require rather tedious computations involving matrix derivatives. The score statistic is

$$\frac{\partial \check{l}(\gamma; y)}{\partial\gamma} = \frac{(n-p)}{2}\,\frac{y'Q_0'V_0^{-1}V_1V_0^{-1}Q_0y}{y'V_0^{-1}Q_0y} - \tfrac{1}{2}\,\mathrm{tr}(Q_0V_1V_0^{-1})$$

where $P_0 = X(X'V_0^{-1}X)^{-1}X'V_0^{-1}$ and $Q_0 = I - P_0$ is the complementary projection.

A particular case of special interest arises when $X=1$, $V_0=I_n$ is the identity, and V_1 is block-diagonal with k blocks of sizes $n_1,\ldots,n_k$. We find that the score statistic is

$$\frac{(n-p)}{2}\,\frac{\sum_r n_r^2(\bar{y}_r-\bar{y})^2}{S_T^2} - \frac{n^2-\sum_r n_r^2}{2n}$$

where $\bar{y}$ is the average of all observations, $\bar{y}_r$ is the average in the rth block and S_T^2 is the total sum of squares about $\bar{y}$. The conventional between-blocks sum of squares is $\sum_r n_r(\bar{y}_r-\bar{y})^2$, so the numerator is not a multiple of the between-blocks sum of squares unless the block sizes are equal. That is to say, the locally most powerful test statistic is not a function of the conventional F-ratio.

Note that the score statistic is a homogeneous rational function of degree (2,2) in y, and is unaffected by translation in $\mathcal{X}$, which is to say that it is a function of the scaled residual. It is also unaffected by simultaneous re-scaling of the matrices (V_0, V_1). The null distribution is such that the numerator and denominator may be regarded as independent. In fact, we may take the inner product in the space to be such that the denominator is exactly one when evaluated at $\check{y}$. Slightly longer calculations demonstrate that the null variance is

$$\mathrm{var}(\partial\check{l}/\partial\gamma) = \frac{(n-p)\,\mathrm{tr}[(Q_0V_1V_0^{-1})^2] - \mathrm{tr}^2(Q_0V_1V_0^{-1})}{2(n-p+2)}.$$

The distribution is determined entirely by the non-zero eigenvalues of $Q_0V_1V_0^{-1}$.

3. Example 1: Hyperbaric O_2 treatment

The data for our first example are taken from a study by Dr. George Huang of the Department of Surgery at the University of Chicago on the effect of oxygen on the rate of healing of surgical wounds of diabetic rats. (Diabetics, both human and animal, tend to have more complications following surgery than non-diabetics.) Thirty rats were first given a drug that has the effect of destroying the pancreas, thereby making the rats diabetic. All the rats underwent surgery, during which an incision was made along the entire length of the back. This was immediately sewn up with staples. The treatment group of 15 rats was subjected to hyperbaric O_2 treatment, i.e., a 100% O_2 environment at 2 atmospheres pressure, for 90 minutes per day following surgery. To ensure that the control rats were handled in a manner comparable to the treated rats, the control group also received O_2 treatment for 90 minutes daily, but at normal atmospheric pressure. Six rats had glucose levels that were deemed too low to be considered diabetic, and were excluded from the experiment. After a 24 day recuperation period the 24 rats still participating in the experiment were sacrificed, i.e., killed. Strips of skin were taken from each of five sites on each rat, each site crossing the surgical scar in a right angle. The strips were put on a tensiometer and stretched to breaking point. The observations in the table give the energy required to break the specimen. Unfortunately some specimens slipped out of the clamps for reasons deemed to be unconnected with the strength. For these specimens, no observation could be made: the values are indicated by '--' in the table. Rats 1–14 received the hyperbaric treatment: rats 15–24 were the controls.

We consider a linear model for the log-transformed data containing additive site and treatment effects. Two observations on the same rat are likely to be positively correlated, so we use a variance-components model in which $\Sigma = \sigma^2((1-\gamma)I + \gamma V)$. Here $V(i,j)$ is one if observations i, j are on the same rat, and zero otherwise, so V has rank 24. Missing data are ignored, so I and V are matrices of order $n = 104$. If there were no missing data, a standard two-sample analysis of the 24 rat means would suffice, and would be equivalent to the present analysis.

The marginal log likelihood (4) for γ shown in Fig. 1 is roughly quadratic in the interval $(0.0, 0.9)$, with a maximum near $\hat\gamma = 0.4$. In other words, the between-rat variance $\sigma_r^2 = \gamma\sigma^2$ accounts for approximately 40% of the total variance of each observation. The within-rat variance $\sigma_\epsilon^2 = (1-\gamma)\sigma^2$ accounts for the remaining 60%. An approximate likelihood-based 95% confidence interval calculation for γ gives $(0.2, 0.65)$

With γ set equal to the value 0.4, the weighted least squares estimator of the systematic effects is obtained from $\hat\beta = (X'WX)^{-1}X'W(\log y)$ with $W^{-1} = 0.6I + 0.4V$. The weighted residual mean square is $s^2 = 0.3546$ on 98 degrees of freedom, so the asymptotic variance matrix of $\hat\beta$ is $s^2(X'WX)^{-1}$. The parameter estimates and their standard errors are as

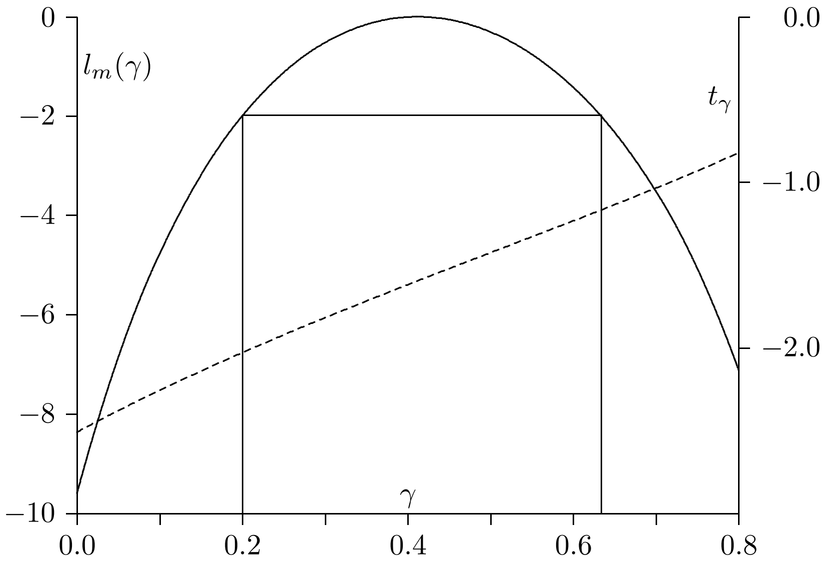

Fig. 1: Marginal log likelihood for the variance-component ratio. The dashed line is the standardized treatment statistic t_γ.

follows.

Parameter	Estimate	*S.E.*
site 1	0.000	0.00
site 2	0.158	0.14
site 3	0.182	0.14
site 4	0.261	0.15
site 5	−0.270	0.15
Treat	−0.294	0.18

The range of values of the variance ratio that are reasonably consistent with the observed data is the likelihood interval $(0.20, 0.63)$ shown in Fig. 1. The standardized treatment statistic, i.e. the treatment t-ratio, increases roughly linearly over this range, from -2.03 at $\gamma = 0.2$ to -1.18 at $\gamma = 0.63$.

It was anticipated that the treatment effect, if present at all, would be positive, so the observed negative value was a distinct disappointment. Even so, a strong negative value would be of biological interest. However, the observed treatment effect is in the range that could reasonably be attributed to random variation alone, so there is insufficient evidence to overturn the null value. If there were no missing data, the preceding analysis using $\hat{\gamma}$ is equivalent to reducing the design to the rat means, and analyzing these as a two-sample problem with 24 independent observations. This makes good sense because the rats, not the sites, are the experimental units.

4. Field experiments

4.1 *A conformal Gaussian model*

It is assumed in this section that the observations are indexed by a finite set of plots, not necessarily rectangular in shape and not necessarily arranged in a regular lattice. The plots are assumed to have finite positive area, and for convenience they are assumed to be of equal area. It is also assumed that the observation is made on an extensive variable such as yield, so the value $Y(A)$ on plot A is an integral over the set

$$Y(A) = \int_A Y(x)\, dx$$

where dx is planar Lebesgue measure. This condition implies as a matter of arithmetic that $Y(A \cup A') = Y(A) + Y(A')$ for non-overlapping plots.

The model for treatment effects is assumed to be of the usual linear from $E(Y) = X\beta$, where X is the model matrix for all relevant treatment effects and other incidental effects connected with ploughing, harvesting, drainage and unevenness of topography. For the covariance function, we focus primarily on the stationary isotropic self-similar Gaussian model in which $Y = \bigl(Y(A_1), \ldots, Y(A_n)\bigr)$ with covariances

$$\begin{aligned} \operatorname{cov}(Y) &= \sigma_0^2 |A|\, I_n + \sigma_1^2 |A|^2\, V \\ &= \sigma^2 \bigl[(1-\gamma) I_n + \gamma |A|\, V\bigr] \\ V_{ij} &= -\operatorname{ave}\bigl(\log |x - x'|\bigr), \qquad (x \in A_i; x' \in A_j) \end{aligned} \tag{7}$$

where $|A|$ is the plot area in suitable units, $\sigma^2 = (\sigma_0^2 + \sigma_1^2)|A|$, and $\gamma = \sigma_1^2/(\sigma_0^2 + \sigma_1^2)$. This family is closed under convolution, and includes white noise and the de Wijs processes as extreme points at $\gamma = 0$ and $\gamma = 1$ respectively.

Henceforth, the term ‘conformal model’ refers to any Gaussian model having this family of covariance functions. For non-overlapping circular plots, V_{ij} is exactly the negative log of the distance between centres. If the plots are non-overlapping squares or approximate squares, V_{ij} is approximately the negative log of the distance between centres. For square plots of side r, the diagonal elements are $V_{ii} = 0.8051 - \log(r)$.

4.2 *Example 2: strawberries*

In the introduction to their wide-ranging paper on th analysis of agricultural field experiments, Besag and Higdon (1999) have made their data freely available to encourage discussion and to stimulate further research. In a sense, the conformal model (7) is a special case of the Besag-Higdon formulation corresponding to a very particular choice of interaction coefficients in their Markov model for fertility effects. The derivation, however, is fundamentally different because (7) is derived purely by consideration of conformal transformations, which makes the nebulous concept of fertility entirely unnecessary.

Consider the strawberry variety trial described in section 6 of the Besag-Higdon paper, a randomized blocks design with eight varieties grown in each of four blocks. The systematic part of the model is thus the indicator matrix X of rank 8 for the varieties. A complication in the design is the presence of a hedge along one side of the field, which depresses the yields in the four adjacent plots to a noticeable degree. Accordingly, we include in the model an indicator factor for proximity to the hedge, so X has rank 9. Lee and Nelder (2001) use the same adjustment in their model 'Lc', but they also include a linear trend across the columns.

No information is recorded on the plot geometry, so we proceed here as if the plots were unit squares, taking V_{ij} to be the negative log of the distance between plot centres, and $V_{ii} = 0.805$. In practice, there may well have been fallow guard strips or paths separating the plots to allow access and to minimize cross-contamination. So this is at best an approximation to the physical geometry of the design. Taking the covariance matrix to be of the form

$$\Sigma = \sigma^2\big((1-\gamma)I_n + \gamma V\big)$$

the marginal likelihood for γ shown in Fig. 2 has a maximum at $\hat{\gamma} = 0.35$. The marginal likelihood ratio test statistic for the hypothesis that $\gamma = 0$ is 5.44 on one degree of freedom, for a p-value of 2%. Taking $\gamma = 0.290$, the weighted least squares estimates of the variety contrasts is computed in the usual way, giving the following values and standard errors.

Table 2: Variety estimates in a strawberry trial

Variety	E	F	G	M	P	Re	Rl	V	hedge
Estimate	0.00	−0.88	0.02	−0.29	0.57	−0.98	−1.70	−0.16	−2.57
s.e.	0.00	0.52	0.54	0.56	0.55	0.53	0.60	0.51	0.62
N-L est	0.00	−0.68	0.00	−0.15	0.73	−0.79	−1.54	−0.19	−2.15
B-H I	0.00	−0.97	−0.31	−0.81	0.03	−1.15	−2.51	−0.21	
B-H II	0.00	−1.02	−0.44	−0.46	0.17	−1.31	−1.99	−0.30	

Similar estimates given by Lee and Nelder (2001) and by Besag and Higdon (1999) are shown for comparison, though these are perhaps not strictly comparable. Nelder and Lee use a non-spatial model with a hedge effect plus a linear trend across columns. Besag and Higdon recognize the hedge effect, but aim to accommodate it by using a more flexible non-isotropic model for fertility effects. Both of these papers report estimates from a range of models, only a few of which are reproduced above. Our estimates of the variety effects, and the ranking of the varieties PGEVM followed by F, Re, Rl, are in good agreement with Nelder an Lee, differing only in the inversion of V and M. Besag and Higdon's ranking PEVGM is in agreement with the conformal sub-model in which the hedge effect is omitted.

The root mean variance of pairwise variety contrasts in our model is 0.54, which is fairly close to the values reported by Lee and Nelder. Contrasts with Rl have higher variance on account of the unfavourable location of this variety in two plots adjacent to the hedge. The large standard error means of course that the rankings are not firmly established.

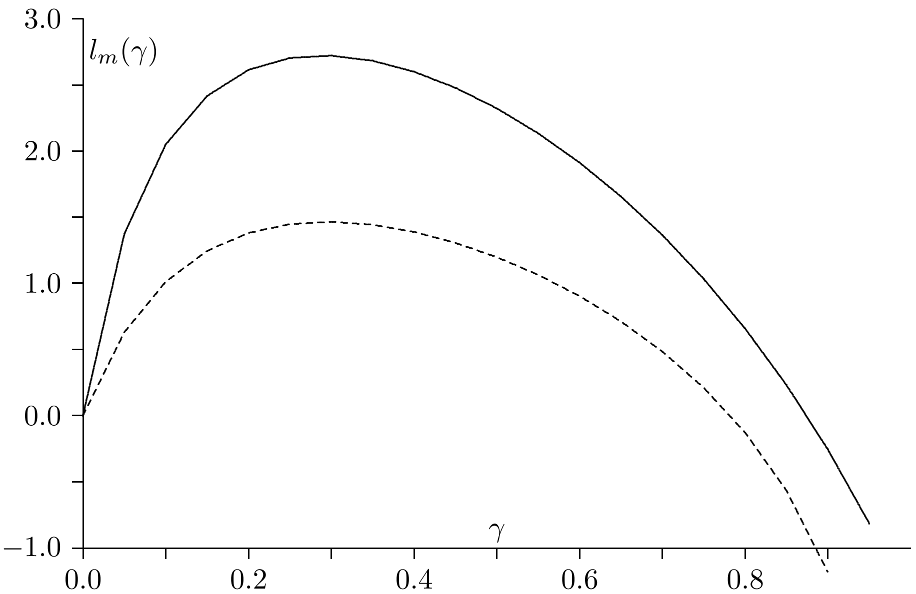

Fig. 2: Marginal log likelihood for the variance-component ratio in the strawberry variety trial: the solid line is the conformal model; the dashed line is the random block model.

This analysis entirely ignores the block structure. It is thus of interest to investigate the extended model in which

$$\Sigma = \sigma^2\bigl(\gamma_0 I_n + \gamma_1 V + \gamma_2 V_b\bigr),$$

where V_b is the equivalence function or matrix for blocks, and the non-negative coefficients γ add to one. That is to say $V_b(i, j) = 1$ if plots i, j are in the same block, and zero otherwise. The marginal log likelihood with γ_1 fixed at zero corresponds to the dashed line in Fig 2, and has a maximum at $\gamma_2 = 0.295$. In the extended model, the likelihood is maximized at $\gamma_1 = 0.29, \gamma_2 = 0.0$, on the boundary of the parameter space corresponding to the conformal model. In this example, the conformal model is more effective for absorbing spatial variation than the random block model. The difference in maximized log likelihoods is 1.26.

The evidence for spatial dependence in this example is virtually indistinguishable from a linear trend across the columns. If such a trend is included among the systematic effects, the evidence for spatial dependence disappears, and the marginal log likelihood has its maximum at $\gamma_1 = 0$.

4.3 *Example 3: Fruit and nut plantations*

Batchelor and Reed (1918) published extensive data on the yields of individual fruit trees taken from four citrus groves and one nut plantation in California, and one apple orchard in Utah. A brief description of all six data sets follows.

The first data set consists of yields of individual 24-year old trees from the navel-orange grove at Arlington Station in 1915–16. The grove consists of 20 rows from north to south, with 50 trees in a row, planted $22'$ by $22'$. In the analysis described here, the grove was split into the eastern and western halves, consisting of 500 trees each in a grid 20×25. This procedure greatly reduces the computational effort, and also provides a check on the adequacy of the model.

The second data set consists of yields from 480 ten-year old navel orange trees in 1916 taken from a grove called Antelope Heights located at Naranjo. There are 15 rows of 33 trees, also planted $22' \times 22'$.

The third data set consists of Valencia orange yields from 15-year old trees at Villa Park in 1916. There are 12 rows of 20 trees, planted $21.5' \times 22.5'$. In the analysis reported here, the grid is taken as $22'$ square.

The fourth data set consists of yields from 23-year old Eureka lemon trees at Upland, California in 1916. There are 364 trees in 14 rows of 26 trees in each row, planted in a square grid of side $24'$.

The fifth data set consists of yields from 320 walnut trees from a 24-year old Santa Barbara softshell seedling grove. This planting is laid out 10 trees wide and 28 trees long, The trees are planted on a square lattice of side $50'$. In the analysis reported here, each tree is assumed to occupy a square cell of side $25'$, so there is an additional $25'$ space between rows and columns.

The last data set consists of yields of apples from a 10-year old Jonathan apple orchard at Providence, Utah. There are eight rows of 28 trees, planted $16'$ apart, east and west, and $30'$ apart north and south. In the analysis reported here, each tree is assumed to occupy a cell of side $16'$, so there is a $14'$ vacant space in one direction.

In preparing the data for the computer, a few minor discrepancies were uncovered, where the values in the table did not tally with the published row and column totals. In some instances, these were corrected, though this could seldom be done with much certainty. To enable interested readers to duplicate or improve on the author's analyses, all data used here are available electronically from the author's web site, together with functions written in R for computing the marginal likelihood function.

In calculating and plotting the marginal log likelihood for γ in the conformal model (7), it was found to be convenient to take the unit of area to be 100 square metres. This choice has no substantive effect on the conclusions, but it does avoid extremely small or extremely large values of γ. For plots of this area, the two terms in (7) contribute roughly equally to the total variance so γ is roughly one half. In all cases, additive row and column effects are eliminated from the analysis whether there are significant variations or not. The marginal log likelihoods for γ in the conformal model (7) are displayed in Fig 3 for each of the six data sets, with two parallel plots for the two halves of the Arlington navel orange grove. In the summary statistics listed below, the log likelihood is measured relative to the model with $\gamma = 0$.

Plantation	$\hat\gamma$	s.e.$(\hat\gamma)$	$\check{l}(\hat\gamma)$
Arlington I	0.299	0.082	26.24
Arlington II	0.295	0.091	10.92
Antelope	0.480	0.086	26.89
Valencia	0.378	0.120	10.70
Eureka	0.302	0.091	16.72
Walnuts	0.216	0.103	5.17
Jonathan	0.000	—	0.00

With the sole exception of the apple orchard at Providence, Utah, the evidence for spatial dependence is overwhelming. For example, the log likelihood ratio statistic for the null hypothesis of no spatial dependence among the walnut trees is 2×5.17 on one degree of freedom for a p-value of 0.0013. For the Providence data, the log likelihood is fairly flat for small γ, so all that can be said is that the data are not consistent with values of γ in excess of about 0.4. Any reasonable spatial covariance model is sufficient to detect spatial dependence in these data, so these analyses provide no evidence that the spatial dependence follows the conformal model (7). Comparisons with alternative covariance models are reported in section 4.6.

It is quite remarkable that these six plantations, three of oranges, one of lemons, one of walnuts, and one of apples, are fairly consistent with each other so far as the conformal coefficient γ is concerned. The estimated common value, obtained by maximizing the sum of the seven marginal log likelihood functions, is $\hat\gamma = 0.327$, or $\hat\gamma/(1-\hat\gamma) = 0.485$ per unit area of $100m^2$, and the log likelihood at that point is 93.19. The likelihood ratio test statistic for the hypothesis of a common value is $2(96.66-93.19) = 6.94$ on six degrees of freedom. This rough and ready analysis treats the two halves of Arlington as independent which they are not. Even if the degrees of freedom are reduced to five to compensate, the evidence against a common value is not strong.

4.4 *Example 4: Great Knott wheat*

Regardless of its original purpose, an extensive uniformity trial presents the opportunity to test the proposed conformal Gaussian model in various ways. The Mercer and Hall data on wheat yields taken from a 20×25 grid of plots in Great Knott field at Rothamsted in the summer of 1910 is ideal in many ways. These data have been used by later authors (Whittle 1954, Besag 1974, McBratney and Webster 1981) for a wide range of purposes. Although most analyses focus on the grain yields, the data consist of both grain and straw yields on a regular grid of 20 rows, each 2.59 m wide, and 25 columns each 3.30 m wide. Details of the experiment taken from the Mercer-Hall paper are reproduced in Andrews and Herzberg (1985) together with the data.

McBratney and Webster (1981) remark that detailed records have not neen kept, and that the plots might not have been contiguous, the practice at the time being to separate plots by discard strips, which would have been $3'$–$6'$ wide. Discard strips are not mentioned in the original paper, and the total area is known to be one acre. The total recorded grain yield of 1974 lbs is equivalent to 2.1 tonnes per hectare in the absence of discard strips, This seems very low by modern standards of roughly 8 tonnes/ha, but it is in good agreement with Broadbalk yields in the same year as reported by Andrews and Herzberg (1985, section 5). If $3'$ discard strips had been used, the yield would have been equivalent to 4.5 tonnes per hectare. By comparison, the best nine of eighteen Broadbalk plots in 1910 yielded in the range 1.8–2.2, the largest being 2.19 tonnes/ha. Mercer and Hall remark only that the one-acre area selected 'promised to be a fair crop for the season'. In other words, this is strong evidence that the plots were contiguous and that discard strips were not used. Thus, the plot area is 8.547 m^2.

Since there are noticeable column effects, attributed by McBratney and Webster (1981) to an earlier ridge and furrow system, we eliminate additive row and column effects in our analyses, and seek to estimate the coefficients in the spatial model. Row effects are not necessary, but are included nonetheless. Although we study the data one variable at a time, we consider the bivariate spatial model with covariance function

$$\mathrm{cov}(Y) = I_n \otimes \Sigma_0 + V \otimes \Sigma_1$$

where Y is a matrix of order $n \times 2$, Σ_0, Σ_1 are 2×2 covariance matrices, and V is the average negative log distance between plots. Then each linear combination Yl gives rise to a univariate conformal model (7) with variance components $\sigma_0^2 = l'\Sigma_0 l$ and $\sigma_1^2 = l'\Sigma_1 l$. In computing and plotting the marginal log likelihood, it was found to be helpful to take the unit of area to be ten square metres. The marginal log likelihoods for the coefficient γ in the conformal model (7) are plotted for straw and grain in Fig. 4. The point estimates are $\hat\gamma = 0.393$ for grain and $\hat\gamma = 0.366$ for straw, both with a standard error of 0.08. That is to say the estimated variance-component ratios $\hat{\gamma}/(1-\hat{\gamma})$ are 0.65 per unit area for grain and 0.58 for straw. For both straw and grain, the evidence for spatial dependence after eliminating additive row and column effects is overwhelming. In itself, this is not surprising: any reasonable model for spatial dependence will detect this. The real surprise lies in the evidence that both processes have the same coefficient γ, suggesting that the variance-component matrices are proportional, i.e. $\Sigma_2 \propto \Sigma_1$.

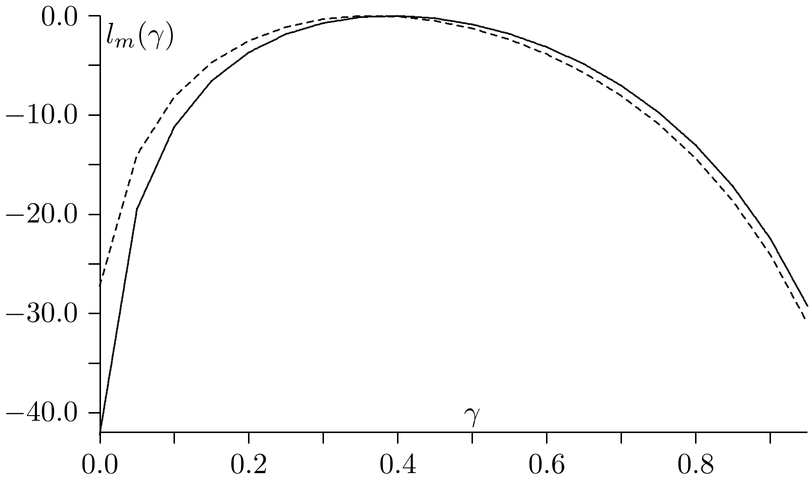

Fig. 4: Marginal log likelihood for the variance-component ratio. The solid line is for grain, the dashed line for straw.

To say that the variance-component matrices are proportional is to say that $\Sigma_0^{-1}\Sigma_1$ is a scalar, a multiple of the identity matrix. This hypothesis can easily be checked by considering various linear combinations $Y^\theta = Y^1 \cos\theta + Y^2 \sin\theta$ of straw and grain. For each $0 \le \theta < \pi$, the conformal model is fitted to Y^θ, and the variance-component ratio $\hat{\gamma}(\theta)$ obtained. If $\Sigma_0^{-1}\Sigma_1$ is a scalar then $\gamma(\theta)$ is constant, and $\hat{\gamma}(\theta)$ should be approximately constant. We already know that $\hat{\gamma}(0)$ is approximately equal to $\hat{\gamma}(\pi/2)$. However, Fig 5 shows conclusively that $\gamma(\theta)$ is not constant, but ranges from 0.07 to 0.50. Then $0.07/0.93 = 0.075$ and 1.0 are the estimated minimum and maximum eigenvalues of $\Sigma_0^{-1}\Sigma_1$. Even for the least favourable linear combination ($\theta \simeq 140°$), the evidence of spatial dependence is very strong with a maximized log likelihood of around 4.8.

As a check on the model, it is helpful to consider four separate quadrants, each 10×12 and one quarter acre in area, omitting the last column for computational convenience. This data splitting allows us to examine whether the variance-component ratio parameter γ in the conformal model is constant or varies across the region. The values of $\hat{\gamma}$ obtained in this manner are as follows

quadrant	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
grain	0.455	0.492	0.324	0.387
straw	0.539	0.229	0.436	0.305

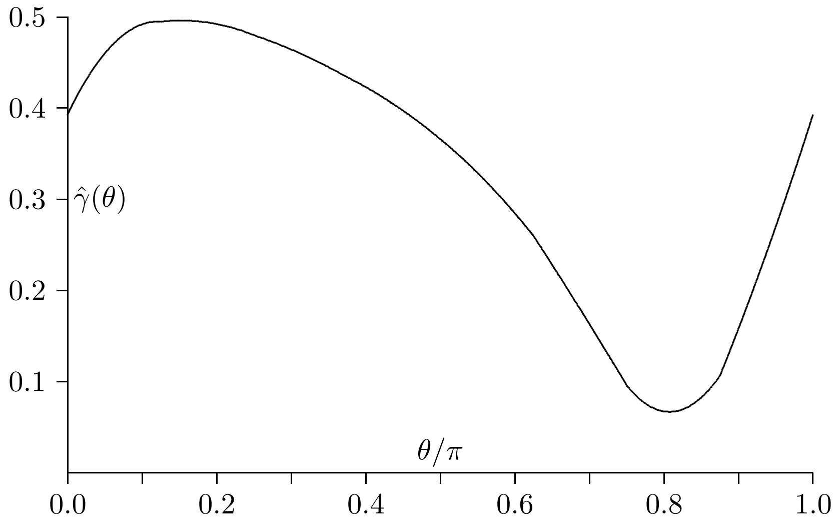

Fig. 5: Fitted conformal coefficient $\hat{\gamma}(\theta)$ as a function of the linear combination $Y^1 \cos\theta + Y^2 \sin\theta$ of grain and straw yields.

The individual standard errors are around 0.18–0.20, so there is no evidence of trend or excess variation. In all cases, the evidence against the null value $\gamma = 0$ is at least moderately strong.

4.5 *Alternative Gaussian models*

The most commonly used Gaussian models are the stationary isotropic models, which are invariant under the group of rigid motions, so that the covariance function $K(z, z')$ depends only on spatial separation $x = |z - z'|$. The family $\{K(\lambda x) : \lambda > 0\}$ is then closed under the similarity group generated by planar translations, rotations and dilations. Three such families suffice for illustrative purposes

$$\sigma^2 \exp(-\lambda x), \qquad \sigma^2(\lambda x)K_1(\lambda x), \quad \text{and} \quad \sigma^2 x^{-\lambda}/\lambda.$$

In the second of these, derived by Whittle (1954) specifically for planar processes, K_1 is the modified Bessel function of order 1. The third expression, the Newtonian potential in Euclidean space of dimension $\nu = 2 + \lambda$, is interpreted as $-\log(x) + \text{const}$ when $\lambda = 0$. The function $x^{-\lambda}/\lambda$ is a covariance function for spatial contrasts in Euclidean space of dimension d provided that $\lambda < d$. Other examples of covariance functions can be found in the literature on spatial models and random fields (Stein, 1999; Chilès and Delfiner 199? Matheron 199?). In particular, Stein recommends using the flexible Matérn class of covariance functions $\sigma^2 x^\nu K_\nu(\lambda x)$, where $\nu > 0$ and K_ν is the modified Bessel function of order ν. One attraction of the Matérn class is that it includes Whittle's model at $\nu = 1$ and the exponential model at $\nu = 1/2$. On the negative side, however, it may be more flexible than is necessary for specific applications such as field trials. The index ν governs the mean square smoothness.

In applications to agricultural processes, it is essential to include a white noise contribution, so the smallest non-trivial family has the form

$$\sigma_0^2\delta_{z-z'} + \sigma_1^2 K(\lambda x),$$

with two variance parameters and a range parameter. The conformal model described in the preceding section has this form, but there is no range parameter because $K(x) = -\log(x)$. The power family also does not have a range parameter.

We will not consider these models in great detail except by way of a check on the adequacy of the conformal model. In the marginal log likelihood, each of the alternative models has two free parameters whereas the conformal model has only one. The conformal model is a sub-model of the power family, so the difference in log likelihoods constitutes the standard test. It is not a sub-model of the other two, so only an informal comparison is available. Table ? gives the maximized marginal log likelihoods for the conformal model and each of the three models described above for nine data sets.

Table ?. Maximized marginal log likelihoods for various models.

	Spatial covariance model			
Data set	Conformal	Exponential	Whittle	Power
Arlington I	26.24			
Arlington II	10.92			
Antelope	26.89			
Valencia	10.70			
Eureka	16.72			
Walnuts	5.17			
Jonathan	0.00			
M-H grain	42.20			
M-H straw	27.22			

5. A Gaussian model for cluster analysis

5.1 *Configuration*

Banfield and Raftery (1993) remark that cluster analysis has developed mainly through the invention and empirical investigation of ad hoc methods in isolation from more formal statistical procedures. Algorithms for clustering are plentiful in the literature, but statistical models are scarce. See Gower (1967) for a comparison of the algorithms then in use for taxonomic purposes. In this section, we construct a Gaussian model for cluster analysis and illustrate its application. We do not, however, go so far as developing an algorithm for finding the clusterings or partitions that have highest posterior probability.

Suppose that we wish to group the n units into a small number of homogeneous clusters based on the observed values $y_i \in \mathcal{R}^q$, and that this is to be done without benefit of covariate information. In principle, all clusterings must be examined and compared. It is assumed that any clustering model is, or should be, unaffected by invertible affine transformation of the n response points in $\mathcal{R}^q$. In other words, the clustering model is unaffected by any transformation in which each y_i is sent to

$$y_i \mapsto a + By_i$$

where $a \in \mathcal{R}^q$ and B is invertible (Friedman and Rubin 1967). Another way of saying the same thing is that the likelihood function should depend on the data only through the configuration statistic, and this argument leads naturally to the marginal log likelihood in section 2.3.

A clustering of the observed units is a partition of the set $\mathbf{n} = \{1, \ldots, n\}$ into disjoint non-empty subsets whose union is $\mathbf{n}$. The clusters are unlabelled, so a clustering is not to be confused with a classification of the units. For example, if there are four units $\mathbf{n} = \{1, 2, 3, 4\}$, the two classifications

$$\begin{aligned} \varphi &= \{1 \mapsto a, \quad 2 \mapsto b, \quad 3 \mapsto a, \quad 4 \mapsto b\} \\ \varphi' &= \{1 \mapsto b, \quad 2 \mapsto a, \quad 3 \mapsto b, \quad 4 \mapsto a\} \end{aligned}$$

are different functions. The function φ determines two classes, so we may talk of the class a, i.e. the set $\varphi^{-1}a = \{1, 3\}$, and the class b or the set $\varphi^{-1}b = \{2, 4\}$. In the classification φ', the labels are reversed. So far as clusterings are concerned, there is only one partition $13|24$, an abbreviation for $\{\{1, 3\}, \{2, 4\}\}$, and the blocks are unlabelled. This distinction may seem pedantic, but it is crucial for modelling. It means that we cannot begin the modelling exercise by saying 'let x denote the class labels' as in Scott and Symons (1971) or Banfield and Raftery (1993). The difference is also crucial from a Bayesian perspective in which one must begin with a prior distribution, either on classifications or on partitions, the two being very different even for moderate vaues of n. The present approach is close in spirit to Friedman and Rubin (1967), who use a number of invariant classification criteria but with no explicit statistical model.

5.2 *Exchangeable random partitions*

A clustering of the observed units is a partition of the set $\mathbf{n} = \{1, \ldots, n\}$ into disjoint non-empty subsets whose union is $\mathbf{n}$. To say the same thing with less risk of ambiguity, a partition E is an equivalence relation on $\mathbf{n}$, a Boolean function $E: \mathbf{n} \times \mathbf{n} \to \{0, 1\}$ that is reflexive, symmetric and transitive. In other words, E is a symmetric binary matrix, and if the units are suitably ordered, E is a block-diagonal matrix, the blocks being the blocks of the partition. The set $\mathcal{E}_n$ of all equivalence relations on $\mathbf{n}$ is the finite set of partitions of $\mathbf{n}$, also called the partition lattice. Accordingly, it is natural to consider a linear Gaussian model for the conditional distribution of Y given E in which $X = 1$ consists of the intercept only. The family of covariance matrices has the form

$$\operatorname{cov}(Y \mid E) = I_n \otimes \Sigma_0 + E \otimes \Sigma_1 \tag{31}$$

in which Σ_0, Σ_1 are positive definite $q \times q$ covariance matrices, and $E \in \mathcal{E}_n$.

The data on which our analysis will be based is the scaled residual, $\check{y}$, whose distribution depends only on the matrix $\Sigma_0^{-1}\Sigma_1$ together with the partition E. To simplify the calculations in the analysis that follows, this matrix is assumed to be a scalar, i.e. $\operatorname{cov}(Y \mid E) = \big((1-\gamma)I_n + \gamma E\big) \otimes \Sigma$. The probability of the data is thus

$$p(\check{y}; \gamma) = \sum_{E \in \mathcal{E}_n} p_n(E)\, p(\check{y} \mid E; \gamma), \tag{32}$$

where the second factor is given by (6) with suitable modification of notation. To complete the specification, we require a distribution or family of distributions on the set of partitions. Since the units are regarded as exchangeable, the sequence of distributions $\{p_n(E)\}$ determines an infinitely exchangeable process. This statement is interpreted in the standard de Finetti sense as follows. First, for each set $\mathbf{n}$, the distribution $p_n(\cdot)$ on $\mathcal{E}_n$ is invariant under permutation of units. Second, $p_n(\cdot)$ is the marginal distribution of $p_{n+1}(\cdot)$ under selection of units. Equivalently, if K is a random partition of $\{1, \ldots, n+1\}$ with distribution $p_{n+1}(\cdot)$ on $\mathcal{E}_{n+1}$, deletion of one unit yields a random partition with distribution $p_n(\cdot)$ on $\mathcal{E}_n$.

Ewens's distribution with scalar parameter $\lambda > 0$ is the simplest infinitely exchangeable process on partitions. To each set partition $E \in \mathcal{E}_n$ there corresponds a number partition $N(E)$ such that $N(E) = 1^{e_1} 2^{e_2} \cdots n^{e_n}$, this being the maximal invariant under permutation. In other words, E contains e_r blocks of size r, so that $e_{\bullet} = \sum e_r$ is the number of blocks, and $n = \sum r e_r$. The Ewens distribution on $\mathcal{E}_n$ with parameter $\lambda > 0$ is

$$p_n(E; \lambda) = \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{r=1}^{n} [\lambda (r-1)!]^{e_r} = \frac{\Gamma(\lambda)\, \lambda^{e_{\bullet}}}{\Gamma(n+\lambda)} \prod_r [(r-1)!]^{e_r}. \tag{33}$$

This is the first factor on the right in (32). For each probability distribution π on the positive numbers we write

$$p_n(E; \pi) = \int p_n(E; \lambda)\, d\pi(\lambda)$$

for the π-mixture. As usual in the theory of infinite exchangeability, every mixture of exchangeable processes is also an exchangeable process.

The choice of the Ewens distribution on partitions is not so arbitrary as it might appear. There are indeed other infinitely exchangeable random partitions, but there appear to be no others that have the sort of limiting behaviour that one would demand for clustering applications. In particular, the number of clusters should tend slowly to infinity as $n \to \infty$, and the distribution of cluster fractions should have a suitable non-degenerate limit. In particular, as Scott an Symons (1971) note, it is unreasonable that the block sizes should be approximately equal for large n. There is no satisfactory alternative to the Ewens family or Ewens mixtures.

For each fixed n, the Ewens family is an exponential family on the partition lattice in which the number of blocks is the canonical sufficient statistic. The cumulant function is $K(\theta) = \log\Gamma(n + e^\theta) - \log\Gamma(e^\theta)$, where $\lambda = e^\theta$, so the mean number of blocks is $\lambda\bigl(\psi(n+\lambda) - \psi(\lambda)\bigr)$, which is approximately $\lambda\log(1+n/\lambda)$. The probability that E consists of a single block is $\Gamma(\lambda)\Gamma(n)/\Gamma(n+\lambda)$, or roughly $n^{-\lambda}$. The number of blocks is roughly Poisson, but the variance is a little greater than the mean. For applications to clustering, values near $\lambda = 1$ are not unreasonable. If a Bayesian analysis is required, we may choose a suitable non-degenerate prior distribution on λ.

Suppose first that the parameters γ and λ are known, for example that both are equal to one. Application of Bayes's theorem gives

$$p(E \mid \check{y}, \lambda, \gamma) = \frac{p(\check{y} \mid E; \gamma)\, p_n(E; \lambda)}{\sum_{\mathcal{E}_n} p(\check{y} \mid E'; \gamma)\, p_n(E'; \lambda)}$$

for the posterior distribution on $\mathcal{E}_n$. More generally, if λ and γ are distributed independently, the posterior is

$$p(E \mid \check{y}, \pi_\lambda, \pi_\gamma) \propto p(\check{y} \mid E; \pi_\gamma)\, p_n(E; \pi_\lambda)$$

proportional to the product of the mixtures. In the example described below, we take $\lambda = 1$ to be known, γ uniformly distributed on $(0, 1)$, and use a Laplace approximation for the first factor. Here γ is the ratio of the between-blocks variance to the total variance.

5.3 *Illustration*

The method described in the preceding section was applied to the data shown in Fig. ? to see if there is evidence of one or more clusters. These data are in fact the log petal length and log petal width of *Iris versicolor* and *Iris virginica* in Fisher's data, so the actual classification is known. In the left panel, the 42 elements labelled '2' are *versicolor*; the 41 labelled '0' are *virginica*. Of the 17 units labelled '1', eight are *versicolor* and nine are *virginica*. Among these 17 units, there are only 13 distinct values.

Given the identifying labels, the two types are reasonably well separated, so one might expect that the partition should be identifiable from the measurements alone. This intuition

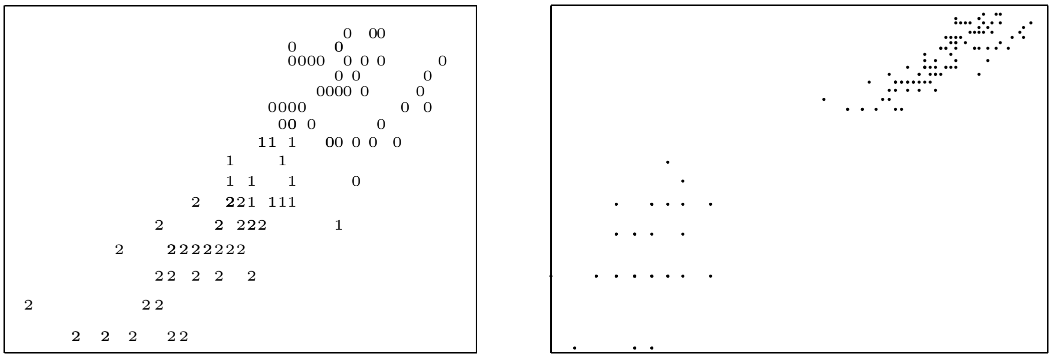

Fig. ?: Scatterplot of log petal length and width for *Iris versicolor* and *I. virginica* (left panel): Right panel includes *I. setosa*.

is simply wrong. The evidence for clusters exists in the labels, not in the measurements. That is to say, if the labels are ignored, little evidence exists for clusters.

To simplify the analysis, we consider only a few of the potential clusterings into two groups, in which the units labelled '0' and '2' are kept in different classes, and the remaining units are assigned to one class or the other. There are 2^{17} partitions of this type, one of which is the actual partition determined by the recorded varieties. One additional partition is included in the analysis for comparative purposes, namely the partition into a single class of size 100. The labels are otherwise ignored, and the prior is exchangeable with $\lambda = 1$, so the expected prior number of blocks is 5.2. In this simplified analysis, the marginal likelihood is computed for each of $2^{17} + 1$ partitions and integrated with respect to the uniform prior distribution on the variance-component ratio γ.

Take the trivial partition with one block as a baseline for comparisons. For each of the 2^{17} partitions, the log likelihood has a maximum near $\hat{\gamma} = 0.7$, and the maximized log likelihood ranges from roughly 50 to 60 above the baseline. The second derivative at $\hat{\lambda}$ ranges from 19 to 24. For the observed partition, the values are $\hat{\gamma} = 0.70$, $\check{l}(\hat{\gamma}) = 54.8$, and $-l''(\hat{\gamma}) = 21.8$. On the basis of the likelihood function alone, it seems that there is strong evidence in favour of two-block partitions. However, the prior on each of these partitions is less than the prior on the baseline partition by a factor of roughly $99!/(49!)^2$, or $\exp(70)$. In order for the posterior to be concentrated on a small mumber of partitions into two roughly similar blocks, the maximized log likelihoods need to be roughly 70 units greater than the baseline. No partitions come close to that level in the present example, so we conclude that no evidence can be found for variety clusters. The posterior probability of all 2^{17} partitions combined is less than 1% of the posterior probability for the baseline.

When the 50 observations for *setosa* are included in the analysis, the conclusion regarding the existence of clusters is very different. The *setosa* cluster is visually well separated from the other two. The two-block partition achieves a log likelihood that is 194.6 units in excess of the baseline, and the three-block partition is 240.6 units above the baseline. The second derivatives are 313.4 and 235.2 respectively. The prior distributions are lower than the baseline by 96.3 and 166.3 units on the log scale, but this is insufficient to offset the log likelihood. Consequently, the posterior distribution on partitions places almost all of its mass on the partition containing two blocks, one block for *setosa* and one for the remainder. The posterior probability for this partition relative to the baseline is approximately 10^{41}, and 10^{11} relative to the three-block partition. The evidence in the measurements points to two clusters, not three.

5.4 *algebraic structure*

The precise nature of what is required of a statistical model for cluster analysis is most clearly revealed when all components of the system are stated in terms of categories, functors and natural transformations (McCullagh 2002). The entire structure rests on the category $\mathcal{C}$ of clusterings or partitions. Each object E in $\mathcal{C}$ is a partition of a finite set $\mathcal{U}$:

$$E \in \mathrm{Obj}(\mathcal{C}): \qquad \mathcal{U} \times \mathcal{U} \xrightarrow{E} \{0,1\}$$

a Boolean function that is reflexive, symmetrric and transitive. Let $E'{:}\,\mathcal{U}' \times \mathcal{U}' \to \{0,1\}$ be another partition in $\mathcal{C}$ possibly of a different finite set. A morphism of E into E' is an injective map $\varphi{:}\,\mathcal{U} \to \mathcal{U}'$ that preserves Boolean relationships: for each $x, y \in \mathcal{U}$, the partition E' satisfies $E'(\varphi x, \varphi y) = E(x, y)$. In terms of a category diagram, the square in the diagram below commutes.

$$\begin{array}{ccccc} E & \qquad & \mathcal{U}\times\mathcal{U} & \xrightarrow{E} & \{0,1\} \\ \varphi\big\downarrow & & \varphi\big\downarrow\ \varphi\big\downarrow & & \| \\ E' & & \mathcal{U}'\times\mathcal{U}' & \xrightarrow{E} & \{0,1\} \end{array}$$

The projection that sends each partition E to its set $\mathcal{U}$, and each morphism $\varphi{:}\, E \to E'$ to the map $\varphi{:}\,\mathcal{U} \to \mathcal{U}'$ on sets is a functor $\mathcal{C} \to \mathbf{Set}$. It is faithful (injective on arrows).

Consider now the functor $\mathcal{E}$ that sends each finite set $\mathcal{U}$ to the set $\mathcal{E}_{\mathcal{U}}$ of partitions of $\mathcal{U}$, and each map $\varphi{:}\,\mathcal{U} \to \mathcal{U}'$ to a map $\mathcal{E}\varphi{:}\, \mathcal{E}_{\mathcal{U}'} \to \mathcal{E}_{\mathcal{U}}$ on partitions defined by

$$(\mathcal{E}\varphi)(x, y) = E'(\varphi x, \varphi y)$$

for each $x, y \in \mathcal{U}'$. Composition of morphisms is preserved, but the order of composition and the direction of arrows is reversed, so $\mathcal{E}$ is a contravariant functor $\mathbf{Set} \to \mathbf{Set}$. In statistical applications connected with clustering, the functor $\mathcal{E}$ necessarily serves as one component of the parameter space. In order for this object to be usable as a component in a Bayesian model, there must be a prior distribution, or process, on $\mathcal{E}$. Since $\mathcal{E}_{\mathcal{U}}$ is finite for each $\mathcal{U}$, it is natural to take the power set as the σ-algebra of events on which probabilities are defined. Let $\P(\mathcal{E}_\mathcal{U})$ be the set of all probability distributions on $E_\mathcal{U}$. A (prior) distribution, or process, is a function π that associates with each equivalence relation $E\colon \mathcal{U}\times\mathcal{U}\to\{0,1\}$ in $\mathcal{C}$ a distribution π_E on $\mathcal{E}_\mathcal{U}$ in such a way that, for each $\varphi\colon E\to E'$ in $\mathcal{C}$, the square in the diagram below commutes.

$$
\begin{tikzcd}
E \arrow[d, "\varphi"'] & \mathcal{E}_\mathcal{U} & \{0\} \arrow[r, "\pi_E"] \arrow[d, equal] & \P(\mathcal{E}_\mathcal{U}) \\
E' & \mathcal{E}_{\mathcal{U}'} \arrow[u, "\mathcal{E}\varphi"] & \{0\} \arrow[r, "\pi_{E'}"] & \P(\mathcal{E}_{\mathcal{U}'}) \arrow[u, "\P\mathcal{E}\varphi"']
\end{tikzcd}
$$

Here, φ is a morphism from one partition into another, $\mathcal{E}\varphi$ is the composition map from one partition lattice into another, and $\P\mathcal{E}\varphi$ is the corresponding map on probability distributions, sending each probability distribution on the lattice $\mathcal{E}_{\mathcal{U}'}$ to a marginal distribution on the lattice $\mathcal{E}_\mathcal{U}$.

For each $\lambda > 0$, the sequence of Ewens distributions satisfies the exchangeability condition, and is thus a suitable choice for a prior distribution. As we remarked ealier, there are others.

5.5 *How much separation?*

References

Andrews, D.A. and Herzberg, (1985) *Data.* Springer.

Bartlett, M.S. (1978) Nearest neighbour models in the analysis of field experiments (with discussion). *J. R. Statist. Soc.* B **40**, 147–174.

Batchelor, L.D. and Reed, H.S. (1918) Relation of the variability of yields of fruit trees to the accuracy of field trials. *J. Agricultural Research* **12**, 245–283.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B **36**, 192–236.

Besag, J. and Kooperberg, C. (1995) On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746

Besag, J. and Higdon, D. (1999) Bayesian analysis of field experiments using neighbouring plots (with discussion). *J. R. Statist. Soc.* B **61**, 691–746.

Diggle, P. (1988) An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–971.

Fisher, R.A. (1936) The use of multiple measures in taxonomic problems. *Ann. Eugenics* **8**, 376–386.

Friedman, H.P. and Rubin, R. (1967) On some invariant criteria for grouping data. *J. Am. Statist. Assoc.* **62**, 1159–1178.

Geisser, S. (1964) Posterior odds for multivariate normal classifications. *J. R. Statist. Soc.* B **26**, 69–76.

Gower, J. (1967) A comparison of some methods of cluster analysis. *Biometrics* **23**, 623–637.

Harville, D.A. (1977) Maximum-likelihood approaches to variance-component estimation and to related problems. *J. Am. Statist. Assoc.* **72**, 320–338.

Harville, D.A. (1978) Alternative formulations and procedures for the two-way mixed model. *Biometrics* **34**, 441–453.

Lee, Y. and Nelder, J.A. (2001) Modelling and analysing correlated non-normal data. *Statistical Modelling* **1**, 3–16.

McBratney, A.B. and Webster, R. (1981) Detection of a ridge and furrow pattern by spectral analysis of crop yield. *Int. Statist. Rev.* **49**, 45–52.

Mercer, W.B. and Hall, A.D. (1911) The experimental error of field trials. *J. Agric. Sci.* **4**, 107–132.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Patterson, H.D. and Thompson, R. (1974) Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference* 197–207.

Banfield, J.D. and Raftery, A.E. (1993) Model-based non-Gaussian clustering. *Biometrics* **49**, 803–821.

Scot, A.J. and Symons, M.J. (1971) Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–397.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika* **49**, 305–314.

Zimmerman, D.L. and Harville, D.A. (1991) A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics* **47**, 223–239.