The Annals of Mathematical Statistics
1972, Vol. 43, No. 4, 1317–1319

REGRESSION OPTIMALITY OF PRINCIPAL COMPONENTS

BY R. L. OBENCHAIN

Bell Telephone Laboratories

Consider $p \geqq 2$ random variables, and let $A_1, \cdots, A_p$ denote the hyperplanes corresponding to the linear regression of each variable onto the other $(p-1)$ variables. Let A_0 denote the hyperplane which passes through the centroid of the distribution and is spanned by the direction vectors defining the first $(p-1)$ principal components. A new optimality property of A_0 is established; A_0 is the best single approximation to $A_1, \cdots, A_p$ when each regression hyperplane is given a certain weighting inversely proportional to the variability associated with its orientation and its prediction rescaling. When $p > 2$ and $k = 1, \cdots, p-2$, certain k-dimensional linear subspaces of A_0 are also shown to have regression optimality properties.

1. Introduction. We adopt the notation of Okamoto (1969). Thus we let $\mathbf{x}$ be a random $p \times 1$ vector with mean $\boldsymbol{\mu} = E(\mathbf{x})$ and covariance $\boldsymbol{\Sigma} = V(\mathbf{x}) = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'$. Let $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p$ be the eigenvalues of $\boldsymbol{\Sigma}$, and let $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p$ be a corresponding set of orthonormal eigenvectors of $\boldsymbol{\Sigma}$. Then, for $i = 1, \cdots, p$, the random variable $\xi_i = \boldsymbol{\gamma}_i'(\mathbf{x} - \boldsymbol{\mu})$ will be called the ith principal component of $\mathbf{x}$. Only the case $V(\xi_p) = \lambda_p > 0$ will be considered in this paper. The principal components of a set of points are also defined as in Okamoto (1969), so the details of this special case of the above formulation will not be repeated here.

Let A_0 be the hyperplane passing through $\boldsymbol{\mu}$ and spanned by the first $(p-1)$ principal component directions, $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{p-1}$. It follows that $A_0 = \{\mathbf{y} \mid \boldsymbol{\gamma}_p'(\mathbf{y} - \boldsymbol{\mu}) = 0\}$, and A_0 is uniquely determined if and only if $\lambda_{p-1} > \lambda_p$.

Let $\boldsymbol{\alpha}$ be a non-null $p \times 1$ vector, and let $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}/(\boldsymbol{\alpha}'\boldsymbol{\alpha})^{\frac{1}{2}}$ denote the unit vector in the (positive) direction of $\boldsymbol{\alpha}$. Consider the linear combination of random variables, $\boldsymbol{\alpha}'\mathbf{x} = (\boldsymbol{\alpha}'\boldsymbol{\alpha})^{\frac{1}{2}}(\boldsymbol{\alpha}^{*\prime}\mathbf{x})$, and note that $V(\boldsymbol{\alpha}'\mathbf{x}) = (\boldsymbol{\alpha}^{*\prime}\boldsymbol{\Sigma}\boldsymbol{\alpha}^*)(\boldsymbol{\alpha}'\boldsymbol{\alpha})$. Thus the variance of $\boldsymbol{\alpha}'\mathbf{x}$ is the product of two factors: $(\boldsymbol{\alpha}^{*\prime}\boldsymbol{\Sigma}\boldsymbol{\alpha}^*)$ is the variance associated with the direction of $\boldsymbol{\alpha}$ (orientation factor), and $(\boldsymbol{\alpha}'\boldsymbol{\alpha})$ is the effect of the scale chosen along the direction of $\boldsymbol{\alpha}$ (rescaling factor).

2. The linear regression hyperplanes. The linear regression of x_i onto $\mathbf{x}_{(-i)} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)'$ is expressed by the equation

$$\hat{x}_i = \mu_i + \boldsymbol{\beta}_i'(\mathbf{x}_{(-i)} - \boldsymbol{\mu}_{(-i)}), \tag{2.1}$$

where $\boldsymbol{\beta}_i$ is a $(p-1) \times 1$ vector of regression coefficients. Whatever the distribution of $\mathbf{x}$, $\boldsymbol{\beta}_i$ is defined as if the joint distribution of $\mathbf{x}$ were multivariate normal with moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In this case, $\hat{x}_i$ of (2.1) is the conditional ex-

Received September 20, 1971; revised November 24, 1971.

AMS 1970 *subject classifications*. Primary 62H25; Secondary 62J05.

Key words and phrases. Principal components, linear regression, hyperplane.

pected value of x_i given $\mathbf{x}_{(-i)}$, and the corresponding conditional variance will be denoted by σ_{ii}^*.

Now (2.1) is rewritten by noting that, given $\mathbf{x}_{(-i)}$, $\hat{x}_i$ is the value of x_i such that

$$(2.2) \qquad \boldsymbol{\zeta}_i'(\mathbf{x} - \boldsymbol{\mu}) = (x_i - \mu_i) - \boldsymbol{\beta}_i'(\mathbf{x}_{(-i)} - \boldsymbol{\mu}_{(-i)}) = 0 \, ,$$

where $\zeta_{ii} = +1$. The hyperplane, A_i, corresponding to (2.1) and (2.2) is

$$(2.3) \qquad A_i = \{\mathbf{y} \mid \boldsymbol{\zeta}_i^{*\prime}(\mathbf{y} - \boldsymbol{\mu}) = 0\} \, ,$$

where $\boldsymbol{\zeta}_i^* = \boldsymbol{\zeta}_i/(\boldsymbol{\zeta}_i'\boldsymbol{\zeta}_i)^{\frac{1}{2}}$. Note that the *prediction equation*, (2.1) or (2.2), requires a specific scaling along the direction $\pm\boldsymbol{\zeta}_i^*$ which defines A_i.

We now state a point which will be known to some readers: the elements of $\boldsymbol{\zeta}_i$ are simply related to those of the ith column (or row) of $\boldsymbol{\Sigma}^{-1}$. The conditional variance, σ_{ii}^*, of x_i given $\mathbf{x}_{(-i)}$ is the reciprocal of the (i, i)th element of $\boldsymbol{\Sigma}^{-1}$, and the ith column of $\boldsymbol{\Sigma}^{-1}$ is $\boldsymbol{\zeta}_i/\sigma_{ii}^*$. Finally, note that $\sigma_{ii}^* = \boldsymbol{\zeta}_i'\boldsymbol{\Sigma}\boldsymbol{\zeta}_i$.

3. The relationship between A_0 and $A_1, \cdots, A_p$. Consider a direction $\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^{*\prime}\boldsymbol{\alpha}^* = 1$, and note that the ith element of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}^*$ is $\boldsymbol{\zeta}_i'\boldsymbol{\alpha}^*/(\boldsymbol{\zeta}_i'\boldsymbol{\Sigma}\boldsymbol{\zeta}_i)$, which is proportional to cosine of the angle, θ_i, between $\boldsymbol{\alpha}^*$ and $\boldsymbol{\zeta}_i$. Now, if the hyperplane passing through $\boldsymbol{\mu}$ and orthogonal to $\boldsymbol{\alpha}^*$ is to approximate all p regression hyperplanes, $A_1, \cdots, A_p$, all of the angles, $\theta_1, \cdots, \theta_p$, should be made as close as possible to zero or $\pm\pi$. Specifically, it is reasonable to maximize, by choice of $\boldsymbol{\alpha}^*$, a weighted sum of the absolute values or squares of the cosines. The ith term in the summation could be weighted in inverse proportion to the variability associated with the regression of x_i on $\mathbf{x}_{(-i)}$.

In accordance with the above considerations, we note that

$$(3.1) \qquad \boldsymbol{\alpha}^{*\prime}\boldsymbol{\Sigma}^{-2}\boldsymbol{\alpha}^* = \sum_{i=1}^{p} \frac{\cos^2\theta_i}{(\boldsymbol{\zeta}_i^{*\prime}\boldsymbol{\Sigma}\boldsymbol{\zeta}_i^*)^2(\boldsymbol{\zeta}_i'\boldsymbol{\zeta}_i)}$$

is a reasonable criterion to be maximized. Note, in particular, that the weight given to the ith term of the summation is more sensitive to the orientation variance factor, $\boldsymbol{\zeta}_i^{*\prime}\boldsymbol{\Sigma}\boldsymbol{\zeta}_i^*$, associated with A_i than to the rescaling variance factor, $\boldsymbol{\zeta}_i'\boldsymbol{\zeta}_i$, associated with the ith prediction equation, (2.1) or (2.2).

THEOREM. (Regression Optimality of Principal Components.) *A_0 is the optimal approximation to $A_1, \cdots, A_p$ in the sense that this choice maximizes* (3.1). *In particular,*

$$(3.2) \qquad \boldsymbol{\alpha}^{*\prime}\boldsymbol{\Sigma}^{-2}\boldsymbol{\alpha}^* \leqq \lambda_p^{-2} \, ,$$

and the maximum is achieved if and only if $\boldsymbol{\alpha}^* = \boldsymbol{\gamma}_p$. The solution is not unique when $\lambda_{p-1} = \lambda_p$.

PROOF. $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p)$, implies that $\boldsymbol{\Sigma}^{-2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Gamma}'$. Thus the eigenvector, $\boldsymbol{\gamma}_p$, corresponding to the smallest eigenvalue, λ_p, of $\boldsymbol{\Sigma}$ also corresponds to the largest eigenvalue of $\boldsymbol{\Sigma}^{-2}$. The theorem thus follows from a well-known lemma, cf. Okamoto (1969), Lemma 2.2.

COMMENT. It should be clear that the choice $\boldsymbol{\alpha}^* = \boldsymbol{\gamma}_p$ maximizes $\boldsymbol{\alpha}^{*\prime}\boldsymbol{\Sigma}^{-k}\boldsymbol{\alpha}^*$ for any positive integer k. However, only when $k = 2$ does this criterion appear to have a simple geometric interpretation, that of (3.1).

4. Concluding remark. In analogy with the three types of optimality properties of principal components given by Okamoto (1969), it would be interesting to display, for $k = 1, \cdots, p$, a regression optimality property of the k-dimensional linear subspace passing through $\boldsymbol{\mu}$ and spanned by the first k principal component directions, $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_k$. Rather than introduce the notation needed to formally present such a characterization, the following argument shows that such an extension is straightforward. A k-dimensional linear subspace passing through $\boldsymbol{\mu}$ is orthogonal to $(p - k)$ mutually orthogonal directions. To maximize its fit to $A_1, \cdots, A_p$, each of the orthogonal directions should be taken to be, as close as is possible, parallel to $\pm\boldsymbol{\zeta}_1^*, \cdots, \pm\boldsymbol{\zeta}_p^*$. The goodness-of-fit criterion would be the sum of $(p - k)$ terms like (3.1), and the optimal orthogonal directions could be chosen to be $\boldsymbol{\gamma}_{k+1}, \cdots, \boldsymbol{\gamma}_p$.

REFERENCES

OKAMOTO, M. (1969). Optimality of principal components. In *Multivariate Analysis*, **2** (P. R. Krishnaiah, ed.). 673–685. Academic Press, New York.