Revista Brasileira de Probabilidade e Estatística (1995), **2**, pp.177–196.

STATISTICAL APPLICATIONS OF THE INVERSE GRAM MATRIX: A REVISITATION

Guido E. del Pino and Hector Galaz
Department of Statistics
Pontificia Universidad Católica de Chile
Casilla 306, Santiago 22, Chile

Summary
Gram matrices are implicit in many statistical settings and their inverses admit interesting geometric interpretations. The potential insights to be gained are often lost, because they get clouded by the particularities of each application. The purpose of this paper is twofold. The first is to serve as a reference for the general results, which seem hard to find in abstract form given here. The second is to illustrate their applications in the areas of linear regression, multivariate analysis, prediction theory, and time series. Of particular interest is a general definition of inverse correlation, which is shown to be related to the concept of inverse autocorrelation in time series analysis.

Key words: Gram matrix; inverse correlation; prediction theory; multivariate analysis; linear regression; time series.

1 Introduction

That the covariance of two random variables may be interpreted as the inner product between two vectors is widely known. With this interpretation, the analog of the covariance matrix for an arbitrary inner product space (i.e a vector space where an inner product $\langle\ ,\ \rangle$ has been defined) is the $n \times n$ matrix, whose entries are $A_{ij} = \langle v_i, v_j \rangle$. By definition $A = \text{Gram}(v_1, \ldots, v_n)$ is the *Gram matrix* of these vectors and this paper is concerned with the geometric properties of its inverse and their relevance in statistics. From a purely mathematical point of view this inverse is perhaps most closely associated with dual bases, a subject briefly reviewed in Section 2.3.

The inverses of Gram matrices appear in disguise in different statistical settings, and it turns out that many particular results may be viewed as special cases of a single abstract result that can be stated in terms of an arbitrary Gram matrix. As a consequence all these results admit common geometric interpretations, in terms of simple concepts such as length, angle, orthogonality, and projections. These geometric interpretations are stated here as Theorems 2.1 and 2.2, and their proofs are given in the Appendix. The inclusion of these elementary results is justified by the difficulty of finding them in print at this level of abstraction. Particular versions are readily available in the literature in particular contexts, like regression, where they appear in connection with updating regression models and with the theory behind stepwise techniques. The main purpose of this paper to illustrate how a straightforward application of these general theorems to linear regression (Section 3), covariance structures and linear prediction theory (Section 4), and time series (Section 5) recovers a large number of important known facts and, as a side benefit, provides useful geometric interpretations. The general strategy is to recognize that some matrix B, which is relevant in a particular context, is actually the inverse of a positive definite matrix A. By identifying A with $\mathrm{Gram}(v_1,\ldots,v_n)$, some aspects of B can be geometrically interpreted in terms of the vectors $v_1,\ldots,v_n$.

In regression, $A = X'X$, where X is the design matrix. From Theorem 2.1 the entries of $(X'X)^{-1}$ are related to the coefficients of the best approximation (in the least squares sense) of the column $\boldsymbol{X}_i$ of X by a linear combination of the remaining columns. Theorem 2.2 shows that the entries $(X'X)^{-1}$ are also associated with the (empirical) correlations between suitable residual vectors, provided the columns are centered. On the other hand $(X'X)^{-1}$ is perhaps mainly associated with the covariance matrix of the least squares estimator of the parameter vector. The associated probabilistic aspects, e.g. the variance inflation factor, are numerically determined by the behavior of the fixed matrix X.

The covariance matrix and its inverse appear also in many areas of multivariate analysis, like the Mahalanobis distance in discriminant analysis or the quadratic form associated with the multivariate normal density. Although the natural vector space, which consists of all random variables with mean zero and finite variance, together with the covariance inner product, is infinite dimensional, in the applications mentioned above we deal with only a finite number of random variables, say $Z_1,\ldots,Z_n$, at a time. Therefore we can just as well work instead with the subspace M formed by all linear combinations of these n random variables $Z_1,\ldots,Z_n$, still using the covariance inner product. Then $\mathrm{Gram}(Z_1,\ldots,Z_n)$ is just the covariance matrix of the random vector $\boldsymbol{Z} = (Z_1,\ldots,Z_n)'$ and orthogonal projections become best linear predictors. In this context we will give a nice interpretation of dual bases, and will illustrate its application to parameter estimation using the regression model and the Fisher information matrix. Since inverse covariance matrices appear very often in statistics, the results discussed here potentially provide useful geometric insights in many areas. A systematic account is given by Whittaker (1990), since inverse covariance are basic elements in graphical models. His corollaries 5.8.1 and 5.8.2 can be understood as particular cases of our equations (2.3) and (2.6) respectively, but they are there directly proved by applying a so called inverse variance lemma. Graphical models for Gaussian data are also discussed by Wermuth (1976) and Lauritzen and Wermuth (1989). In these models the existence of zero entries in the inverse covariance matrix implies the conditional independence of the corresponding random variables given the remaining ones. In our approach zero entries in the inverse Gram matrix can be interpreted in terms of orthogonality of some vectors.

Based on the inverse covariance we suggest a definition of *inverse correlation* in Section 4. Theorem 2.2 shows its numerical equality to the negative of the corresponding partial correlation. An interesting application of these ideas, which in fact provided the original motivation for this paper, appears in the context of time series, where time ordering gives content to concepts such as future, past, next, last, in between, interpolation, extrapolation, one step ahead, etc. Theorem 2.1 immediately implies that the elements of the inverse covariance matrix are closely associated with the problem of *interpolation* and the geometric insights obtained lead to a simple heuristic derivation of the interpolation formulae (Section 5) in either the time or the frequency domains. After their introduction by Cleveland (1972) and Chatfield (1977) for model identification, inverse autocorrelations have been systematically studied by Bhansali in a number of papers. This author and many others, including e.g. Brubacker and Wilson (1976) and Maravall and Peña (1988), apply them to the estimation of missing values in time series. Theorems 2.1 and 2.2 provide a heuristic justification for this application and, at the same time, gives us a warning about their limitations when used for finite samples. Bhansali (1990) gives analogs of Theorems 2.1 and 2.2 in the context of infinite time series. The use of inverse autocorrelations for the identification of the order of autoregressive models, is based on the fact that, for an $AR(p)$ model, all inverse autocorrelations beyond the p–th one are zero. By looking at the inverse covariance matrix associated to n consecutive observations, this condition translate into that of zero entries below the p–th subdiagonal, a particular case of the zero entries in the inverse Gram Matrix.

2 Basic results

2.1 Definitions and an abstract approach

Let H be an inner product space. For any finite family $(v_1,\ldots,v_n)$ of H, its Gram matrix $\mathrm{Gram}(v_1,\ldots,v_n)$ is the $n\times n$ matrix with entries $\langle v_i,v_j\rangle$ $i,j=1,\ldots,n$. This matrix essentially contains all geometric information relative to the family $(v_1,\ldots,v_n)$, automatically disregarding the nature of the elements of H (e.g. p–tuples, random variables, deterministic functions, etc.). The possible infinite dimension of H does not create any special problems, since we can always replace it implicitly by the subspace M formed by all linear combinations of the vectors $v_1,\ldots,v_n$. The geometric character of our results mean that we can always think as if H were equal to $\mathbb{R}^p$ and the inner product were euclidean, i.e. $\langle \boldsymbol{v},\boldsymbol{w}\rangle = \boldsymbol{v}'\boldsymbol{w}$. Unless stated otherwise we always assume that this is the innerproduct chosen.

Given a problem posed in some concrete space, say H_0, the general strategy is the following: First check that the problem can be stated solely in terms of a particular inner product on H_0. This inner product is seldom given from the outset and it must be discovered. The second step is to pose and solve the problem in the context of an arbitrary inner product space H and the third is to translate the solution in terms of the original space H_0. For an intuitive understanding of the problem it is usually convenient to pose the problem and interpret the solution in $\mathbb{R}^3$. Sometimes a direct solution is possible in this space and if an analisis of it shows that no particular properties of $\mathbb{R}^3$ are there used, it essentially provides the proof that is valid in the general case.

As an illustration of this general strategy we show how it can be employed to prove the well known result that the zero mean random variables $(Z_1, \ldots, Z_n)$ are linearly independent with probability one, if and only if their covariance matrix is non singular. In the first step we recognize that in terms of the inner product $\langle U, V \rangle =$ Cov (U, V) (see Section 4 for more details) the problem is to show that $(v_1, \ldots, v_m)$ are linearly independent if and only if Gram$(v_1, \ldots, v_m)$ is nonsingular. In the second step we prove this using the equality

$$\boldsymbol{a}'\text{Gram}(v_1, \ldots, v_n)\boldsymbol{a} = \| \sum_{i=1}^{n} a_i v_i \|^2,$$

so that Gram$(v_1, \ldots, v_n)$ is always nonnegative definite and furthermore it is positive definite unless some linear combination $\sum_{i=1}^{n} a_i v_i$ has length 0.

Applying the general result to $\mathbb{R}^n$ implies that the $n \times m$ matrix X is of full rank, if and only if $X'X$ is nonsigular, a useful result for linear models. Further geometric intuition can be obtained from the fact that $\det(X'X)$ is the m-dimensional volume spanned by the m column vectors in $\mathbb{R}^n$, so that volume zero is equivalent to the property that all vectors be contained in a lower dimensional subspace. By the way, a direct proof in $\mathbb{R}^n$ is based on $\boldsymbol{a}'(X'X)\boldsymbol{a} = (X\boldsymbol{a})'(X\boldsymbol{a})$, which is nothing but a restatement of the abstract equality above.

2.2 Main theorems

The orthogonal projection of v_i onto the subspace spanned by $(v_j, j \neq i)$, which we denote by $\hat{v}_i$, is fully determined by Gram$(v_1, \ldots, v_n)$ and so are $\tilde{v}_i = v_i - \hat{v}_i$ and $\| \tilde{v}_i \|^{-2}$. Since this paper discusses the inverse Gram matrix, we assume throughout that $(v_1, \ldots, v_n)$ are linearly independent (and so are a basis for their spanned subspace M). A consequence of this assumption is that the coefficients $b_{ij}, j = 1, \ldots, n$ in

$$\hat{v}_i = \sum_{i \neq j} b_{ij} v_j \tag{2.1}$$

are uniquely determined. Note that the coefficients of $\tilde{v}_i$ in this basis are 1 for v_i and $-b_{ij}$ for v_j, $j \neq i$.

Our first theorem establishes a one to one correspondence between the n numbers $(b_{ij}, j \neq i, \| \tilde{v}_i \|^{-2})$ and the i-th row of the inverse Gram matrix.

Theorem 2.1 *Let* $A = \mathrm{Gram}(v_1, \ldots, v_n)$ *be nonsingular. The entries* $A^{ij}, i,j = 1, \ldots, n$ *of* A^{-1} *satisfy the following conditions:*

$$A^{ii} = \| \tilde{v}_i \|^{-2} \tag{2.2}$$

$$A^{ij} = -b_{ij} \| \tilde{v}_i \|^{-2}, j \neq i, i,j = 1, \ldots, n. \tag{2.3}$$

From (2.2) the diagonal element A^{ii} is the reciprocal of the squared distance between v_i and the subspace spanned by $v_j, j \neq i$. We may then use the number $\frac{1}{\sqrt{A^{ii}}}$ as an indication on how closely may v_i be approximated by a linear combination of the remaining vectors. In the extreme case where the distance is 0, A^{ii} would be infinitely large, but this cannot arise if A is nonsingular.

If we multiply each v_i by a scalar α, $\frac{1}{\sqrt{A^{ii}}}$ will also be multiplied by α, which shows this distance to be scale dependent. A natural scale-free measure is

$$\frac{\| \tilde{v}_i \|^2}{\| v_i \|^2} = (A_{ii} A^{ii})^{-1}.$$

In Theorem 2.1 the inner product was used to define right angles but it also provides a general definition of angle. In the context of this theorem, it seems natural to consider the α_{ij} formed by $(\tilde{v}_i, \tilde{v}_j)$. In the study of partial correlations and many other areas one encounters another kind of angle, which in the abstract framework can be stated as follows. For any $i \neq j$ and any vector v denote by $v^*\{i,j\}$ the part of v that is orthogonal to the subspace spanned by $(v_r, r \neq i,j)$ and by γ_{ij} the angle formed $(v_i^*\{i,j\}, v_j^*\{i,j\})$. In terms of the inner products these angles are best represented by their cosines:

$$c_{ij} = \cos \alpha_{ij} = \frac{\langle \tilde{v}_i, \tilde{v}_j \rangle}{\| \tilde{v}_i \| \| \tilde{v}_j \|}, \qquad d_{ij} = \cos \gamma_{ij} = \frac{\langle v_i^*\{i,j\}, v_j^*\{i,j\} \rangle}{\| v_i^*\{i,j\} \| \| v_j^*\{i,j\} \|}. \tag{2.4}$$

With this notation we are ready to state our second theorem, which interprets the entries of the inverse Gram matrix in terms of these two families of angles. Note that unlike lengths, angles are invariant under scale transformations.

Theorem 2.2 *With c_{ij} and d_{ij} given by (2.4), the entries A^{ij} of the inverse Gram matrix $A = \text{Gram}(v_1, \ldots, v_n)$ satisfy*

$$\frac{A^{ij}}{(A^{ii}A^{jj})^{1/2}} = c_{ij} \;\; i,j = 1,\ldots,n \tag{2.5}$$

$$= -d_{ij},\; i,j = 1,\ldots,n. \tag{2.6}$$

The lefthandside of (2.5) can be identified with a correlation if one identifies the positive definite matrix A^{-1} with a covariance matrix. From (2.2) and (2.3) the cosines c_{ij} and the coefficients b_{ij} are related to each other thhrough

$$c_{ij} = -\frac{\| \tilde{v_j} \|}{\| \tilde{v_i} \|} b_{ij}. \tag{2.7}$$

In particular $c_{ij} = -b_{ij}$ if and only if $\| \tilde{v_j} \| = \| \tilde{v_i} \|$. This situation arises approximately in the interpolation of stationary time series (See Section 5).

2.3 The dual basis

Let $v_1, \ldots, v_n$ be linearly independent vectors in an arbitrary vector space. Then $(v_1, \ldots, v_n)$ is a basis for the subspace M spanned by these vectors. Its dual basis is an array of linear functionals $(L_1, \ldots, L_n)$ defined on this space and satisfying $L_i(v_j) = 1$ or 0, according to $i = j$ or $i \neq j$ respectively. For an inner product space, L_i may be represented by the vector w_i satisfying $L_i v_j = \langle w_i, v_j \rangle$ and $(w_1, \ldots, w_n)$ is also called a dual basis. A direct definition of the dual basis $(w_1, \ldots, w_n)$ of $(v_1, \ldots, v_n)$, is as a solution of $\langle v_i, w_j \rangle = 1$ for $i = j$ and $\langle v_i, w_j \rangle = 0$ for $i \neq j$, where the w_i belongs to M. Dual bases are well known in numerical analysis and optimization, but not so in statistics. Some key properties of the dual basis are

1. For $v \in M, i = 1, \ldots, n$,

$$v = \sum \alpha_i v_i \text{ implies } \langle w_i, v \rangle = \alpha_i. \tag{2.8}$$

2. If the orthogonal projection of an arbitrary v onto M is $\hat{v} = \sum \hat{\alpha}_i v_i$, then

$$\hat{\alpha}_i = \langle w_i, v \rangle. \tag{2.9}$$

3. The Gram matrix of the dual basis coincides with the inverse of the Gram matrix of the original basis, i.e.

$$\text{Gram}(w_1, \ldots, w_n) = \text{Gram}^{-1}(v_1, \ldots, v_n). \tag{2.10}$$

An explicit construction of the dual basis is given by the inverse Gram matrix:

$$\begin{aligned} w_i &= \sum_{j=1}^{n} A^{ij} v_j \qquad (2.11) \\ &= A^{ii} \tilde{v}_i. \qquad (2.12) \end{aligned}$$

Since w_i only differs from $\tilde{v}_i$ by a scalar constant, it can be characterized by its length. Two equivalent conditions are $\| w_i \| = \| \tilde{v}_i \|^{-1}$ and $\| w_i \|^2 = A^{ii}$.

3 Application to regression

Regression analysis and linear models provide many interesting examples of Gram matrices. For any full rank design matrix X one has the positive definite matrices $X'X, C'(X'X)^{-1}C$, and the augmented Gram matrix generated by the columns of X and the data vector $\boldsymbol{y}$, e.g.

$$\begin{bmatrix} X'X & X'\boldsymbol{y} \\ \boldsymbol{y}'X & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix}.$$

This matrix condenses all the second order information needed for computing the estimators, their covariance matrix, and the various sums of squares used to build ANOVA tables. The effect of adding or deleting variables, as in stepwise selection, can be interpreted as operations performed on this matrix. The addition or deletion of observations, as done for computing regression diagnostic, is more related to updates on $X'X$. From a computational viewpoint, Beaton's SWEEP operator is particularly important and the reader is referred to Goodnight (1979) for an excelent tutorial. In what follows, we concentrate only in those aspects more directly connected with the general results of this paper.

3.1 Uncentered regressors

Consider the standard linear regression model, where uncorrelated observations of a response variable Y are made for N combinations of values of k regressors $X_1, \ldots, X_k$. In matrix terms $E(\boldsymbol{Y}) = \boldsymbol{\mu} = X\boldsymbol{\beta}$, Var $(\boldsymbol{Y}) = \sigma^2 I$, where X is a $N \times k$ full rank design matrix. To apply Theorem 2.1 we just interpret $(X'X)^{-1}$ as the inverse of the Gram matrix $X'X$ and make the follwing identifications: v_i is the i–th column $\boldsymbol{X}_i$ of the design matrix $X, H = \mathbb{R}^N$, and M = column space of X. From (2.2) $((X'X)^{ii})^{-1/2}$ is the distance between the point represented by $\boldsymbol{X}_i$ and the subspace spanned by the remaining columns. Furthermore (2.3) shows that the coefficients of the linear combination of the columns $\boldsymbol{X}_j, j \neq i$, which best approximates

$\boldsymbol{X}_i$ in the least squares sense are obtained dividing the i-th row of $(X'X)^{-1}$ by $(X'X)^{ii}$ and changing signs.

Let us now turn our attention to Theorem 2.2 . In the regression context the vectors $\tilde{\boldsymbol{X}}_r$ and $\boldsymbol{X}_r^*$ both contain the residuals from some suitable regressions. To avoid notational complexities we state the results for $i = 1, j = 2$:

(a) $\tilde{\boldsymbol{X}}_1$ contains residuals of X_1 vs. $X_2, X_3, \ldots, X_k$;

(b) $\tilde{\boldsymbol{X}}_2$ contains residuals of X_2 vs. $X_1, X_3, \ldots, X_k$;

(c) $\boldsymbol{X}_1^*\{1,2\}$ contains residuals of X_1 vs. $X_3, \ldots, X_k$;

(d) $\boldsymbol{X}_2^*\{1,2\}$ contains residuals of X_2 vs. $X_3, \ldots, X_k$.

Let us now translate the properties (2.8)–(2.12) of the dual basis into the full rank regression context. The columns of X form a basis for the column space of X. Applying (2.8) to $\boldsymbol{\mu} = X\boldsymbol{\beta}$ shows that the dual basis is given by the ordered columns of $Z = X(X'X)^{-1}$. This is also a direct consequence of (2.11) and may be obtained directly from the identity $I = (X'X)(X'X)^{-1} = X'Z$. Then (2.9) reduces to the standard formula for the least squares estimator of $\boldsymbol{\beta}$ and (2.10) corresponds to $Z'Z = (X'X)^{-1}$. Finally (2.12) shows that the i–th element of the dual basis is obtained dividing the residual vector $\tilde{\boldsymbol{X}}_i$ by its squared length. This links the estimated component $\hat{\beta}_i$ with the residual vector of the i–th column of X vs. the remaining ones, and therefore with two step least squares procedures.

3.2 Centered variables

In many applications the column $\boldsymbol{X}_i$ is derived from some other column $\boldsymbol{T}_j$ through centering, i.e. by subtracting from its element of $\boldsymbol{T}_i$ its average. Geometrically $\boldsymbol{X}_i$ is the orthogonal complement of $\boldsymbol{X}_i$ with respect to the subspace spanned by $(1, 1, \ldots, 1)'$. The value $(A_{ii}A^{ii})$ in (2.5) then coincides with a well known measure of collinearity called the variance inflation factor (see e.g. Theil (1971), Chatterjee and Price (1977)). Furthermore c_{ij} and d_{ij} are empirical correlation coefficients computed between some suitable residual vectors.

3.3 Estimator covariances

If we now concentrate on linear estimation, we recognize $(X'X)^{-1}$ as the matrix appearing in the formula for the covariance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}.$$

This implies $c_{ij} = \mathrm{corr}(\hat{\beta}_i, \hat{\beta}_j)$. Denoting by $\overline{\beta_i}$ the estimator of β_i obtained when only X_i is included in the model (i.e. when all $\beta_j, j \neq i$ are assumed to be zero), the variance inflation factor is just the ratio $\mathrm{Var}(\hat{\beta}_i)/\mathrm{Var}(\overline{\beta_i})$.

3.4 Generalized Least Squares

The only change needed to handle generalized least squares, with weight matrix W is to chose the inner product as $\langle \boldsymbol{a} \ , \ \boldsymbol{b} \rangle = \boldsymbol{a}'W\boldsymbol{b}$, which results in the new Gram matrix $X'WX$.

4 Application to covariance structures

The variance matrix and its inverse appear in many areas of multivariate analysis, e.g. the Mahalanobis distance in discriminant analysis. But perhaps the first encounter with an inverse covariance matrix is in the multivariate normal density. To analyze inverse covariances choose the inner product Cov (U, V) on the infinite dimensional space H of all random variables *with mean zero* and finite variance. The general framework is now: Let $Z_1, \ldots, Z_n$ be zero mean random variables with non singular covariance matrix V and let M be the set of all linear combination of the random variables $Z_1, \ldots, Z_n$. The covariance matrix can be identified with $\mathrm{Gram}(Z_1, \ldots, Z_n)$. Below we show how to get information on the inverse covariance matrix, based on a direct application of Theorems 2.1 and 2.2.

Note first that for the covariance inner product, orthogonal projections may be interpreted as best linear predictors (BLP). From Theorem 2.1, (2.2) provides the coefficients of the BLP of Z_i given $Z_j, j \neq i$ and $\tilde{Z}_i$ is the corresponding prediction error. From Theorem 2.2, the element c_{ij} is the correlation of the prediction errors $\tilde{Z}_i$ and $\tilde{Z}_j$. Although this is not a standard a rather natural name for the c_{ij} is *inverse correlations.* It will be seen later that this terminology is in agreement with that implicit in inverse autocorrelations, a concept arising in time series.

On the other hand, the coefficient d_{ij} coincides with the partial correlation coefficient of Z_i and Z_j (given the remaining variables). Theorem 2.2 shows that the inverse correlation coincides with the negative of the partial correlation.

4.1 Multivariate normal distribution

If every linear combination of the random variables Y_i has a univariate normal distribution, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is said to have a multivariate normal distribution. This distribution is determined by its mean $\boldsymbol{\mu}$ and its covariance matrix V, and is denoted by $N(\boldsymbol{\mu}, V)$. If V is singular, there exists a linear combination of the Y_i that is constant with probability one, and there is no density. Using the Choleski factorization of V it is easily shown that there exists a density f and it is given by $\log f(\boldsymbol{y}) = c - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'V^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \boldsymbol{y} \in \mathbb{R}^n$.

For a given vector $\boldsymbol{\mu}$ the inverse covariance matrix completely determines the probabilistic properties of $\boldsymbol{Y}$. In particular V is diagonal if and only if the Y_i are statistically independent. Since it is V^{-1} rather than V that appears in this density, the conditional distributions are easier to describe in terms of V^{-1}. In particular the conditional variance of Y_i given the values of all other variables is $\frac{1}{V^{ii}}$ and $V^{ij} = 0$ is equivalent to the conditional independence of Y_i and Y_j given the remaining variables.

4.2 Dual bases

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ and let $(T_1, \ldots, T_n)$ be the dual basis of $(Z_1, \ldots, Z_n)$. Writing $\boldsymbol{T} = (T_1, \ldots, T_n)'$. From (2.11) it follows that

$$\boldsymbol{T} = (\text{ Var } \boldsymbol{Z})^{-1}\boldsymbol{Z}.$$

From (2.10) and (2.12) we get

$$T_i = V^{ii}\tilde{Z}_i$$

and $\text{Var }(T_i) = (\text{ Var } \tilde{Z}_i)^{-1} = V^{ii}$.

Two applications to the problem of estimation are:

(a) **Regression Analysis:**
Consider the linear model, $E(\boldsymbol{Y}) = X\boldsymbol{\beta}$, $\text{Var }(\boldsymbol{Y}) = V$, where X is of full rank, and take $\boldsymbol{Z} = X'V^{-1}\boldsymbol{Y}$. Then $\boldsymbol{T} = (X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{Y}$ becomes the best linear unbiased estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Theorem 2.1 shows that $\hat{\beta}_i$ is a multiple of the BLP error of $\boldsymbol{x}_i'V^{-1}\boldsymbol{Y}$ vs. $\boldsymbol{x}_j'V^{-1}\boldsymbol{Y}, j \neq i$.

(b) **Fisher Information Matrix:**
For every $i = 1, \ldots, k$ choose Z_i to be the i-th score, i.e. the partial derivative S_i of the log-likelihood with respect to θ_i. The covariance matrix $V = \mathcal{I}(\boldsymbol{\theta})$ is called Fisher information. In this context $T_i = U_i$ is sometimes called the i–th efficient score.

Theorem 2.1 shows that U_i is a multiple of the BLP error of the S_i vs. $S_j, j \neq i$.

In a large sample situation the score vector $\boldsymbol{S}$ follows approximately a multivariate normal distribution. The representation of U_i in terms of prediction errors may then be replaced by an interpretation in terms of the conditional distribution of S_i given $S_j, j \neq i$.

5 Application to time series

5.1 Finite time series

When the random variable Z_i is interpreted as the value of an observation at "time i", $Z_1, \ldots, Z_n$ is called a (finite) time series. In this context the BLP of Z_i given $Z_j, j \neq i$, becomes the *best linear interpolator* of Z_i and $\tilde{Z}_i$ becomes the corresponding *interpolation error*. The relationship between inverse correlations and interpolation for infinite time series are well known. Here we briefly develop the basic intuitions to show that these relationships are essentially a direct consequence of Theorems 2.1 and 2.2, emphasizing finite time series. In this finite case these theorems link the inverse covariance matrix with the best linear interpolators.

From a matrix viewpoint the covariance matrix of a stationary time series is Toeplitz, i.e. it is constant along subdiagonals. The properties of inverse Gram matrix then tell us something about the inverse of a Toeplitz matrix.

5.2 Infinite time series

Linear prediction theory is easily extended to an infinite number of predictors. All that is needed is a general concept of orthogonal projectors in the Hilbert space setting (see e.g. Brockwell and Davis (1987)). Knowledge about time series prediction may be employed to get an intuitive feeling for the meaning of the inverse of an (infinite) covariance matrix. Despite a number of technical problems involved in making this formal, it is instructive to interpret theoretical results in this way.

Although this does not hold in general, assume that the interpolation error $\tilde{Z}_i$ is expressible as an infinite series $\sum_{-\infty}^{\infty} g_{im} Z_{i-m}$ and denote its variance by d_i. Intuitively the infinite array $(g_{im}, -\infty < m < \infty)$ and the scalar d_i should determine the i-th row of the inverse covariance matrix, if such a matrix can be made meaningful in the infinite case. For stationary time series, as well as for others that may be reduced to stationarity by a simple linear filter (e.g. ARIMA models), the following *time invariance property* holds: the coefficients g_{im} and the variances d_i do not depend upon i.

It then makes sense to talk of *the* interpolation variance and *the* coefficient of the best interpolator corresponding to a given (either positive or negative) lagged value, without having to specify the time position of the random variable being interpolated. Unfortunately this invariance property cannot be expected to hold in the finite case, because the support S_i formed by all m such that $g_{im} \neq 0$ will depend upon i. For an illustration take $n = 5, i = 5$. Since the best linear predictor of Z_5 is a linear combination of Z_{5-4}, Z_{5-3}, Z_{5-2}, and Z_{5-1}, it follows that S_5 is contained in $\{0, 1, 2, 3, 4\}$. A similar argument shows that S_3 and S_1 are contained in

$S_3 = \{-2,-1,0,1,2\}$, and $S_5 = \{0,-1,-2,-3,-4\}$ respectively. For an infinite series S_i would be $\{-2,-1,0,1,2\}$, for any value of i.

An intuitive approach to the study of infinite time series is to let $n = 2t - 1$ and consider the interpolation of the middle observation Z_t. The coefficient g_{tm} of Z_{t-m} in the interpolation is expected to converge to a constant g_m when m is fixed and the length n of the series tends to ∞. But in the limit any observation may be thought to be in the middle of the series and this provides a heuristic proof of the invariance property.

For a stationary time series the time reversability property implies that the inverse covariance matrix of a finite segment of this series is symmetric with respect to the second diagonal. This implies that $d_i = d_{n-i+1}$ and $g_{ij} = g_{n-i+1,n-j+1}$. On intuitive grounds $d_1 \geq d_2 \geq \cdots \geq d_m$, where m is the smallest equal to or exceeding $\frac{n}{2}$. These properties may be translated into abstract statements about matrices. For instance, symmetry with respect to the second diagonal holds for the inverse of any symmetric Toeplitz. This condition is called persymmety by Golub and Van Loan (1989), who prove that it actually holds for arbitrary Toeplitz matrices, without the condition of symmetry.

Intuitively speaking, the inverse of a Toeplitz matrix is then approximately Toeplitz, except for border effects, which disappear as n tends to ∞. Therefore, in the limit infinite case the Toeplitz property for the inverse holds exactly, with the constant diagonal value being the reciprocal of the interpolation variance of Z_t given all past and future values. From (2.7) it follows that, in the limit, inverse correlations coincide with the negative of the coefficients of the best linear interpolator. Since the inverse correlations c_{ij} depend only on $\mid i - j \mid$, it is natural to call $c_{i,i+k}$ the k-th inverse autocorrelation or the inverse autocorrelation of order k. It turns out that this definition is equivalent to the standard ones in the literature (see Cleveland (1972) and the discussion below) in the infinite case. The point of view taken here has the advantage of not involving asymptotic approximations and gives us some feeling about the difficulties involved when attempting to use the asymptotic results in the finite case.

To convert the heuristics into a rigorous formulation it is better to bypass infinite matrices, and to use instead the inversion of a linear operators. It turns out that computing the inverse of these operators by diagonalization is directly related to the spectral properties of the time series. It is precisely within this framework that the interpolation problem for an infinite stationary time series was first solved by Kolmogorov in 1941. Although this classical result is available in many theoretical books (e.g. Grenander and Rosenblatt (1957), Hannan (1960, pp.22-24; 1970, pp.163-168), Lamperti (1977), Rozanov (1967), Whittle (1983), and Yaglom (1962)), a reference to it is often missing in most applied books. Kolmogorov (1941a,b) expresses the error variance and the coefficients of the optimal interpolator in terms of the Fourier series of the reciprocal of the spectral density function. He shows that if the reciprocal of the spectral density function $f(\lambda)$ is square integrable, with Fourier series

$$\frac{1}{f(\lambda)} = \sum_{-\infty}^{\infty} g_k e^{2\pi i k \lambda} \quad \lambda \in [-\pi, \pi],$$

then g_0^{-1} is the error variance of the optimal interpolator and $-g_k/g_0$ is the coefficient of Z_{t+k} in the optimal interpolator of Z_t. The superficial resemblance of these equations with (2.2) and (2.3) respectively has a deeper root. The key fact is that the covariance matrix V of a stationary time series is approximately diagonalized by the "finite Fourier transform", that is $V = PDP^*$ holds approximately, with

$$P_{tj} = \frac{1}{\sqrt{n}} h_j(t)$$

and D is a diagonal matrix with elements $d(t) = 2\pi f(\lambda_t)$, where $\lambda_j = \frac{2\pi j}{n}$ and $h_j(t) = e^{i\lambda_j t}$. But then $V^{-1} = PD^{-1}P^*$, so that the approximate equalities hold:

$$\begin{aligned} V^{r,r+m} &= \frac{1}{n} \sum_t h_r(t) h_{r+m}(t)/d(t) \\ &= \frac{1}{4\pi^2} \frac{2\pi}{n} \sum h_m(t) f(\lambda_t). \\ &= \frac{1}{4\pi^2} \int_{-\pi}^{\pi} e^{i\lambda m} \frac{1}{f(\lambda)} d\lambda \\ &= \frac{1}{4\pi^2} g_m. \end{aligned}$$

To this degree of approximations $V^{r,r+m}$ is independent on r, i.e. V^{-1} is a Toeplitz matrix. The inverse correlation associated to Z_r, Z_{r+m} should then be close to g_m/g_0. This spectral approach underlies the definition of inverse autocorrelations by Cleveland (1972), who originally proposed them as tools for identifying ARMA models. This concept is also treated by Chatfield (1977). Many researchers have recognized its importance in the context of interpolation (see e.g. Brubacker and Wilson (1976), Maravall and Peña (1988), Bhansali (1990)).

In closing, we would like to stress the fact that most of the literature implicitly discuss interpolation and inverse autocorrelations in the context of infinite time series. Based on the general theorems for inverse Gram matrices one can take the finite case as the starting point and then introduce the infinite case as a convenient theoretical approximation, which in qualitative terms will be worse the closer the observation being interpolated is to the extremes of the series and the slower the inverse autocorrelations tend to zero.

6 Zeroes on the inverse Gram matrix

6.1 General interpretation

A zero value for A^{ij} in the inverse Gram matrix is geometrically equivalent to the orthogonality of the vectors $\tilde{v}_i$ and $\tilde{v}_j$, that is to $d_{ij} = 0$. From (2.5) and (2.6) $d_{ij} = -c_{ij}$ and hence this is equivalent to $c_{ij} = 0$. Therefore the following conditions are equivalent:

(a0) $\tilde{v}_i \perp \tilde{v}_j$;

(b0) $v_i^*\{i,j\} \perp v_j^*\{i,j\}$;

(c0) The coefficient of v_j in the orthogonal projection of v_i onto $span(v_j, j \neq i)$ is null;

(d0) The coefficient of v_i in the orthogonal projection of v_j onto $span(v_i, i \neq j)$ is null.

This equivalence of four conditions in the abstract setting of inner product spaces has a number of interesting consequences in particular statistical applications, which we discuss next.

6.2 Application to regression

Consider the regression model $E(\boldsymbol{Y}) = \boldsymbol{\mu} = X\boldsymbol{\beta}$, Var $(\boldsymbol{Y}) = \sigma^2 I$ discussed in Section 3. To illustrate the ideas we choose X to be a $N \times 4$ full rank matrix and analyze the meaning of the condition $(X'X)^{14} = 0$. To analyze the meaning of $(X'X)^{14} = 0$ consider the following (artificial) least squares fits, where both the dependent and independent variables are some columns of X :

(i) $\boldsymbol{X}_1$ vs. $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$;

(ii) $\boldsymbol{X}_4$ vs. $\boldsymbol{X}_1$,$\boldsymbol{X}_2$, $\boldsymbol{X}_3$;

(iii) $\boldsymbol{X}_1$ vs. $\boldsymbol{X}_2$, $\boldsymbol{X}_3$;

(iv) $\boldsymbol{X}_4$ vs. $\boldsymbol{X}_2$, $\boldsymbol{X}_3$.

From each of these regressions we form the corresponding residual vectors, denoting them by $\boldsymbol{e}(i)$–$\boldsymbol{e}(iv)$ respectively. Conditions (a0) and (b0) become (a1) and (b1) below. Looking at the coefficients in (i) and (ii), conditions (c1) and (d1) below are a direct consequence of (c0) and (d0) respectively.

(a1) The empirical correlation between $\boldsymbol{e}(i)$ and $\boldsymbol{e}(ii)$ is 0;

(b1) The empirical correlation between $\boldsymbol{e}(iii)$ and $\boldsymbol{e}(iv)$ is 0;

(c1) The coefficient of $\boldsymbol{X}_4$ in (i) is 0;

(d1) The coefficient of $\boldsymbol{X}_1$ in (ii) is 0.

6.3 Application to correlation and linear prediction

From the discussion of Section 4, the general conditions (a0)–(d0) become:

(a2) Corr $(\tilde{Z}_i, \tilde{Z}_j) = 0$;

(b2) The partial correlation of Z_i and Z_j, given $Z_r, r \neq i, j$, is null;

(c2) The coefficient of Z_j in the BLP of Z_i given $Z_r, r \neq i$, is null;

(d2) The coefficient of Z_i in the BLP of Z_j given $Z_r, r \neq j$, is null.

The fact that i and j may be interchanged in (c2) to get (d2) comes from the symmetry property $c_{ij} = c_{ji}$, which proofs the validity of the implication $c_{ij} = 0 \Rightarrow c_{ji} = 0$.

6.4 Application to time series

Many time series models are defined in terms of a time evolution. It is a useful fact that formulae for the BLP of Z_t given $(Z_s, s < t)$, together with the variance of its prediction error, for every $t = 1, \ldots, n$ completely determines the covariance structure of a finite time series $Z_1, \ldots, Z_n$. This property is equivalent to the Choleski decomposition of the covariance matrix. From the discussion of Section 5, zero entries in the inverse covariance matrix are associated with zero coefficients in the BLP. We are concerned here with the situation where at most p coefficients are non zero, where p is a fixed number, which does not depend on the number n of observations. For an infinite stationary time series, this leads to the *autoregressive model of order* p, denoted by $AR(p)$; for a finite one this condition means that all subdiagonals of the lower triangular matrix L that appears in the Levinson-Durbin algorithm, beyond the p-th one are null.

The standard definition of an AR(p) model treats the future and the past in an asymmetric way. It may be shown however that for this model the best linear interpolator of X_t involves only the values of X_s with $0 <| t - s |\leq p$. Hence the subdiagonals and superdiagonals of the inverse covariance matrix, whose distances to the main diagonal are larger than p, are null. This last condition provides a way for extending the definition of AR(p) to a finite and nonstationary time series. By omitting the stationarity requirement one gets the so called antedependent process of order p.

An application of the general conditions (a2)–(d2) shows that an infinite stationary time series follows an $AR(p)$ model if and only if any of the following conditions holds:

(a3) The inverse autocorrelations of order k are zero for $k > p$;

(b3) The partial autocorrelations of order k are zero for $k > p$;

(c3) The coefficient of Z_{t_k} in the best linear interpolator of Z_t is null for $|k| > p$;

(d3) The coefficient of Z_{t_k} in the best linear predictor of Z_t is null for $k > p$.

One of the major advantages of looking at the empirical inverse autocorrelations rather than at the empirical partial autocorrelations, is the superiority of the former for identifying models with structural zero coefficients for some lags, e.g. for seasonal models. This property is linked to the theoretical behavior of inverse and partial autocorrelations in an AR model of the form

$$X_t - \sum_{i=1}^{M} \phi_{j(i)} X_{t-j(i)} = a_t.$$

It turns out that the error associated to the best linear interpolator has the representation

$$X_t - \sum_{i=1}^{M} \omega_{j(i)} (X_{t-j(i)} + X_{t-j(i)}).$$

From (a2) and (c2) it follows that the inverse autocorrelations are zero for any lag k not included in the set $\{j(1), \ldots, j(M)\}$. For partial autocorrelations, this holds only for $j(i) = i, i = 1, \ldots, M$, i.e. for the unrestricted AR(M) model.

6.5 Gaussian case and conditional independence

When Y has a multivariate normal distribution $N(\mu, V)$, Y_i and Y_j are *conditionally independent* given the remaining components if and only if $V^{ij} = 0$. This condition is also equivalent to that of zero partial or inverse correlation. Thus a geometric property translates into a probabilistic one. Conditional independence is a key concept behind many statistical models, e.g. Markov Models, Covariance Selection Models (Wermuth (1976)), and their extensions to Graphical Models (Lauritzen y Wermuth (1989), Whitakker (1990)). The associated diagrams are a very useful tool and can be adapted to express the appearance of zeroes in the inverse Gram matrix.

7 Summary

We have attempted to illustrate the fruitfulness of singling out the inverse of an arbitrary Gram matrix as an abstract object worth of being studied on its own right. Furthermore we have stressed a geometric viewpoint, since it gives us a powerful way of depicting the results in particular applications. Applications have been made to areas as different as linear regression, multivariate analysis, prediction theory, Gaussian graphical models, and time series (particularly through inverse correlations and autocorrelations). Although these abstract objects appear in disguise, once they are uncovered a number of particular results may be immediately derived from their general properties. Of particular interest are zero entries, which admit an interesting geometrical interpretation in terms of orthogonality and are automatically equivalent to conditional independence relationships in the Gaussian case.

These ideas are particularly important in linear prediction theory, since the inverse covariance matrix is typically much more relevant than the covariance matrix itself. This means that large changes in the covariances often affect most predictions only slightly.

8 Appendix: proofs

Proof of Theorem 1

Let D be the $n \times n$ matrix, whose entries are given by the right hand side of (2.2) and (2.3). We need only show that $AD = I$, where I denotes the identity matrix. This means that the entries D_{ij} must satisfy the equations

$$\begin{aligned}\sum_{j=1}^{n} D_{ij} A_{jr} &= 1, \text{if } j = i \\ &= 0, \text{if } j \neq i.\end{aligned}$$

The linear independence of $v_1, \ldots v_n$ implies that $\| \tilde{v}_i \| > 0$, and hence (2.2) is well defined. But

$$\begin{aligned}\sum_{j=1}^{n} D_{ij} A_{jr} &= D_{ii} A_{ir} + \sum_{j \neq i}^{n} D_{ij} A_{jr} \\ &= D_{ii} \left[A_{ir} - \sum_{j \neq i}^{n} b_{ij} A_{jr} \right] \\ &= D_{ii} \left[\langle v_i, v_r \rangle - \sum_{j \neq i}^{n} b_{ij} \langle v_j, v_r \rangle \right] \\ &= D_{ii} \left[\langle v_i - \sum_{j \neq i}^{n} b_{ij} v_j, v_r \rangle \right] \\ &= D_{ii} \langle \tilde{v}_i, v_r \rangle.\end{aligned}$$

The result then follows from $\langle \tilde{v}_i, v_r \rangle = 0, i \neq r$ and $\| \tilde{v}_i \|^2 = \langle \tilde{v}_i, v_i \rangle$.

Proof of Theorem 2

To simplify the notation we write here v_i^* and v_i^* for $v_i^*\{i,j\}$ and $v_i^*\{i,j\}$ respectively.

(i) By expressing $\tilde{v}_i$ in terms of $v_1, \ldots, v_n$, forming the inner product with $\tilde{v}_j$, and using the equality $\langle \tilde{v}_j, v_r \rangle = 0, r \neq j$, we get

$$\begin{aligned} \langle \tilde{v}_i, \tilde{v}_r \rangle &= \frac{A^{ir} \langle \tilde{v}_r, v_r \rangle}{A^{ii}} \\ &= \frac{A^{ir}}{A^{ii} A^{rr}} \end{aligned}$$

and (2.5) follows from (2.2).

(ii) Without loss of generality we may assume $\| v_r^* \| = 1, r = i, j$. Let $\langle v_i^*, v_j^* \rangle = \lambda$. But (v_i^*, v_j^*) span the orthogonal complement N of the subspace spanned by $(v_s, s \neq i, j)$ and clearly $\tilde{v}_i, \tilde{v}_j \in N$. It is then easy to get the relationships

$$\tilde{v}_i = v_i^* - \lambda v_j^* \qquad \tilde{v}_j = v_j^* - \lambda v_i^*.$$

From $\| v_r^* \|^2 = \langle \tilde{v}_r, v_r^* \rangle$, $r = i, j$ and $\langle \tilde{v}_i, \tilde{v}_j \rangle = \lambda - 2\lambda + \lambda^3$ we get

$$c_{ij} = \frac{-\lambda(1 - \lambda^2)}{1 - \lambda^2}$$

and hence $c_{ij} = -\lambda = -d_{ij}$.

Acknowledgements

This research has been partially supported by Fondo Nacional de Investigación Científica y Tecnológica under Grants FONDECYT 91-808 and 1930503.

(Received July, 1995. Revised October, 1995.)

References

Bhansali, R.J. (1990). On a relationship between the inverse of a stationary covariance matrix and the linear interpolator. *J. Appl. Prob.*, **27**, 156-170.

Brockwell, P.J. and Davies,R.A. (1987). *Time series: theory and methods.* New York, Springer Verlag.

Brubacker, S., Wilson, G. (1976). Interpolating time series with applications to the estimation of holidat effects on electricity demand. *J. Roy. Statist. Soc C* , **25**, 107-116.

Chatfield, C. (1977). Inverse autocorrelations. *J. R. Statist. Soc. A*,**142**, 363-377.

Chatterjee, S., Price, B. (1977). *Regression analysis by example.* New York, John Wiley.

Cleveland, W.S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics*, **14**, 277-293.

Golub, G.H., Van Loan, C.F.(1989). *Matrix computations.* 2nd.ed. Baltimore, The John Hopkins University Press.

Goodnight, J. (1979). A tutorial on the SWEEP operator. *The American Statistician*, **33**, 149-158.

Grenander, U., Rosenblatt, M.(1957). *Statistical analysis of stationary time series.* New York, John Wiley.

Hannan, E.J. (1960). *Time series analysis.* London, Methuen.

Hannan, E.J. (1970). *Multiple time series.* New York, John Wiley.

Kolmogorov, A.N. (1941a). Stationary sequences in Hilbert space. *Bul. Moscow State Univ.* 2(6).

Kolmogorov, A.N. (1941b). Interpolation and extrapolation of stationary random sequences. *Izv. Akad. Nauk. SSSR, Ser. Mat.*, **2** (3).

Lamperti, J. (1977). *Stochastic processes.* New York, Springer Verlag.

Lauritzen, S.L., Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, **17**, 31-57.

Maravall, A., Peña, D. (1988). Missing observations in time series and the dual autocorrelation function. Madrid, Banco de España. Doc. Trabajo 8803.

Rozanov, YU. A. (1967). *Stationary random processes.* San Francisco, Holden Day. 96-104.

Theil, H. (1971). *Principles of cconometrics.* New York, John Wiley.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection models. *Biometrics*, **32**, 253-264.

Whittaker, J. (1990). *Graphical models in applied multivariate analysis.* Chichester, John Wiley.

Whittle, P. (1983). *Prediction and regulation.* 2nd.ed. Ed. Minnesota, U. of Minnesota Press.

Yaglom, A.M. (1962). *An introduction to the theory of stationary random functions.* New York, Dover.