Quotient spaces and statistical models

Peter McCullagh
Department of Statistics
University of Chicago

Summary

The purpose of this paper is to draw attention to the widespread occurrence of quotient spaces in statistical work. Quotient spaces are intrinsic to probability distributions, residuals, interaction, test statistics, and incomplete observations. The theme is that explicit recognition of the quotient-space can offer surprising conceptual simplification. The advantages of working directly with the quotient space are hard to describe in general. As the examples demonstrate, the answer lies partly in directness of approach.

1. Probability distributions

1.1 *The probability simplex*

Let $\Omega = \{\omega_1, \ldots, \omega_k\}$ be a finite set containing k elements. The vector space $\mathcal{V} = \mathcal{R}^\Omega$ is the set of real-valued functions on Ω: $\mathcal{V} = \{f : \Omega \mapsto \mathcal{R}\}$. Clearly, $\mathcal{V}$ is a vector space of dimension k. One particularly important subset, denoted here by $\mathbf{1}$, is the set of functions that are constant on Ω:

$$\mathbf{1} = \{f \in \mathcal{V} : f(\omega_1) = \cdots = f(\omega_k)\}.$$

It is evident that $\mathbf{1}$ is a subspace of dimension 1, isomorphic with $\mathcal{R}$.

The probability simplex on Ω is that subset of $\mathcal{V}$ consisting of non-negative real-valued functions whose sum is one:

$$\mathcal{S} = \{p \in \mathcal{V} : p(\omega) \geq 0; \sum p(\omega) = 1\}.$$

Each point p in the simplex is a probability distribution on Ω.

The simplex is a bounded set, and thus not a vector space. Nevertheless, there exists a very natural association of $\mathcal{S}$ with a vector space. We associate with any point $\eta \in \mathcal{V}$, a probability distribution $p \in \mathcal{S}$ as follows:

$$p(\omega) = p_\omega = \exp(\eta_\omega) \Big/ \sum \exp(\eta_\omega). \tag{1}$$

It is clear that if we replace each η_ω by $\eta_\omega + c$ in the preceding expression, the probability distribution is unaffected. In other words, (1) is a function on $\mathcal{V}$ that is constant on the cosets of $\mathbf{1}$. To say the same thing in another way, (1) is a function from the quotient space $\mathcal{V}/\mathbf{1}$ into the simplex. Further, this function is 1–1. In other words, distinct cosets are mapped to distinct points in the simplex.

A slight complication occurs here in that every point in $\mathcal{V}/\mathbf{1}$ is mapped to an *interior* point of the simplex: every interior point p in S has an inverse image η in $\mathcal{V}/\mathbf{1}$ given by

$$\eta(\omega) = \log(p(\omega)) + c, \tag{2}$$

where c is any arbitrary constant. Boundary points on the simplex have no inverse image in $\mathcal{V}/\mathbf{1}$, except as limit points.

As with all generalized linear models, the purpose of transformation is to associate a certain vector space with the set of probability distributions. In the sense of isomorphism, any $(k-1)$-dimensional vector space can be used for this purpose. For example, we might have chosen the subspace of $\mathcal{V}$ consisting of vectors η whose components sum to zero, while retaining (1). The choice between these is not a choice between right and wrong, but between what is natural and what is not. The theme of this paper is that, in this context, the quotient space is a more natural choice than any subspace of $\mathcal{V}$.

1.2 *Log likelihood function*

In this section Ω is interpreted as a parameter space, and $\mathcal{V} = \mathcal{R}^\Omega$ as before. To keep matters simple, Ω is finite, or at least countable. Let $p(y;\omega)$ be the probability of observing y when the parameter is ω. In probability calculations, p is considered primarily as a function of y for fixed ω. In likelihood calculations, the roles are reversed: for inferential purposes, given the particular data y observed, $l(\omega;y) = \log p(y;\omega)$ is a function on Ω. It may be the case that, for the particular y observed, $p(y;\omega)$ is constant on Ω, in which case this observation is uninformative for selecting among the possible values in Ω. For inferential purposes, any function that is constant on Ω is equivalent to zero. The log likelihood is a function on Ω defined by

$$l(\omega;y) = \log p(y;\omega),$$

together with the equivalence relation $l_1 \sim l_2$ if $l_1 - l_2 \in \mathbf{1}$. To say the same thing in another way, the log likelihood is a vector in the quotient space $\mathcal{V}/\mathbf{1}$. Any additive constant is irrelevant even if it depends on y.

1.3 *Bayes's theorem*

Let Ω be the parameter space, and let π be a prior distribution on Ω. Let y denote the observed data, and let $p(y;\omega)$ be the probability of y for parameter ω. Both the prior distribution π and the posterior distribution $\pi(\omega \,|\, y)$ are points in the simplex, having inverse images $\log \pi + \mathbf{1}$ and $\log \pi(\cdot \,|\, y) + \mathbf{1}$, both in the quotient space $\mathcal{V}/\mathbf{1}$. Likewise, the log likelihood $l(\cdot;y) + \mathbf{1}$ is a point in the same space.

The usual expression for Bayes's theorem

$$\pi(\omega \,|\, y) = \frac{p(y;\omega)\,\pi(\omega)}{\sum p(y;j)\,\pi(j)}.$$

can be interpreted as vector addition in the quotient space

$$\log \pi(\omega \,|\, y) + \mathbf{1} = (l(\omega;y) + \mathbf{1}) + (\log \pi(\omega) + \mathbf{1}).$$

This may be abbreviated to the more familiar form

$$\log \pi(\omega \,|\, y) = l(\omega;y) + \log \pi(\omega),$$

in which it is understood that these are functions in $\mathcal{V}/\mathbf{1}$, equivalent modulo constant functions. Bayes's theorem is thus interpreted as vector addition in $\mathcal{V}/\mathbf{1}$.

By associating the prior with a point in $\mathcal{V}/\mathbf{1}$, it is implicitly assumed that π lies in the interior of the simplex. In other words, no element of Ω has zero prior probability. This difficulty can be avoided by eliminating from Ω all points having zero prior probability. Alternatively, for computational purposes, a value $-\infty$ can be assigned to $\log \pi$ for those elements having zero prior probability.

1.4 *Log-linear and multinomial response models*

In a multinomial response model, it is essential to distinguish between the response factor with levels Ω, and all other factors with composite levels Ω_A. The principal aim of a multinomial response model is to study how the response vector π in $\mathcal{S}$, or the corresponding vector $\log \pi$ in $\mathcal{V}/\mathbf{1}$, depends on experimental or design conditions as encoded in the levels Ω_A.

Let A be the vector space of real-valued functions on Ω_A. In other words, a vector $\alpha \in A$ is a function assigning to each $i \in \Omega_A$ a real number α_i.

Let π be a list of n probability vectors in $\mathcal{S}$, each vector in the list being associated with an element of the set Ω_A. In other words, for any $i \in \Omega_A$,

$$\pi_i = (\pi_{i1}, \ldots, \pi_{ik})$$

is a probability vector in the simplex $\mathcal{S}$, and $\log \pi$ is a function from Ω_A into the quotient space $\mathcal{V}/\mathbf{1}$. The ordered list of n vectors $(\log \pi_1, \ldots, \log \pi_n)$, indexed by Ω_A, determine a point in the direct sum space

$$(\mathcal{V}/\mathbf{1}) \oplus \cdots \oplus (V/1) \cong \mathcal{V}^{\Omega_A}/(1^{\Omega_A}) \cong (\mathcal{V} \otimes A)/A.$$

Every point in this set corresponds to a list of multinomial probabilities. Conversely, every list of multinomial response probabilities corresponds to a point or limit in this set.

Each element of $\mathcal{V}^{\Omega_A}$ is a list, indexed by Ω_A, of functions from Ω into $\mathcal{R}$. In other words, $f = (f_1, \ldots, f_n)$ where $f_i = (f_{i1}, \ldots, f_{ik})$, with $i \in \Omega_A$. Thus there is a natural isomorphism between $\mathcal{V}^{\Omega_A}$ and the tensor product space $\mathcal{V} \otimes A$ of real-valued functions on $\Omega \times \Omega_A$. The set of multinomial response probabilities is thus in 1–1 correspondence with the quotient space $(\mathcal{V} \otimes A)/(1 \otimes A)$, also written as $(\mathcal{V} \otimes A)/A$.

Let μ be any non-negative function on $\Omega \times \Omega_A$ and let $\eta = \log \mu$. Any Poisson response model of the form $\eta \in \mathcal{M}$ is called log-linear if $\mathcal{M}$ is a subspace of $\mathcal{V} \otimes A$. For present purposes, however, $\mathcal{M}$ is an arbitrary subset of $\mathcal{V} \otimes A$, not necessarily a subspace. The conditions under which $\mathcal{M}$ corresponds to a multinomial response model are easily expressed as follows: $\mathcal{M}$ is a multinomial response model if and only if $\mathcal{M}$ is a set of cosets of A. Equivalently, $\mathcal{M} = \mathcal{M} + A$, in which it is understood that $A \equiv A \otimes 1$. Generally speaking, in models for ordinal responses (McCullagh, 1980), or models containing multiplicative effects (Anderson, 1984), $\mathcal{M}$ is not a subspace.

For log-linear models in which $\mathcal{M}$ is a subspace, the condition $\mathcal{M} = \mathcal{M} + A$ reduces to $(1 \otimes A) \subset \mathcal{M}$, which is the familiar condition given by Palmgren (1981).

1.5 *Sampling*

Let Ω be the population of N units, Ω_0 the sample of n units, and Ω_1 the unsampled units. A statistical variate y is thus a vector in $\mathcal{V} = \mathcal{R}^\Omega$; the observed variate y_0, the restriction of y to the subset Ω_0, is a vector in $\mathcal{R}^{\Omega_0}$. Let $\mathcal{V}_0$ be the subspace of $\mathcal{V}$ consisting of all functions that are zero on Ω_0, and let $\mathcal{V}_1$ be the complementary subspace of functions that are zero on Ω_1. Thus $\mathcal{V}_1$ has dimension n and $\mathcal{V}_0$ has dimension $N - n$. Two vectors y, y' in $\mathcal{V}$ such that $y - y' \in \mathcal{V}_0$ have the property that their restrictions to Ω_0 are equal. Thus, all vectors in the coset $y + \mathcal{V}_0$ give rise to the same observation in $\mathcal{R}^{\Omega_0}$. In other words, each point in $\mathcal{R}^{\Omega_0}$ identifies a coset of $\mathcal{V}_0$ in $\mathcal{V}$. The vector space $\mathcal{R}_0^\Omega$ is thus naturally isomorphic with the quotient $\mathcal{V}/\mathcal{V}_0$.

It should be emphasized that, although the dimensions are equal, the quotient space $\mathcal{V}/\mathcal{V}_0$ is quite different from the subspace $\mathcal{V}_1$. In particular, if $\mathcal{V}$ is an inner product space, $\mathcal{V}_1$ and the quotient $\mathcal{V}/\mathcal{V}_0$ typically have incompatible inner products.

Let $y_0 \in \mathcal{R}^{\Omega_0}$ denote the observed value. This can, be represented by the partitioned vector $(y_0, \star)$, where $\star$ denotes missing or arbitrary values on Ω_1. Suppose that the distribution of the random vector y on $\mathcal{V}$ has zero mean and positive-definite covariance matrix partitioned according to (Ω_0, Ω_1) as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix}.$$

Then the best linear predictor for y given $(y, \star)$ is

$$E\big(y \,|\, (y_0, \star)\big) = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix} \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_0 \\ \star \end{pmatrix}.$$

In the discussion that follows, we regard $\mathcal{V}$ as an inner product space with inner product matrix Σ. The transformation given above is then a self-conjugate projection on $\mathcal{V}$ i.e. an orthogonal projection, with null space $\mathcal{V}_0$ and range $\mathcal{V}_0^\perp$.

2. Linear models

2.1 *Regression coefficients*

Let Ω be the set of statistical units on which observations are made, and let $\mathcal{V} = \mathcal{R}^\Omega$ be the vector space of functions on these units. The response vector y and its expectation $\mu = E(Y)$ are points in $\mathcal{V}$. A linear regression model is an assertion that μ lies in a subspace $\mathcal{X}$ spanned by given vectors $x_1, \ldots, x_p$. This is usually written in the form

$$\mu = x_1\beta_1 + \cdots + x_p\beta_p,$$

or $\mu = X\beta$, where β is a vector of coefficients to be estimated. The least-squares estimate of μ is $\hat{\mu} = Py$, where P is the orthogonal projection on to $\mathcal{X}$. Alternatively, in the appropriate metric, $y - \hat{\mu}$ is orthogonal to $\mathcal{X}$.

Suppose now that we wish to extend the model by adding a further covariate z. The extended model is thus $\mu \in \mathcal{X} + \mathcal{Z}$, or

$$\mu = x_1\beta_1 + \cdots + x_p\beta_p + \gamma z.$$

The least-squares estimate $\hat{\gamma}$ is a function $\hat{\gamma}(y, z)$ having the following properties for each $x \in \mathcal{X}$:

1. $\hat{\gamma}(y + x, z) = \hat{\gamma}(y, z)$.
2. $\hat{\gamma}(y_1 + y_2, z) = \hat{\gamma}(y_1, z) + \hat{\gamma}(y_2, z)$.
3. $\hat{\gamma}(y, z + x) = \hat{\gamma}(y, z)$.

Properties 1 and 2 state that $\hat{\gamma}$ is a linear function of y that is constant on cosets of $\mathcal{X}$. In other words, $\hat{\gamma}$ is a linear function on the quotient space $\mathcal{V}/\mathcal{X}$. So far as the second argument is concerned, $\hat{\gamma}$ is a non-linear function on the quotient space. These properties are apparent from the explicit matrix expression

$$\hat{\gamma} = \frac{z^T W Q y}{z^T W Q z}$$

where W is the matrix of the inner product, and $Q = I - P$ is the orthogonal projection on to $\mathcal{X}^\perp$. The matrix expression $z^T W Q y$ is in fact the quotient-space inner product $\langle z + \mathcal{X},\, y + \mathcal{X} \rangle$ in $\mathcal{V}/\mathcal{X}$.

2.2 *Interaction*

Consider an experimental design with two factors, A with levels Ω_A and B with levels Ω_B, and a real-valued response y. By additivity of effects, we mean that the expected response μ is an additive function of the levels of the two factors. In other words, for some functions $\alpha \in A$ and $\beta \in B$

$$\mu_{ij} = \alpha_i + \beta_j.$$

It is irrelevant here whether the design is complete or balanced. To say the same thing in another way, the vector μ lies in the additive subspace $A + B$ in $A.B$. Interaction is what remains in $A.B$ after additive effects have been eliminated or ignored. The definition is as follows.

(i) μ has zero interaction if and only if $\mu \in A + B$.
(ii) Interaction is additive: $\mathcal{I}(\mu_1 + \mu_2) = \mathcal{I}(\mu_1) + \mathcal{I}(\mu_2)$.

There are various ways of defining interaction to satisfy these conditions. One way is to define $\mathcal{I}(\mu) = Q\mu$ as the projection along $A+B$ on to any subspace complementary to $A+B$ in $A.B$. Since no inner product is given, there is no reason to prefer one complementary subspace over another. The most direct and effective way is to define the interaction of μ as the coset $\mu + (A + B)$, i.e. μ modulo additivity. The interaction space is thereby identified with the quotient space $A.B/(A + B)$. Each point in the interaction space is a coset of $A + B$ in the space $A.B$, the zero coset being the additive subspace $A + B$.

It should be emphasized that the quotient space $A.B/(A + B)$ is not at all the same as the subspace of functions

$$\mathcal{W} = \{(i, j) \mapsto f_{ij} : \sum_i f_{ij} = \sum_j f_{ij} = 0\},$$

the usual complement of $A + B$ in $A.B$. In particular, unless the design is balanced, the norm of the projection of y on to the subspace is not the same as the norm of the projection on to the quotient space. For numerical comparison, suppose that A and B each have two levels, and that the inner product matrix is diagonal with components $W = \mathrm{diag}\{2, 1, 1, 1\}$, corresponding to replicate observations in the $(1, 1)$-cell only. The squared norm of the orthogonal projection of y on to the subspace is given by

$$\|P_{\mathcal{W}} y\|^2 = (2y_{11} - y_{12} - y_{21} + y_{22})^2/5.$$

The squared norm of the projection on to the quotient space is

$$\|P_{A.B}y\|^2 - \|P_{A+B}y\|^2 = 2(y_{11} - y_{12} - y_{21} + y_{22})^2/7.$$

It is clear from the definition and from the example that the property of zero interaction is unaffected by the choice of inner product. Zero interaction does not imply a orthogonality with $\mathcal{W}$.

2.3 *Residuals in linear models*

A *linear* model is a statement of the form $\mu \in \mathcal{X}$, where $\mathcal{X}$ is a subspace of $\mathcal{V} = \mathcal{R}^\Omega$, and Ω is the set of statistical units. The residual is any departure of the observed y from this subspace having the following properties.

(i) y has zero residual if and only if $y \in \mathcal{X}$.
(ii) Residuals are additive.

In the absence of a preferred complementary space, these conditions are equivalent to defining the residual space as the quotient space $\mathcal{V}/\mathcal{X}$. In other words, the residual associated with y is the coset $y + \mathcal{X}$.

This definition has a number of advantages, as well as disadvantages, over the conventional definition, which is the orthogonal projection of y on to $\mathcal{X}^\perp$. The main advantage is most evident in likelihood calculations concerning variance components, where a range of inner products is under consideration. It is frequently convenient in probability calculations to take $\mathcal{V}$ to be an inner product space in which the components of the inner product matrix are the components of the *inverse* covariance matrix Σ^{-1}. In probability calculations, the standard definition of residual, $R = Qy$, as the orthogonal projection to $\mathcal{X}^\perp$ is perfectly satisfactory. In likelihood calculations, however, this definition poses serious conceptual difficulties. Since $Q = I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$ depends on Σ, the residual also depends on Σ, and may not be observable. What is even more perplexing, the space $\mathcal{X}^\perp$ itself depends on Σ, and thus on the variance components. The usual definition of likelihood and likelihood ratio (as a Radon-Nikodym derivative) is therefore inapplicable.

By contrast, the quotient-space definition does not require $\mathcal{V}$ to be an inner product space, and the observed residual remains fixed at $y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$ as the range of inner products is considered. Derivation of the residual likelihood is thus conceptually straightforward (McCullagh, 1996), leading quite directly to the REML likelihood (section 4). As a consequence, if Y is normally distributed with mean $\mu \in \mathcal{X}$ and covarince matrix Σ, the marginal log likelihood for Σ based on the residual $y + \mathcal{X}$ is

$$-\tfrac{1}{2}\|Qy\|^2 - \tfrac{1}{2}\log\det\Sigma - \tfrac{1}{2}\log\det(X^T\Sigma^{-1}X),$$

where Q is the orthogonal projection, with respect to Σ^{-1}, on to $\mathcal{X}^\perp$, and the columns of X form a basis in $\mathcal{X}$.

The main disadvantage of the quotient-space definition is that a coset is not a function on the units. The coset definition is thus not suitable for plotting purposes.

3. Covariance functions and variograms

Let $\mathcal{V}$ be the vector space of real-valued functions $A \mapsto \mathcal{R}$, where A is a given region of the plane, or real 3-space. Let Y be a stochastic process on A, so each realization of Y is an element in $\mathcal{V}$. The mean function $\mu = E(Y)$ is a vector in $\mathcal{V}$, sometimes assumed to lie in 1, the subspace of constant functions. The covariance function

$$\sigma(u, v) = \mathrm{cov}(Y(u), Y(v))$$

is a vector in $\mathcal{V}^{\otimes 2}$, in the symmetric subspace, $\mathrm{sym}^2(\mathcal{V})$. The variogram defined by

$$\gamma(u, v) = \mathrm{var}(Y(u) - Y(v)),$$

is also a function in $\mathrm{sym}^2(\mathcal{V})$. It is clear that any information in the variogram can be obtained from the covariance function by the relation

$$\gamma(u, v) = \sigma(u, u) + \sigma(v, v) - 2\sigma(u, v).$$

But the converse is not true because, for any real-valued random variable Z, constant on A, the variogram of the process $Y + Z$ is identical with the variogram of Y. Further, if Z does not have finite second moments, the covariance function of $Y + Z$ does not exist. Nevertheless, the relation between the two methods of specifying second-moment properties is quite close.

The statement that the variogram of the process $Y + Z$ is the same as that of Y for any $Z \in 1$ is equivalent to saying that the variogram is a function in the quotient space $(\mathcal{V}/\mathbf{1})^{\otimes 2}$. Now, by isomorphism,

$$(\mathcal{V}/\mathbf{1})^{\otimes 2} \cong \mathcal{V}^{\otimes 2}/(\mathcal{V} \otimes 1 + 1 \otimes \mathcal{V})$$
$$\mathrm{sym}^2(\mathcal{V}/\mathbf{1}) \cong \mathrm{sym}^2(\mathcal{V})/\,\mathrm{sym}^2(\mathcal{V} \otimes 1 + 1 \otimes \mathcal{V})$$

The space $\mathrm{sym}^2(\mathcal{V} \otimes 1 + 1 \otimes \mathcal{V})$ is the space of symmetric additive functions $\alpha(u) + \alpha(v)$, isomorphic with $\mathcal{V}$, on $A \times A$. To say the same thing in another way, if the covariance function σ is regarded as an element in this symmetric quotient space, the set of covariance functions

$$\sigma'(u, v) = \{\sigma(u, v) + \alpha(u) + \alpha(v) : \alpha \in \mathcal{V}\}$$

constitute the coset of equivalent covariance functions. The elements in this set are equivalent as a specification of the variance of any contrast $\theta \in 1^0$:

$$\sum_{u,v} \sigma'(u, v)\theta(u)\theta(v) = \sum_{u,v} \sigma(u, v)\theta(u)\theta(v)$$

for all linear functionals θ such that $\theta(1) = \sum \theta(u) = 0$. Provided that the covariance function is regarded as a partial specification modulo this equivalence relation, the covariance and the variogram are equivalent, incomplete, specifications of the second moments of the process.

It is necessary, of course, that a covariance function should be positive definite, or at least semi-definite, on the dual space of linear functionals on $\mathcal{V}$. If σ is positive definite on this space, the coset σ' is positive definite on the dual of $\mathcal{V}/\mathbf{1}$, which is the space of contrasts 1^0.

More generally, let μ be a point in $\mathcal{X}$, a subspace of $\mathcal{V}$, the residual $y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$, and let $\mathcal{X}^0$ be that subspace of linear functionals on $\mathcal{V}$ taking the value zero on $\mathcal{X}$. Define the subspace $\mathrm{sym}(\mathcal{V} \otimes \mathcal{X})$ by

$$\mathrm{sym}(\mathcal{V} \otimes \mathcal{X}) = \{(u, v) \mapsto \alpha(u)x(v) + \alpha(v)x(u) : x \in \mathcal{X}, \alpha \in \mathcal{V}\}.$$

For any covariance function σ in $\mathrm{sym}^2(\mathcal{V})$, the coset $\sigma' = \sigma + \mathrm{sym}(\mathcal{V} \otimes \mathcal{X})$ is the set of functions

$$\sigma' = \{(u, v) \mapsto \sigma(u, v) + \alpha(u)x(v) + \alpha(v)x(u) : x \in \mathcal{X}, \alpha \in \mathcal{V}\}.$$

This coset is a point in the space $\mathrm{sym}^2(\mathcal{V})/\mathrm{sym}(\mathcal{V} \otimes \mathcal{X})$, which is isomorphic with $\mathrm{sym}^2(\mathcal{V}/\mathcal{X})$. If $\sigma \in \mathrm{sym}^2(\mathcal{V})$ is the covariance function of Y, then $\sigma' \in \mathrm{sym}^2(\mathcal{V}/\mathcal{X})$ is the covariance function of the residual $Y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$.

So far as the residual space $\mathcal{V}/\mathcal{X}$ is concerned, the relevant set of linear functionals is the dual space of $\mathcal{V}/\mathcal{X}$, which is $\mathcal{X}^0$. It is easy to check that for any linear functional $\theta \in \mathcal{X}^0$, all points in the coset σ' give the same variance:

$$\sum \sigma'(u, v)\theta(u)\theta(v) = \sum \sigma(u, v)\theta(u)\theta(v).$$

In other words, a covariance function, regarded as a function in the quotient space $\mathrm{sym}^2(\mathcal{V})/\mathrm{sym}(\mathcal{V} \otimes \mathcal{X})$, determines the covariance of all $\mathcal{X}$-contrasts, linear functionals taking the value zero on $\mathcal{X}$. Note that if $\mathcal{V}$ has dimension n and $\mathcal{X}$ has dimension p, then $\mathrm{sym}^2(\mathcal{V})$ has dimension $n(n+1)/2$, $\mathcal{V} \otimes \mathcal{X}$ has dimension np, and $\mathrm{sym}(\mathcal{V} \otimes \mathcal{X})$ has dimension $p(n-p)+p(p+1)/2$, where $p = \dim(\mathcal{X})$. So the quotient space has dimension $(n-p)(n-p+1)/2$.

4. Miscellaneous points

4.1 *Inner product on a quotient space*

Let $\langle \cdot, \cdot \rangle$ be an inner product in $\mathcal{V}$. This automatically determines an inner product on each subspace of $\mathcal{V}$. It is slightly less obvious, however, what we mean by a compatible inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}/\mathcal{V}_0}$ in $\mathcal{V}/\mathcal{V}_0$. In principle, the inner product must be defined on cosets of $\mathcal{V}_0$, but it is often simpler to define it on $\mathcal{V}$ with the condition that it be constant on each coset of $\mathcal{V}_0$. In addition to symmetry and bi-linearity, the following conditions must be satisfied.

$$\begin{aligned} \langle u_1 + v,\, u_2 \rangle_{\mathcal{V}/\mathcal{V}_0} &= \langle u_1,\, u_2 \rangle_{\mathcal{V}/\mathcal{V}_0} \quad \text{for} \quad v \in \mathcal{V}_0 \\ \langle u,\, u \rangle_{\mathcal{V}/\mathcal{V}_0} &\geq 0 \\ \langle u,\, u \rangle_{\mathcal{V}/\mathcal{V}_0} = 0 &\iff u \in \mathcal{V}_0. \end{aligned}$$

These conditions are satisfied by a large number of bi-linear functions unrelated to the inner product in $\mathcal{V}$. The most natural choice to ensure compatibility is to make $\langle \cdot, \cdot \rangle_{\mathcal{V}/\mathcal{V}_0}$ coincide with $\langle \cdot, \cdot \rangle$ on the subspace $\mathcal{V}_0^\perp$. In other words, $\langle u,\, v \rangle_{\mathcal{V}/\mathcal{V}_0} = \langle Qu,\, Qv \rangle$, where Q is the orthogonal projection on to $\mathcal{V}_0^\perp$. To say the same thing in another way,

$$\langle u,\, v \rangle_{\mathcal{V}/\mathcal{V}_0} = \langle u,\, v \rangle - \langle Pu,\, Pv \rangle,$$

where P is the orthogonal projection on to $\mathcal{V}_0$ (Tjur 1974, section 11). In particular, the squared quotient-space norm is

$$\|v\|^2_{\mathcal{V}/\mathcal{V}_0} = \|v\|^2 - \|Pv\|^2.$$

4.2 *Sampling and prediction*

Let Ω be the population of N units, Ω_0 the sample of n observed units, and Ω_1 the unsampled units. A statistical variate y is thus a vector in $\mathcal{V} = \mathcal{R}^{\Omega}$; the observed variate y_0, the restriction of y to the subset Ω_0, is a vector in $\mathcal{R}^{\Omega_0}$. What is the relationship between the vector spaces $\mathcal{R}^{\Omega}$ and $\mathcal{R}^{\Omega_0}$?

Let $\mathcal{V}_0$ be the subspace of $\mathcal{V}$ consisting of all functions that are zero on Ω_0, and let $\mathcal{V}_1$ be the complementary subspace of functions that are zero on Ω_1. Thus $\mathcal{V}_1$ has dimension n and $\mathcal{V}_0$ has dimension $N - n$. Two vectors y, y' in $\mathcal{V}$ such that $y - y' \in \mathcal{V}_0$ have the property that their restrictions to Ω_0 are equal. Thus, all vectors in the coset $y + \mathcal{V}_0$ give rise to the same observation in $\mathcal{R}^{\Omega_0}$. In other words, each point in $\mathcal{R}^{\Omega_0}$ identifies a coset of $\mathcal{V}_0$ in $\mathcal{V}$. The vector space $\mathcal{R}^{\Omega_0}$ is thus naturally isomorphic with the quotient $\mathcal{V}/\mathcal{V}_0$.

It should be emphasized that, although the dimensions are equal, the quotient space $\mathcal{V}/\mathcal{V}_0$ is quite different from the subspace $\mathcal{V}_1$. In particular, if $\mathcal{V}$ is an inner product space, $\mathcal{V}_1$ and the quotient $\mathcal{V}/\mathcal{V}_0$ typically have incompatible inner products.

Let $y_0 \in \mathcal{R}^{\Omega_0}$ denote the observed value. This can, be represented by the partitioned vector $(y_0, \star)$, where $\star$ denotes missing or arbitrary values on Ω_1. Suppose that the distribution of the random vector y on $\mathcal{V}$ has zero mean and positive-definite covariance matrix partitioned according to (Ω_0, Ω_1) as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix}.$$

Then the best linear predictor for y given $(y_0, \star)$ is

$$E\big(y \,|\, (y_0, \star)\big) = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix} \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_0 \\ \star \end{pmatrix} = \begin{pmatrix} y_0 \\ \Sigma_{10}\Sigma_{00}^{-1} y_0 \end{pmatrix}.$$

It is appropriate here to regard $\mathcal{V}$ as an inner product space with inner product matrix Σ^{-1}. The quotient-space inner product is $\mathrm{diag}\{\Sigma_{00}^{-1}, 0\}$. The transformation given above is then a self-conjugate linear mapping on $\mathcal{V}$ with null space $\mathcal{V}_0$ and range $\mathcal{V}_0^{\perp}$.

4.3 *Book orthogonality and Tjur systems*

Two subspaces $\mathcal{X}, \mathcal{Z}$ of the inner product space $\mathcal{V}$ are said to be orthogonal if $\langle x, z\rangle = 0$ for every $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. The subspaces that occur in linear models are usually overlapping because they ordinarily contain at least the subspace of constant functions. Consequently they cannot be orthogonal in the usual sense. Geometric orthogonality (Tjur, 1984), also called book orthogonality in unpublished lecture notes by M. Wichura, is the condition that the subspaces $\mathcal{X}$ and $\mathcal{Z}$ be orthogonal modulo their intersection, i.e. $\langle \mathcal{X}, \mathcal{Z}\rangle_{\mathcal{V}/\mathcal{X}\cap\mathcal{Z}} = 0$. This is a useful concept thought slightly counter-intuitive in some respects. For example $\mathcal{X}$ is book-orthogonal to itself and to all subspaces.

A collection $\mathcal{L}$ of subspaces of $\mathcal{V}$ is a Tjur system if three conditions are satisfied:

(i) $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{L}$ implies $\mathcal{X}_1$ and $\mathcal{X}_2$ are book orthogonal.
(ii) $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{L}$ implies $\mathcal{X}_1 \cap \mathcal{X}_2 \in \mathcal{L}$.
(iii) $\mathcal{V}$ is in $\mathcal{L}$.

The importance of a Tjur system is that it gives rise to a unique analysis-of-variance decomposition as follows. To each $\mathcal{X} \in \mathcal{L}$ we associate the vector space

$$\mathcal{X}' = \sum_{\mathcal{Z} \subset \mathcal{X}} \mathcal{Z},$$

the span of all proper subspaces of $\mathcal{X}$ in $\mathcal{L}$. A Tjur system is not ordinarily closed under vector spans, so $\mathcal{X}'$ need not be in $\mathcal{L}$. To each $\mathcal{X} \in \mathcal{L}$ we associate the quotient space $\mathcal{X}/\mathcal{X}'$ and the projection

$$\|P_{\mathcal{X}/\mathcal{X}'} y\|^2 = \|P_{\mathcal{X}} y\|^2 - \|P_{\mathcal{X}'} y\|^2.$$

Then the total sum of squares is the sum

$$\|y\|^2 = \sum_{\mathcal{X} \in \mathcal{L}} \|P_{\mathcal{X}/\mathcal{X}'} y\|^2$$

of independent components.

When a Tjur system is extended to a distributive lattice by the inclusion of vector spans, each complete lattice chain

$$\mathcal{V} \equiv \mathcal{X}_n \supset \mathcal{X}_{n-1} \supset \cdots \supset \mathcal{X}_0 \equiv 0$$

gives rise to a decomposition indexed by the quotients $\mathcal{X}_i/\mathcal{X}_{i-1}$. Book orthogonality ensures that all such decompositions are numerically equivalent, differing only in the order of terms.

4.4 *Lebesgue measure*

A basis $\{e_1, \ldots, e_n\}$ in $\mathcal{V}$ defines an association, a linear transformation from the point $(x^1, \ldots, x^n)$ in $\mathcal{R}^n$ to the point

$$v = x^1 e_1 + \cdots + x^n e_n \tag{3}$$

in $\mathcal{V}$. This transformation takes the unit cube $[0,1]^n$ in $\mathcal{R}^n$ into the set

$$(0, e_1) \times \cdots \times (0, e_n)$$

in $\mathcal{V}$. The measure, or n-dimensional volume of this set is $\det^{1/2}\{e_1, \ldots, e_n\} = |G|^{1/2}$, where $g_{rs} = \langle e_r, e_s \rangle$ is the matrix of inner products of the basis vectors. Thus, we write

$$dv = \det^{1/2}\{e_1, \ldots, e_n\}\, dx = |G|^{1/2}\, dx$$

giving the Jacobian of the transformation (3).

Let $\mathcal{V}_0$ be the subspace spanned by $\{e_1, \ldots, e_p\}$. Then $\{e_{p+1}, \ldots, e_n\}$ is a basis in some complementary space, and $\{e_{p+1} + \mathcal{V}_0, \ldots, e_n + \mathcal{V}_0\}$ is a basis in $\mathcal{V}/\mathcal{V}_0$. Using properties of determinants, we find

$$\begin{aligned} dv = |G|^{1/2}\, dx &= \det^{1/2}\{e_1, \ldots, e_p, e_{p+1}, \ldots, e_n\}\, dx^1 \cdots dx^n \\ &= \det^{1/2}\{e_1, \ldots, e_p, Qe_{p+1}, \ldots, Qe_n\}\, dx^1 \cdots dx^n \\ &= \det^{1/2}\{e_1, \ldots, e_p\}\, dx^1 \cdots dx^p \times \det^{1/2}\{Qe_{p+1}, \ldots, Qe_n\}\, dx^{p+1} \cdots dx^n \\ &= dv_0 \times \det^{1/2}\{e_{p+1} + \mathcal{V}_0, \ldots, e_n + \mathcal{V}_0\}\, dx^{p+1} \cdots dx^n \end{aligned}$$

where Q is the orthognal projection on to $\mathcal{V}_0^\perp$. The Jacobian of the transformation from R^{n-p} into $\mathcal{V}/\mathcal{V}_0$ associated with the transformation

$$v + \mathcal{V}_0 = x^{p+1}e_{p+1} + \cdots + x^n e_n + \mathcal{V}_0$$

is thus

$$\det{}^{1/2}\{e_{p+1} + \mathcal{V}_0, \ldots, e_n + \mathcal{V}_0\} = |G|^{1/2}/\det{}^{1/2}\{e_1, \ldots, e_p\}.$$

In matrix notation, if $\{e_1, \ldots, e_p\}$ are the columns of X, $dv_0 = |X^T G X|^{1/2}\, dx^1, \ldots, dx^p$, so $|X^T G X|^{1/2}$ is the Jacobian of the transformation from $\mathcal{R}^p$ to $\mathcal{V}_0$. The Jacobian of the transformation from R^{n-p} into $\mathcal{V}/\mathcal{V}_0$ is $|G|^{1/2}/|X^T G X|^{1/2}$.

4.5 *Normal density*

In order to construct a density in $\mathcal{V}$, it is necessary first to have a measure in $\mathcal{V}$. This is taken to be the Lebesgue measure associated with the given basis. The standard normal density at v in $\mathcal{V}$ is

$$(2\pi)^{-n/2}\, \exp(-\|v\|^2/2)\, dv = (2\pi)^{-n/2}\, \exp(-\|v\|^2/2)\, |G|^{1/2}\, dx$$

Associated with any subspace $\mathcal{V}_0$ there is a factorization into two parts. First, in the exponent, we have

$$\|v\|^2 = \|Pv\|^2 + \|Qv\|^2 = \|Pv\|^2 + \|v + \mathcal{V}_0\|^2_{\mathcal{V}/\mathcal{V}_0}.$$

Second, Lebesgue measure factors as shown above. Omitting the powers of 2π, the joint density becomes

$$\exp(-\|Pv\|^2/2)|\bar{G}|^{1/2}\, dx^1 \cdots dx^p \times \exp(-\|Qv\|^2/2)|G|^{1/2}|\bar{G}|^{-1/2}\, dx^{p+1} \cdots dx^n,$$

where $\bar{G} = \det\{e_1, \ldots, e_p\}$. The first factor gives the density of PY at v in $\mathcal{V}_0$, which is standard normal on that subspace. The second factor gives the density at $v + \mathcal{V}_0$ of $Y + \mathcal{V}_0$, also standard normal, but in the quotient space. The second factor is also known as the residual likelihood: For an alternative derivation, see Patterson and Thompson (1971), Harville (1974, 1977) or Searle, Casella and McCulloch (1992, section 6.6).

5. Test statistics

The following is a list of steps frequently used for testing a composite null hypothesis against a specific class of alternatives. The recipe is not intended to be completely general: only hypotheses concerning the mean response, $\mu = E(Y)$ in $\mathcal{V}$ are considered.

1. The null hypothesis is formulated as a model $\mu \in \mathcal{M}_0$, where $\mathcal{M}_0$ is a smooth manifold, often a vector subspace or a non-linear transformation of a vector subspace, in $\mathcal{V}$.
2. An alternative model $\mu \in \mathcal{M}_1$ is formulated in which $\mathcal{M}_1$ is a manifold containing $\mathcal{M}_0$.
3. The fitted value $\hat{\mu}_0 \in \mathcal{M}_0$ is computed.
4. For any point $\mu \in \mathcal{M}_0$, the tangent space of $\mathcal{M}_1$ at μ contains the tangent space of $\mathcal{M}_0$ at μ. These spaces $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}$ are computed at $\hat{\mu}_0$.

5. The total sum of squares is decomposed into three components

$$\begin{aligned}\|y\|^2 &= \|P_0 y\|^2 + (\|P_1 y\|^2 - \|P_0 y\|^2) + (\|y\|^2 - \|P_1 y\|^2) \\ &= \|P_0 y\|^2_{\mathcal{V}_0} + \|P_1 y\|^2_{\mathcal{V}_1/\mathcal{V}_0} + \|y\|^2_{\mathcal{V}/\mathcal{V}_1}\end{aligned}$$

   where P_0 and P_1 are the orthogonal projections on to $\mathcal{V}_0$ and $\mathcal{V}_1$ respectively. In the second expression the norms are on the spaces $\mathcal{V}_0$, $\mathcal{V}_1/\mathcal{V}_0$ and $\mathcal{V}/\mathcal{V}_1$ respectively.

5. If the residual variance is known, the Rao score statistic is $\|P_{\mathcal{V}_1/\mathcal{V}_0} y\|^2$, which has approximately the χ^2 distribution with degrees of freedom equal to $\dim(\mathcal{V}_1/\mathcal{V}_0)$. Otherwise, the residual variance is estimated using the residual mean square, and the ratio

$$F = \frac{\|P_1 y\|^2_{\mathcal{V}_1/\mathcal{V}_0} \big/ \dim(\mathcal{V}_1/\mathcal{V}_0)}{\|y\|^2_{\mathcal{V}/\mathcal{V}_1} \big/ \dim(\mathcal{V}/\mathcal{V}_1)}$$

   is used instead.

For a definition of the tangent space, see Kass and Voss (1997, p. 312).

It is true that mutually orthogonal subspaces can be constructed such that the decomosition given above is expressible as the sum of projections on to subspaces. The three subspaces are

$$\mathcal{V}_0, \quad \mathcal{V}_1 \cap \mathcal{V}_0^\perp, \quad \text{and} \quad \mathcal{V}_1^\perp.$$

The decomposition by orthogonal subspaces is, at best, a source for confusion. In most applications, there is no reason to single out one complement of $\mathcal{V}_0$ in $\mathcal{V}_1$ for special consideration. Only the quotient spaces matter.

6. Affine spaces

6.1 *Tangent spaces*

In the study of dependence, it is conventional to regard the response y as a point in the vector space $\mathcal{V} = \mathcal{R}^\Omega$, and in much of the preceding discussion we have done so without comment. In many cases this choice is inappropriate, but it is inappropriate in benign ways that have little effect and can be overlooked. However, there are instances in which a more appropriate formulation is called for. The most common problem is that most physical measurements have no preferred origin, whereas every vector space has a definite zero point. Temperature, for example, can be recorded on various scales for which the origins do not coincide. A better model for statistical purposes is often an affine space, which is not closed under addition, and has no origin. The standard geometrical construction of an affine space is a translation of a subspace, i.e. a coset of the subspace (Birkhoff and MacLane, 1948, section 13).

An affine space $\mathcal{A}$ has two defining properties that are relevant for statistical purposes. First, $\mathcal{A}$ is closed under averages. For any points $u_1, \ldots, u_k$ in $\mathcal{A}$, and for any real numbers $a_1, \ldots, a_k$ such that $\sum a_j = 1$, the combination $\sum a_j u_j$ is a point in $\mathcal{A}$. Second, $\mathcal{A}$ has a tangent space, $\mathcal{T}(\mathcal{A})$,

$$\mathcal{T}(\mathcal{A}) = \{u - v : u, v \in \mathcal{A}\}$$

consisting of differences of the elements in $\mathcal{A}$. The tangent space is a vector space. If $\mathcal{A}$ is a coset of a subspace, $\mathcal{V}_0$, then $\mathcal{T}(\mathcal{A}) = \mathcal{V}_0$ is the subspace. Real physical space in the Newtonian sense is the most transparent example of an affine space: it makes no sense to add positions. Displacement, relative position, velocity, relative velocity, force, momentum and acceleration are vectors in the tangent space.

In the case of statistical models, if the response Y lies in the affine space $\mathcal{A}$, then $\mu = E(Y)$ also lies in $\mathcal{A}$ and the difference $Y - \mu$ lies in the tangent space, $\mathcal{T}(\mathcal{A})$. A model for μ specifies that $\mu \in \mathcal{M}$, where $\mathcal{M}$ is a subset of $\mathcal{A}$. In the case of linear models, $\mathcal{M}$ is an affine subspace. For generalized linear models a specified transformation $g(\mu)$ lies in an affine set. More generally, in regular problems, $\mathcal{M}$ is a smooth manifold in $\mathcal{A}$. In all cases, the tangent space of $\mathcal{M}$ at μ_0 is a subspace of $\mathcal{T}(\mathcal{A})$.

6.2 *Orthogonality and least-squares projection*

Suppose that $\mathcal{T}(\mathcal{A})$ is an inner product space. The least squares projection of y on to $\mathcal{M}$ is a point $\hat{\mu} \in \mathcal{M}$ such that

$$\|y - \hat{\mu}\|^2 \leq \|y - \mu\|^2$$

for all $\mu \in \mathcal{M}$. The definition involves only points in the tangent space. It follows that the residual vector $y - \hat{\mu}$ is orthogonal to the tangent space of $\mathcal{M}$ at $\hat{\mu}$:

$$\langle \hat{d}_r,\, y - \hat{\mu} \rangle = 0,$$

where $\{\hat{d}_r\}$ are vectors spanning the tangent space at $\hat{\mu}$. Reverting to matrix notation, let $\mu = \mu(\beta)$ be a parameterization of $\mathcal{M}$, and let $d_r = \partial\mu/\partial\beta_r$ be the derivative vectors, which span the tangent space of $\mathcal{M}$ at μ. The orthogonality condition then becomes

$$\hat{D}^T W (y - \hat{\mu}) = 0, \tag{4}$$

in which W is the matrix of the inner product, and $\{d_r\}$ are the columns of D.

The preceding derivation, which is entirely geometrical, assumes implicitly that W is a known matrix. In practice, however, statistical criteria determine the choice of inner product matrix, usually the inverse covariance matrix of Y. Unfortunately, Σ often depends on μ, in which case we do not have an inner product space in the conventional sense. Rather, each tangent space is a different inner product space, and the orthogonality condition becomes $\langle \hat{d}_r,\, y - \hat{\mu} \rangle_{\hat{\mu}} = 0$, orthogonality in the tangent space at $\hat{\mu}$ with respect to the inner product on that space. In matrix notation, we have

$$\hat{D}^T \hat{W} (y - \hat{\mu}) = 0, \tag{5}$$

in which $\hat{W} = \Sigma^{-1}(\hat{\mu})$ is the matrix of the inner product in the tangent space at $\hat{\mu}$.

Analogy in mathematics can make the progression from the least squares equation (4) to the quasi-likelihood equation (5) seems obvious and trivial. This deception is justified to the extent that the answer is correct. It should be pointed out, however, that (5) does not ordinarily satisfy the 'least squares', or minimum chi-squared criterion

$$\|y - \tilde{\mu}\|^2_{\tilde{\mu}} \leq \|y - \mu\|^2_{\mu}$$

for all $\mu \in \mathcal{M}$. Under fairly general conditions that are not easy to codify satisfactorily in the form of a theorem, $\hat{\mu}$ is consistent but $\tilde{\mu}$ is not (McCullagh, 1984; Heyde, 1997).

7. Incomplete observation

Suppose that the complete response vector y lies in the vector space $\mathcal{V} = \mathcal{R}^\Omega$, and that the model for the mean of Y is $\mu \in \mathcal{M}$, where $\mathcal{M}$ is a smooth manifold in $\mathcal{V}$. The incomplete response is such that the value of y is not observed for every individual unit: only totals for certain combinations of units are observed. One way to express this relationship is to say that $y' = Ly$ is a point in a new vector space $\mathcal{V}' = L\mathcal{V}$ obtained by some linear transformation L. The model for the transformed mean $\mu' = L\mu$ is $\mu' \in \mathcal{M}' = L\mathcal{M}$, a manifold in the transformed space. While this formulation is conceptually straightforward, it does have the disadvantage that the new vector space is not directly connected with the statistical units in Ω.

Another equivalent way to express the relation in vector-space terms is to say that y' is a coset of $\mathcal{V}_0$ in $\mathcal{V}$. The subspace $\mathcal{V}_0$ is the null space of L, the set of vectors satisfying $Lv = 0$. The observation is thus regarded as a point in $\mathcal{V}' = \mathcal{V}/\mathcal{V}_0$. The least-squares orthogonality condition becomes

$$\langle \hat{d}_r,\, y' - \hat{\mu} \rangle_{\mathcal{V}/\mathcal{V}_0} = 0, \tag{6}$$

where d_r are vectors spanning the tangent space of $\mathcal{M}$ at μ. Ordinarily, unless $\mathcal{M}$ is an affine set, this equation must be solved numerically to obtain $\hat{\mu}$.

Details of the usual algorithm for finding $\hat{\mu}$ are most easily expressed in matrix notation in which W is the matrix of the inner product, Q is the orthogonal projection on to $\mathcal{V}_0^\perp$, and the columns of D are the vectors d_r. Beginning at a point $\mu_0 = \mu(\beta_0)$, a variation of Newton's method gives the iterative scheme

$$D^T W Q D(\hat{\beta} - \beta_0) = D^T W Q (y' - \mu_0)$$

in which $\mu(\beta)$ is a parameterization of $\mathcal{M}$, and $D = \partial\mu/\partial\beta$ at μ_0. At convergence, $\beta_0 = \hat{\beta}$, and equation (6) is satisfied. This algorithm usually converges at near-quadratic rate.

In the preceding expressions, y' is a point in the quotient space, $\mathcal{V}$ modulo $\mathcal{V}_0$. If it is more convenient to work directly in $\mathcal{V}$, however, any point in the coset $y' + \mathcal{V}_0$ can be used in place of y'. In this connection, it is sometimes convenient to split the equation into two parts, an 'E' step and an 'M' step. Given parameter values β_0 and fitted values $\mu_0 = \mu(\beta_0)$, the 'E' step generates a complete-data vector $\tilde{y}$, a point in $\mathcal{V}$, from the incomplete data vector. The 'M' step asserts that $\tilde{y} - \hat{\mu}$ is orthogonal to the tangent space of $\mathcal{M}$ at $\hat{\mu}$, as if the data were complete.

$$\begin{aligned} \tilde{y} - \mu_0 &= Q(y' - \mu_0) \\ \langle \hat{d}_r,\, \tilde{y} - \hat{\mu} \rangle &= \hat{D}^T W (\tilde{y} - \hat{\mu}) = 0. \end{aligned} \tag{7}$$

In the first part, Q is a linear transformation from $\mathcal{V}/\mathcal{V}_0$ into the subspace $\mathcal{V}_0^\perp$. In the second part, the orthogonal projection of $\tilde{y} - \mu_0$ on to the tangent space at μ_0 is $D(D^T W D)^- D^T W (\tilde{y} - \mu_0)$. One step of Newton's method gives

$$D(\hat{\beta} - \beta_0) = D(D^T W D)^- D^T W (\tilde{y} - \mu_0) = P(\tilde{y} - \mu_0).$$

Iteration between the two parts is required to obtain the solution.

The preceding discussion assumes that the space $\mathcal{V}$ is Euclidean, i.e. that the matrix of the inner product is fixed and known, at least up to a multiplicative constant. In statistical problems, however, $W \propto \Sigma^{-1}$, the inverse covariance matrix of the response vector, which often depends

on μ, and perhaps on additional parameters. We consider here the simplest case in which, apart from a possibly unknown scalar multiple, Σ depends on μ only. To express this in purely algebraic terms, each point $\mu \in \mathcal{V}$ has its own tangent space with its own distinctive inner product. Each of these tangent spaces is isomorphic with $\mathcal{V}$, contains $\mathcal{V}_0$, and thus the quotient space $\mathcal{V}/\mathcal{V}_0$. Each subspace and quotient space inherits its inner product from the tangent space.

With this in mind, equation (6) and the associated iterative scheme are immediately generalizable. The inner product in each case is an inner product in the tangent space at the current value of μ. The matrices W and Q are computed at the current value μ_0, which is then updated as new estimates become available. Similar comments apply to the two-step E-M scheme in (7).

In order that (7) coincide with the E-M algorithm for maximum likelihood calculation, it is necessary that the first step coincide with conditional expectation, and the second step with maximization of the complete-data likelihood. In the context of exponential-family models or exponential dispersion models, the 'M' step is maximization. If the complete data are jointly normally distributed, the 'E' step is exactly the conditional expected value given the observed data. Apart from a few other special cases, the first step is not conditional expectation, but only a linear approximation to conditional expectation. Nevertheless, (7) gives consistent estimates under fairly general conditions. Identifiability requires $\mathcal{V}_0$ and the tangent space at μ to be non-overlapping vector subspaces. The approximate covariance matrix of $\hat{\beta}$ is then proportional to the inverse of $\hat{D}^T\hat{W}\hat{Q}\hat{D}$.

REFERENCES

Anderson, J.A. (1984) Regression and ordered categorical variables (with discussion). *J. R. Statist. Soc.* B

Birkhoff, G. and MacLane, S. (1948) *A Survey of Modern Algebra.* MacMillan, New York.

Dempster, A.P., Laird, N. and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Harville, D.A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–5.

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems (with discussion). *J. Am. Statist. Assoc.* **72**, 320–40.

Heyde, C.C. (1997) *Quasi-likelihood and its Application.* Springer series in Statistics.

Heyde, C.C. and Morton, R. (1996) Quasi-likelihood and the EM algorithm. *J. R. Statist. Soc.* B **58**, 317–27.

Kass, R.E. and Voss, P. (1997) *Geometrical Foundations of Asymptotic Inference.* New york, Wiley.

McCullagh, P. (1980) Regression models for ordinal data. *J. R. Statist. Soc.* B **42**, 109–42.

McCullagh, P. (1984) Generalized linear models. *European J. Operational Research* **16**, 285–92.

McCullagh, P. (1997) Linear models, vector spaces and residual likelihood. In *Modelling Longitudinal and Spatially correlated Data.* Springer Lecture Notes in Statistics, **22**, 1–10.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models.* Chapman and Hall, London.

Palmgren, J. (1981) The Fisher information matrix for log-linear models, arguing conditionally on the observed explanatory variables. *Biometrika* **68**, 563–6.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–54.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance Components*. J. Wiley & Sons, New York.

Tjur, T. (1974) *Conditional Probability Distributions*. University of Copenhagen lecture notes in Statistics, vol 2.

Tjur, T. (1984) Analysis of variance models in orthogonal designs (with discussion). *International Statistical Review* **52**, 33–81.

Wichura, M.J. (198?) Unpublished lecture notes on linear models.