NOAA Technical Memorandum ERL PMEL-71

THE PRINCIPAL DISCRIMINANT METHOD OF PREDICTION: THEORY AND EVALUATION

Rudolph W. Preisendorfer

Curtis D. Mobley
Joint Institute for the Study of the Atmosphere and Ocean
University of Washington
Seattle, Washington

Tim P. Barnett
Climate Research Group
Scripps Institute of Oceanography
La Jolla, California

Pacific Marine Environmental Laboratory
Seattle, Washington
May 1987

UNITED STATES DEPARTMENT OF COMMERCE

Malcolm Baldrige, Secretary

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Anthony J. Calio, Administrator

Environmental Research Laboratories

Vernon E. Derr, Director

NOTICE

Mention of a commercial company or product does not constitute an endorsement by NOAA Environmental Research Laboratories. Use for publicity or advertising purposes of information from this publication concerning proprietary products or the tests of such products is not authorized.

Contribution No. 931 from NOAA/Pacific Marine Environmental Laboratory

For sale by the National Technical Information Service, 5285 Port Royal Road
Springfield, VA 22161

CONTENTS

Preface

This report was prepared by the junior authors subsequent to the untimely death of Dr. Rudolph Preisendorfer. Part I is based on the unpublished notes of Preisendorfer (Preisendorfer, 1984), to whom all credit is acknowledged for the conceptual framework of the Principal Discriminant Method. Part II is based on the Master's Degree Theses of Christine Elias and Steve Fatjo. Their work was performed at the Naval Postgraduate School in close association with Dr. Preisendorfer. Part III is the work of the junior authors.

The Principal Discriminant Method of Prediction: Theory and Evaluation

Rudolph W. Preisendorfer[1]
Curtis D. Mobley[2]
Tim P. Barnett[3]

ABSTRACT. The Principal Discriminant Method (PDM) of prediction employs a novel combination of principal component analysis and statistical discriminant analysis. Discriminant analysis is based on the construction of discrete category subsets of predictor values in a multidimensional predictor space. A category subset contains those predictor values which give rise to a predictand (or observation) in that particular category. A new predictor value is then assigned to a particular category (i.e., a forecast is made) through the use of probability distribution functions which have been fitted to the category subsets. The PDM uses principal component analysis to define the multidimensional probability distribution functions associated with the category subsets. Because of its underlying discriminant nature, the PDM is also applicable to problems in data classification.

After presenting the theory of the PDM, it is subjected to four analyses. The first uses actual data to forecast discrete values of horizontal visibility over the ocean, using the PDM in a Model Output Statistics (MOS) setting. The second analysis is also in an MOS setting, but this time artificially constructed data sets are used, with predetermined levels of noise and inherent predicability. In each study the PDM is compared with other forecast methods (such as linear regression). The third analysis uses the PDM to forecast the onset of the 1982-83 El Niño, as expressed by sea surface temperature anomalies, using wind anomalies as the predictors. In the fourth analysis, sea level pressures over the North Pacific are used to predict surface air temperatures over North America.

It is found that when applied to artificial data, the PDM shows forecast skills which are comparable to other standard forecast techniques. However, when applied to actual data sets, the PDM is generally outperformed by other forecast techniques. It is concluded that the failure of the PDM in these situations is a consequence of the noisy nature of the data sets, which prevents the PDM from adequately defining the category subsets. If the input data sets are suitably smoothed or filtered in order to increase the signal-to-noise ratio, then the PDM is once again comparable in skill to other forecast techniques. The underlying concepts of the PDM do, therefore, appear sound, and it is felt that the PDM shows considerable promise.

---

[1] NOAA/Pacific Marine Environmental Laboratory, 7600 Sand Point Way NE, Seattle, WA, 98115-0070

[2] Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, AK-40, Seattle, WA 98195.

[3] Climate Research Group, Scripps Institute of Oceanography, La Jolla, CA 92093

PART I. THEORY OF THE PRINCIPAL DISCRIMINANT METHOD

1. Introduction

Discriminant methods in general, and the Principal Discriminant Method (PDM) in particular, can be applied to forecasting problems in which it is desired to forecast a discrete state of the atmosphere or ocean. An example is the forecasting of seasonal temperatures as one of the three discrete states "above average," "average," or "below average." Because of its underlying discriminant nature, the PDM can also be used in data classification. An example is the assignment of the observed state of the atmosphere to one of several discrete "climate types." A further application of the PDM is the linking of the output of a General Circulation Model (GCM) of the atmosphere with observed fields in order to produce Model-Output Statistic (MOS) schemes of prediction. Our description of the PDM shows its essential form so as to facilitate applications to any of the problems just mentioned.

The successful construction of category subsets in a multidimensional predictor space is a *sine qua non* of any discriminant method, along with the fitting of versatile probability density functions to these subsets. The modifier "principal" in the name of the present method derives from the fact that, for multiple-predictors, essential use is made of Principal Component Analysis (PCA) in order to determine appropriate probability density functions for the category subsets. Another feature of the PDM is that of self-evaluation of predictive skill. This is supplied by three indices of skill: the potential predictability, the potential 0-class error and the potential 1-class error in the predictand categories. These indices along with their critical values, supplied by a Monte Carlo technique, help the user decide how

much confidence to place on a given prediction made by the PDM. Also, during the construction of the PDM's working parts, provision is made to test the method on an independent data set. This testing gives another indication of how well a data set is constituted to allow predictions of its variables' future states.

The exposition of the PDM will be made in two parts. The first part treats the case of a single predictor, in which case the PDM reduces to a classical discriminant method. In real applications the single-predictor mode can yield much information about the potential predictability of a predictand by a given predictor, along with some information about the skill of the predictions. The single-predictor mode of the PDM can therefore stand as an independent, preliminary prediction method. The second part treats the case of multiple predictors. It is expected that the predictability will increase when a single predictor is joined by several more predictors, and when the category subsets in the resultant multidimensional predictor space can be carved out of the swarm of data points there. It is in this mode that the PDM realizes its full power, via its application of Principal Component Analysis to the multidimensional swarm of data points.

The reader desiring an elementary discussion of discriminant analysis in its conventional statistical formulation is referred to Lachenbruch (1975). This compact text also contains a bibliography of 579 references showing an amazing diversity of problems amenable to solution by discriminant analysis. A pioneering application of classical discriminant analysis in meteorology will be found in Miller (1962).

2. The Single-Predictor Stage

It is assumed that we have available a data set consisting of simultaneous observations of both predictors and predictands. Such a data set is required in order to construct the PDM model. After the model has been constructed, it is capable of making forecasts when given new predictor values.

A. The Predictor-Predictand Pair. Let X(J,K) denote the value of the Kth *predictor* X at time J. It is convenient to standardize the predictor in time so that the time series X(J,K), J = 1,2,...,N, has zero mean and unit variance for each K, K = 1,2,...,NK. Let Y(J) denote the value of the *predictand* Y at the same time J. For example, in a Model Output Statistic setting, the various predictors X(J,K) might be the sea surface temperature (K = 1), the surface pressure (K = 2), the relative humidity (K = 3), etc., all at the same spatial location, and a particular predictand Y(J) might be the horizontal visibility at the same time J and at the same or a different location.

B. The Time-lagged Predictor-Predictand Pair. In order to fully use the predictive capabilities of the PDM, we introduce a time lag NTAU into Y(J), so as to pair Y(J + NTAU) with X(J,K), NTAU $\geq$ 0. For simplicity, it will be assumed that NTAU has been introduced into Y(J), and we will retain the notation "X(J,K)" and "Y(J)" for the lagged predictor-predictand pair, where now J = 1,2,...,NT labels the common range of times of the lagged pair. (Our notation is designed to facilitate the coding of the associated computer programs.) Thus X(J,K) might denote a wind anomaly for month J and region K of the ocean, and Y(J) might denote an SST anomaly for a later month J + NTAU and the same or a different region of the ocean.

C. Q-tiling the Predictand. Divide the range of predictand values $\{Y(J): J = 1,\ldots,NT\}$ into Q intervals. By judicious choice of the boundary values $B_1, B_2, \ldots, B_{Q-1}$ between these intervals, we can "Q-tile" the predictand Y(J) into Q discrete categories. Let NY(J) denote the value of the discrete category to which Y(J) belongs; thus NY(J) = M if Y(J) falls into category M, $1 \leq M \leq Q$. Figure 1 illustrates these ideas for the case of Q = 3, called a *tercile* categorization. In the figure we define

$$NY(J) \equiv 1 \quad \text{if} \quad Y(J) < B_1,$$

$$NY(J) \equiv 2 \quad \text{if} \quad B_1 \leq Y(J) < B_2,$$

$$NY(J) \equiv 3 \quad \text{if} \quad B_2 \leq Y(J),$$

for J = 1,...,NT. There is no requirement that the Q categories be equally populated after the Q-tiling of the predictand: Fig. 1c shows five points in category 1, nine in category 2, and seven in category 3.

D. The Discriminant Set. The time series for the Kth predictor X(J,K) (Fig. 1a) and the Q-tiled predictand NY(J) (Fig. 1c) can be combined to form a single diagram, called the *discriminant diagram*. Figure 2 shows the discriminant diagram corresponding to Fig. 1. In this example diagram, one sees at a glance that large, positive predictor values tend to be associated with terciled predictand values in category 1; predictor values near zero are associated with category 2 predictand values; and large, negative predictor values tend to correspond to predictand values in category 3. The *discriminant set* consists of the NT pairs of points [X(J,K),NY(J)], J = 1,2,...,NT.

E. Training and Testing Sets. The discriminant set of NT points is randomly split into two subsets of predetermined sizes NTRN and NTST. The

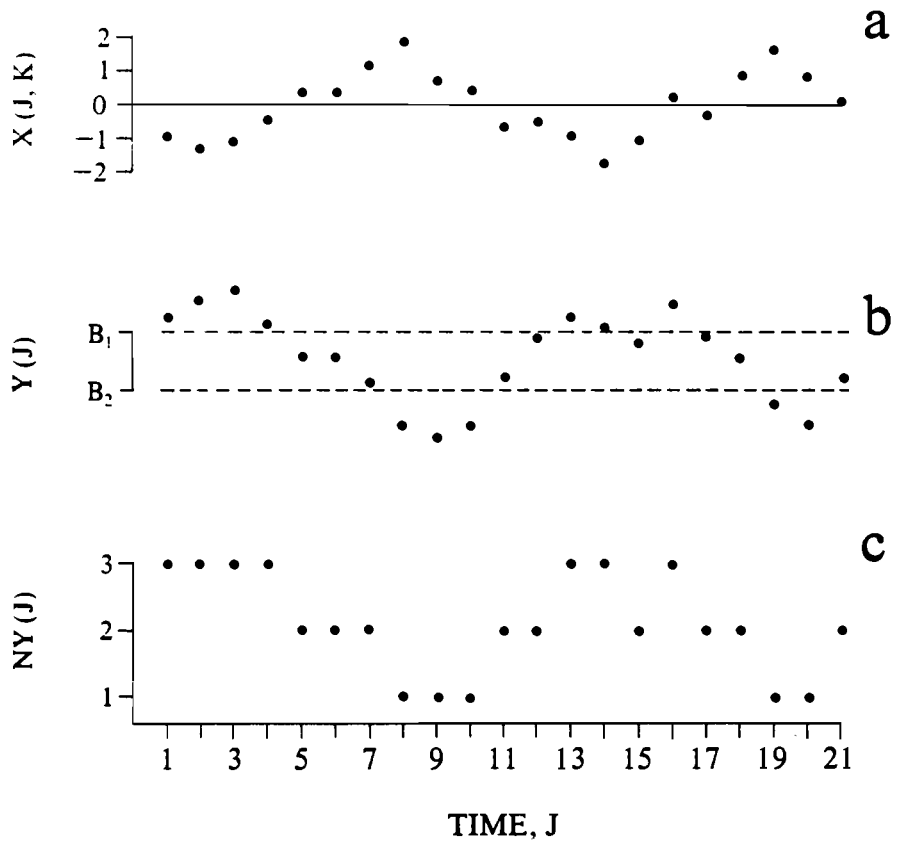

Figure 1.--Illustration of a predictor-predictand pair and a tercile categorization. Panel a shows a standardized predictor time series X(J,K), J = 1,...,NT = 21 and K fixed. Panel b shows the corresponding time series of the predictand values, Y(J), J = 1,...,NT; boundary values B_1 and B_2 are indicated. The terciled values of the predictand, NY(J), are shown in panel c.

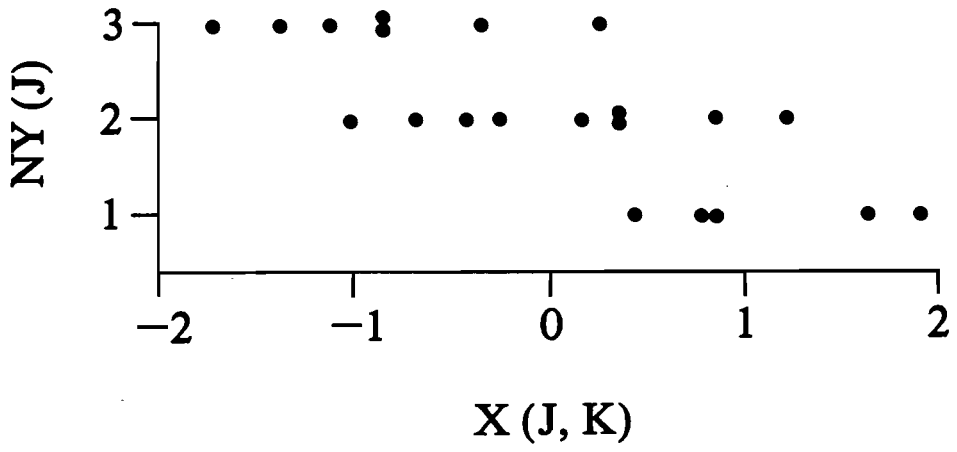

Figure 2.--The discriminant diagram corresponding to Fig. 1a and Fig. 1c. K is held fixed as J runs from 1 to NT = 21.

subset containing NTRN points is the *training set*, and the subset containing NTST points it the *testing set*. Typically we choose NTRN = 2·NTST, so that two-thirds of the NT available data points can be used to "train," or construct, the PDM; and one-third of the points can be used to "test" the PDM. Figure 3 shows a possible partition of the points of Fig. 2 into training and testing sets. Let TRNX(I,K), I = 1,2,...,NTRN, denote those values of X(J,K), J = 1,...,NT, which fall into the training set. Likewise, let NTRNY(I), I = 1,...,NTRN, denote the corresponding values of NY(J). Those points of the discriminant set which have been randomly assigned to the testing set are denoted by [TSTX(I,K), NTSTY(I)], I = 1,2,...,NTST.

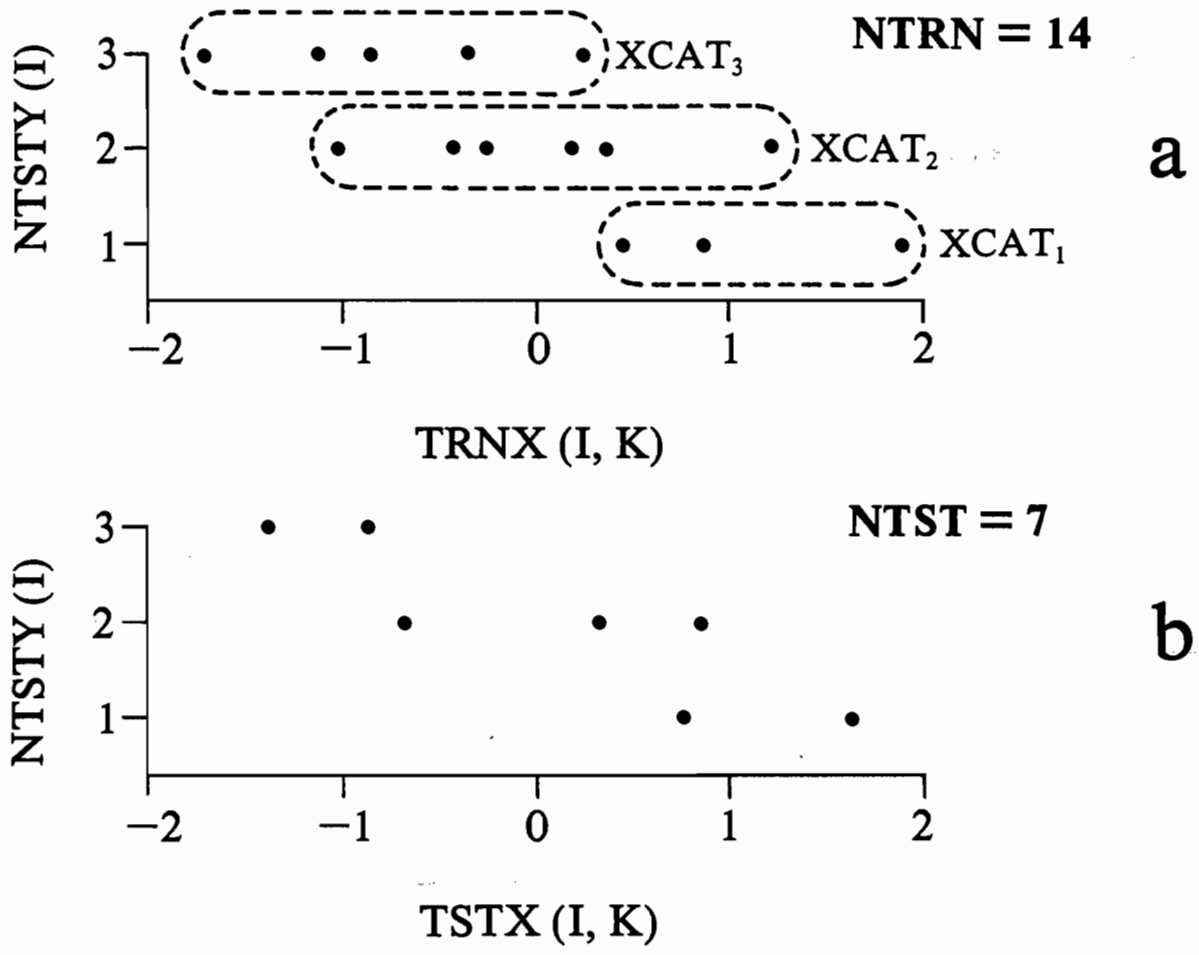

Figure 3.--A partitioning of the discriminant set shown in Fig. 2 into a training set (panel a) and a testing set (panel b). The category subsets of the training set are indicated.

F. Category Subsets of Predictor Space. The subset of predictor points in the training set which are associated with category M of predictand values is termed the Mth *category subset of the predictor space*, denoted by $XCAT_M(I,K)$, $I = 1,2,\ldots,NCAT_M$, $M = 1,2,\ldots,Q$. That is, if NTRNY(I) = M, the corresponding TRNX(I,K) is an element of $XCAT_M(I,K)$. Figure 3 shows the three category subsets for the illustrated training set: $XCAT_1$ with $NCAT_1 = 3$, $XCAT_2$ with $NCAT_2 = 6$, and $XCAT_3$ with $NCAT_3 = 5$. The category subsets form the heart of the discriminant structure of the PDM.

G. Fitting the Probability Density Functions. Once the category subsets of predictor points have been obtained, any discriminant method, including the PDM, requires the fitting of probability density functions to these category subsets. A decisive point in the discriminant method can arise when choosing the specific form of the probability density function (pdf) to be fitted to the category subsets. To be specific, we will choose the gaussian distribution for this exposition. However, it may occasionally be worthwhile to use a pdf specifically tailored to a given data set. The form of the gaussian pdf for category M is

$$\phi_M(X) = (2\pi\sigma_M^2)^{-\frac{1}{2}} \text{EXP}\left[-\frac{(X - \bar{X}_M)^2}{2\sigma_M^2}\right]$$

where

$\bar{X}_M$ is the average over I of the Mth category subset $\{XCAT_M(I,K)$: $I = 1,\ldots,NCAT_M\}$

and

σ_M^2 is the variance of this set of points.

Note that although the original data set X(J,K), J = 1,...,NT, was standardized to zero mean and unit variance, the category subsets $XCAT_M(I,K)$ in general have nonzero means and non-unit variances. Figure 4 shows the fitted gaussian pdf's, $\phi_1(X)$, $\phi_2(X)$ and $\phi_3(X)$, for the category subsets of Fig. 3a. Once the $\phi_M(X)$, M = 1,...,Q, have been determined, the construction (or training) of the PDM model is complete. Observe that implicit in the $\phi_M(X)$ is the fact that they were constructed for a particular realization of the training set. A different partition of the discriminant set into training and testing sets would yield somewhat different $\phi_M(X)$ functions.

H. Making a Prediction. Suppose a new predictor realization X' occurs for predictor K; i.e. X' = X(J,K) for some time J. We wish to use the PDM model constructed above in order to make a predictand forecast for the new predictor value X'. Various strategies can be adopted regarding the manner in which the pdf's $\phi_M(X)$ are employed in making a forecast.

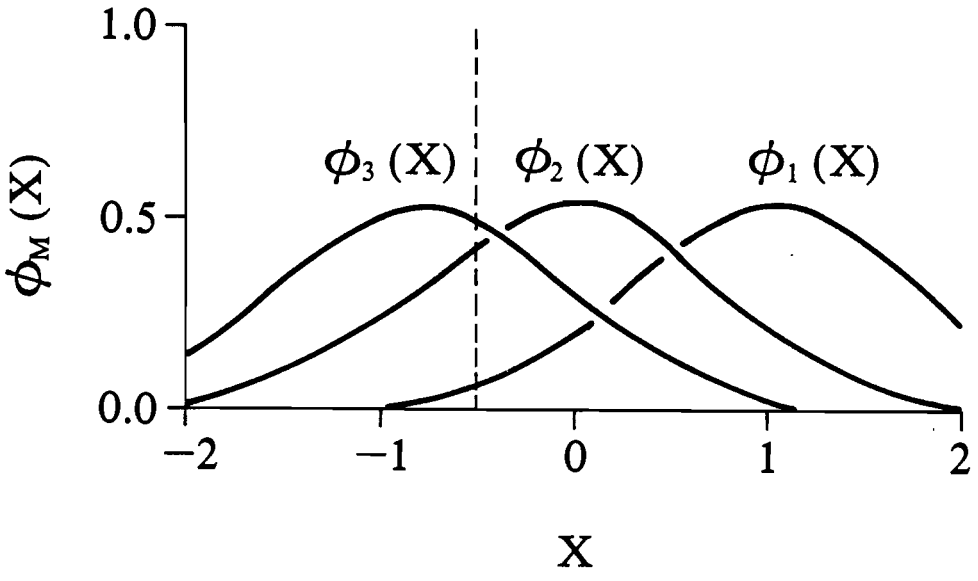

Figure 4.--The pdf's $\phi_1(X)$, $\phi_2(X)$ and $\phi_3(X)$ for the category subsets of Fig. 3a.

i. Maximum Probability Strategy. Given a predictor value X', we compute $\phi_M(X')$ for each category M = 1,...,Q and note which M value, call it M', has the *maximum pdf value*. The prediction is then that NY(J) = M'. In Fig. 4 we see, for example, that X' = -0.5 would yield a prediction of NY in category 3, X' = 0.0 would predict NY = 2, and so on.

ii. Bayesian Strategy. The maximum probability strategy is easily interpreted and computationally simple; however, it may not make the best use of the available information. The method of Bayesian inference is perhaps better suited to the problem at hand.

Strictly speaking, the $\phi_M(X)$ pdf's relate to *conditional* probabilities: namely, $\phi_M(X)$ gives the pdf of X *given* that category M is observed. To fix this idea, let us write $\phi(X|M) = \phi_M(X)$. What we really need in order to make a forecast is the probability that category M occurs *given* that a specific value of X occurs; let us denote this by P(M|X). The category, call it M', with the greatest probability P(M|X) for the given value of X = X' is then the category forecasted by the PDM when X' is observed. Since the Q predictand categories are mutually exclusive and exhaustive, Bayes' theorem (see, for example, Box and Tiao, 1972)

$$P(M|X) = \frac{P(X|M)P(M)}{\sum_{M=1}^{Q} P(X|M)P(M)} = \frac{\phi(X|M)P(M)}{\sum_{M=1}^{Q} \phi(X|M)P(M)}$$

can be used to obtain the desired P(M|X) values. Here P(X|M) is the *probability* of X given M, which is just $\phi(X|M)$. (Note that $\sum_{M=1}^{Q} P(X|M) < 1$ in general.) P(M), known as the *a priori* probability of category M occurring, lies at the heart of Bayesian inference. P(M) is a measure of our knowledge about what forecast category will occur, *before* the predictor value X is

obtained. The selection of appropriate P(M) values is a task which falls on the user of the Bayesian strategy and is an extra computation above those required for the maximum probability strategy. If we were making a *random* forecast of NY(J), it would be reasonable (but not necessary) to make the probability of randomly choosing category M proportional to the number of points of the training set which fall in category M. So a reasonable choice of P(M) is

$$P(M) \equiv \frac{NCAT_M}{NTRN} \quad , \quad M = 1,\ldots,Q \ .$$

It should be understood that in making this choice of P(M) we are allowing information about the relative distribution of points in the category subsets to influence the PDM's forecast of the predictand when given a new predictor value X'. This is the whole point of the Bayesian strategy. Another choice of P(M) could lead to an entirely different forecast being made for the same X' value. If we wish to make no use of our knowledge about the distribution of points in the category subsets, we can pick P(M) = 1/Q for all M. This is the case of equally likely *a priori* distributions, for which the Bayesian strategy reduces to the maximum probability strategy.

An example of the difference in the Bayesian and maximum probability strategies can be obtained from Figs. 3 and 4. From the category subsets of the training set of Fig. 3a and the above choice for P(M) we get

$$P(1) = \frac{3}{14} \quad , \quad P(2) = \frac{6}{14} \quad , \quad P(3) = \frac{5}{14} \ .$$

From the pdf's of Fig. 4 evaluated at X' = -0.5 we get the $P(X'|M)$ values

$$P(X'|1) = .06 \quad , \quad P(X'|2) = .42 \quad , \quad P(X'|3) = .48.$$

Then from Bayes' theorem we get

$$P(1|X'=-0.5) = \frac{(.06)\frac{3}{14}}{(.06)\frac{3}{14} + (.42)\frac{6}{14} + (.48)\frac{5}{14}} = 0.0353$$

$$P(2|X'=-0.5) = 0.4941$$

$$P(3|X'=-0.5) = 0.4706.$$

We see that $P(M|X'=-0.5)$ is largest for M = 2, and therefore the forecast is NY(J) = 2. But recall that the maximum probability strategy gave NY(J) = 3 for this X value! Clearly the use of the additional information contained in the P(M) has had a profound influence on the forecast. For a further discussion of philosophical matters relating the choice of the *a priori* distributions see, for example, the work of Box and Tiao (1972).

One can devise other forecast strategies than the two discussed here. Which is best can be determined, if at all, only by trial and error evaluation of the competing strategies. It is quite possible that for one type of problem (e.g. forecasting in a noisy environment), one forecast strategy will prove superior, whereas for another type of problem (e.g. data classification in a noise-free environment), another strategy will yield more accurate forecasts.

I. Potential Predictability. The PDM as it now stands is ready to make predictions by whichever strategy is chosen in the previous paragraph. However, it is of great interest also to compute some measure of confidence in these predictions. When the pdf's $\phi_M(X)$ are not well separated, then the predictions have low skill, no matter what prediction strategy we choose.

Note, for example, in Fig. 4 that for predictor values X' near 0.5 it is nearly equally probable that the predictand is in category 1 or 2, if we use the maximum probability strategy. Conversely, if the $\phi_M(X)$ are well separated, then the PDM has no difficulty in determining which pdf has the maximum value for a given X', and we have greater confidence that the predictions will be correct. Therefore a measure of our confidence in the predictions can be obtained via a measure of how well separated are the pdf's. One measure of this separation is given by the *potential predictability* index, PP, now to be defined.

Let

$$S(I) \equiv \sum_{M=1}^{Q} \phi_M(TRNX(I,K))$$

for I = 1,...,NTRN and K held fixed. Then define

$$P'(I,M) \equiv \frac{\phi_M(TRNX(I,K))}{S(I)}$$

for M = 1,...,Q and I = 1,...,NTRN. Note that $\sum_{M=1}^{Q} P'(I,M) = 1$. If the pdf's are identical, P'(I,M) = 1/Q. Thus a measure of how far the pdf's are from being identical is $\sum_{M=1}^{Q} [P'(I,M) - \frac{1}{Q}]^2$. Moreover, if the pdf's are perfectly separated, then

$$\sum_{M=1}^{Q} [P'(I,M) - \tfrac{1}{Q}]^2 = \underbrace{(1 - \tfrac{1}{Q})^2}_{\substack{\text{one such} \\ \text{term when} \\ P(I,M) = 1}} + \underbrace{(0 - \tfrac{1}{Q})^2 + \cdots + (0 - \tfrac{1}{Q})^2}_{\substack{Q-1 \text{ such terms} \\ \text{when } P(I,M) = 0}}$$

or

$$\sum_{M=1}^{Q} [P'(I,M) - \frac{1}{Q}]^2 = \frac{Q-1}{Q} .$$

Thus we are led to define

$$PP(I) \equiv \frac{Q}{Q-1} \sum_{M=1}^{Q} [P'(I,M) - \frac{1}{Q}]^2 .$$

Clearly PP(I) = 1 if the pdf's are perfectly separated and PP(M) = 0 if the pdf's are identical. Finally we define the potential predictability, PP, as

$$PP \equiv \frac{1}{NTRN} \sum_{I=1}^{NTRN} PP(I) .$$

Thus PP has the property $0 \leq PP \leq 1$ and is a measure of how distinct the pdf's are: PP approaches zero as the pdf's become identical (and our confidence in a prediction decreases), and PP approaches 1 as the pdf's become widely separated (and our confidence in a prediction increases). This definition for PP is consistent with the choice of the maximum probability strategy for making a forecast, as discussed in §H.i above. If the Bayesian strategy of §H.ii is chosen, the definition must be modified slightly by using

$$P'(I,M) \equiv P(M|X = TRNX(I,K)).$$

$$= \frac{\phi_M(TRNX(I,K))P(M)}{\sum_{M=1}^{Q} \phi_M(TRNX(I,K))P(M)} ,$$

which reduces to the previous definition of P'(I,M) if the *a priori* distributions P(M) are chosen to be equally likely, i.e., P(M) = 1/Q.

PP is implicitly indexed by K for the particular predictor X(J,K) in question. Moreover, PP depends on the particular partition of the discriminant set into training and testing sets. Thus one should make several (say NR) random partitions of the discriminant set, and compute PP for each. Then in the final tally, the average PP over all partitions should be taken:

$$\text{AVGPP}(K) = \frac{1}{NR} \sum_{JR=1}^{NR} \text{PP}(K,JR) \ ,$$

where we now explicitly show the predictor (K) and partition (JR) indices. When comparing two possible predictors for a given predictand, the one with the higher AVGPP will represent the higher predictability, on average.

J. Monte Carlo Significance Test for PP. However, while one predictor may have a higher potential predictability than another, for a given predictand, it is possible that neither is significant in the statistical sense.

Recall Fig. 1. The situation there indicates a correlational (and perhaps a causal) connection between X(J,K) and NY(J). A random relation between predictor and predictand would occur if, for example, the category (1, 2 or 3 in Fig. 1) at time J were assigned to NY(J) in a random way. Thus, for the Monte Carlo tests to be devised here, let a random number generator choose, at each time J, a class M and define a new array NRANY(J) = M, J = 1,$\cdots$,NT. NRANY is thus a random version of NY. The probability of randomly assigning a particular M value to NRANY(J) should be made proportional to the relative frequency of occurrence of the Mth category in the Q-tiling of the original data set, so that the Monte Carlo test will simulate as closely as possible the real experiment.

We can now use the given predictor set X(J,K) and the newly defined random predictand NRANY(J) to produce training and testing sets (as in paragraph E), and carry through all the subsequent steps to obtain a value of PP. This entire process can then be repeated after generating a new realization of the random predictand NRANY, to obtain another value of PP for a random relation between predictor and predictand. This process can be repeated to generate, say, 100 values of PP for random predictor-predictand connections. These 100 values can be ordered from smallest to largest; call them PP(1) for the smallest to PP(100) for the largest. The 5% critical value for PP is then determined from the 96th smallest PP value, PP(96). Thus the probability that a *randomly produced* PP value will equal or exceed PP(96) is approximately 0.05. Therefore, if the PP value determined for the *actual* predictor-predictand pair satisfies

$$PP \geq PP(96),$$

we will say that PP is significant at the 5% level.

If one wants to establish a critical value for AVGPP(K), then the Monte Carlo simulation is conducted so as to mimic the generation of AVGPP(K), as described in paragraph I. Thus one randomly produces NR realizations of PP(K,JR), finds their average, and goes through this average-finding procedure 100 times in all. The 96th smallest randomly generated AVGPP value then gives the 5% initial value for AVGPP.

We note also that there are other measures of separation of the category swarms. For example, Hotellings T^2 test (the multivariate generalization of Student's t test) can be used to test for significant separation of a pair of category means $\overline{X}_M$. However, such tests often depend on assumptions of

normality or independence of events. The potential predictability measure of separation was developed in an attempt to have a nonparametric test.

K. Class Errors. The potential predictability gives us one measure of how well a particular predictor can be expected to forecast predictand values. Another straightforward indicator of how well a prediction method is doing, when predicting categories, is to count the number of predictions that are correct (0-class errors) and the number of predictions that are off by one category (1-class errors). In the PDM we shall do this two ways: we will determine the *potential* 0- and 1-class errors, PA0 and PA1 respectively, using the *training* set; and we will determine the *actual* 0- and 1-class errors, A0 and A1, using the *testing* set.

i. PA0 and PA1. Recall the probabilities P'(I,M) which were defined when developing the PP index (using either the maximum probability or Bayesian strategies). For each I value, find the maximum of the Q probabilities, {P'(I,M): M = 1,...,Q}, and let M'(I) be the M value for which P(I,M) is a maximum. We now define the *potential 0-class error* as

$$PA0 \equiv \frac{1}{NTRN} \sum_{I=1}^{NTRN} P'(I,M'(I)) \ .$$

Note that as the pdf's $\phi_M(X)$ become well separated, P'(I,M'(I)), and consequently PA0, approach one. As the pdf's become identical, P'(I,M'(I)) and PA0 approach the value 1/Q. PA0 is therefore another measure, based on the pdf's $\phi_M(X)$, of how confidently we can expect the PDM to make a correct category forecast.

But even if the PDM makes an incorrect forecast, it is clearly better to have a forecast that misses by only one category than to have a forecast that misses by two or more categories. For example, if category 1 is observed, a

forecast of category 2 is closer to the truth than is a forecast of category 3. Thus it is useful to have a measure of how likely it is that the PDM will err by only one category, if it indeed makes an incorrect forecast. Toward this end, we define

$$\begin{aligned} AP(I,1) &\equiv 0 \\ AP(I,2) &\equiv P'(I,1) \\ AP(I,3) &\equiv P'(I,2) \\ &\vdots \\ AP(I,Q+1) &\equiv P'(I,Q) \\ AP(I,Q+2) &\equiv 0. \end{aligned}$$

The idea here is to have $P'(I,M'(I)-1) = 0$ if $M'(I) = 1$ and $P'(I,M'(I)+1) = 0$ if $M'(I) = Q$. Then define

$$PA1 \equiv \frac{1}{2}\,\frac{1}{NTRN}\sum_{I=1}^{NTRN}\left[AP(I,M'(I)) + AP(I,M'(I)+2)\right] .$$

A moment's reflection shows that PA1 is a measure of the probability that a category one less or one greater than the correct forecast category will be selected, if indeed the M'(I) value gives a false forecast. As the pdf's $\phi_M(X)$ become well separated, PA1 approaches 0; and as the pdf's become identical, PA1 approaches $\frac{1}{Q}$. Thus we have

$$0 \le PA1 \le \frac{1}{Q} \le PA0 \le 1 .$$

The *larger* is PA0, the better *may* X(J,K) predict NY(J), and the *smaller* is PA1, the better *may* X(J,K) predict NY(J).

ii. A0 and A1. After the PDM has been constructed, or trained, using the training set [TRNX(I,K), NTRNY(I)], we can apply the PDM to the testing set predictors, TSTX(I,K), and verify the predictions it makes against the actual observations for the testing set, NTSTY(I). Each time the PDM makes a correct forecast, we tally one to the 0-class error score, and each time the PDM forecast errs by one category we tally one to the 1-class error score. Then define

$$A0 \equiv \frac{1}{NTST} \text{ [number of 0-class errors]}$$

$$A1 \equiv \frac{1}{NTST} \text{ [number of 1-class errors] .}$$

A0 and A1 satisfy

$$0 \leq A0 \leq 1$$

$$0 \leq A1 \leq 1 \text{ .}$$

The *larger* A0, the better has the PDM forecasted the testing set values, and the *smaller* A1, the better has the PDM performed. Unlike PP, PA0 and PA1, which are based on the fitted pdf's defining the PDM model, A0 and A1 are actual forecast scores made by the PDM when applied to an *independent* testing set. Our studies of the PDM in the following Parts II and III will make use of the training and testing sets in the manner just discussed: the PDM will be defined using the training set, and its performance will then be evaluated using the testing set. The A0 and A1 scores are a convenient means of presenting forecast skill when discrete forecast categories are used. See, for example, Preisendorfer and Mobley (1984) for the use of A0 and A1 in scoring seasonal climate forecasts.

L. Significance Tests for Class Errors. The Monte Carlo procedure, used in paragraph J to determine the 5% critical value for potential predictability, is equally applicable to the determination of critical values for PA0, PA1, A0 and A1. For each of the 100 realizations of the random data set NRANY, we can compute PA0 and PA1 from the associated training set, and we can compute A0 and A1 scores from the associated testing set. We then determine the 5% *upper* critical levels PA0(96) and A0(96), and the 5% *lower* critical values PA1(05) and A1(05). Significantly good predictions will have PA0 and A0 scores that equal or *exceed* PA0(96) and A0(96), respectively. Significantly good predictions will have PA1 and A1 scores that equal or *are less than* PA1(05) and A1(05), respectively.

M. Ranking and Screening Single Predictors. The net result of this section is the ability to individually rank (for a given predictand Y(J)) the predictors X(J,K), K = 1,...,NK, in terms of their PP, PA0, PA1, A0 and A1 scores. Those predictors that have significant potential predictability and class-error scores become candidates for further consideration in the multiple predictor stage. Predictors that have non-significant scores as single-predictors of a predictand are unlikely to add useful information if they are combined with other predictors in the multiple-predictor stage, and therefore can be dropped from further consideration.

3. The Multiple-Predictor Stage

After performing the single-predictor analyses of the previous section on each predictor X(J,K), K = 1,...,NK, we have, for a fixed predictand Y(J), a set of predictors ordered by their potential predictability scores. We drop from further consideration any predictors which did not have statistically significant PP scores in the single-predictor stage, so that NP ≤ NK

predictors remain. We now turn our attention to the task of constructing a PDM model which has more than one predictor for a given predictand.

We choose the predictor with the highest potential predictability score as the first predictor to be included in the multiple-predictor PDM model. We then must screen the remaining NP-1 predictors in order to select those which, when combined with the first predictor, yield a multiple-predictor model which is, in some sense, optimum.

A. Correlational Screening of Predictors. Suppose we have already selected L-1 predictors, L = 2,...,NP-1. Let these selected predictors be X(J,KX), KX = 1,...,L-1. Let the remaining set of unselected predictors be denoted by W(J,KW), KW = 1,...,NW; NW+L-1 = NP. Let "Cor[KW,KX]" or "Cor[W(·,KW),X(·,KX)]" denote the correlation between the indicated predictors. The number

$$C(KW) \equiv \underset{KX = 1,\ldots,L-1}{\text{Max}}\{|\text{Cor}[KW,KX]|\}$$

is a measure of the distance between the KWth unselected predictor W(J,KW) and the set of L-1 previously selected predictors X(J,KX). The larger C(KW) is, the closer W(J,KW) is to {X(J,KX), KX = 1,...,L-1} as a whole.

When choosing a new candidate predictor for addition to the previously selected predictors, we choose that predictor W(J,KW) which has the *minimum* correlation magnitude, Cor[KW,KX]. In so doing, we are selecting that predictor which is least correlated with the existing predictors and therefore most likely to add *new information* to the model. If LW is the value of KW giving the minimum C(KW), then we set X(J,L) = W(J,LW), J = 1,...,NT. This correlational screening is the first step in choosing the Lth predictor. Whether or not this candidate predictor is retained in the PDM model will

depend on its effect on the PP, PA0 and PA1 scores, to be discussed in paragraph K below.

B. The L-dimensional Discriminant Set. Having added a candidate Lth predictor, we now have a set of L predictors which at each time J form a vector $\underline{X}(J) \equiv [X(J,1),X(J,2),\ldots,X(J,L)]$ in euclidean L-space E_L. As the time index J varies, $\underline{X}(J)$ moves about in E_L. The category-valued predictand NY(J) concurrently changes with J. The set of all ordered pairs $[\underline{X}(J),NY(J)]$, J = 1,...,NT, constitutes the L-dimensional discriminant set.

C. L-dimensional Training and Testing Sets. The L-dimensional discriminant set is randomly split into two parts, exactly as in §1.E. The result is a set of L-component vectors TRNX(I), I = 1,...,NTRN, containing those elements of $\underline{X}(J)$ randomly falling into the training set, and another set of vectors TSTX(I), I = 1,...,NTST, containing the remaining elements of $\underline{X}(J)$. The associated sets of predictands NTRNY(J) and NTSTY(J) are defined just as before.

D. Category Subsets of Predictor Space. We can now define subsets of E_L, the setting of the predictor space, that are associated with each of the Q predictand categories. The logic of this definition is the same as that of §1.F. Thus we set $\underline{XCAT}_M(I) = \underline{TRNX}(I)$ if NTRNY(I) = M; the number of points tallied to $\underline{XCAT}_M(I)$ is $NCAT_M$.

It is to the Q subsets of E_L, $\underline{XCAT}_M(I)$, M = 1,...,Q, that we will eventually fit L-dimensional probability density functions. However, before fitting the pdf's, we perform a preliminary analysis of the L-dimensional category subsets using Principal Component Analysis (PCA). *It is in this application of PCA that the PDM parts company with classical discriminant analysis.*

E. Binary PCA Decomposition of Category Subsets. Let us consider, for didactic purposes, the case of two predictors (L = 2) and a terciled predictand (Q = 3). Figure 5 shows three swarms of points in E_2, representing the three category subsets. These sets of points were artificially generated for the purpose of illustrating this section. In classical discriminant analysis, each category subset would be fitted with a bivariate normal pdf. For a point swarm shaped like that of category 2, the bivariate normal pdf would probably be quite satisfactory: Fig. 6 shows the category 2 swarm and the best-fit binormal pdf. However, for an irregularly shaped swarm, such as category 1 of Fig. 5, the bivariate normal pdf is clearly a poor representation of the actual shape of the category subset. Figure 7 shows the best-fit bivariate normal pdf for category 1. Since discriminant methods depend upon having pdf's which accurately delineate the category subsets, we could not expect accurate forecasts from a model based on fits as poor as that of Fig. 7. (See Fig. 7 of Miller, 1962, for an example of a binormal pdf being forced upon a category subset which is clearly bimodal in E_2).

Principal Component Analysis (PCA) enables us to systematically and objectively subdivide an arbitrarily shaped category swarm into a number of smaller point swarms in E_L. If each of the smaller swarms is then roughly elliptical in shape (in terms of hyperellipses in E_L), then a multinormal pdf can be well-fitted to each smaller swarm. The pdf describing the original, irregularly shaped category swarm can then be constructed as a weighted sum of the multinormal pdf's of the smaller swarms. *This is the central idea of the PDM.*

Let $\underset{\sim}{X}_M \equiv \{\underline{XCAT}_M(I),\ I = 1,\ldots,NCAT_M\}$, $M = 1,\ldots,Q$, represent the Mth category subset. $\underset{\sim}{X}_M$ is viewed as an $NCAT_M$ by L data matrix. A general outline of the $\underset{\sim}{X}_M$ category swarm for E_2 is illustrated in Fig. 8. PCA of the category swarm $\underset{\sim}{X}_M$ provides the following:

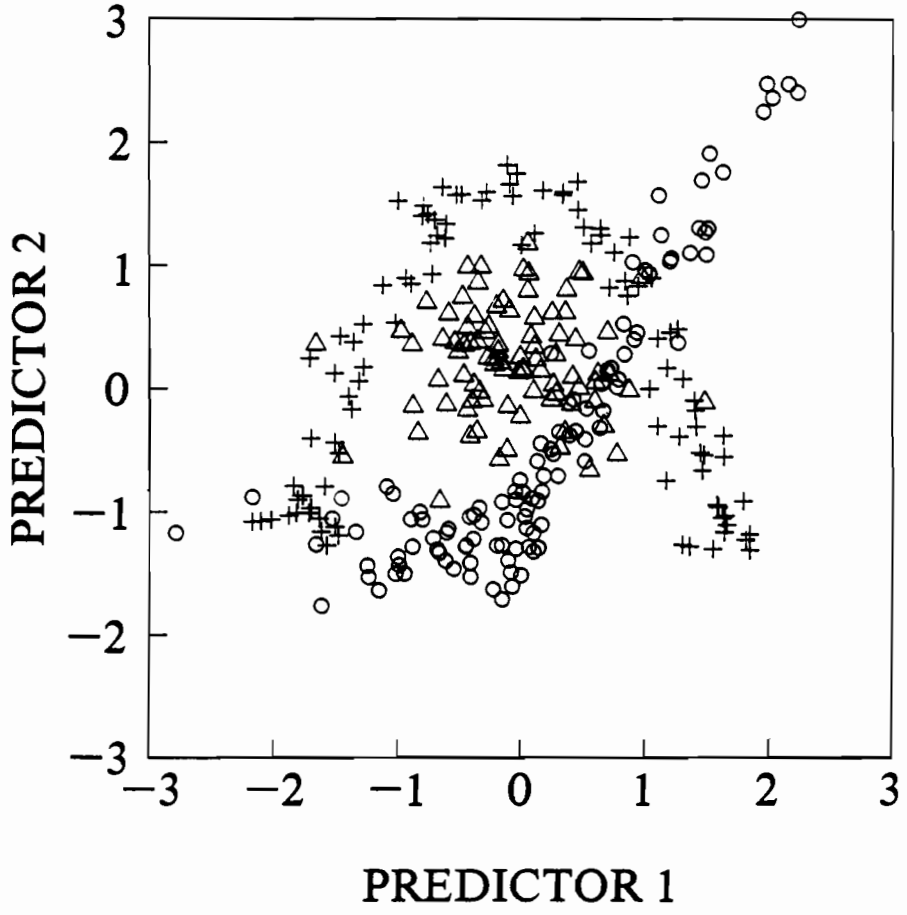

Figure 5.--An illustration of three category swarms $\underline{XCAT}_1$, ("+" symbols, $NCAT_1$ = 99 points), $\underline{XCAT}_2$ ("Δ" symbols, $NCAT_2$ = 89), and $\underline{XCAT}_3$ ("0" symbols; $NCAT_3$ = 112) in E_2.

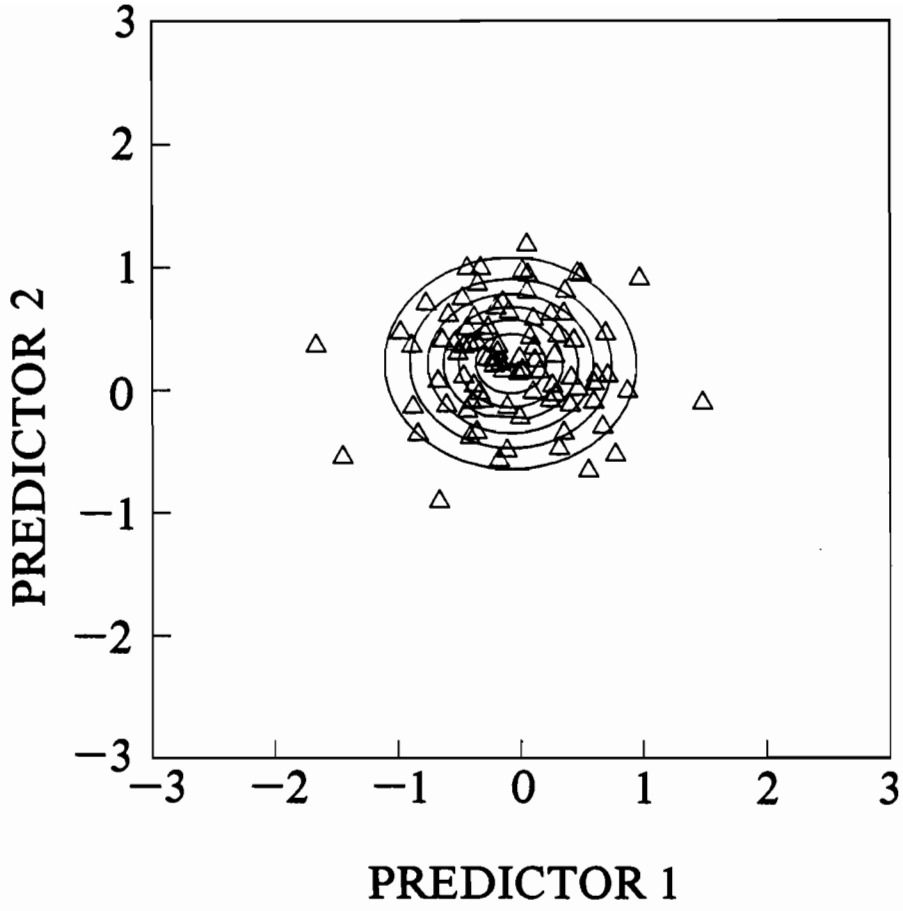

Figure 6.--The category 2 point swarm of Fig. 5 and the probability contours of the best-fit bivariate normal pdf.

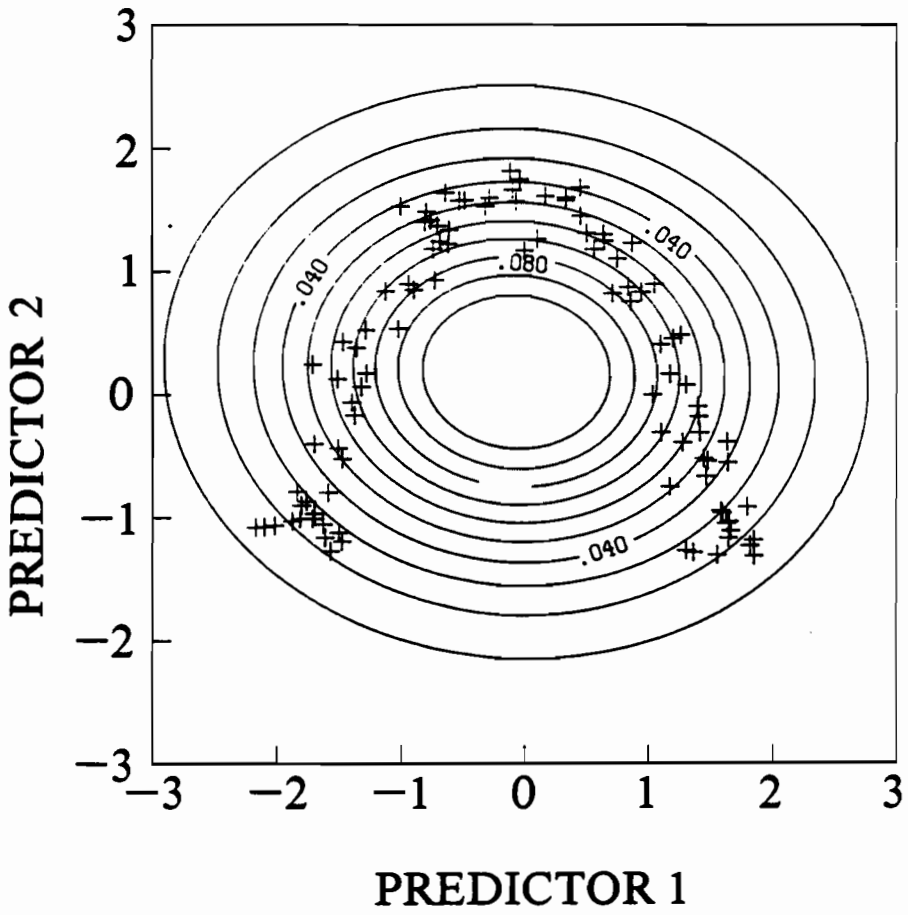

Figure 7.--The category 1 point swarm of Fig. 5 and the probability contours of the best-fit bivariate normal pdf.

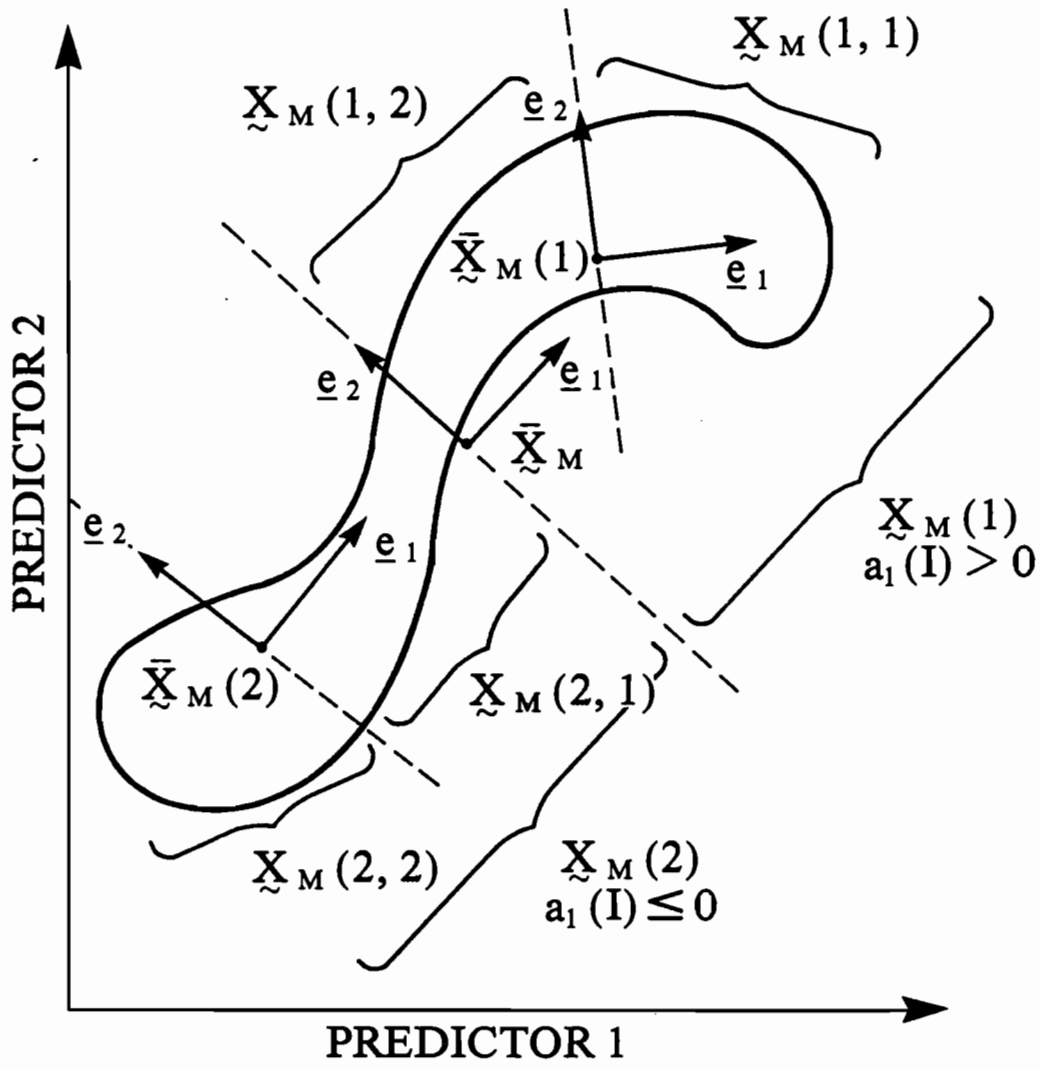

Figure 8.--A general category swarm $\underset{\sim}{X}_M$ in E_2, represented by the heavy line. The eigenvectors $\underline{e}_1$ and $\underline{e}_2$ resulting from the PCA of $\underset{\sim}{X}_M$ are shown at the centroid of the swarm $\bar{\underset{\sim}{X}}_M$. Level 1 subswarms $\underset{\sim}{X}_M(1)$ and $\underset{\sim}{X}_M(2)$ are identified, along with their centroids and eigenvectors. Four level 2 subswarms $\underset{\sim}{X}_M(\alpha_1,\alpha_2)$ are also labeled.

$\bar{\underset{\sim}{X}}_M$: the centroid in E_L of the point swarm $\underset{\sim}{X}_M$

$\underline{e}_1,\ldots,\underline{e}_L$: a set of L-dimensional orthonormal eigenvectors at $\bar{\underset{\sim}{X}}_M$

$\ell_1,\ldots,\ell_L$: a set of eigenvalues associated with the eigenvectors. ℓ_j is the scatter (= $(NCAT_M-1)$ times the variance) of the swarm in the $\underline{e}_j$ direction. The eigenvalues are ordered so that $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_L \geq 0$, with an associated ordering of the eigenvectors.

$\underline{a}_1$: the first principal component of $\underset{\sim}{X}_M$; $\underline{a}_1$ is an $NCAT_M$-dimensional vector defined by $\underline{a}_1 \equiv (\underset{\sim}{X}_M - \bar{\underset{\sim}{X}}_M)\underline{e}_1$.

The steps of PCA are reviewed in the Appendix.

The signs of the $NCAT_M$ elements of $\underline{a}_1$ divide the category swarm $\underset{\sim}{X}_M$ into two subswarms separated by a hyperplane in E_L which passes through the centroid $\bar{\underset{\sim}{X}}_M$ perpendicular to $\underline{e}_1$. Thus

if $a_1(I) > 0$, place the point $\underline{XCAT}_M(I)$ in subswarm 1, denoted by $\underset{\sim}{X}_M(1)$, and

if $a_1(I) \leq 0$, place the point $\underline{XCAT}_M(I)$ in subswarm 2, denoted by $\underset{\sim}{X}_M(2)$.

It is this elegant property of PCA that allows us to use it as a tool for dividing sinuous category swarms into smaller, and hopefully more symmetrical, point swarms. The subswarms $\underset{\sim}{X}_M(1)$ and $\underset{\sim}{X}_M(2)$ individually may be close to an elliptical shape, so that a multivariate normal pdf adequately fits each subswarm. If either of the subswarms is still too distorted in shape, it can be further subdivided by another application of PCA to that subswarm. This subdivision process can continue until the original category swarm has been reduced to a number of smaller swarms, each of which is roughly elliptical in shape in E_L.

These successive subdivisions are conveniently displayed as a tree. The general notation is that $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$, $\ell = 1,2,\ldots$, represents a "parent"

swarm, which is itself the result of ℓ binary subdivisions of the original category set $\underset{\sim}{X}_M$. If $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ is subdivided again by PCA, the subswarms are denoted by $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell,1)$ and $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell,2)$. The α-indices thus take on the value 1 or 2 at each binary decomposition. Figure 8 illustrates the division of category swarm $\underset{\sim}{X}_M$, with centroid at $\overline{\underset{\sim}{X}}_M$, into subswarms $\underset{\sim}{X}_M(1)$ and $\underset{\sim}{X}_M(2)$, with centroids at $\underset{\sim}{X}_M$ (1) and $\overline{\underset{\sim}{X}}_M(2)$, respectively. The figure also shows the further subdivision of $\underset{\sim}{X}_M(1)$ into $\underset{\sim}{X}_M(1,1)$ and $\underset{\sim}{X}_M(1,2)$, and of $\underset{\sim}{X}_M(2)$ into $\underset{\sim}{X}_M(2,1)$ and $\underset{\sim}{X}_M(2,2)$. Figure 9 shows the tree representation of the decompositions of Fig. 8.

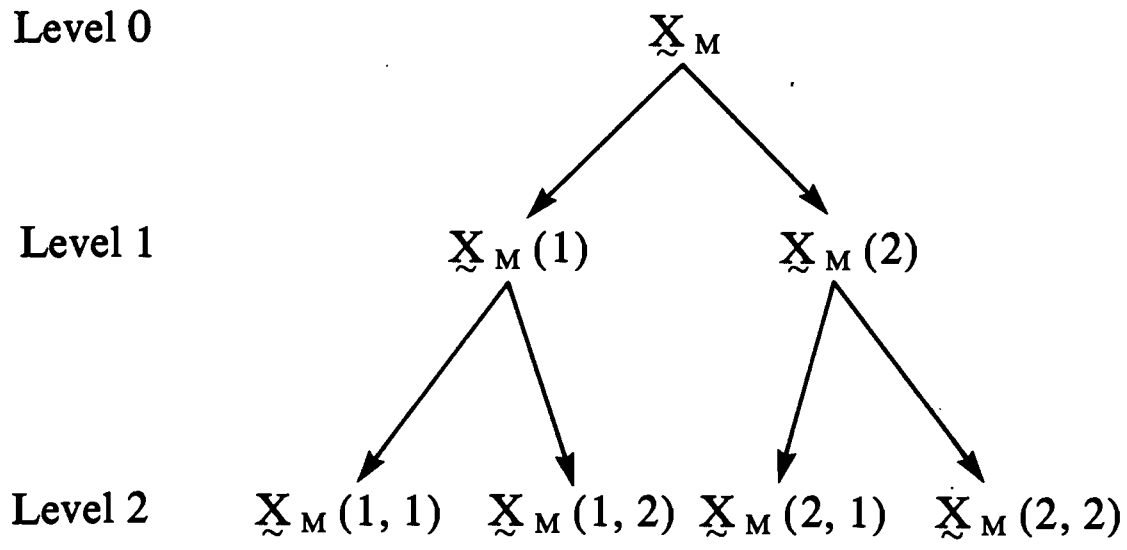

Figure 9.--A tree showing the level 2 subdivision of a category swarm $\underset{\sim}{X}_M$. This figure corresponds to the diagram of Fig. 8.

F. Termination of the PCA Decomposition Process. The question now arises: When is it necessary to subdivide a point swarm, or equivalently, how can we determine when a point swarm is sufficiently close to an elliptical shape? This problem of deciding whether or not to subdivide a point swarm is not a trivial one, and no really satisfactory solution has yet been found. However, several proposed solutions have been investigated, and we will discuss these briefly.

Let $N_M(\alpha_1,\cdots,\alpha_\ell)$ be the number of points in $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$. In order to perform PCA on $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$, $N_M(\alpha_1,\cdots,\alpha_\ell)$ must satisfy

$$N_M(\alpha_1,\cdots,\alpha_\ell) > L \ .$$

This requirement merely assures us that there are enough points to give a non-trivial PCA. If there are too few points in the swarm, we make $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ a *terminal node* of the PCA decomposition process; i.e., we declare the swarm $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ ready to be fit with a multivariate normal pdf. However, if sufficient points are available, we can proceed as follows.

i. Strategy 1. Perform a PCA on $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$. From the resulting ordered eigenvalues, $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_L \geq 0$, compute

$$\lambda \equiv \frac{\ell_1}{\ell_2 + \ell_3 + \cdots + \ell_L} \ .$$

One can now invoke a Monte Carlo strategy to see if λ is significantly large. In practice, however, such an approach is both computationally expensive and overly strenuous, in the sense that category swarms are subdivided just because they are non-*spherical*. Swarms can deviate greatly from a spherical shape and still be adequately fit by a multivariate normal

pdf; it is the sinuous shapes (cf. Fig. 8) and multimodal (or clustered) point swarms that need to be decomposed.

ii. Strategy 2. Perform a PCA on $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$. From the resulting ordered eigenvalues, compute

$$\lambda \equiv \frac{\lambda_1}{\frac{1}{(L-1)} \sum_{j=2}^{L} \ell_j} .$$

This λ is a measure of the largest eigenvalue relative to the average of the other eigenvalues, or in geometric terms, it is a measure of the scatter of the point swarm along the direction of greatest variance relative to the average scatter in the other directions. If the swarm is spherical, $\lambda = 1$; λ increases as the swarm becomes elongated. We then test λ against some ad hoc value λ_0 ($\lambda_0 = 2$, say) to see if the swarm is too far from spherical to be acceptable. Thus if $\lambda \leq \lambda_0$, declare the swarm to be a terminal node; if $\lambda > \lambda_0$, subdivide $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$.

This criterion is easily applied, but is sometimes too lenient in forcing a further subdivision of $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$. Recall, for example, the "C-shaped" swarm of category 1 in Fig. 5. Clearly, this category swarm needs to be subdivided. However, when PCA is performed on this initial swarm, $\ell_1 = 1.63$ and $\ell_2 = 1.09$, since PCA detects comparable variances in any direction about the centroid. Thus $\lambda = 1.5$ and no subdivision is called for by the test $\lambda > \lambda_0 = 2$. For category 2, $\lambda = 1.4$ and the decision to make no further subdivision is reasonable. The elongated category 3 swarm has $\lambda = 14.6$, consistent with its clear need for further subdivision.

iii. Strategy 3. We can force subdivision of the category swarms to continue until one or both of the subswarms $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell,1)$ and $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell,2)$ of $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ fail to satisfy the requirements

$$N_M(\alpha_1,\cdots,\alpha_\ell,1) > L$$
$$N_M(\alpha_1,\cdots,\alpha_\ell,2) > L$$

on the minimum number of points necessary for performing further PCA. The parent swarm $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ can then be declared terminal. This procedure works quite well at reducing the original swarm to a (sometimes large) number of subswarms, each of which can be fitted with a multivariate normal pdf. However, we must suspect that at some level of this lengthy decomposition, the resulting pdf's are being influenced more by the noise in the original point swarm than by the signal. That is to say, referring to the "C-shaped" category 1 swarm of Fig. 5, we want a final pdf which describes the overall "C-shape," but which is not unduly influenced by the exact positions of the individual points of the swarm. Having too fine a resolution of the category swarms may in fact degrade the class error scores A0 and A1 of the PDM model when it is applied to independent data, even though the PP, PA0 and PA1 scores have all been improved by the finer resolution of the testing set. (This phenomenon will be seen in part III, below.)

One simple way to terminate the PCA subdivision process is to simply force all initial category swarms $\underset{\sim}{X}_M$ to undergo a fixed number of subdivisions, say to level 2, as shown in Fig. 9. This procedure seems to work fairly well in practice. If a category swarm $\underset{\sim}{X}_M$ is nearly spherical to begin, as is category 2 of Fig. 5, little harm is done in decomposing it into, say, the four subswarms of a level 2 decomposition. If $\underset{\sim}{X}_M$ is sinuous, as are categories 1 and 3 of Fig. 5, then a 2-level decomposition goes a long way toward generating a reasonable resolution of the original swarm, but without getting too near the noise level.

G. Fitting pdf's to the Terminal Nodes. Let us suppose that the Mth category subset $\underset{\sim}{X}_M$ has been decomposed into a number of terminal nodes $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$. Let $\underset{\sim}{T}_M(t)$ denote the t^{th} terminal node $\underset{\sim}{X}_M(\alpha_1,\cdots,\alpha_\ell)$ of $\underset{\sim}{X}_M$, and let NT_M be the number of terminal nodes of $\underset{\sim}{X}_M$; $t = 1,2,\ldots,NT_M$. Thus $NT_M = 1$ for the case of no decomposition of the original category subset, $NT_M = 4$ for a level 2 decomposition like that of Figs. 8 and 9, and so on. Let $N_M(t)$ denote the number of points $N_M(\alpha_1,\cdots,\alpha_\ell)$ in the t^{th} terminal node; $\sum_{t=1}^{NT_M} N_M(t) = NCAT_M$. The centroid of $\underset{\sim}{T}_M(t)$ is located at $\overline{\underset{\sim}{T}}_M(t)$. Finally, let $\underset{\sim}{C}_M(t)$ be the L×L covariance matrix of $\underset{\sim}{T}_M(t)$ (cf. the Appendix), with determinant $DETC_M(t)$ and inverse $\underset{\sim}{C}_M^{-1}(t)$.

The best-fit multivariate normal pdf for the t^{th} terminal node $\underset{\sim}{T}_M(t)$ is then

$$\phi_M(t,\underline{X}) = (2\pi)^{-L/2}(DETC_M(t))^{-\frac{1}{2}} \times$$
$$EXP\{-0.5[\underline{X} - \overline{\underset{\sim}{T}}_M(t)]^T \underset{\sim}{C}_M^{-1}(t) [\underline{X} - \overline{\underset{\sim}{T}}_M(t)]\} .$$

(It is assumed that $DETC_M(t) \neq 0$, so that $\underset{\sim}{C}_M^{-1}(t)$ exists; if this is not the case, the PCA decomposition leading to this terminal node is not made, and the parent swarm is declared terminal. Alternatively, the swarm can be discarded.) $\underline{X}$ is an arbitrary point in E_L. As noted in Appendix A, $\underset{\sim}{C}_M^{-1}(t)$ is readily obtained from the eigenvalues and eigenvectors obtained in the PCA of $\underset{\sim}{T}_M(t)$, viz.

$$\underset{\sim}{C}_M^{-1}(t) = (NCAT_M - 1) \sum_{j=1}^{L} \frac{1}{\ell_j} \underline{e}_j \underline{e}_j^T .$$

H. Assembling the pdf's. A multivariate normal pdf is fitted to each terminal node $\underset{\sim}{T}_M(t)$ of $N_M(t)$ points, $t = 1,\ldots,NT_M$. We define a weighting

function $W_M(t) = N_M(t)/NCAT_M$, so that $\sum_{t=1}^{NT_M} W_M(t) = 1$. The probability distribution function for the Mth category subset is then taken to be

$$\Phi_M(\underline{X}) = \sum_{t=1}^{NT_M} W_M(t)\ \phi_M(t,\underline{X})$$

for M = 1,···,Q, $\underline{X}$ in E_L. *These pdf's $\Phi_M(\underline{X})$ define the desired PDM model.*

Figure 7 showed the binormal pdf for the category 1 point swarm of Fig. 5; this is the case of $NT_M = 1$, or no PCA decomposition of the category set. Figure 10 shows the contours of $\Phi_1(\underline{X})$ when determined by a level 2 decomposition, as illustrated in Figs. 8 and 9 and discussed in the latter part of §3.F above. This pdf is clearly a much more realistic description of the category 1 swarm than is the pdf of Fig. 7. If the PCA decomposition is allowed to proceed until just before the minimum point requirement $N_M(t) > L$ is violated, the category 1 point swarm of Fig. 5 is reduced to 23 terminal nodes. Figure 11 shows the tree diagram of this maximum possible decomposition. Figure 12 shows the $\Phi_1(\underline{X})$ contours determined from the terminal nodes of Fig. 11. This pdf gives a very sharp delineation of the category subset, but the finestructure of the probability contours is clearly being determined by the individual points of the category subset, which may be undesirable, as discussed in §3.F.

I. Making a Prediction. Just as in the single predictor case, we must choose a prediction strategy (maximum probability, Bayesian, or another) for using the pdf's $\Phi_M(\underline{X})$ to make a prediction. If the maximum probability strategy is chosen, then given a new predictor realization $\underline{X}'$ (now an L-dimensional vector), we evaluate $\Phi_M(\underline{X}')$, M = 1,...,Q. The prediction is then that the predictand falls into category M', where M' is the M value corresponding to the maximum $\Phi_M(\underline{X}')$, M = 1,...,Q. If the Bayesian strategy is

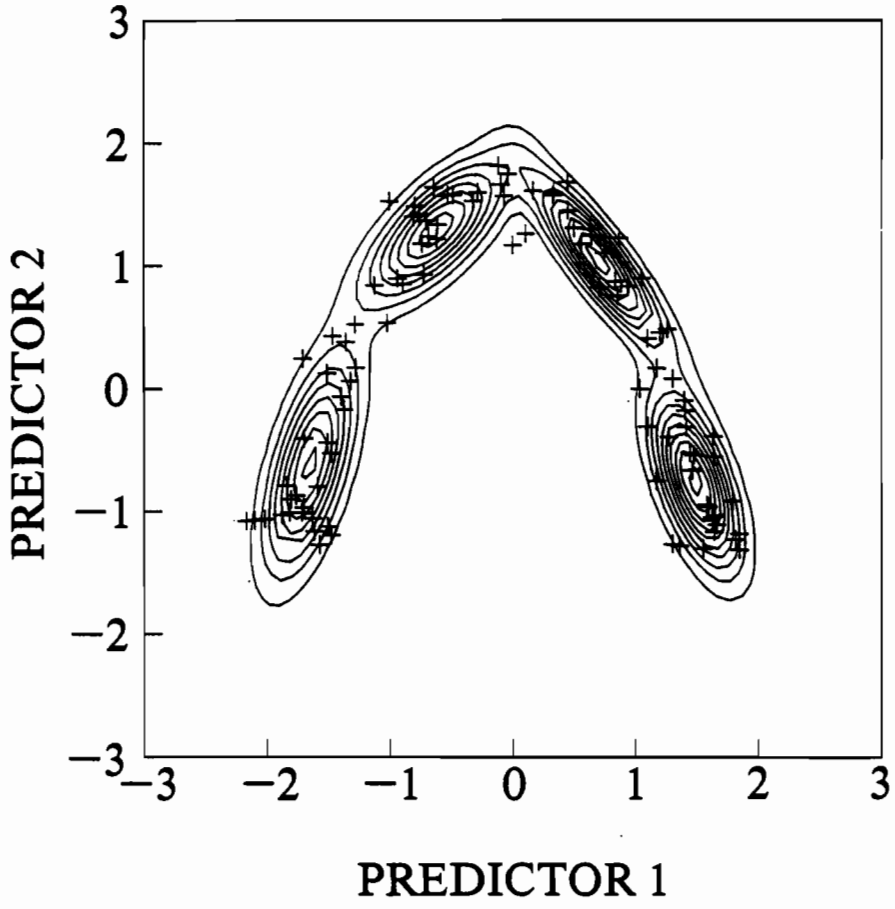

Figure 10.--The category 1 point swarm of Fig. 5 and the probability contours of $\Phi_1(\underline{X})$ as determined by a level 2 PCA decomposition.

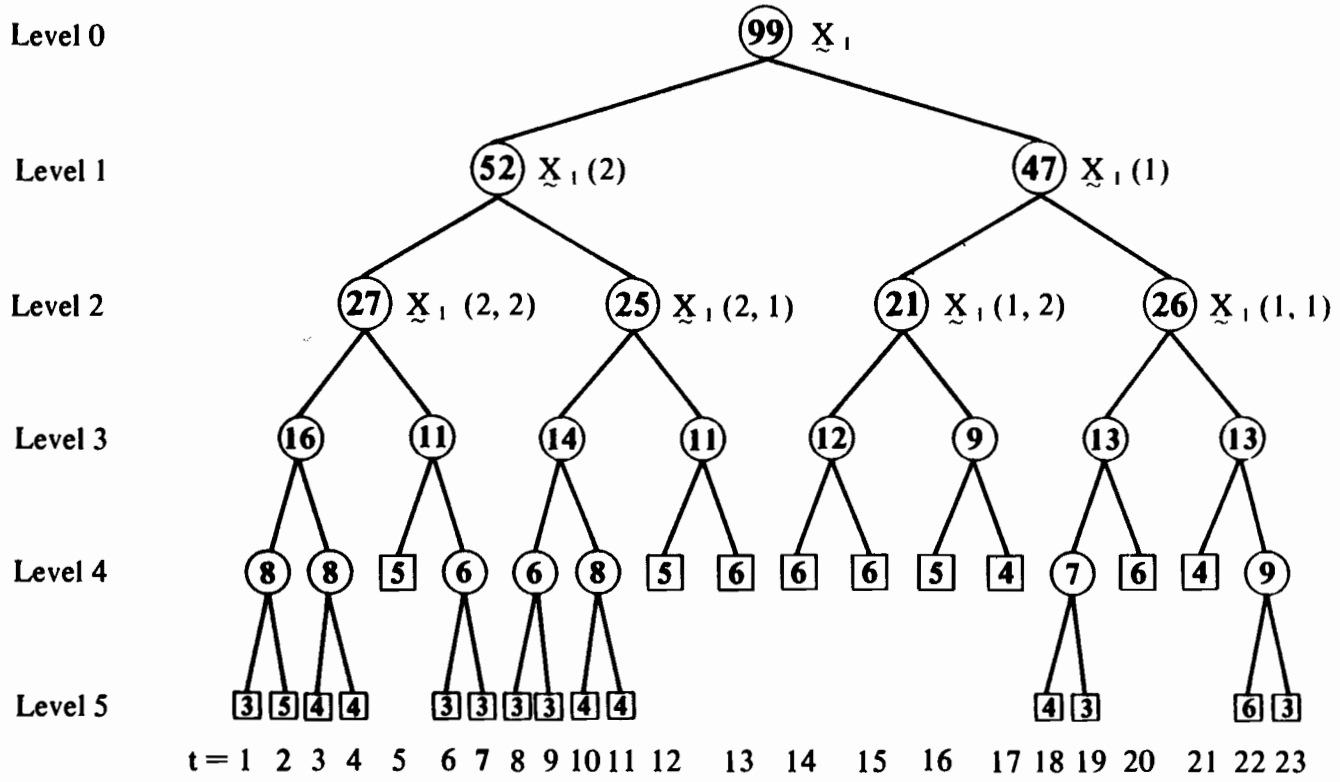

Figure 11.--The tree diagram showing the maximum possible decomposition of the category 1 subset of Fig. 5. The circles represent the $\underset{\sim}{X}_1(\alpha_1,\cdots,\alpha_\ell)$ subsets, and the numbers within the circles give the number of points in the subswarm, $N_1(\alpha_1,\cdots,\alpha_\ell)$. Terminal nodes $\underset{\sim}{T}_1(t)$ are represented by boxes; the enclosed numbers give $N_1(t)$.

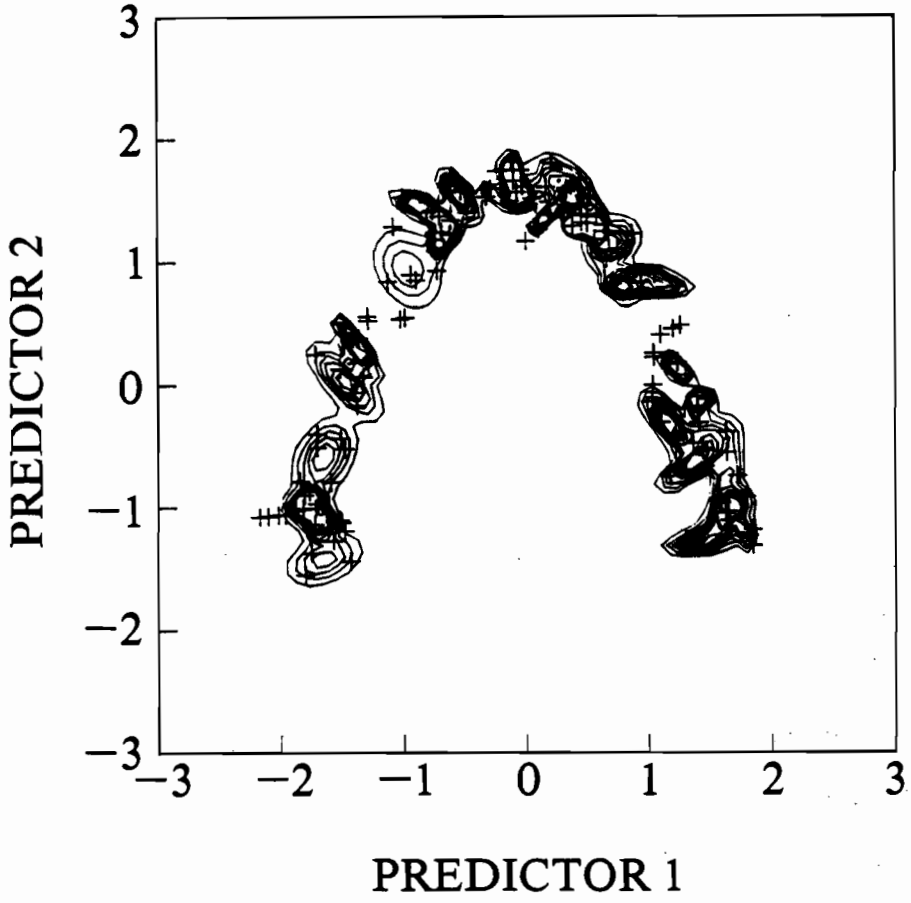

Figure 12.--The category 1 points of Fig. 5 and the $\Phi_1(\underline{X})$ probability contours as constructed from the 23 terminal nodes of Fig. 11.

chosen, the *a priori* probabilities can be set to $P(M) = NCAT_M/NTRN$, as in the single-predictor case, and the pdf's $\Phi_M(\underline{X}) \equiv \Phi(\underline{X}|M)$ are used in Bayes' formula.

J. Potential Predictability, Class Errors, and Significance Tests. These matters all proceed in exact analogy to the single predictor case. Thus in computing the potential predictability index, PP, for the maximum probability strategy we first compute

$$S(I) = \sum_{M=1}^{Q} \Phi_M(\underline{TRNX}(I))$$

and

$$P'(I,M) = \frac{1}{S(I)} \Phi_M(\underline{TRNX}(I))$$

for M = 1,...,Q and I = 1,...,NTRN. The only difference from the single-predictor case is that we are now using the L-dimensional training set values $\underline{TRNX}(I)$ in the multivariate pdf's $\Phi_M(\underline{X})$. Subsequent formulas leading to PP or AVGPP are unchanged. Likewise, the modifications required for the Bayesian strategy are trivial.

The potential predictability is now measuring the separation of pdf's in an L-dimensional space. Figures 13-15 show three sets of pdf's as determined for the example point swarms of Fig. 5, where L = 2. Figure 13 (reproducing parts of Figs. 6 and 7) shows in superposition contours of equal probability of the three best-fit binormal pdf's, $\phi_M(\underline{X})$, as would be obtained in classical discriminant theory. The potential predictability for these pdf's is PP = 0.39, when using the maximum probability forecast strategy. Figure 14 shows the pdf's, $\Phi_M(\underline{X})$, as obtained by level 2 PCA, as illustrated in Figs. 8-10. The eye can now easily distinguish the three pdf's determined from the three point swarms of Fig. 5, and the potential predictability has

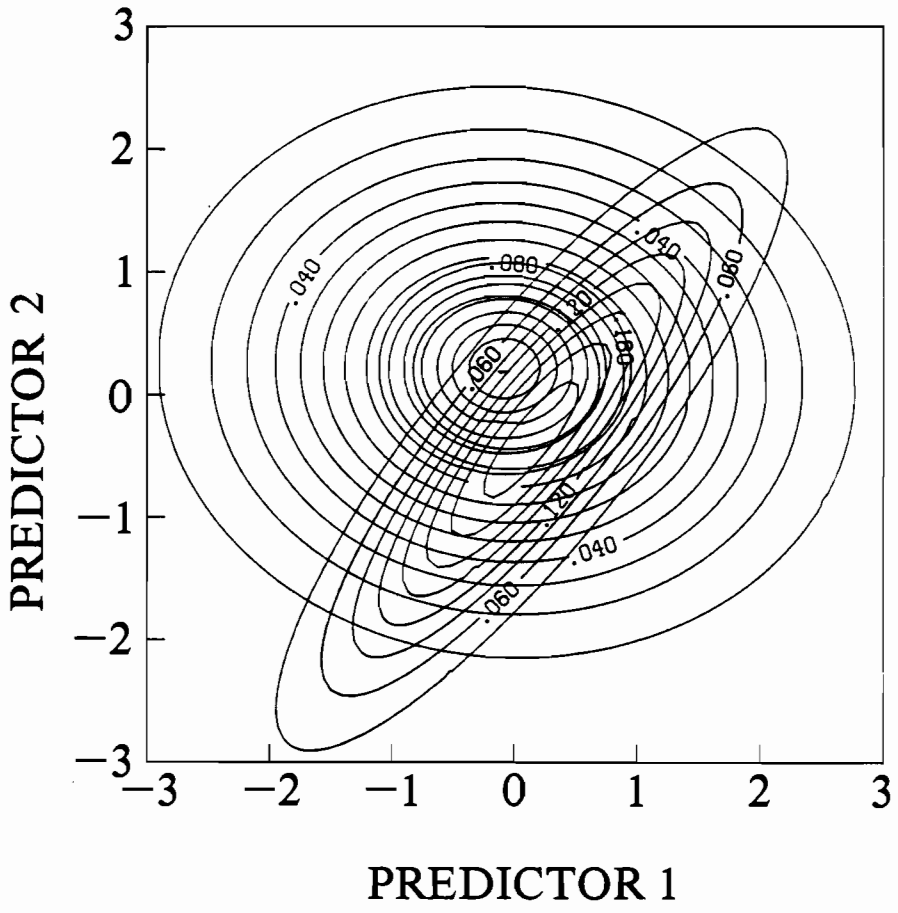

Figure 13.--Contours of equal probability of the three binormal pdf's $\Phi_M(\underline{X})$, M = 1,2,3, fitting the three category subsets of Fig. 5. The contour interval is different for each of the three pdf's.

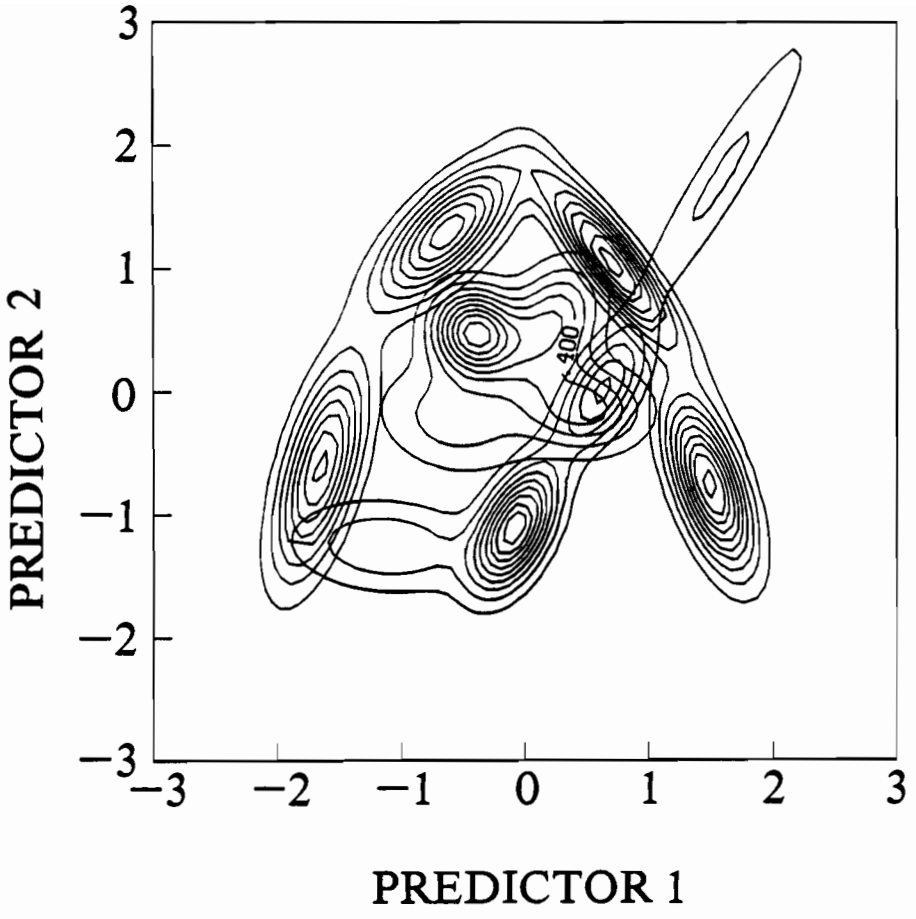

Figure 14.--Contours of equal probability of the three pdf's $\Phi_M(\underline{X})$ as determined from a level 2 PCA decomposition of each category subset of Fig. 5. Contour intervals vary.

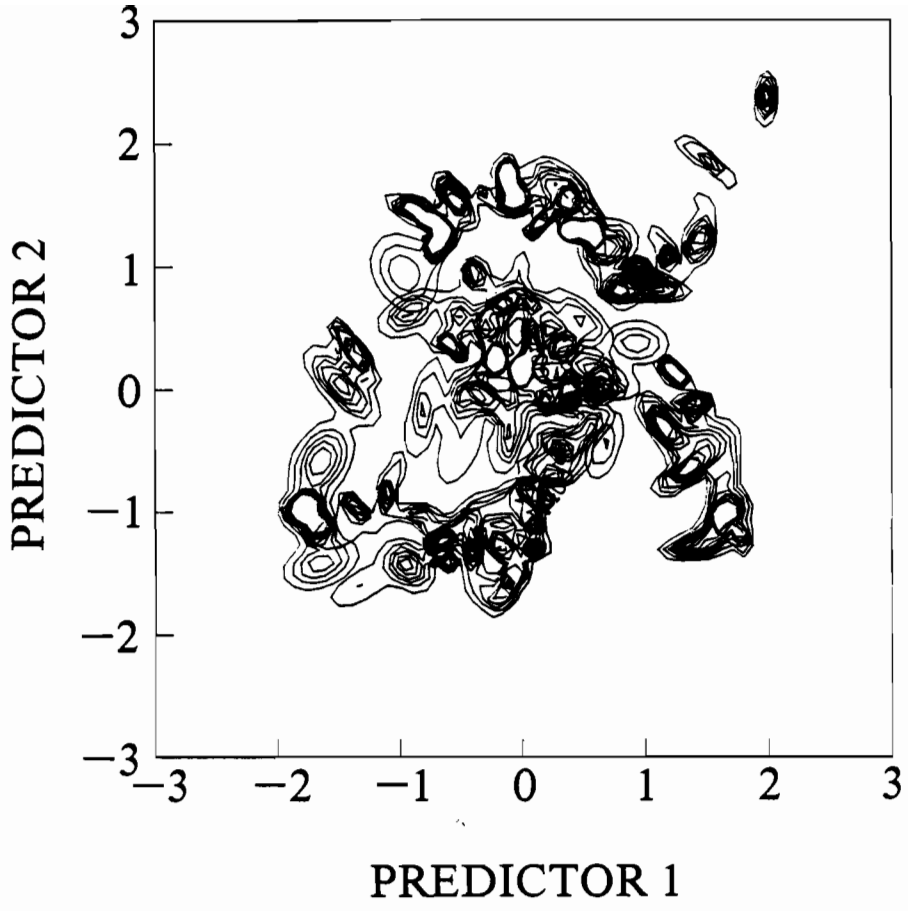

Figure 15.--Contours of equal probability of the three pdf's $\Phi_M(\underline{X})$ as determined from the maximum possible PCA decomposition of the category subsets of Fig. 5. Contour intervals vary.

risen to PP = 0.77. Figure 15 shows the pdf's as determined from the maximum possible PCA decomposition of the category swarms, as shown in Figs. 11 and 12. These pdf's show even better separation, as verified by their PP value of PP = 0.87, but the noise in the data (i.e., the positions of the individual points) has clearly affected the pdf's themselves. If the pdf's of Fig. 15 were used for actual forecasting, it might often occur that predictor values X' would "fall into the gaps" of these irregularly shaped pdf's in such a manner as to cause the point to be ascribed to the wrong pdf, thus giving an incorrect forecast.

Given the probabilities P'(I,M), the potential class errors PA0 and PA1 are immediately available. The actual class errors A0 and A1 are now computed from the multipredictor testing set TSTX(I), I = 1,...,NTST.

Monte Carlo experiments for determining 5% significance levels on PP, PA0, PA1, A0 and A1 proceed, in principal, as before. Now, however, when the randomly generated predictand NRANY(I) is analyzed using the multivariate predictors, it is necessary to perform a full PCA decomposition in order to get the needed pdf's (as described in paragraphs E-H above). This PCA analysis becomes prohibitively expensive when it must be repeated 100 times in a Monte Carlo experiment. Thus, in practice, the 5% significance levels may not be available.

K. Final Screening of the Candidate Predictor. We recall from paragraph A that we have admitted a candidate Lth predictor to the PDM model, based upon the correlation screening described there. We now may use the information gathered in the previous paragraph to decide whether or not to keep the candidate predictor in the model. Let PA0(L-1) and PA1(L-1) denote the PA0 and PA1 scores obtained from the PDM model before the candidate Lth predictor was admitted (if L = 2, we have the single predictor potential class errors

available). Let PP(L), PA0(L) and PA1(L) be the scores obtained after the candidate Lth predictor was admitted. Moreover, let PP(96;L), PA0(96;L) and PA1(05;L) be the appropriate 5% critical values as determined by Monte Carlo simulations. We then accept the candidate Lth predictor X(J,L) into the PDM model if the following conditions hold:

(i) $PP(L) \geq PP(96;L)$

(ii) $PA0(L) > PA0(L-1)$ and $PA1(L) \leq PA1(L-1)$

(iii) $PA0(L) \geq PA0(96;L)$ and $PA1(L) \leq PA1(05;L)$.

If these three conditions are not satisfied, we delete the candidate predictor from the model and return to paragraph A above to select the next candidate predictor. We continue in this manner until all possible predictors have been examined, at which time the PDM model is complete.

Condition (i) is simply the requirement that the model have a statistically significant potential predictability. Condition (ii) is the requirement that the addition of the Kth predictor improves the potential class error scores, and condition (iii) expresses the requirement that the model's potential class error scores be statistically significant. Conditions (i) and (iii) can be relaxed by using, say, a 10% significance level instead of the 5% level shown. Condition (ii) cannot be relaxed.

L. Scoring the PDM Model. Once the PDM model is complete, we can compute the actual class errors A0 and A1, using the testing set TSTX(I), I = 1,...,NTST, generated during the examination of the final predictor which was admitted to the model. These A0 and A1 scores, together with the information shown in (i), (ii), (iii) of the previous paragraph, are the data by which we measure the PDM model's actual and potential skills.

4. Appendix. PCA of the Point Swarm $\underset{\sim}{X}_M$.

Let the category swarm $\underset{\sim}{X}_M$ as defined in §3.E be regarded as an $NCAT_M$ by L matrix: the L columns of the matrix correspond to the L predictors of the PDM model; the rows of the matrix correspond to the points (times) of the training set. Thus

$$\underset{\sim}{X}_M = \begin{bmatrix} XCAT_M(1,1) & XCAT_M(1,2) & \cdots & XCAT_M(1,L) \\ XCAT_M(2,1) & XCAT_M(2,2) & \cdots & XCAT_M(2,L) \\ \vdots & & & \vdots \\ XCAT_M(NCAT_M,1) & \cdots & & XCAT_M(NCAT_M,L) \end{bmatrix}$$

The centroid of the swarm $\underset{\sim}{X}_M$ is located at

$$\overline{\underset{\sim}{X}}_M = [\overline{XCAT}_M(\cdot,1), \overline{XCAT}_M(\cdot,2),\cdots,\overline{XCAT}_M(\cdot,L)]$$

where

$$\overline{XCAT}_M(\cdot,L) = \frac{1}{NCAT_M} \sum_{I=1}^{NCAT_M} XCAT_M(I,L) \ .$$

Recall that the category swarms are not centered in time, even though the original data set was standardized.

The first step in the PCA of $\underset{\sim}{X}_M$ is to center its columns in time. Let $\underset{\sim}{Z}$ denote the time centered $\underset{\sim}{X}_M$:

$$\underset{\sim}{Z}(I,KX) = \underset{\sim}{X}_M(I,KX) - \overline{XCAT}_M(\cdot,KX)$$

where $I = 1,\ldots,NCAT_M$ and $KX = 1,\ldots,L$.

We then define the L by L *scatter matrix*, $\underline{S}$, by

$$\underset{\sim}{S} \equiv \underset{\sim}{Z}^T \underset{\sim}{Z} \ ,$$

where $\underset{\sim}{Z}^T$ denotes the transpose of $\underset{\sim}{Z}$. Let ℓ_j and $\underline{e}_j$, $j = 1,\ldots,L$, be the eigenvalues and eigenvectors, respectively, of the scatter matrix. Since $\underset{\sim}{S}$ is a real, symmetric matrix, the eigenvalues ℓ_j are non-negative and can be ordered by size:

$$\ell_1 \geq \ell_2 \geq \ell_3 \geq \cdots \geq \ell_L \geq 0 \ .$$

The correspondingly ordered eigenvectors $\underline{e}_j$ have their directions fixed by requiring that the first element of each eigenvector be positive: $\underline{e}_j(1) > 0$. (The $\underline{e}_j$ are sometimes referred to as the *empirical orthogonal functions* of $\underset{\sim}{Z}$.) The *principal components* of $\underset{\sim}{Z}$, denoted by $\underline{a}_j$, are vectors of length $NCAT_M$ defined by

$$\underline{a}_j = \underset{\sim}{Z}\, \underline{e}_j \ .$$

Only the first principal component of $\underset{\sim}{Z}$, namely

$$\underline{a}_1 = (\underset{\sim}{X}_M - \overline{\underset{\sim}{X}}_M)\underline{e}_1$$

or

$$a_1(I) = \sum_{KX=1}^{L} [XCAT_M(I,KX) - \overline{XCAT}_M(\cdot,KX)]e_1(KX) \quad , \quad I = 1,\ldots,NCAT_M \ ,$$

is needed for our decomposition of the category swarm.

We also note that the *covariance matrix*, $\underset{\sim}{C}$, of the data set $\underset{\sim}{X}_M$ is given by

$$\underset{\sim}{C} = \frac{1}{NCAT_M - 1}\,\underset{\sim}{S} \ .$$

Therefore $\underset{\sim}{C}$ has eigenvalues

$$\lambda_j = \frac{1}{NCAT_M - 1}\,\ell_j \ ,$$

where ℓ_j are the eigenvalues of $\underset{\sim}{S}$, and $\underset{\sim}{C}$ has the same eigenvectors as $\underset{\sim}{S}$. The determinant of $\underset{\sim}{C}$ is

$$\det(\underset{\sim}{C}) = \prod_{j=1}^{L} \lambda_j = (NCAT_M-1)^{-L} \prod_{j=1}^{L} \ell_j = (NCAT_M-1)^{-L} \det(\underset{\sim}{S}) \ .$$

Moreover, since $\underset{\sim}{S}$ has the representation

$$\underset{\sim}{S} = \sum_{j=1}^{L} \ell_j \underline{e}_j \underline{e}_j^T$$

and $\underline{e}_j^T \underline{e}_k = \delta_{jk}$, it follows that

$$\begin{aligned} \underset{\sim}{C}^{-1} &= (NCAT_M-1)\underset{\sim}{S}^{-1} \\ &= (NCAT_M-1) \sum_{j=1}^{L} \frac{1}{\ell_j}\, \underline{e}_j \underline{e}_j^T \\ &= \sum_{j=1}^{L} \frac{1}{\lambda_j}\, \underline{e}_j \underline{e}_j^T \ . \end{aligned}$$

Thus no explicit matrix inversion is required to obtain $\underset{\sim}{C}^{-1}$ after the PCA of $\underset{\sim M}{X}$ has been performed. This result is of great use in fitting the multivariate normal pdf's.

PART II. EVALUATION OF THE PDM IN A MODEL OUTPUT STATISTICS SETTING

1. Forecasting Visibility

The first application of the PDM was made by Elias (1985) in a Master's Thesis study. This section briefly summarizes this work.

The problem studied was the forecasting of horizontal atmospheric visibility over selected regions of the North Atlantic Ocean. The potential predictors were the output quantities of the Fleet Numerical Oceanography Center's Navy Operational Global Atmospheric Prediction System (NOGAPS). Examples of these potential predictors are the air temperature at the surface and at the 1000, 925, 850, 700 and 500 mb levels; geopotential height, vapor pressure, and wind components, all at the same levels; surface pressure, surface moisture flux, and cloud parameters. Other potential predictors were derived from the NOGAPS output; e.g. relative humidity and vertical gradients of temperature, geopotential height, vapor pressure, and winds. Visibility observations were obtained from ship reports. The predictor-predictand data set was constructed by interpolating the gridded NOGAPS output to the locations of the reporting ships. The reported visibilities first were terciled as follows: category 1, visibility less than 2 km; category 2, visibility between 2 and 10 km; category 3, visibility greater than or equal to 10 km. A later part of the study used only two categories of visibility: category 1 for visibility less than D_o amd category 2 for visibility greater than or equal to D_o, where D_o was either 2 km or 4 km.

The PDM initially was implemented with these options:

(1) The maximum probability strategy of forecasting was used, as described in §I.2.H.i.

(2) A point swarm was split by PCA if $\lambda > \lambda(96)$, as described in §I.3.F.i.

(3) Hotelling's T^2 test was applied pairwise to the category swarms in order to determine the separation of each pair of category swarms (i.e. categories 1 and 2, 1 and 3, 2 and 3 were compared in the case of 3 categories). If the average separation of all category pairs was significant at the 5% level, then the category swarms were considered to be significantly separated.

(4) The 5% significance levels for PAO and AO were found from standard statistical procedures based on the assumption that PAO and AO are normally distributed.

Other characteristics of the PDM, such as randomly splitting the entire data set into two-thirds training set and one-third testing set, were all as described in part I of this report.

Forecasts made by the PDM were compared with the corresponding forecasts made by three other MOS forecast models. These other models also were proposed by Preisendorfer (1983a,b,c) and previously had been investigated by Karl (1984), Diunizio (1984), and Wooster (1984) in Master's Theses. These competing models are all characterized by the discrete Q-tiling of both predictor and predictand values (unlike the PDM which Q-tiles only the predictand values). The training set is used to define discrete conditional probabilities for the predictand category given the predictor category. The three models differ in how these discrete conditional probabilities are used to make a forecast when given a new predictor value. In the multiple predictor stage, these predictor values are of course vectors in E_L, just as with the PDM. The previous studies found that the discrete conditional probability techniques are comparable in skiII to multiple linear regression when used as an MOS forecast model. Both the PDM and the discrete conditional probability models are expected to have an advantage over linear regression whenever the relation between predictor and predictand is non-linear.

Forecasts were made for three regions of the North Atlantic. A total of 2200 to 4500 predictor-predictand pairs was available for analysis, depending on the region. Time lags of 0, 24 and 48 hours were studied, in the sense that the NOGAPS forecast valid at a given time 0, 24 or 48 hours in the future was used to forecast the visibility at the *same* time. It was found in the original terciling of the visibilities that most observations fell into category 3 (good visibility) and that categories 1 and 2 were not well separated. This observation led to the two-category classification mentioned above.

The two main conclusions of the Elias study are as follows.

(1) The PDM model, as initially implemented, was outperformed by all competing models in all measures of skill (e.g. A0 and A1 scores) for all regions and time lags.

(2) When the PDM was reprogrammed to use the point swarm splitting criterion $\lambda > \lambda_0 = 2$, as described in §I.2.H.ii, the skill of the PDM approached that of the other models. However, only one such case study was made.

2. Artificial Data

After the disappointing performance of the PDM in the Elias study, it was felt that a comparison of the PDM and its competitors should be made using an artificial data set, constructed with known properties. Such a study was made in a Master's Thesis by Fatjo (1986).

The first hurdle of such a study is the construction of an artificial data set which realistically simulates an actual MOS forecast situation. This is not an easy problem, since it requires the simulation of four data sets: (1) the natural primary fields (e.g. winds and temperatures as produced by nature), (2) the modeled primary fields (e.g. winds and temperatures as

predicted by the NOGAPS general circulation model), (3) the natural secondary fields (e.g. visibility as produced by nature), and (4) the observed secondary fields (e.g. visibility as recorded by a human observer or instrument). The time series of the natural primary fields are to be constructed with prescribed autocorrelations, cross correlations and signal-to-noise ratios. The associated natural secondary fields must have the desired connections with the natural primary fields, so that they are in principle predictable from the primary fields. The modeled primary fields must simulate the inherent imperfections of a general circulation model as a predictor of the natural primary fields. And finally, the observed secondary fields must simulate the errors made by an observer when measuring the natural secondary fields. Once the desired data sets are available, a MOS model (e.g. the PDM) is constructed using a part of the modeled primary fields and observed secondary fields (the training set). Then the remainder of the modeled primary fields (the testing set) is used as input to the MOS model, and its predictions of the secondary fields are scored against the natural secondary fields. The mathematical techniques for constructing these data sets are found in Preisendorfer (1985).

Fatjo used the Preisendorfer (1985) technique to generate two natural data sets with 1200 time values for each of 8 predictors and one predictand. One natural data set, termed the "easy set" had a signal-to-noise ratio of 4, thus making it relatively easy to predict the secondary fields from the primary fields. The other data set, termed the "hard set," had a signal-to-noise ratio of 1, making it relatively hard to discern the relation between predictors and predictand. Two general circulation models were also simulated: a "good model," which did a relatively good job of "predicting" the natural primary field (95% of the original field was reproduced), and a

"bad model," which did a poorer job of simulating nature (only 50% of the natural field was reproduced). And finally, three observers were simulated: a "perfect observer," who never made a mistake in recording the natural secondary field to make the observed secondary field; a "good observer," who occasionally made a wrong observation (87% correct observations), but never by more than one predictand category; and a "bad observer," who more often made incorrect observations, sometimes by more than one category (69% correct observations).

Three models were used for comparison with the PDM. The first was the most successful of the three discrete conditional probability models mentioned in the discussion of the Elias study. The second was classical discriminant analysis, and the third competitor was multiple linear regression. The PDM was used with these options:

(1) The predictand was terciled so that each category subset had 400 points, and nearly the same variance.

(2) The Bayesian strategy for forecasting was used, as described in §I.2.H.ii.

(3) The $\lambda > 2$ criterion for PCA splitting of point swarms was used, as described in §I.3.F.ii.

(4) The potential predictability, PP, was used to measure the separation of the category swarms, but the 5% significance levels of PP were not determined because of computational expense (recall the comments at the end of §I.3.J).

(5) In lieu of using the algorithm of §I.3.K (which requires significance levels) to determine the final PDM model, the final model was taken to be the three predictors which together gave the highest PP value. That is, all combinations of three predictors were tested, and their PP scores were ranked.

The four MOS forecast models were applied to the various combinations of easy and hard data sets; good and bad general circulation models; and perfect, good, and bad observers. The resulting skill scores (A0, A1, etc.) of the MOS forecast models were compared using standard Analysis of Variance (ANOVA) techniques, with the following general conclusions:

(1) As expected, all MOS forecast models gave their best results when applied to the simulated data from the "easy" data sets, "good" GCM, and "perfect" observers. All MOS forecast models then performed with less and less success until the "hard" data, "bad" GCM, and "bad" observer case was reached, at which time all MOS forecasters produced their lowest skill.

(2) The transition from "easy" to "hard" data sets of the natural primary fields had a much larger effect on the forecast skills than the transitions from "good" to "bad" GCMs or "perfect" to "bad" observers had.

(3) There was no statistically significant difference in the PDM, classical discriminant, and linear regression skills for the various cases studied, as determined by ANOVA. The discrete conditional probability method scored better than the other forecasters when scored on the training set (PA0, PA1), but scored poorer than the others on the testing set (A0, A1).

PART III. EVALUATION OF THE PDM IN A CLIMATE FORECAST SETTING

1. Forecasting the El Niño of 1982-83

We have seen in part II that the PDM did not show any advantage over other forecasters when applied in an MOS setting. However, the performance of the PDM may be quite different when it is applied in some other forecast situation. In particular, since the El Niño of 1982-83 displayed such large anomalies, it might be hoped that the PDM can detect the separation between normal and abnormal anomaly categories, and thereby successfully hindcast the 1982-83 event.

Barnett (1984) addressed the problem of statistically forecasting sea surface temperature (SST) anomalies in the equatorial Pacific using wind anomalies as predictors, during the 1982-83 El Niño. His study used a sophisticated regression model which related the SST anomalies in the predictand regions to the prior wind anomalies in the predictor regions. Barnett found, among other things, that it was possible to forecast the onset of El Niño, as measured by SST anomalies in a region off the coast of Peru (his "SST5" region), using wind anomalies from various regions in the central Pacific. These forecasts were quite successful at lead times of up to 4 months. Although the Barnett model did an acceptable job of forecasting the *onset* of the 1982-83 El Niño, it failed to accurately predict the *decline* of the El Niño, for reasons discussed in the 1984 paper. It was felt that a repetition of his study would be another means of evaluating the PDM's forecast ability.

A. Using Unfiltered Predictors. The data set consists of monthly wind and temperature anomalies for the 476 months from January 1947 to August 1986. There are 4 regions of the equatorial Pacific for which u-component (East-

West) wind anomalies are available, and 3 regions for which there are v-component (North-South) wind anomalies. Thus there are 7 possible predictors (labeled 1,...,7 corresponding to Barnett's U1, U2, U3, U4, V1, V2, V3, respectively). It was desired to tercile the predictand SST anomalies so that only the extreme events would fall outside the "normal" category. Inspection of the SST record shows that if boundaries $B_1 = -0.5°C$ and $B_2 = 1.2°C$ are selected (see §I.2.C), then slightly less than one-sixth of the anomalies fall into category 1 (below normal SST), somewhat more than two-thirds fall into category 2 (normal SST), and slightly less than one-sixth fall into category 3 (above normal SST). The above normal category so defined contains only anomalies which are greater than two standard deviations from the mean, which is a reasonable definition of El Niño. The 396 months from January 1947 to December 1979 were taken to be the training set, and the 80 months from January 1980 to August 1986 were taken to be the testing set. The training set contains several El Niños, so that the PDM should have a good opportunity to define the category pdf's. The 1982-83 event stands out prominently in the testing set, as is seen in Fig. 16.

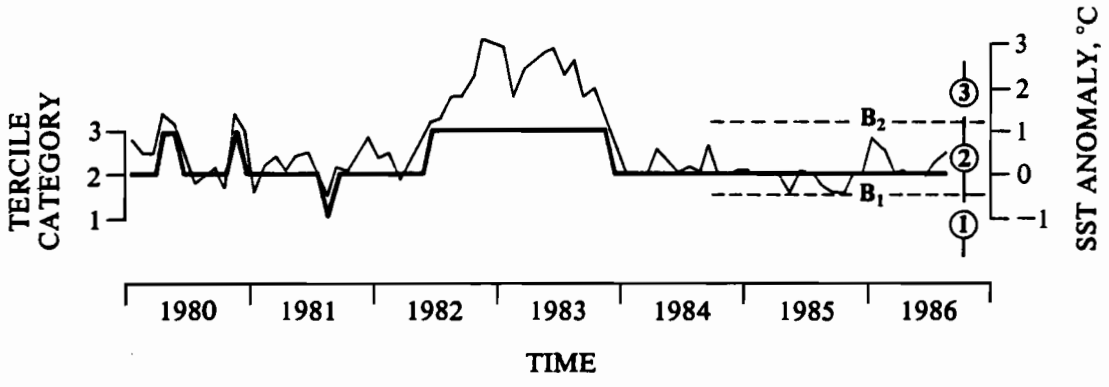

Figure 16.--The testing set for prediction of the 1983 El Niño. The light line and the scale at the right show the actual SST anomalies. The heavy line and the scale at the left show the corresponding tercile category values.

The PDM was applied in various configurations:

(1) Both maximum probability and Bayesian strategies were used. In the Bayesian case, the priors were made proportional to the number of points in the category (cf. §I.2.H).

(2) Category swarms were forced to undergo a predetermined number of PCA subdivisions, either 0 (as seen in Fig. 13), 2 (as seen in Fig. 14), or the maximum possible number (as seen in Fig. 15), as discussed in §I.3.F.iii.

(3) The potential predictability was used to measure the separation of the category pdf's, although the 5% significance levels were computed only in the single predictor cases (owing to computational expense).

(4) The individual predictors were rated by their potential predictability scores in order to select the first predictor. Subsequent predictors were added to the model in the order given by the correlations, as described in §I.3.A. Models containing 1 to 7 predictors were compared.

For a time lag of NTAU = 0, predictor 5 (wind in region V1) has the highest potential predictability score of any individual predictor. If the maximum probability strategy is chosen, this value is PP = 0.196; the 5% significance level is PP(96) = .019, so that PP is significant. For the Bayesian strategy, PP = 0.377 and PP(96) = 0.316, so that PP is once again significant. Predictor 5 thus becomes the first predictor of the PDM model. Predictor 6 (wind in region V2) is least correlated with predictor 5, and therefore becomes the next predictor added to the model. With two or more predictors in the model, we also have the possibility of forecast skills depending on the number of PCA decompositions of the category sets. Figure 17 shows the dependence of the potential predictability on the form of the PDM model. In this figure we note the following behavior of the potential predictability:

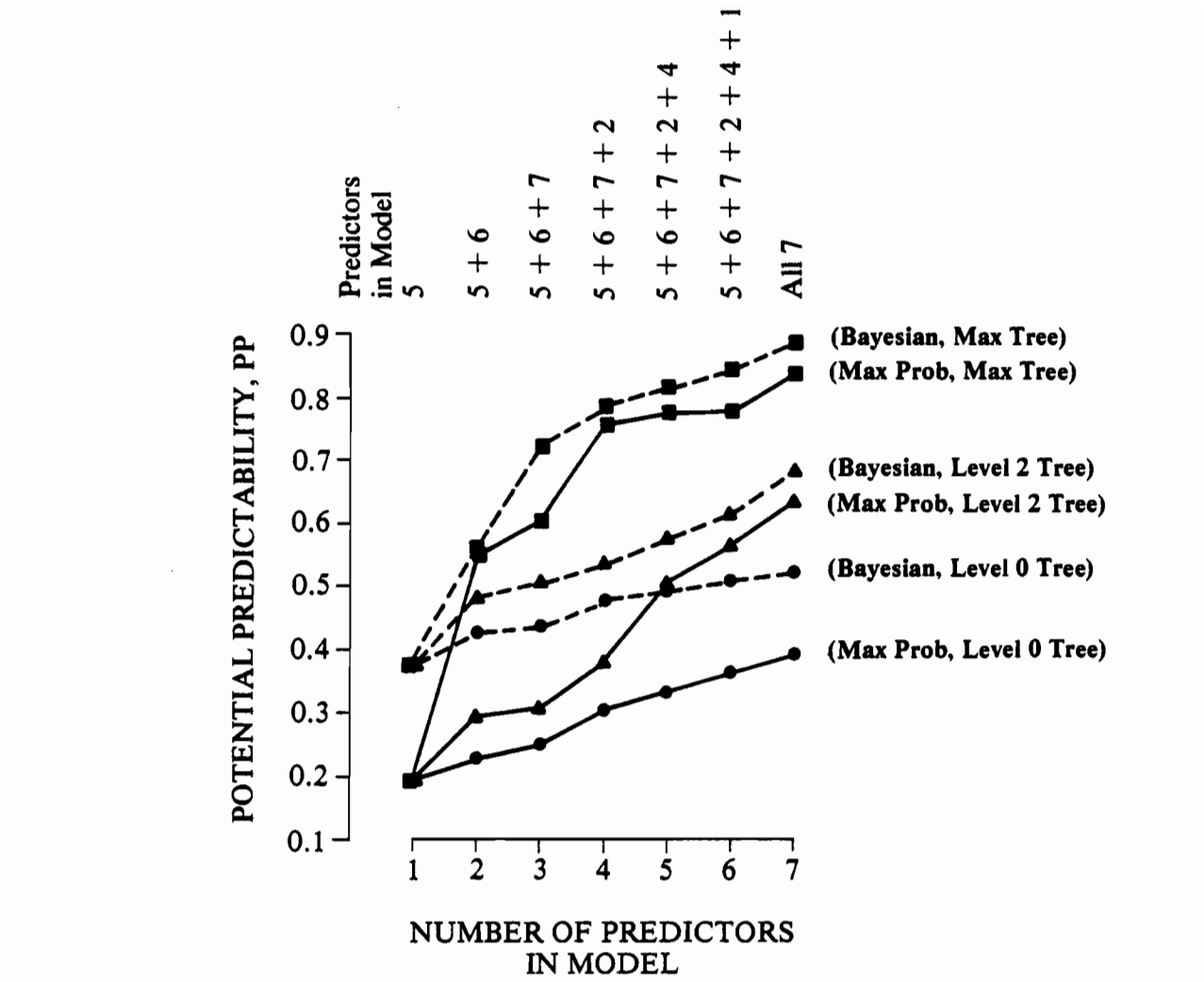

Figure 17.--Potential predictability values for various PDM models (time lag NTAU = 0). The solid curves are for the maximum probability forecast strategy, and the dashed curves are for the Bayesian strategy. Dots are for no PCA decomposition of the category swarms (a level 0 decomposition, equivalent to classical discriminant analysis), triangles are for a level 2 PCA decomposition, and squares identify the curves for which the maximum possible number of PCA decompositions was performed.

1) All else being equal, PP is greater for the Bayesian forecast strategy than for the maximum probability strategy.

2) All else being equal, PP increases as the number of PCA decompositions of the category swarms increases.

3) All else being equal, PP increases as more predictors are added to the model.

Similar results were found for PA0 and PA1, e.g. PA1 decreases (the model becomes better) as predictors are added, all else being equal, and so on. This behavior is consistent with our expectations.

However, when the various PDM models of Fig. 17 are applied to the testing set, the A0 and A1 scores are quite disappointing. Figure 18 shows the A0 scores for the same situations as the PP scores of Fig. 17. We note first of all that, since the terciling of the predictand was designed so that most SST anomalies fall into the "normal" category, climatology (which always predicts normal) is an excellent forecaster, with an A0 score of 0.725. Since climatology gets its high A0 score by virtue of the chosen terciling, it is not valid to compare the PDM's A0 scores with climatology, and of course climatology has no value as forecaster of the onset of an El Niño. Conversely, poor A0 scores of the PDM do not imply that it failed to forecast the onset of the 1982-83 El Niño (the goal of this study), since A0 is a global measure of the PDM's performance.

In Fig. 18 we note that adding more predictors to the model does not always improve the A0 score, all else being equal. Moreover, increasing the number of PCA decompositions of the category swarms generally decreases the A0 scores. We have commented previously (§I.3.F.iii) on the possibility of this sort of behavior, owing to the effects of noise in the data. The Bayesian strategy generally gives better A0 scores than the maximum probability

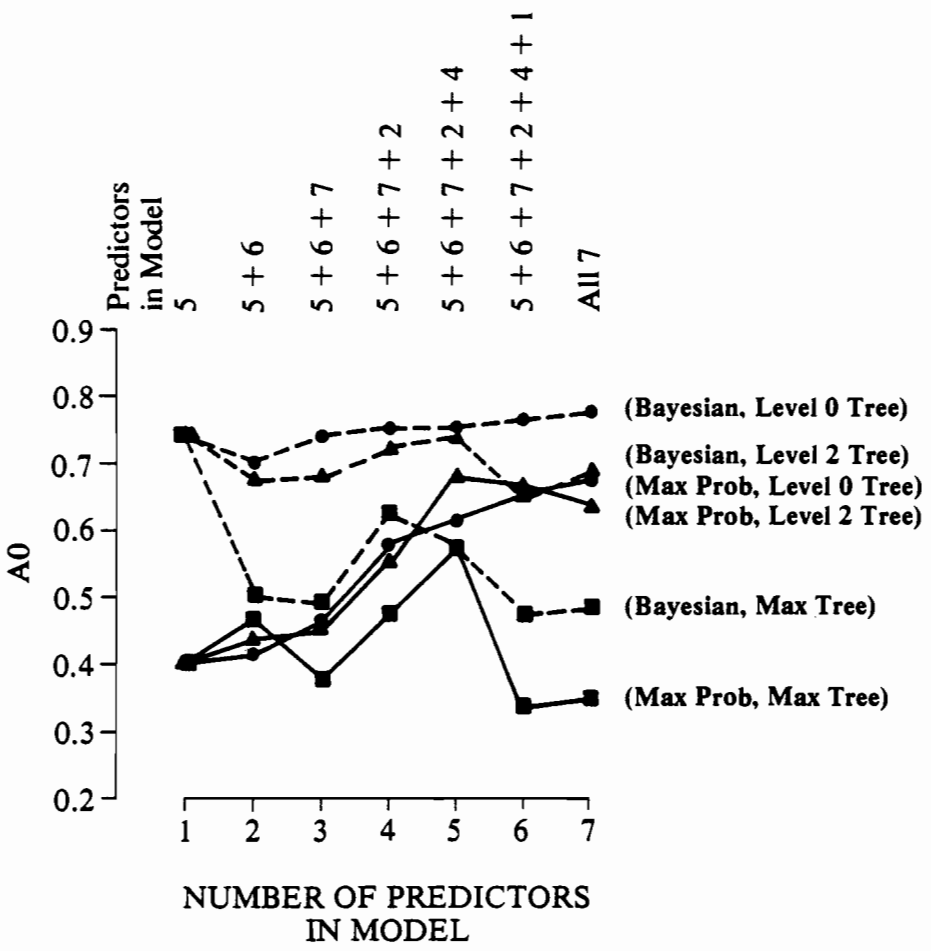

Figure 18.--The A0 scores for the same PDM models discussed in Fig. 17, when the PDM is applied to the testing set. See Fig. 17 for the curve notation.

strategy. The only model that generally outperforms climatology, though not by much, is the Bayesian strategy with no PCA decomposition (i.e. classical discriminant analysis).

The studies summarized in Figs. 17 and 18 were all made for the case of zero time lag between predictor and predictand values; i.e., wind anomalies at a given time were being used to forecast SST anomalies at the same time. But a forecast technique is of real value only if it can predict the future. We therefore now turn to a study of the PDM's behavior for time lags, NTAU, greater than zero, so that wind anomalies at time T are being used to predict SST anomalies at time T + NTAU. We narrow our discussion to PDM models using the Bayesian forecast strategy and a level 2 PCA decomposition of category swarms (even though no PCA decomposition gives somewhat better A0 scores when NTAU = 0). PDM models containing either two or five predictors, as shown in Figs. 17 and 18, were considered. Figure 19 shows the various PP and A0 scores as a function of the time lag NTAU.

We note in Fig. 19 that the PP scores decrease somewhat as NTAU increases from 0 to 4 months for the use of a two-predictor model, but that PP scores are relatively independent of NTAU for the five-predictor model. The A0 scores of the five-predictor model, on the other hand, decrease with NTAU, whereas the A0 scores of the two-predictor model increase at a time lag of 3 months. In addition, Fig. 19 shows the A0 scores as determined by persistence of the predictand. Persistence simply uses the predictand value at time T as the forecast for time T + NTAU. If NTAU = 0, the predictand is used to forecast itself and therefore persistence receives a perfect A0 score. The A0 skill of persistence decreases with NTAU, as expected, but persistence is by far the best forecaster at all time lags. These high A0 scores are of course a consequence of the chosen terciling scheme and the persistent nature of SST anomalies. Persistence cannot forecast the onset of an El Niño.

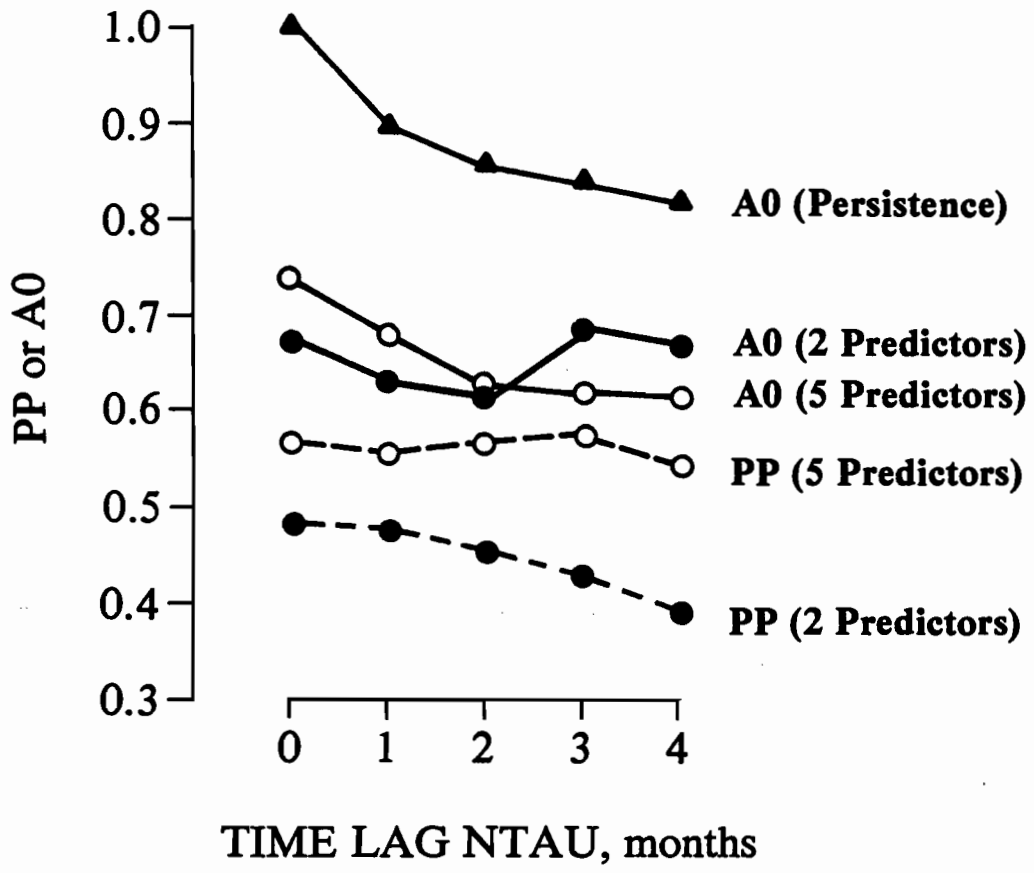

Figure 19.--PP scores (dashed lines) and A0 scores (solid lines) for the two-predictor PDM (solid dots), five-predictor PDM (open dots), and persistence (triangles), all as a function of time lag NTAU.

Figure 19 was generated for two-predictor and five-predictor models in which the particular predictors in the model were held fixed (i.e., predictors 5 and 6 in the first case and predictors 5, 6, 7, 2 and 4 in the second case). In general we would expect that the best predictors for one time lag might not be the best for another time lag. Indeed, for NTAU = 0 or 1, predictor 5 has the highest PP of any single predictor, whereas for NTAU = 2, 3, or 4, predictor 2 (Barnett's U2) has the highest PP. However, for the present data set, this dependence is weak: for NTAU = 4, predictor 2 has PP = 0.336 and predictor 5 has PP = 0.316.

As we have stated above, scores like A0 are overall measures of a predictor's performance. In this study we are, however, particularly interested in forecasting the 1982-83 El Niño. Let us now see how the PDM performed in this task. The upper panel of Fig. 20 shows the SST anomalies for 1981-84, along with the prediction of Barnett's model for a lead time of NTAU = 4 months. As seen in Fig. 20a and as noted previously, the Barnett model does a respectable job of forecasting the onset of the El Niño, although it fails to forecast the decay of the event, since it ignores local forcing. Panel b of Fig. 20 shows the terciled predictand values; these values represent a perfect forecast in terms of the tercile categories. From Fig. 19 we see that the best PDM forecaster for NTAU = 4 is the two-predictor model; Fig. 20c shows the actual tercile forecasts made by this model. Figure 20d shows the forecasts made by the five-predictor PDM model. It is quite clear from these figures that the PDM has failed even to detect the presence of the El Niño, let alone predict the onset of the event. Other variations of the PDM technique all show equally disappointing results.

We must now ask why the PDM makes such a poor showing when applied to actual data, since the basic technique seems so powerful and promising. In

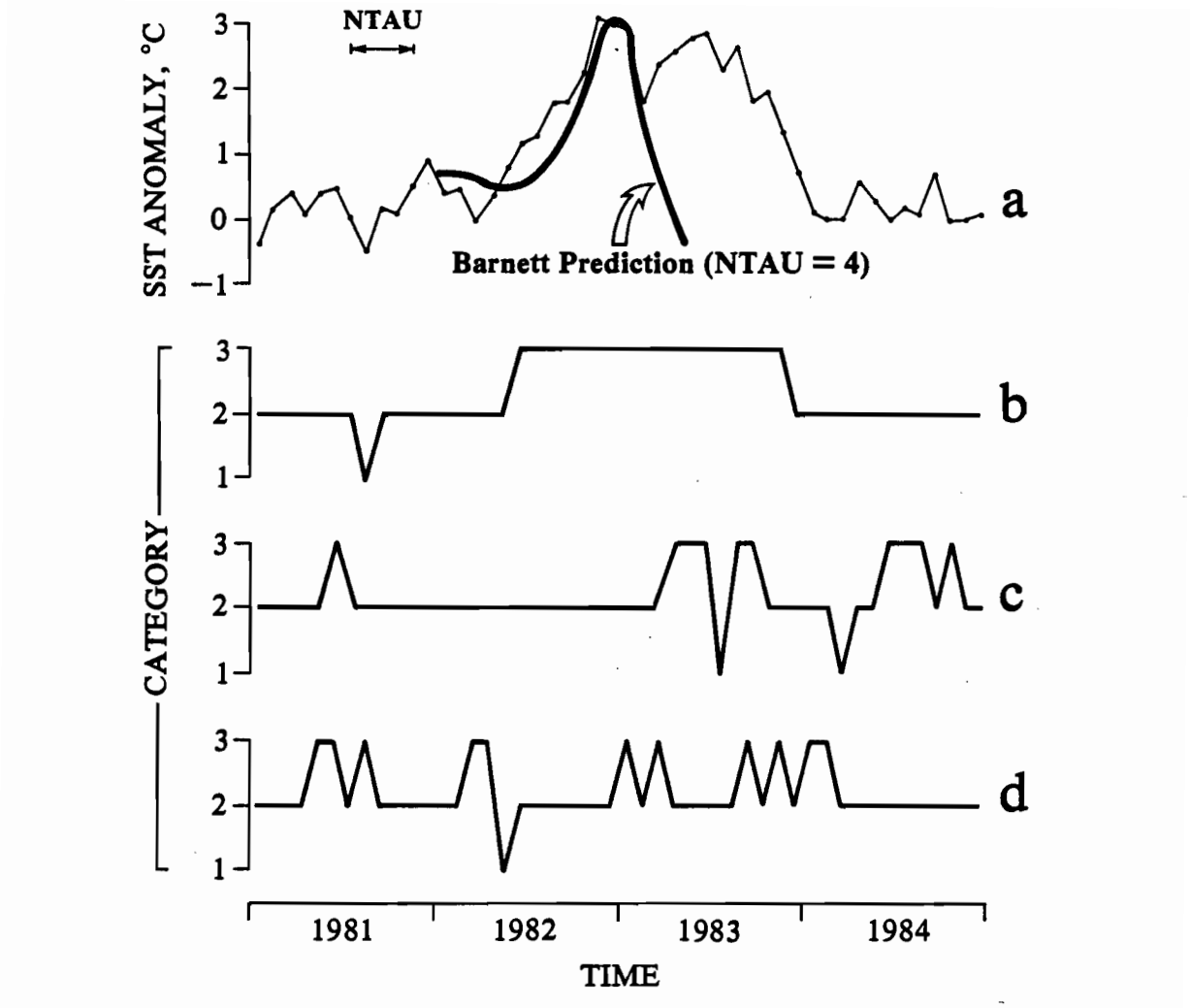

Figure 20.--Predictand time series for 1981-84. Panel (a) shows the actual SST anomalies and the Barnett prediction for a time lag of NTAU = 4 months. Panel (b) shows the terciled SST anomalies (a perfect PDM forecast). Panel (c) shows the PDM forecast made by the two-predictor model, and panel (d) shows the forecast made by the 5-predictor model.

the El Niño study just described, we may speculate that since the event of 1982-83 was so exceptional, it may be that the training set (1947-79) had no events comparable to the one in the testing set (1980-86). If this were the case, the pdf's defining the PDM model might not be able to place the extreme predictor values into the proper category. This idea was tested by pooling the entire 1947-86 data set and then randomly splitting it into two-thirds training set and one-third testing set in the manner described before. The training set then has some points from the 1982-83 event, as does the testing set. However, the forecast scores obtained in this fashion show no improvement over those already discussed.

The cause of the PDM's failure is that the data are so noisy that the category swarms cannot be adequately distinguished. Figure 21 shows the category swarms for the two-predictor model with NTAU = 4. The points for the extreme categories 1 and 3 are nearly lost in the swarm of points for category 2. The associated pdf's, $\Phi_M(\underline{X})$ are correspondingly overlapping. Given such data, neither the PDM nor any similar technique can be expected to show any usable degree of forecast skill. Poor data are also likely the cause of the PDM's poor performance in forecasting visibilities, as described in §II.1. That the Barnett forecast technique was able to make any sense of these data and thereby generate the forecast shown in Fig. 20a speaks highly of his method.

B. Using Filtered Predictors. If the poor performance of the PDM in the El Niño forecast is indeed due to noise in the data, then perhaps filtering or smoothing the raw predictor values will increase the signal-to-noise ratio and thereby allow the PDM to extract the information needed to make its forecast. To investigate this possibility, a series of forecasts was made using two types of filters:

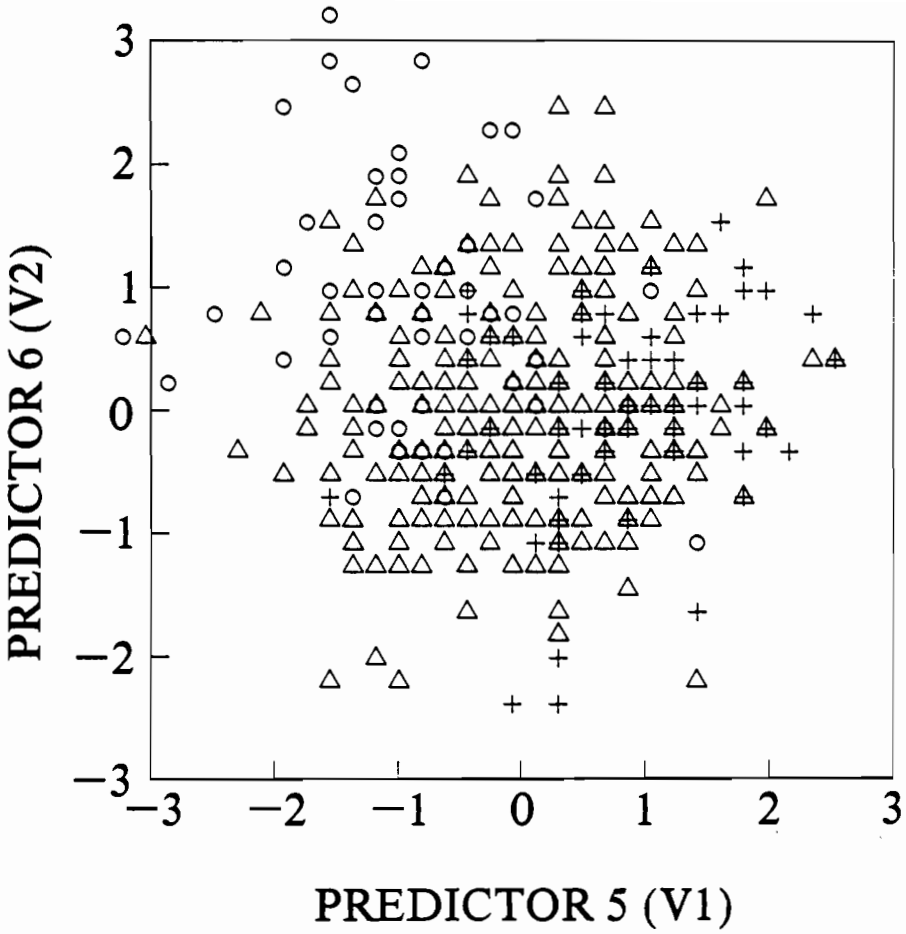

Figure 21.--Category swarms $\underline{XCAT}_M$ for the two-predictor PDM model for NTAU = 4: category 1 ($\underline{XCAT}_1$, "+" symbols, $NCAT_1$ = 65 points); category 2, ($\underline{XCAT}_2$, "Δ" symbols, $NCAT_2$ = 278 points); category 3, ($\underline{XCAT}_3$, "0" symbols, $NCAT_3$ = 53 points).

(1) A 7-point running mean was applied to each predictor time series. Thus each predictor value X(J,K), K = 1,...,NK, was replaced by a smoothed value, XS(J,K), given by

$$XS(J,K) \equiv \frac{1}{7} \sum_{JS=J-3}^{J+3} X(JS,K) \ .$$

The three months at the beginning and end of the 476-month time series were left unsmoothed. The PDM analysis then proceeded as before, but now using the XS(J,K) as predictors.

(2) As before, the training set TRNX was selected to be the first NTRN = 396 months of each of the NK = 7 predictors. A PCA was then performed on the training set (cf. the Appendix in §I) to get

$$\underset{\sim}{A} = \underset{\sim}{TRNX} \cdot \underset{\sim}{E} \ ,$$

where $\underset{\sim}{E} \equiv [\underline{e}_1, \underline{e}_2, \cdots, \underline{e}_7]$ is the 7 × 7 matrix of empirical orthogonal functions (EOFs), and $\underset{\sim}{A} \equiv [\underline{a}_1, \underline{a}_2, \cdots, \underline{a}_7]$ is the 396 × 7 matrix of principal components (amplitudes of the EOFs). The EOFs $\underline{e}_j$ are ordered by the size of the associated eigenvalues ℓ_j, $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_7 \geq 0$. Thus $\underline{e}_1$ is the EOF which explains the most variance of TRNX of any of the $\underline{e}_j$. After performing the PCA, the principal component time series $\underline{a}_j = [a_j(1), \cdots, a_j(NTRN)]^T$, j = 1,...,NK, were used as the predictors in training the PDM, rather than using the original X(J,K) as predictors.

The testing set TSTX was defined as before to be the predictors from 1980-1986. However, before making a forecast using the testing set, we replaced TSTX by amplitudes ATST, defined by

$$\underset{\sim}{ATST} \equiv \underset{\sim}{TSTX} \cdot \underset{\sim}{E} \ ,$$

where $\underset{\sim}{E}$ is the EOF matrix of the training set. We thus performed the same transformation to the training and testing sets, so that the $\underset{\sim}{ATST}$ values can be used in the probability distribution functions Φ_M of the PDM.

It is a property of PCA that the amplitudes $\underline{a}_j$ are uncorrelated: $\underline{a}_j \cdot \underline{a}_k = \ell_j \delta_{jk}$. Therefore the idea of using the correlations between predictors (recall §I.3.A) when constructing the PDM model is no longer valid. However, the ordering of eigenvalues with ℓ_1 largest guarantees us that the associated $\underline{a}_1$ is the best possible predictor of any of the $\underline{a}_j$, in the sense that the most variance is explained by this predictor. The second best predictor is then $\underline{a}_2$, and so on. A two-predictor PDM model would always use $\underline{a}_1$ and $\underline{a}_2$ to forecast the predictand.

A series of runs was made to compare the forecasts made using the filtered predictors with the forecasts seen in paragraph A. The Bayesian forecast strategy and a level 2 decomposition of the category swarms were chosen. Figure 22 shows the A0 scores comparable to those of the two-predictor model (using unfiltered predictors 5 and 6) of Fig. 19. We note first that the A0 scores obtained after applying the 7-point running mean filter to predictors 5 and 6 are in fact lower than the scores obtained using unfiltered predictors 5 and 6. However, if we perform a PCA and then use principal components 1 and 2, the A0 scores are generally higher than the scores of the unfiltered 2-predictor model. These results can be interpreted as follows. The running mean is a low-pass temporal filter which leaves a low frequency, but possibly still random, time series. The spatial correlations between the predictor time series are relatively unchanged by the temporal smoothing. The PCA operation on the other hand is a spatial filter, and the

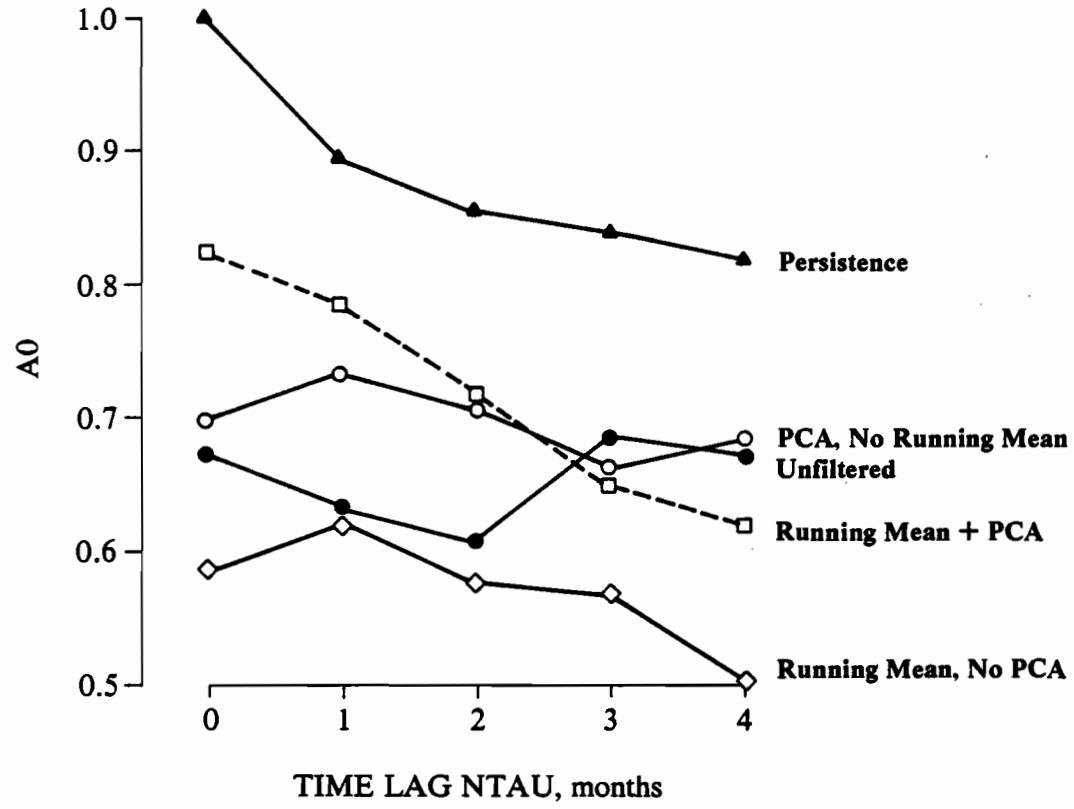

Figure 22.--A0 scores for various two-predictor PDM models: "●" symbols, unfiltered predictors 5 and 6 from Fig. 19; "◇" symbols, predictors 5 and 6 with a 7-point running mean; "○" symbols, principal components 1 and 2; "□" symbols, 7-point running mean, then PCA and using principal components 1 and 2.

resulting time series $\underline{a}_1$ contains spatially coherent information from all of the original predictor regions. Time series $\underline{a}_2$ also contains spatially coherent information from all of the original predictor regions, though of a spatial pattern which is distinct from that of $\underline{a}_1$. Thus merely filtering high frequency noise from the predictor time series does not improve the A0 scores, whereas using the spatially coherent signal from all the original predictor regions does lead to a better set of predictors $\underline{a}_1$ and $\underline{a}_2$. Figure 22 also shows that if we first apply the 7-point running mean to each of the original 7 predictors and then perform a PCA on the smoothed time series, we get greatly improved A0 scores for short time lags, although the A0 scores are degraded for longer time lags.

Figure 23 shows the actual category forecasts made by the two-predictor PDM model using $\underline{a}_1$ and $\underline{a}_2$ as predictors. We now see that the PDM forecasts are similar to that of the Barnett model: a rise to the above-normal category followed by a fall to the below-normal category. Thus with the aid of the preliminary PCA spatial filtering of the noisy wind fields, the PDM has been able to extract the same information from the original data set as did the Barnett linear prediction model. Both models show the same inadequacy of the data set for predicting the latter part of the 1982-83 El Niño.

2. Forecasting Winter Surface Air Temperature over the U.S. Mainland

We next evaluate the PDM's ability to forecast air temperature over the continental United States. The experiment setup was as follows:

(1) The predictor data set was the monthly sea level pressure (SLP) field between 20°N and 80°N and between 140°E and 10°W for the period 1930-80. Each training set of predictors was screened against the training set predictands to define regional predictor averaging areas. The average SLP over these

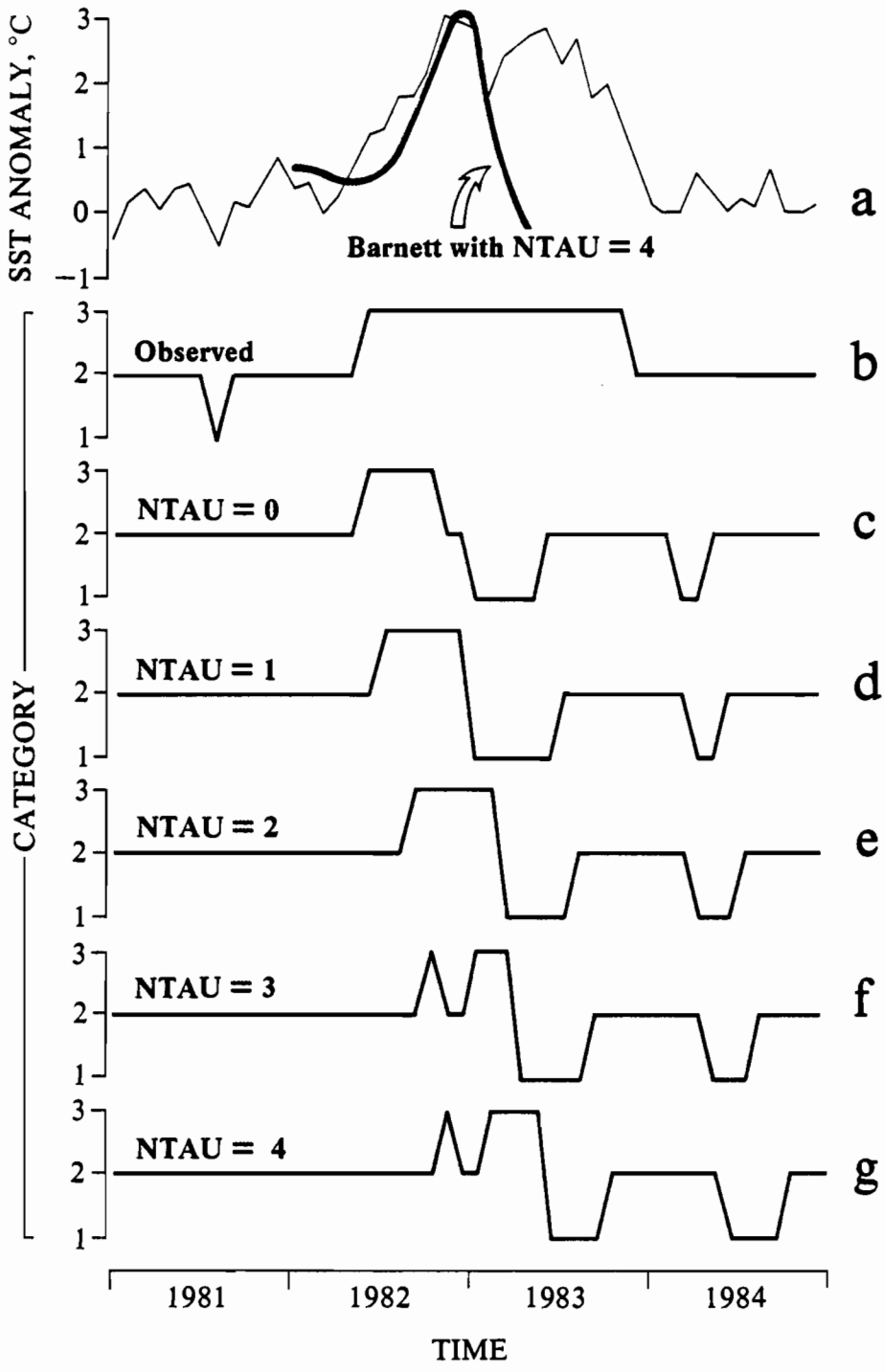

Figure 23.--Category forecasts made using principal components 1 and 2 (" " symbols of Fig. 22). Panels a and b are reproduced from Fig. 20. Panels c-g are for time lags NTAU as shown.

regions became the final set of predictors used in the experiment. The prescreening reduced the size of the predictor set by order 100 and made extensive PDM work feasible.

(2) The predictand set consisted of seasonal surface air temperature anomalies at 33 widely spaced locations in the continental United States. These anomalies were terciled into 3 equally populated classes; the tercile class is the quantity predicted by the PDM.

(3) the PDM configuration was as follows:

i) The maximum probability strategy was used to select the forecast

ii) Category swarms underwent at most one PCA subdivision. The results showed that most of the forecast skill was captured by the first predictor chosen, so high-level splitting gained us nothing of substance.

iii) The potential predictability was used to rank predictors

iv) Monte Carlo simulations were made by replacing the data with "white noise" in order to determine statistical significance levels. These operations were performed on the training sets, and significance of the potential predictability was determined. These models when run on the testing sets gave similar estimates for forecast scores.

v) The PDM model construction was done on ten different realizations of training/testing sets. Thus, different training/testing sets were chosen at random for each realizations. The *average* scores over this ensemble are the results discussed below.

4) Forecasts were made for winter at a lead time of one season. The scores are shown as "percent correct category forecasts" and thus a randomly made forecast has an expected value of 33%. Note that these are actual forecast skills since the test sets in no way entered the predictor screening or PDM pdf construction.

The results of the single-predictor experiments are shown in Fig. 24. Monte Carlo simulations showing forecast skill values (A0) averaged over the entire U.S. in excess of 50% are significant at the 95% level. The highest forecast skills are in the eastern and western thirds of the country; there is a distinct skill minimum in the central region. Both results are in complete accord with earlier studies (cf. Barnett, 1981; Barnett and Preisendorfer, 1987). The levels of skill are also comparable with the former work but are slightly higher than the latter, particularly along the west coast.

Adding an additional predictor and allowing the possibility of a single PCA subdivision of the category subset gives the results shown on Fig. 25. Skill scores are increased typically by 5-10%, particularly in the northeast part of the country. Notice that the central region, where skills were low, exhibits either no change or a decrease in skill.

The above results can be contrasted against the average forecast skills obtained from the Monte Carlo experiments (Fig. 26). The expected value of 33% is indeed realized on average over the entire U.S. The forecasts made by the PDM are clearly much better than random chance.

We conclude that the PDM performance in forecasting winter air temperature over the United States is highly statistically significant. Further, the skill levels are comparable with those obtained by other methods. If it is eventually found that the climate system is "regional" in nature, then the PDM offers one of the few statistical techniques for long range forecasting and diagnostics.

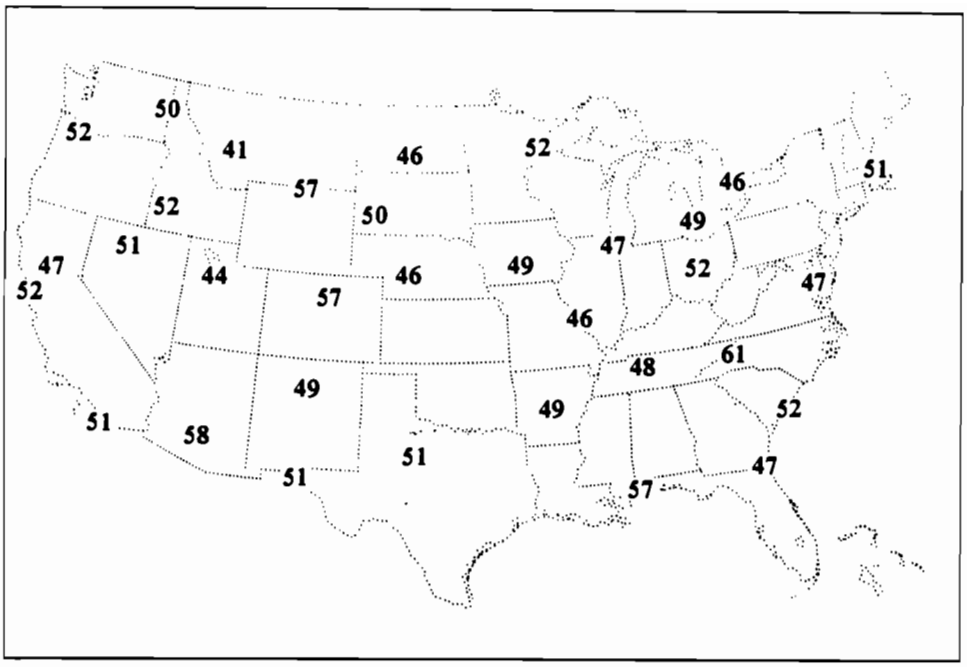

Figure 24.--Winter surface air temperature A0 scores, expressed as percent correct category forecasts, for the best single-predictor PDM. Values greater than 50 are significant at the 95% level.

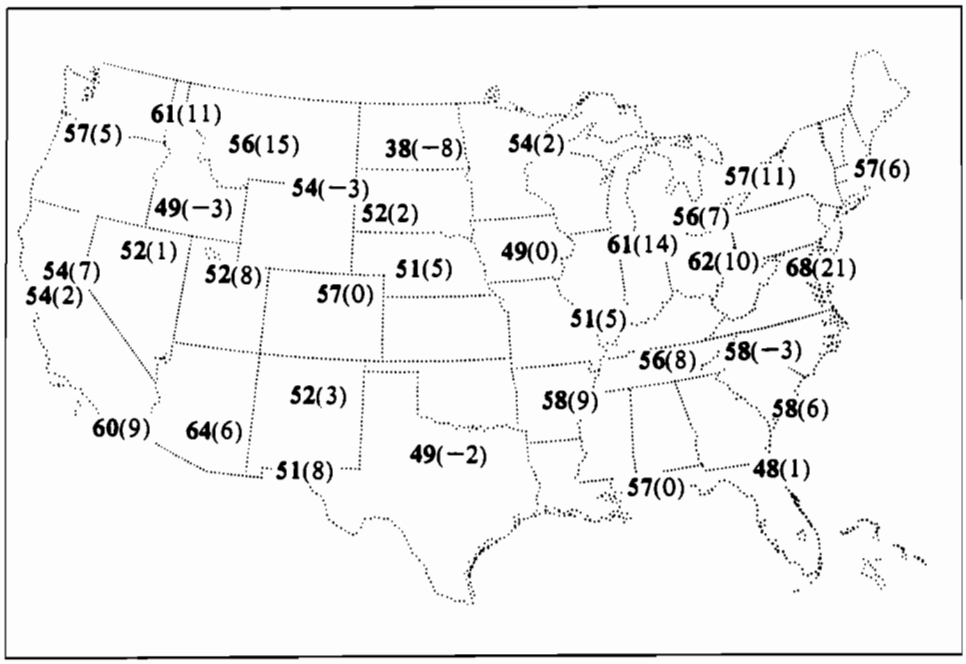

Figure 25.--Percentage A0 scores for the best two-predictor PDM. Values in parentheses show the increase or decrease (negative numbers) in the A0 scores relative to those of the single-predictor PDM, shown in Fig. 24.

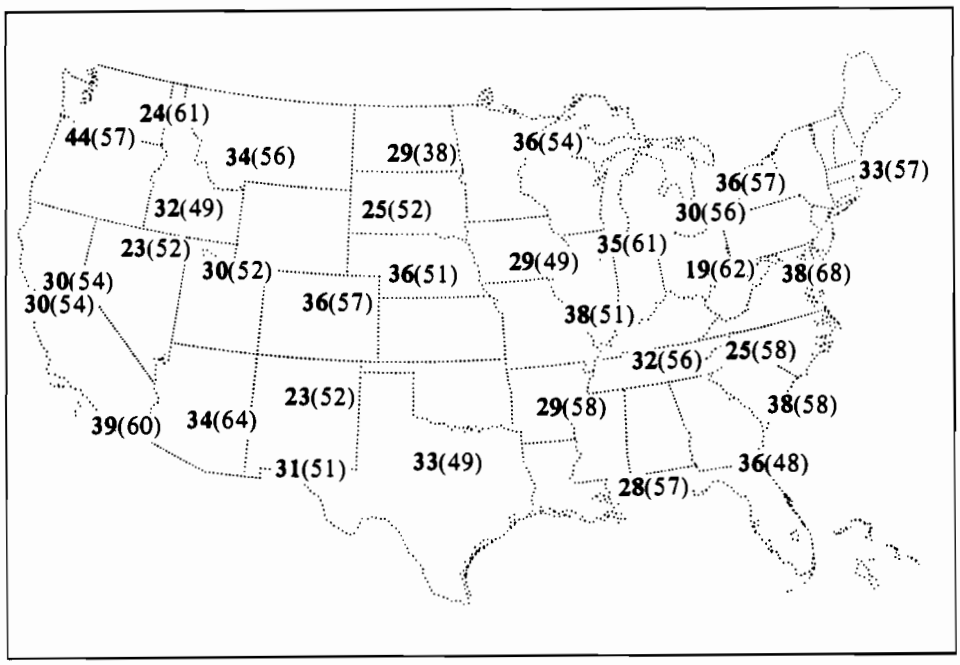

Figure 26.--Percentage A0 scores obtained from Monte Carlo experiments in which the category forecast was made at random. The expected value is 33. The numbers in parentheses show the A0 scores of the two-predictor PDM, from Fig. 25.

Acknowledgments

This work was supported in part by the U.S. Tropical Ocean-Global Atmosphere (TOGA) Program (NOAA/NA85AA-D-AC132), in part by the U.S. Climate Program Office via the Climate Research Group, Scripps Institution of Oceanography (NOAA/NA86-AA-D-CP104) and by the National Science Foundation (ATM85-13713). Word processing was performed by Mr. Ryan Whitney and figures were drawn by Ms. Gini Curl.

References

Barnett, T.P. (1981): Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Weather Rev., 109*, 1021-1041.

_____ (1984): Prediction of the El Niño of 1982-83. *Mon. Weather Rev., 112*, 1403-1407.

Barnett, T.P. and R.W. Preisendorfer (1987): Origins and levels of monthly and seasonal forecast skill for North American surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.*, in press.

Box, G.E.P., and G.C. Tiao (1972): *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA, 588 pp.

Diunizio, M. (1984): An Evaluation of Discretized Conditional Probability and Linear Regression Threshold Techniques in Model Output Statistics Forecasting of Visibility over the North Atlantic Ocean. Master's Thesis (R.J. Renard, advisor), Dept. of Meteorology, Naval Postgraduate School, Monterey, CA, 233 pp.

Elias, K.C. (1985): Forecasting Atmospheric Visibility over the Summer North Atlantic Using the Principal Discriminant Method. Master's Thesis (R.J. Renard, advisor), Dept. of Meteorology, Naval Postgraduate School, Monterey, CA, 112 pp.

Fatjo, S.J. (1986): A Study to Determine the Relative Skill of Four Model Output Statistics Prediction Methods Using Simulated Data Fields. Master's Thesis (R.J. Renard, advisor), Dept. of Meteorology, Naval Postgraduate School, Monterey, CA, 71 pp.

Karl, M.L. (1984): Experiments in Forecasting Atmospheric Marine Horizontal Visibility Using Model Output Statistics with Conditional Probabilities of Discretized Parameters. Master's Thesis (R.J. Renard, advisor), Dept. of Meteorology, Naval Postgraduate School, Monterey, CA, 165 pp.

Lachenbruch, P.A. (1975): *Discriminant Analysis*. Hafner Press, New York, 128 pp.

Miller, R.G. (1962): Statistical Prediction by Discriminant Analysis. *Meteorological Monographs*, Vol. 4, No. 25, 54 pp.

Preisendorfer, R.W. (1983a): Proposed studies of some basic marine atmospheric visibility prediction schemes using model output statistics. Unpublished manuscript, Department of Meteorology, Naval Postgraduate School, Monterey, CA, 28 pp.

_____ (1983b): Maximum-probability and natural regression prediction strategies. Unpublished manuscript, Department of Meteorology, Naval Postgraduate School, Monterey, CA, 10 pp.

_____ (1983c): Tests for functional dependence of predictors. Unpublished manuscript, Department of Meteorology, Naval Postgraduate School, Monterey, CA, 5 pp.

_____ (1984): The principal discriminant method of prediction. Unpublished manuscript, Pacific Marine Environmental Laboratory/NOAA, Seattle, WA, 25 pp.

_____ (1985): Simulation data sets for testing MOS (Model Output Statistics) prediction methods. NOAA Tech. Memo. ERL PMEL-65, Pacific Marine Environmental Laboratory/NOAA, Seattle, WA, 53 pp.

Preisendorfer, R.W., and C.D. Mobley (1984): Climate forecast verifications, United States mainland, 1974-83. *Mon. Weather Rev., 112*, 809-825.

Wooster, M.H. (1984): An Evaluation of Discretized Conditional Probability and Linear Regression Threshold Techniques in Model Output Statistics Forecasting of Cloud Amount and Ceiling Over the North Atlantic Ocean. M.S. Thesis (R.J. Renard, advisor), Dept. of Meteorology, Naval Postgraduate School, Monterey, CA, 187 pp.