SANDIA REPORT

SAND99-1505
Unlimited Release
Printed June 1999

Multivariate Analysis of Seismic Field Data

M. Kathleen Alam

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (703) 605-6000
Web site: http://www.ntis.gov/ordering.htm

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND99-1505
Unlimited Release
Printed June 1999

Multivariate Analysis of Seismic Field Data

M. Kathleen Alam
Materials and Processes Computation and Modeling Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0342

Abstract

This report includes the details of the model building procedure and prediction of seismic field data. Principal Components Regression, a multivariate analysis technique, was used to model seismic data collected as two pieces of equipment were cycled on and off. Models built that included only the two pieces of equipment of interest had trouble predicting data containing signals not included in the model. Evidence for poor predictions came from the prediction curves as well as spectral F-ratio plots. Once the extraneous signals were included in the model, predictions improved dramatically. While Principal Components Regression performed well for the present data sets, the present data analysis suggests further work will be needed to develop more robust modeling methods as the data become more complex.

Acknowledgment

The author thanks Greg Elbring of Dept. 6116, Geophysical Technology for providing the data for this analysis, the programs and help in translating it into the appropriate format, as well as his useful comments upon reviewing this manuscript.

TABLE OF CONTENTS

INTRODUCTION **7**

DATA PRE-PROCESSING **7**

MULTIVARIATE DATA ANALYSIS METHODS: AN INTRODUCTION **8**

PCA 8
PCR 11
SPECTRAL F-RATIO 12

DATA ANALYSIS **12**

EVALUATION OF TRAINING DATA 13
PCR ROBUSTIFICATION – ADDING TEST 1 SAMPLES 22
PCR ROBUSTIFICATION – ADDING TEST1 AND TEST2 SAMPLES 24
USING ROBUSTIFIED MODEL TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 TO PREDICT TEST DATA FROM DIFFERENT LOCATIONS 25
SPECIFIC SITE MODELS PREDICTING TEST2 DATA 27

SUMMARY/CONCLUSIONS **29**

REFERENCES **32**

TABLE OF FIGURES

FIGURE 1 SINGULAR VALUES FOR TRAIN1_A1 AFTER DATA WAS MEAN CENTERED. 13
FIGURE 2 LOADINGS AND SCORES FOR TRAIN1_A1 14
FIGURE 3 SINGULAR VALUE PLOT FOR THE COMBINED SET, TRAIN1_A1 + TRAIN2_A1 15
FIGURE 4 LOADINGS AND SCORES FOR COMBINED SET, TRAIN1_A1 + TRAIN2_A1 16
FIGURE 5 SINGULAR VALUES FOR TEST2_A1 17
FIGURE 6 LOADINGS AND SCORES FOR FACTORS 1-3 OF TEST2_A1. 18
FIGURE 7 LOADINGS AND SCORES FOR FACTORS 4-6 OF TEST2_A1. 19
FIGURE 8 PREDICTED TIME PROFILES OF THE TEST DATA FROM STATION A1 USING PCR MODEL BASED ON TRAIN1_A1 + TRAIN2_A1. 21
FIGURE 9 SPECTRAL F-RATIOS CALCULATED FOR EACH TEST SET.. 21
FIGURE 10 PREDICTION OF TEST DATA FROM STATION A1 USING TRAINING DATA 'ROBUSTIFIED' WITH DATA FROM TEST1_A1 23
FIGURE 11 SPECTRAL F-RATIOS CALCULATED FROM THE SPECTRAL RESIDUALS OF TEST1_A1, TEST2_A1 AND TEST3_A1 USING A TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 MODEL.. 23
FIGURE 12 PREDICTION OF TEST DATA FROM A1 STATION USING TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 MODEL 25
FIGURE 13 SPECTRAL F-RATIOS CALCULATED FROM PREDICTION OF TEST DATA FROM A1 STATION USING TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 MODEL. 25
FIGURE 14 PREDICTION OF TEST2_B1 DATA USING TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 MODEL 26
FIGURE 15 PREDICTION OF TEST2_C1 DATA USING TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 MODEL 26
FIGURE 16 PREDICTION OF TEST2_D1 DATA USING TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 MODEL 27
FIGURE 17 PREDICTION OF TEST2_B4 DATA USING TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 MODEL 27

FIGURE 18 PREDICTION OF THE TEST2_B1 DATA USING A TRAIN1_B1 +TRAIN2_B1 + TEST1_B1 + TEST2_B1 MODEL. 28
FIGURE 19 PREDICTION OF THE TEST2_C1 DATA USING A TRAIN1_C1+TRAIN2_C1+TEST1_C1+TEST2_C1 MODEL. 29
FIGURE 20 PREDICTION OF THE TEST2_D1 DATA USING A TRAIN1_D1 + TRAIN2_D1 + TEST1_D1 + TEST2_D1 MODEL.. 29

TABLES

TABLE 1 8
TABLE 2 28

Introduction

Multivariate methods that have been historically applied to chemical data (e.g., infrared, mass spectrometry, inductively coupled plasma) have been used to evaluate seismic data gathered from a field test. The goal of the field test was to evaluate whether seismic signatures could be used to determine when designated pieces of equipment were being operated. Using the multivariate technique Principle Components Regression (PCR)(Wold, Esbensen et al. 1987), models were developed for two pieces of equipment, using training data collected as the equipment was cycled on and off. The developed models were used to predict three test sets of data. The test data contained the signatures from the two pieces of designated equipment, as well as signatures from additional pieces of equipment.

Data Pre-Processing

Data was received in time-domain format and was transformed to the frequency domain using the MATLAB function **specgram,** which performs a Fourier transform of the data given a windowing function, the number of points in the transform, window size, and sampling rate. The same **specgram** parameters were used to convert all data sets except set D1, which had a different sampling rate. The MATLAB code is shown below:

```
Samprate = 16000;                    %8000 for D1 data
low_freq=1;
order = fix ( samprate / low_freq );
w = window ( 'bar', order, 0 );      %code received from G. Elbring -- 'bar' is
                                     %Bartlett
noverlap=50;
[B,f,t] = specgram ( data, order, samprate, w, noverlap );
```

Five data sets were received: TRAIN1, TRAIN2, TEST1, TEST2, TEST3. Each set of data was obtained at 5 stations (except TEST3, which was obtained at 2 stations). The stations were indexed as A1, B1, C1, D1 and B4. Station A1 provided data with the best signal to noise ratio with B1, C1 and D1 containing progressively decreasing signal levels. Station B4 was at the same location as B1, but at a different azimuth, and provided data with very low signal to noise. The data sets and their dimensions (after

Fourier transformation) are listed in Table 1. In all cases, only frequencies between 0 and 500 Hz were used for analysis.

Table 1

Data Set	Dimension—Frequencies x Number of Samples
TRAIN1	500 x 150
TRAIN2	500 x 150
TEST1	500 x 903
TEST2	500 x 421
TEST3	500 x 622

TRAIN1 contains data collected as Instrument 1 was cycled on and off while TRAIN2 contains data collected as Instrument 2 was cycled on and off. TEST1, TEST2 and TEST3 contained data collected as both Instrument 1 and 2 were cycled on and off. In addition, TEST1, TEST2 and TEST3 contained signals from other pieces of equipment.

Multivariate Data Analysis Methods: An Introduction

PCA

Principle Components Analysis (PCA), also known as factor analysis, refers to the transformation of data into an orthogonal basis set. The variance described by the basis vectors is largest in the first vector and decreases with additional vectors. A data matrix, X, with r samples, each sample containing n points ($r < n$), can be interpreted as an ensemble of r points in n dimensional space. PCA is the process of fitting a series of "lines and planes of closest fit to systems of points in space"(Wold, Esbensen et al. 1987). The "closest fit" in this case is the least squares fit.

The vectors within this new basis set are referred to as PCA loadings and scores. If the noise within the data is randomly varying, it will be contained in later loadings and scores of the PCA decomposition. Estimating the data set by using only the significant PCA loadings can effectively filter random noise.

Mathematically, PCA can be described as the decomposition of a data matrix, X, with rank r, into a series of rank 1 matrices. If X has dimensions r x n then

$$X = M_1 + M_2 + M_3 + ... + M_r$$

Each of the rank 1 matrices can be written as outer products of two vectors: a score vector, t_h, and a loading vector, p'_h, where the apostrophe (') represents the transpose.

$$X = t_1 p'_1 + t_2 p'_2 + t_3 p'_3 + ... + t_h p'_h + t_r p'_r$$

Each vector t_h is r x 1 and each p'_h vector is 1 x n. The equation above can be represented in matrix notation as:

$$X = TP'$$

The loading vectors, p_h, can be thought of as the spectral, or frequency features, while the scores, t_h, can be thought of as the concentration, or amount, of that spectral feature contained in the sample. The insignificant loading and scores can be deleted and the data matrix estimated using only the significant PCA factors, 1:k.

$X = T_{rxk} P'_{nxk} + E$ where $E = T_{(k+1):r} P'_{(k+1):r}$, and represents the error.

$\hat{X} = T_{rxk} P'_{nxk}$ where $\hat{X}$ is the estimate of X.

Many methods exist for calculating scores and loadings. The most readily interpretable method is based on iterative least squares,(Wold 1966), (Geladi and Kowalski 1986) and is called Nonlinear Iterative Partial Least Squares, or NIPALS. The method begins by selecting a vector, $x_{1:r}$, from the data matrix and using it to initiate the score vector, t. The selection of the initial vector is arbitrary, and some authors have used a vector of ones.(Wold 1966) The steps of the procedure are outlined below.

1) $t_{init} = x_{1:r}$

2) The initial least squares estimate of p_{init} is calculated using t_{init}:

   $p'_{init} = (t'_{init} * X)/(t'_{init} * t_{init})$

3) The vector p_{init} is normalized to length 1:

   $p'_{init} = p'_{init}/norm(p'_{init})$

4) The least squares estimate of t is made using the newly estimated p.

   $t_{new} = X * p_{init}/(p'_{init} * p_{init})$

5) Compare t_{new} to t_{init}. If they are the same, the estimation of t and p for step h is complete. If they are not equal, substitute t_{new} for t_{init} and run through steps 2-5 until t_{new} and t_{init} converge.

Conceptually, one can imagine that the first 'plane' calculated (defined by t_1 and p_1) sits at the center of the multidimensional data space. Once the first set of vectors is calculated (h=1), the dot product of t_1*p_1 is calculated and removed from X. The remainder is substituted for X and steps 1-5 are repeated for as many vectors as needed.

From the NIPALS equations, the calculation of the scores, t, and the loadings, p, can be reformulated into the calculation of the eigenvectors of X'X and XX'. If we let $t'_{init}t_{init}$ and $p'_{init}p_{init}$ equal constants C_t and C_p, then step 2 and step 4 above become:

$$p'_{init}=t'_{init}*X/C_t \qquad \text{and} \qquad t_{init}=X*p_{init}/C_p$$

Substituting gives:

$$p'_{init}=(X*p_{init}/C_p)'*X/C_t \qquad \text{and} \qquad t_{init}=X*(t'_{init}*X/C_t)'/C_p$$

Combining C_p and C_t into the general constant C and rearranging gives:

$$C*p'_{init}= p'_{init} *X'* X \qquad \text{and} \qquad C*t_{init}=X* X'* t_{init}$$

These equations can be rearranged to form the classic eigenvector-eigenvalue equations:

$$(C*I-X'*X)*p'_{init}=0 \qquad \text{and} \qquad (C*I-X*X')*t_{init}=0$$

where I is the identity matrix. Thus the loadings and scores are the eigenvectors of $X'*X$ and $X*X'$ respectively. The MATLAB function for the singular value decomposition, **svd**, calculates the eigenvectors of $X'*X$ and $X*X'$ directly, given X. The **svd** function gives two orthonormal matrices, and a diagonal matrix of singular values. One of the orthonormal matrices provides the loadings, and the other, when multiplied by the singular value matrix, provides the scores.

Once the scores and loadings are obtained, they can be used to relate the spectral information to some external value, such as concentration or class. The singular values can also be used to evaluate the significance of the factors. The relative size of the singular values provides an indication of the significance of the factor. For simple data, the number of significant factors can be estimated by plotting the singular values. More rigorous methods for determination of the appropriate number of significant factors to retain involves F-test evaluation(Haaland and Thomas 1988) of the predictive residuals or comparison of the singular values to known noise levels in the system under study(Wold, Esbensen et al. 1987).

PCR

Principle components regression, (PCR), relates the scores and loading matrices obtained in PCA to an external variable in a least squares sense. Once the data matrix is estimated by the appropriate number of PCA loadings,

$$\hat{X} = T_{rxk} P'_{nxk}$$

a least squares solution can be found to relate the vector containing the state of the external variable, y, to $\hat{X}$. Let us begin with the initial assumption, that the spectral data is related to some extraneous variable, in this case the status of the instrument:

$$y \alpha \hat{X}$$

y is typically a vector, with dimensions r x 1, and α is the proportionality symbol. For the present data, y is a vector of ones and zeros, which designated the status of the instrument as 'on' or 'off'. The relationship between y and $\hat{X}$ can be defined as:

$$y = \hat{X} * b + E_y$$

where b is a regression vector relating y to $\hat{X}$ in a least squares sense and E_y is the error vector. So, substituting scores and loadings for $\hat{X}$ gives:

$$y = [T_{rxk} * (P_{nxk})] * b + E_y$$

The regression vector b (n x 1) can be estimated in a least squares sense.

$$b \cong [T_{rxk} * (P_{nxk})]^{-1} * y$$

The matrix $[T_{rxk}*(P_{nxk})']$ is probably not square, thus a traditional inverse can't be calculated. The psuedoinverse is calculated instead. In MATLAB the function is **pinv**, and in the literature is designated by a superscript '+'.(Strang 1980) So,

$$[T_{rxk}*(P_{nxk})']^{+} = [P'_{nxk}]^{+} * [T_{rxk}]^{+}$$

The psuedoinverses of the T and P matrices are their left inverses.

$$[T_{rxk}*(P_{nxk})']^{+} = [(P_{nxk}*(P_{nxk})')^{-1} * P_{nxk}] * [((T_{rxk})' * T_{rxk})^{-1} * (T_{rxk})']$$

Thus b is estimated as:

$$b \cong [T_{rxk} * (P_{nxk})]^{+} * y$$

Once b is estimated using calibration data, the $y_{unknown}$ value of an unknown spectrum, $x_{unknown}$, can be estimated as:

$$y_{unknown} \cong x_{unknown} * b$$

More details on the least squares procedure, orthonormal matrices, and the singular value decomposition can be found in Strang.(Strang 1980)

Spectral F-ratio

The ability of a calibration model to adequately predict an unknown sample depends whether or not the unknown sample resides within the calibration space. While it is not necessary to know all the components or features within the calibration or unknown, the calibration should include those states expected in the unknown samples. A prediction can still be obtained if an unknown sample contains components not contained in the calibration, however, the confidence in that prediction is decreased. Often the presence of new components within the unknown can be detected using the spectral residuals calculated by estimating the spectrum of the unknown sample and comparing it to the true spectrum. The size of the residuals is then compared to the calibration residuals, resulting in a single number called the spectral F-ratio. An F-test can then be performed on the resulting value. Since computing the degrees of freedom can be difficult(Haaland and Thomas 1988), the resulting spectral F-ratios are not used as absolute value for comparison to F-test tables. Rather, they are typically used as a guide to flag possible outliers within the context of the data being examined.

Data Analysis

Prior to PCR analysis, PCA was used to evaluate each data set. Singular value plots, loadings and scores were examined for each data set. Data exploration provided information regarding outliers (spurious signals not consistent with the other data within the set), locations of instrument transition signals, as well as changes in overall intensity.

Evaluation of Training Data

Shown in Figure 1 is a plot of singular values 1-10 for training set TRAIN1_A1, after the mean of the data was removed. The singular values level off beginning at factor 2 suggesting that there is 1 primary factor in this data set. This result is consistent with the information that TRAIN1_A1 data contained only 1 instrument signal varying between two states.

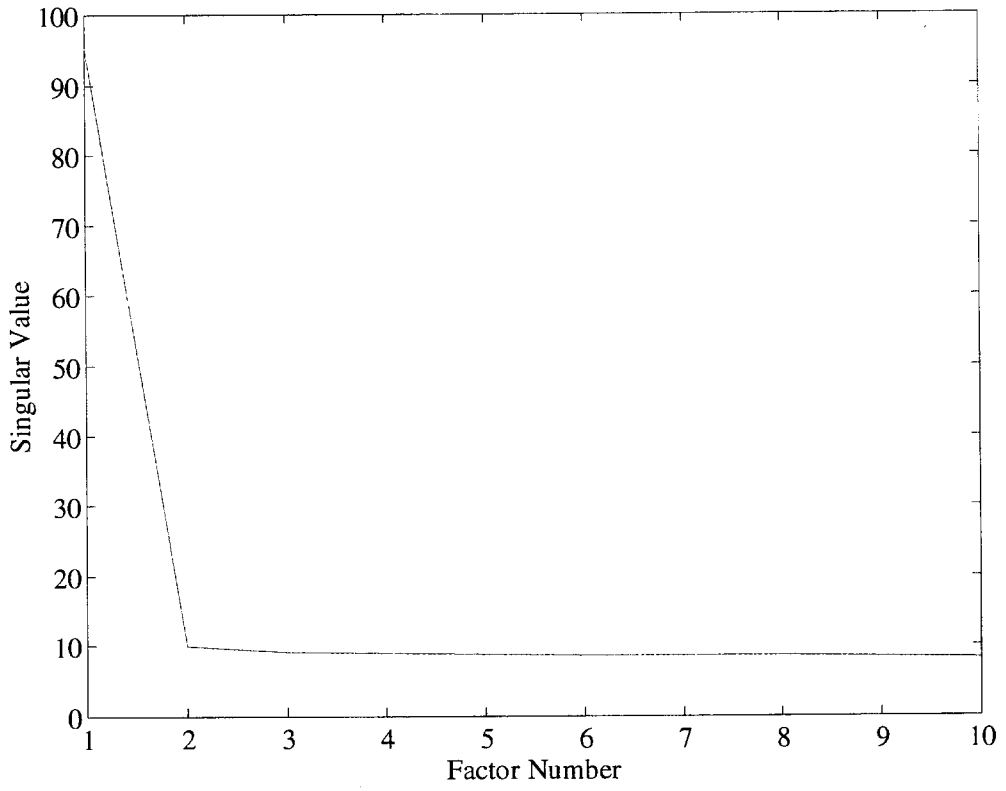

Figure 1 Singular values for TRAIN1_A1 after data was mean centered.

Examination of the scores for factor 1 (see Figure 2) indicates the time at which Instrument 1 was turned on. A very distinct fall and rise of the scores corresponds to the times Instrument 1 was turned on and off. Interestingly, the second and third factors, although only accounting for 0.72% and 0.61% of the information contained in the data, show spikes in the score plots during the 'rise' and 'fall' time for the instrument, as the

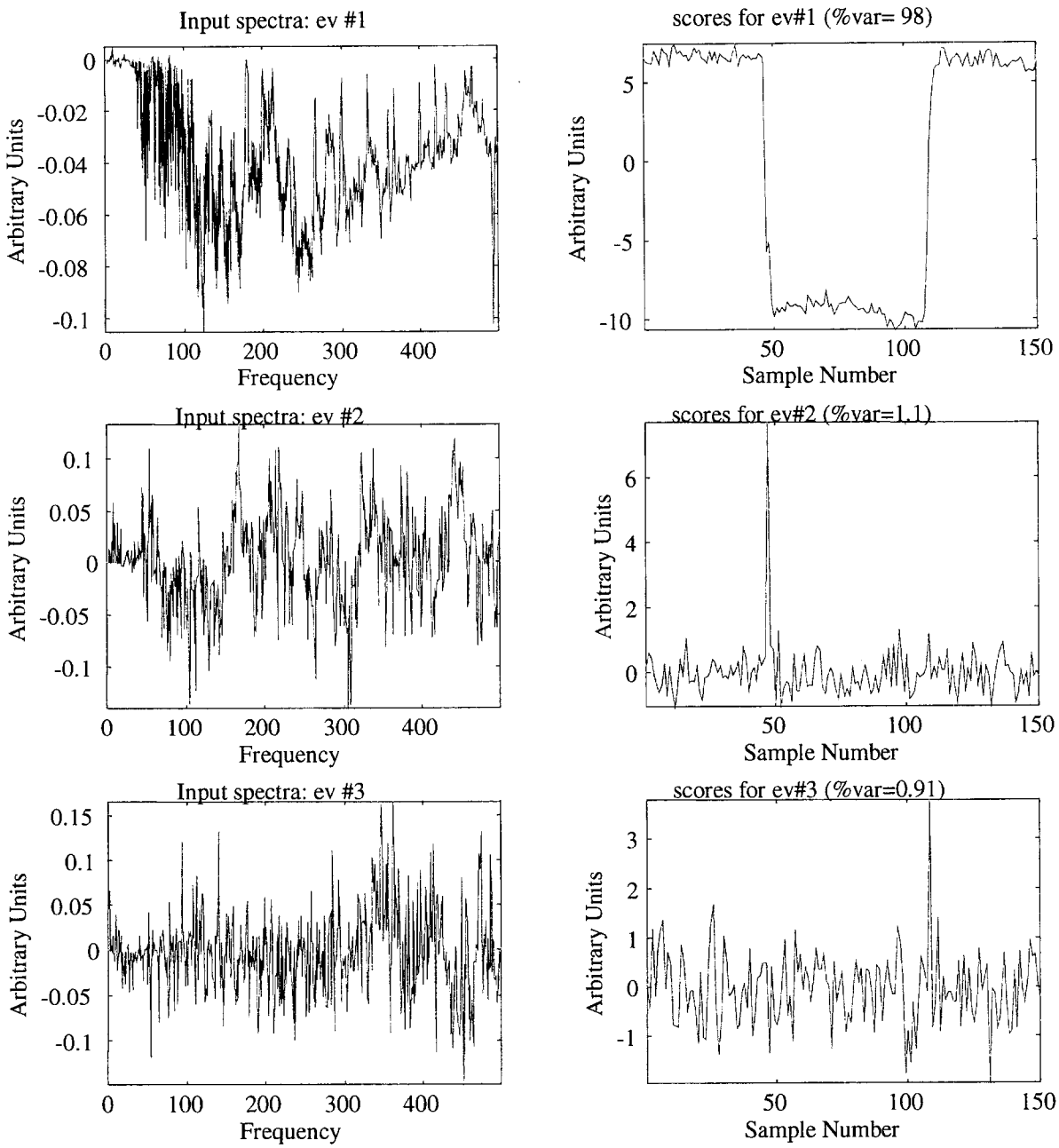

Figure 2 Loadings and Scores for TRAIN1_A1

instrument is being turned on and off. This may indicate that the signal seen during these transient periods is slightly different than that seen at equilibrium. Loading vector 1, noted as ev1 in Figure 2, provides an indication as to those frequencies that are important to identification of the signal from Instrument 1, in the presence of background signal only.

Combining the data from TRAIN1_A1 and TRAIN2_A1 provided the singular value plot shown in Figure 3. In this case, there appear to be two significant factors within the data. Examination of the first two loading and scores vectors for the combined data set (see Figure 4) shows scores for factor 1 that are increasing at those times when either Instrument 1 or Instrument 2 is turned on. Thus, this factor provides information regarding the average intensity of the signal throughout the entire data set. The scores for factor two are positive when Instrument 1 is on and negative when Instrument 2 is on, indicating that factor 2 is providing the difference between the two signals.

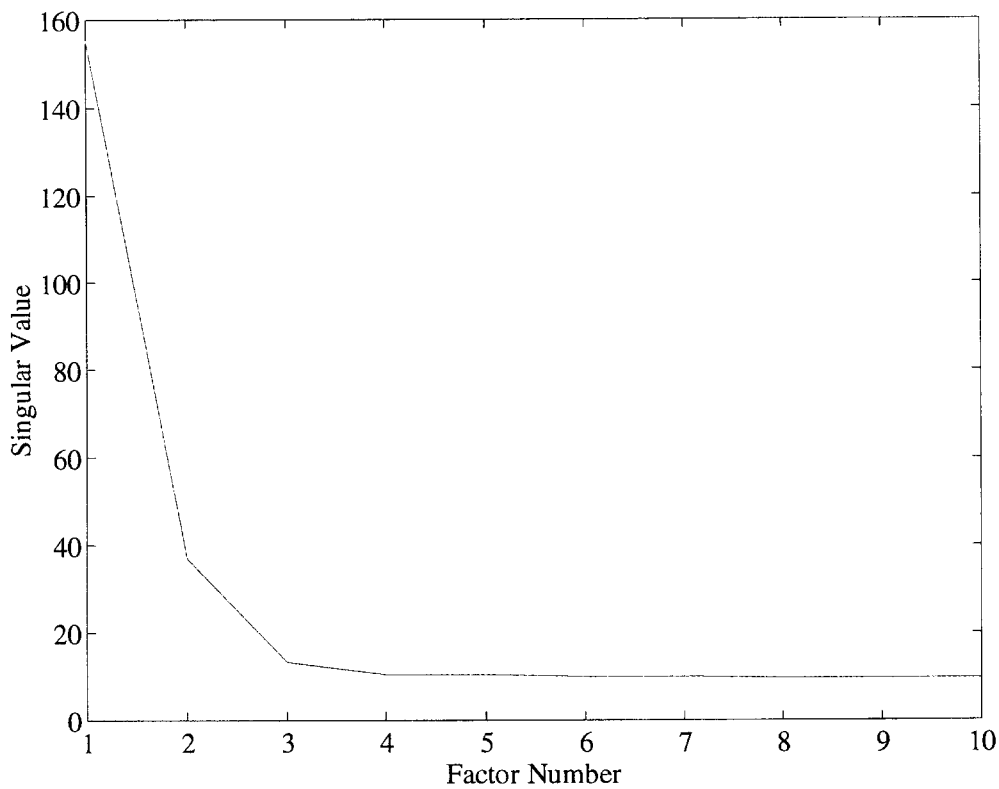

Figure 3 Singular value plot for the combined set, TRAIN1_A1 + TRAIN2_A1

Interpretation of scores and loadings plots is relatively straightforward for simple systems. As the number of signals increase and overlap, interpretation becomes much more difficult. For instance, TEST2_A1 contains a variety of signals, including

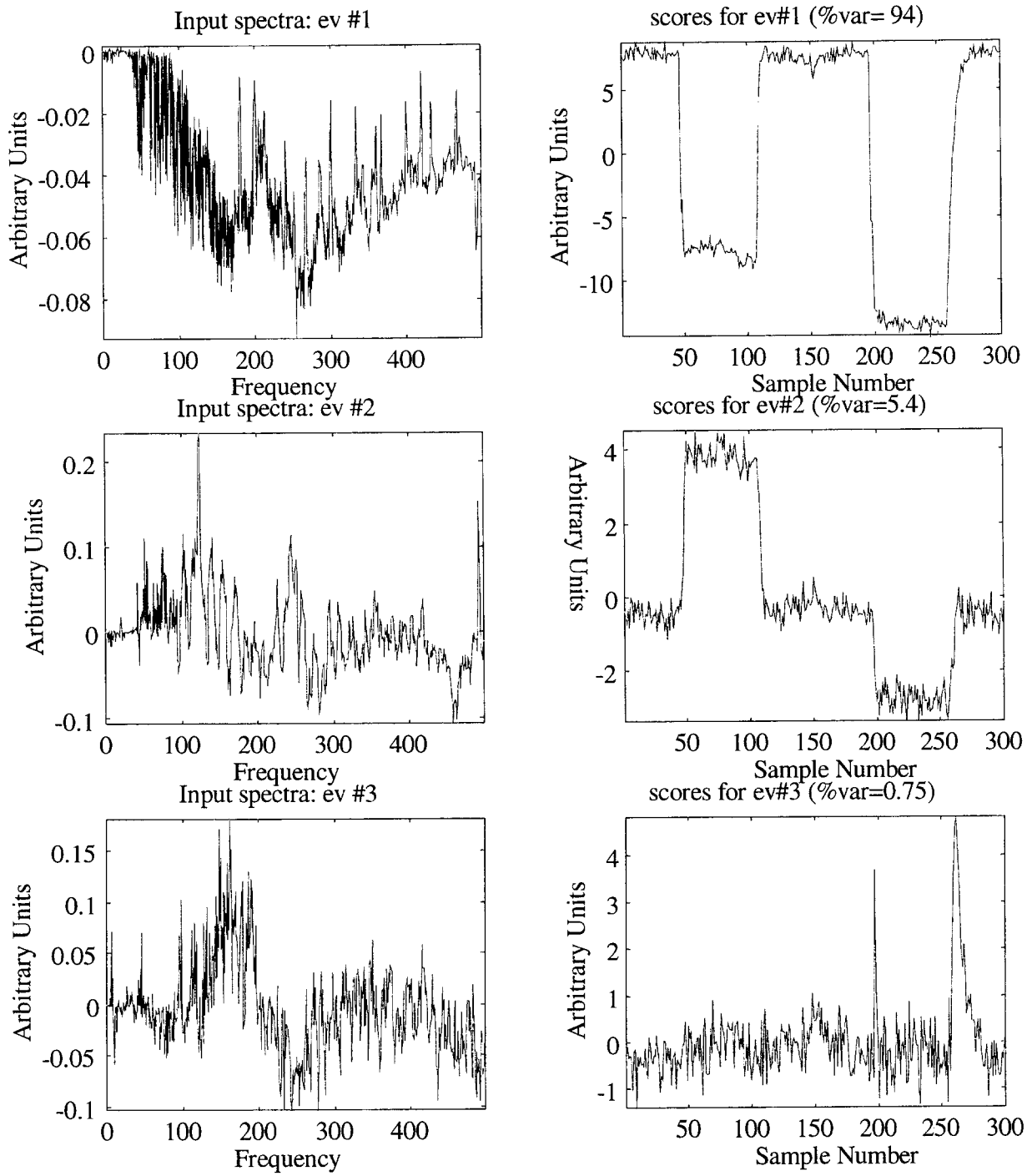

Figure 4 Loadings and Scores for Combined Set, TRAIN1_A1 + TRAIN2_A1

Instrument 1 and Instrument 2. Shown in Figure 5 is the singular value plot for the TEST2_A1 data set. Again, there is a slight increase in the variance explained by the higher factors, suggesting an increase in complexity. Figure 6 and Figure 7show the loadings and scores for the first six factors calculated from TEST2_A1. The scores for the first factor indicate the change in signal intensity as equipment is turned on or off. Interpretation of the scores plots becomes more difficult, however, since the experimental protocol is unknown. Without knowledge of the experiment, one cannot say with certainty whether each step corresponds to a piece of equipment being turned on, since it could be that as one piece of equipment is being turned on, another with similar characteristics is being turned off. Combined with predictive information however, the scores data can aid in interpretation. The other seismic data sets were similarly evaluated using PCA, but will not presented here.

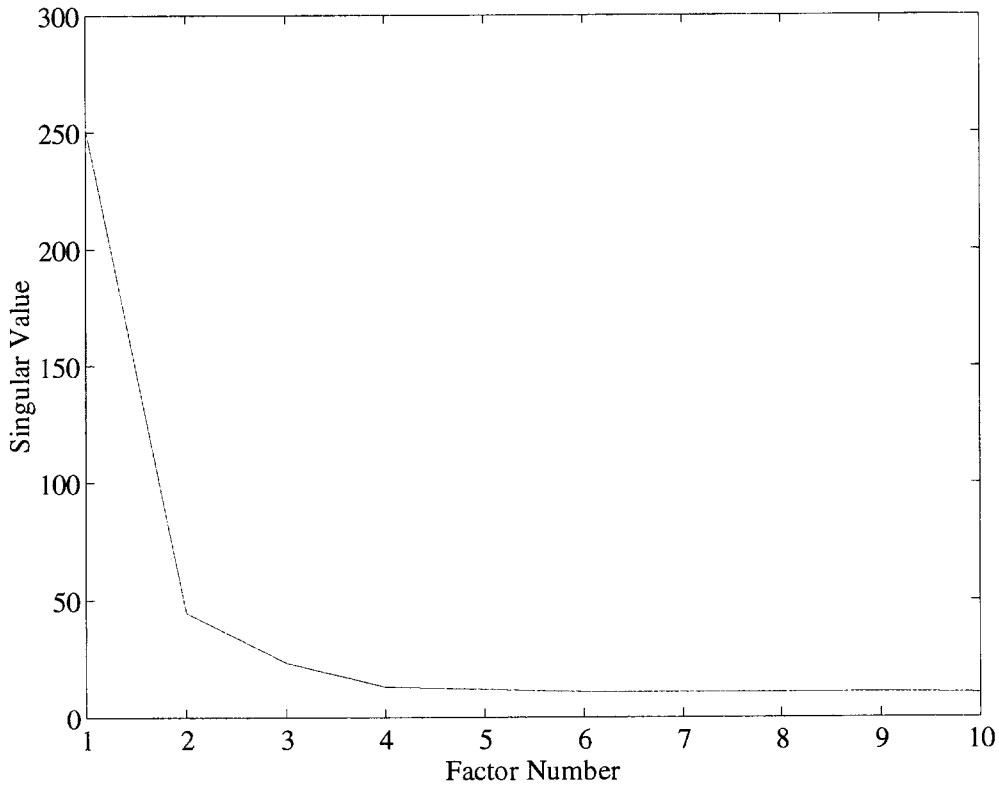

Figure 5 Singular values for TEST2_A1

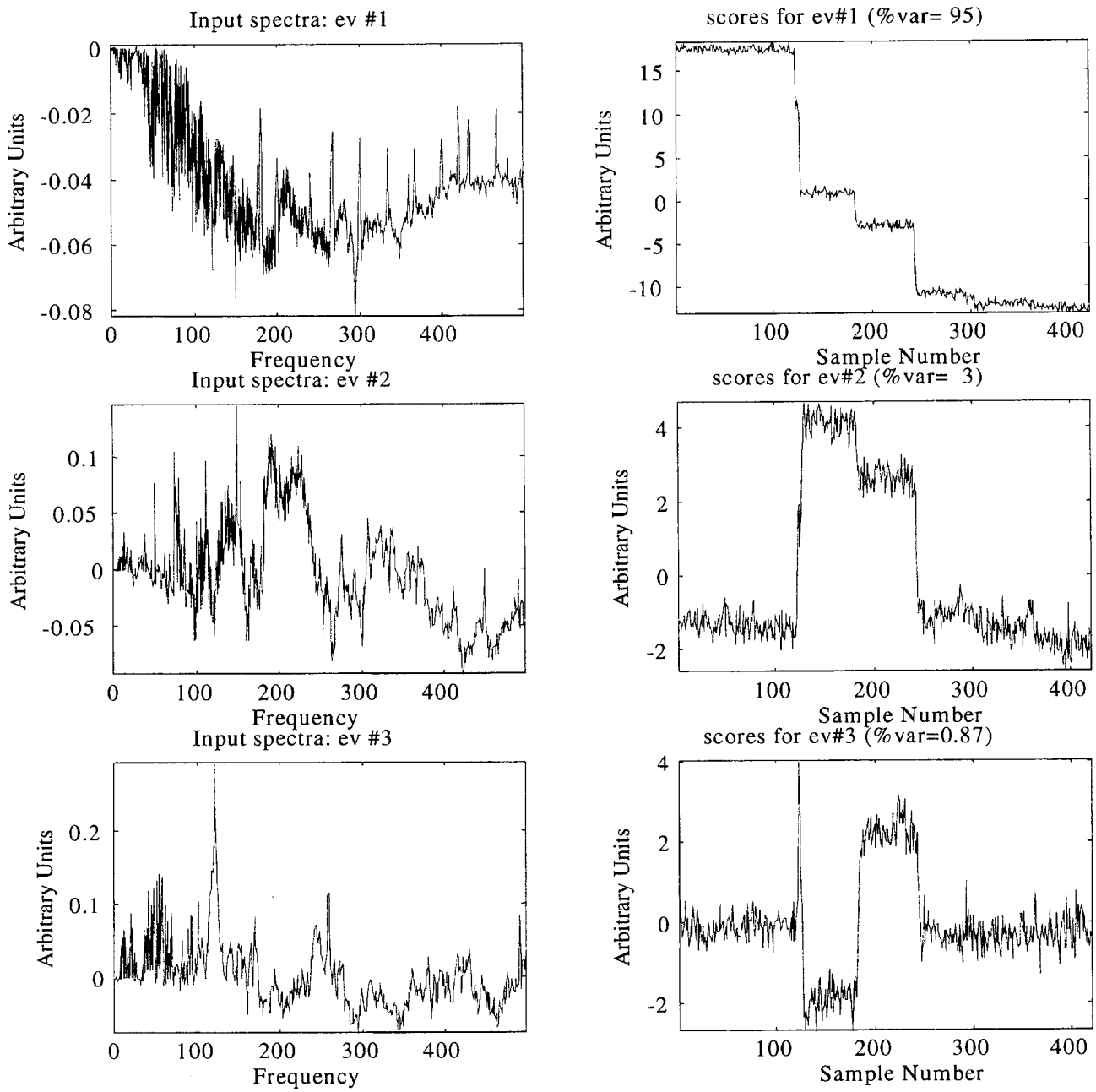

Figure 6 Loadings and scores for factors 1-3 of TEST2_A1.

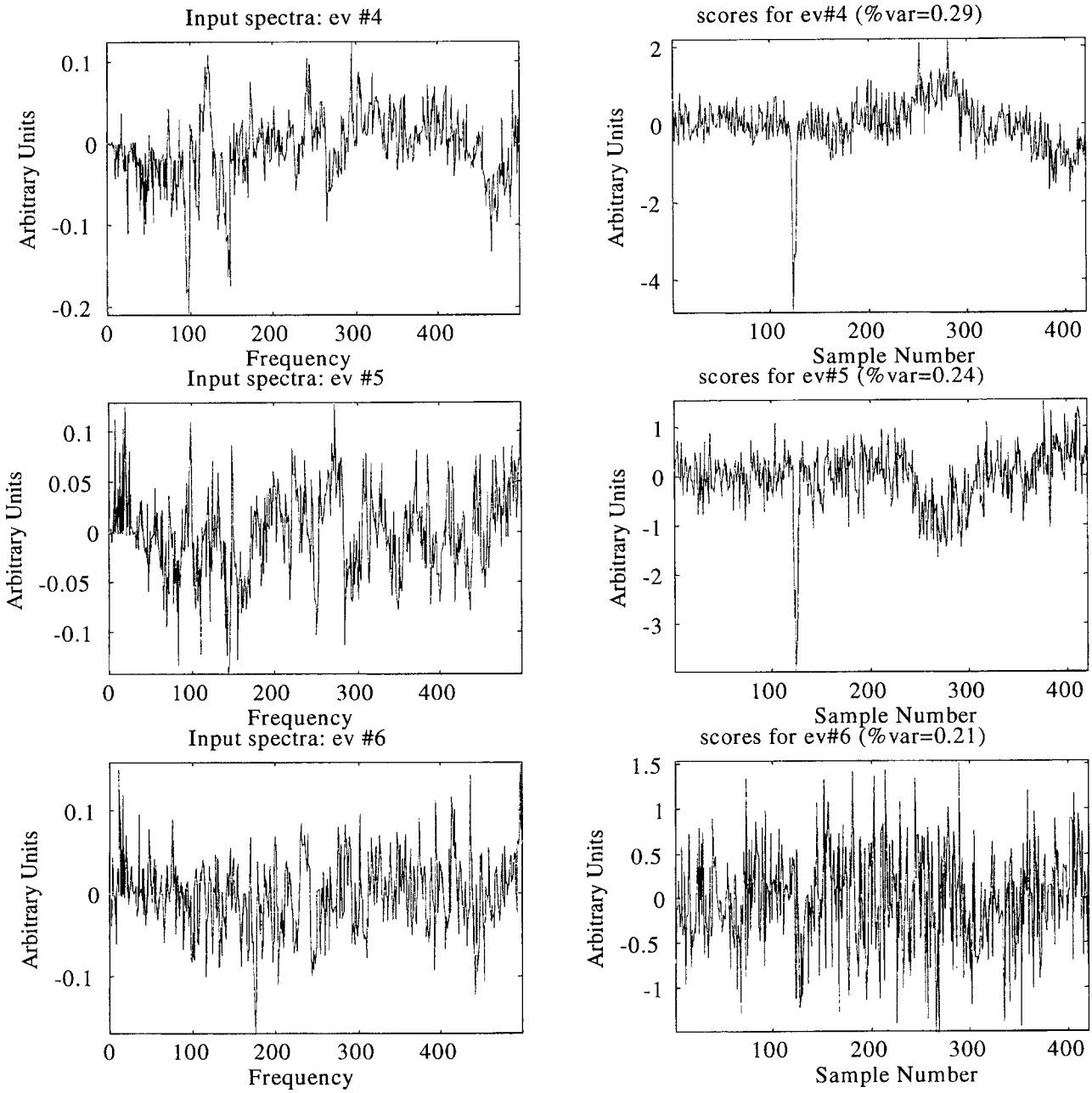

Figure 7 Loadings and scores for factors 4-6 of TEST2_A1.

PCR Models—TRAIN1_A1 + TRAIN2_A1

The combination of data from TRAIN1_A1 and TRAIN2_A1 was used for model building. Instrument status values were assigned as 0 for off, and 1 for on. Only data collected at steady state were used for model building, meaning data comprising the transitions from on state to another were removed from model building. Score plots were used to determine the appropriate data to remove. Single-sample cross validation was used to evaluate model size. Cross validation refers to a leave-one-sample-out procedure, such that each sample is removed sequentially and the remaining data is used to predict the removed sample. Samples that are dependent on one another should be removed as a group and predicted separately. If the samples removed during each cycle are independent, a reasonable estimate of predictive ability can be obtained by the cross validation method, using only calibration data. For the present data, the individual samples were not independent, however given that there were only two states for each instrument (off or on) within the data, cross validation pulling a state out at a time was not feasible. Single-sample cross validation for the present data may have provided models that overfit the calibration data. Final evaluation of the models was not made on the basis of the fit to the calibration data, rather the model evaluation was based on the model predictions of the test data.

The PCR model created for Instrument 1 based on TRAIN1_A1 + TRAIN2_A1 data contained 2 factors, and produced a correlation coefficient (R^2) of 0.98 between the predicted and reference calibration data. The model created for Instrument 2 also contained 2 factors and produced an R^2 of 0.99 between the predicted and reference calibration data. Using these models to predict TEST1_A1, TEST2_A1 and TEST3_A1 provided the time profile curves shown in Figure 8.

Prediction of TEST_A1 data shows an increase in the predicted signal for Instrument 1 immediately, reaching the 'on' state of 1 after approximately 3 samples. The predicted signal for Instrument 1 is noted to decrease at Sample 63, at the point in time where the predicted signal for Instrument 2 is increasing. It is unlikely that the drop in signal intensity to a value other than 1 or 0 is real behavior, since the instruments on/off signals were defined with two states, 1 and 0. The most probable explanation for the intensity change is that the model used to predict the TEST1_A1 data does not have

adequate information to predict Instrument 1 in the presence of Instrument 2. This also implies that the two signals are not strictly additive. The spectral F-ratios calculated from the prediction residuals of TEST1_A1 support this conclusion. Shown in Figure 9, the spectral F-ratios provide an indication of those samples that do not fit the calibration model. For TEST1_A1, a change in the size of the spectra F-ratio is noted at sample 63, indicating that those samples containing signals from both Instrument 1 and Instrument 2 are different from samples within the model.

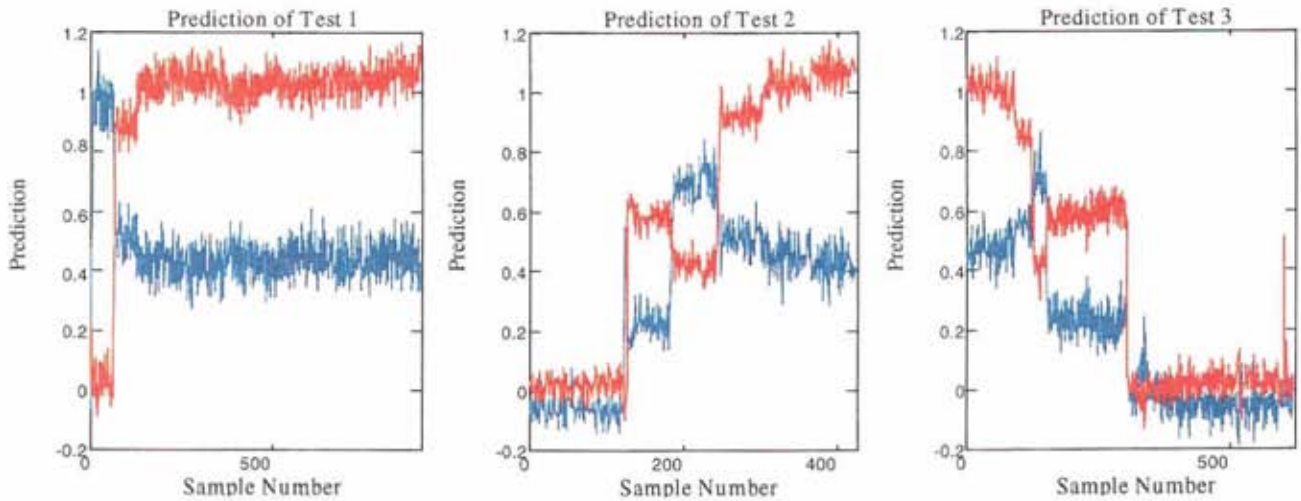

Figure 8 Predicted time profiles of the TEST data from station A1 using PCR model based on TRAIN1_A1 + TRAIN2_A1. Blue indicates prediction of Instrument 1; red indicates prediction of Instrument 2.

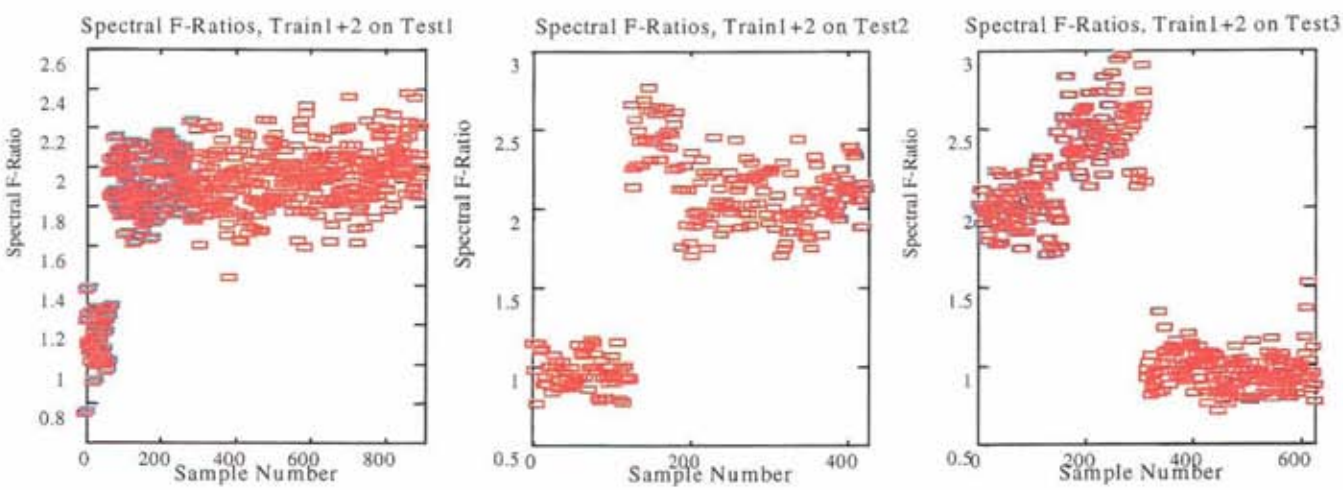

Figure 9 Spectral F-ratios calculated for each TEST set. TRAIN1_A1 + TRAIN2_A1 data was used for the model. Since model sizes for both instruments are the same (2 factors), the spectral F-ratios are identical for both instruments.

The prediction profiles for TEST2_A1 in Figure 8 show step changes to values other than 0 or 1. Again, this may mean that the model does not have enough information to predict all the states present in the TEST2_A1 data. Indeed, examination of the spectral F-ratios for the prediction of TEST2_A1 (Figure 9) indicates a large jump after sample 122, indicating dissimilarity between the model and those samples after 122. Prediction profiles for TEST3_A1 show step changes as well. Spectral F-ratios calculated from the residuals of TEST3_A1 indicate that samples 1-307 are different than those in the calibration.

PCR Robustification – Adding TEST1 samples

In order to improve the predictive ability of the calibration models, data from the test sets were added to the training data. Based on the predictive ability of the TRAIN1_A1 + TRAIN2_A1 models, an educated guess was made as to the instrument status of the test data. In particular, spectral data from TEST1_A1was chosen that appeared to contain signals from both instruments. Samples 800 through 903 from TEST1_A1 were added to the training data. The instrument status of both instruments was set to 1 for the added data. Using this 'robustified' data set, new PCR models were created for both instruments. Each model contained 3 factors, and provided R^2 values of 0.98 and 0.99 between the reference and predicted calibration values for Instrument 1 and Instrument 2 respectively. The increase in the number of factors used for the model indicates the added complexity of the model over that contained in the TRAIN1_A1 + TRAIN2_A1 model. The increase in the number of factors also suggests that the presence of the simultaneous signals within the data is not a simple addition of the individual signals. Predictions and spectral F-ratios using the TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 model are shown in Figure 10 and Figure 11 respectively. The prediction profiles for TEST1_A1 show maxima at 1.0. The predicted signal from Instrument 1 climbs immediately and reaches the maximum value after five samples. The predicted signal from Instrument 2 remains at zero until sample 63. Spectral F-ratios for the prediction residuals from TEST1_A1 are all at approximately the same level, indicating all predicted samples now appear to be within the model space.

Prediction profiles for TEST2_A1 also show a maximum at 1.0; however plateaus are also seen at 0.4 and 0.6 in both prediction profiles for samples 123-244. Spectral F-ratios calculated for the prediction residuals indicate samples 123-244 are different from those in the calibration model, suggesting the presence of new signals within this range. From the prediction plots for TEST2_A1, it appears that Instrument 1 and Instrument 2 are being turned on at 184 and 245 respectively. It is possible, however, that these signals are from other instruments with similar signals. Also notable in the prediction profile of Instrument 2 (red) are two additional steps at samples 303 and 365.

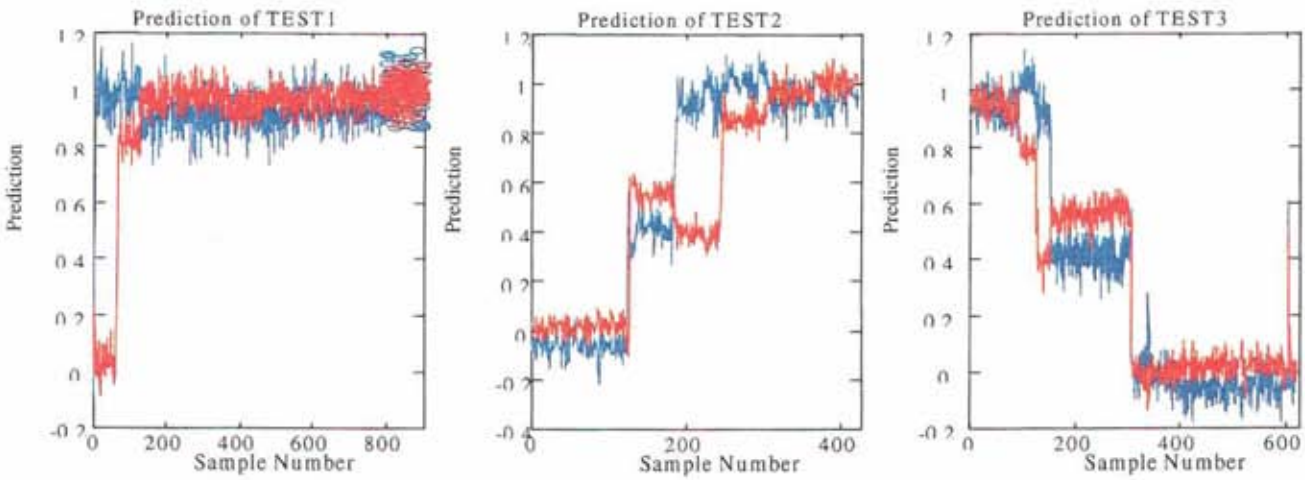

Figure 10 Prediction of TEST data from station A1 using training data 'robustified' with data from TEST1_A1. Blue indicates prediction of Instrument 1; red indicates prediction of Instrument 2. Circles in prediction profile of TEST1_A1 indicate those samples included in calibration model.

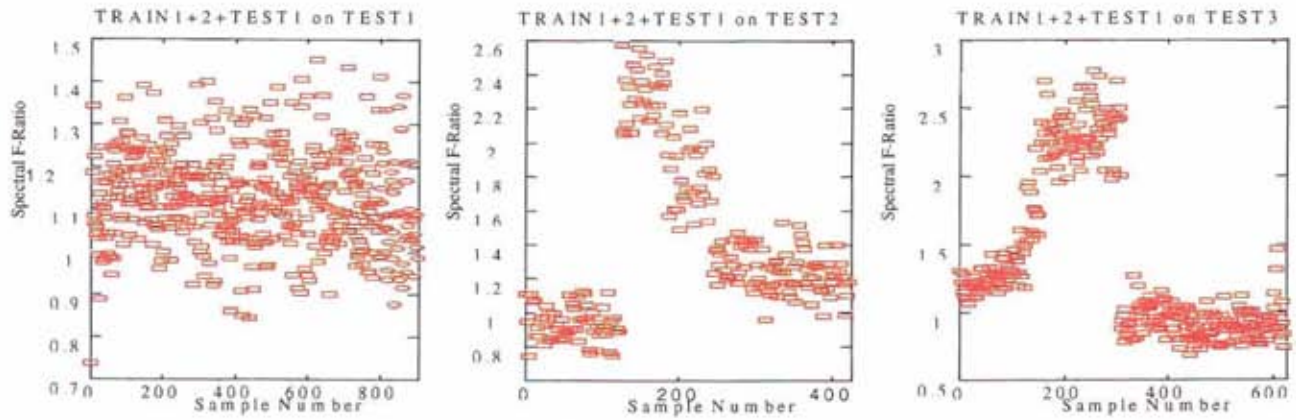

Figure 11 Spectral F-ratios calculated from the spectral residuals of TEST1_A1, TEST2_A1 and TEST3_A1 using a TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 model. Circles indicate those samples used in the calibration.

Prediction of TEST3_A1 shows similar behavior to that seen in the prediction of TEST2_A1. Instrument 1 and 2 are on at the beginning of the test and appear to be turned off at 153 and 124 respectively. Instead of reaching zero, the predicted signals hit at plateau of approximately 0.6. The signals remain at 0.6 until sample 307. At sample 307, both predicted signals go to zero. Spectral F-ratios calculated from the residuals of TEST3 indicate sample 140-307 to be different from those in the calibration.

PCR Robustification – Adding TEST1 and TEST2 samples

Further robustification of the calibration model was done using samples from TEST2_A1. Samples 140-160 and 210-230 were chosen from TEST2_A1 to add to the calibration model. Samples 140-160 were assigned the values 0 and 0 for both instruments (both off). Samples 210-230 were assigned the values 1 and 0, indicating Instrument 1 was on and Instrument 2 was off. These instrument values were estimated using the previous prediction values and spectral F-ratio values as a guide. The PCR model created for Instrument 1 based on TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 data contained 5 factors while the model created for Instrument 2 contained 3 factors. The R^2 values for the calibration data were 0.98 and 0.99 for Instrument 1 and Instrument 2 respectively. The increase in the number of factors for Instrument 1 may indicate a similarity between the new signals within the added TEST2_A1 data and the signal from Instrument 1, and the need for more factors to delineate the difference. Using these models to predict TEST1_A1, TEST2_A1 and TEST3_A1 provided the time profile curves shown in Figure 12. The spectral F-ratio values, shown in Figure 13, indicate the models are now explaining the predicted data adequately. The spectral F-ratios for TEST2_A1 show several samples that are significantly higher than the rest. These do not appear to be associated with spurious signals. Rather, these appear to be associated with the transition of an instrument from one state to another. It is very likely that these signals are different from those reached at steady state, and were not included in the model.

Using robustified model TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 to predict test data from different locations

The robustified model TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 was used to predict test data recorded on a different instrument at a different location. Only prediction of TEST2 data sets will be presented here. Prediction of TEST2_B1, TEST2_C1, TEST2_D1 and TEST2_B4 using the robustified A1 model are presented in Figure 14-17 along with the resulting spectral F-ratio values.

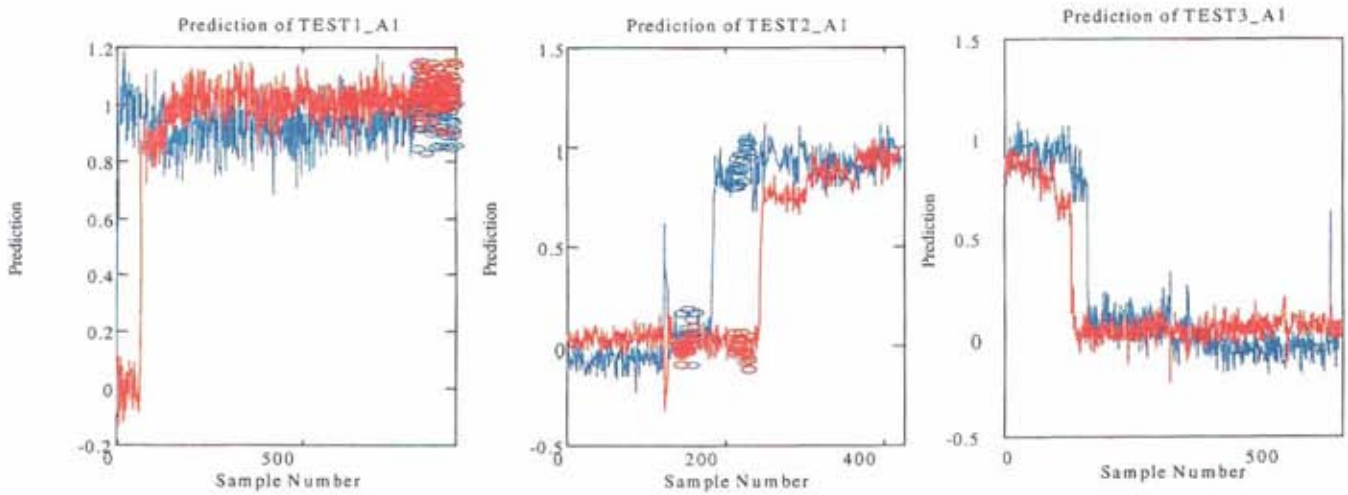

Figure 12 Prediction of TEST data from A1 station using TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 model. Circles indicate those samples used in the calibration.

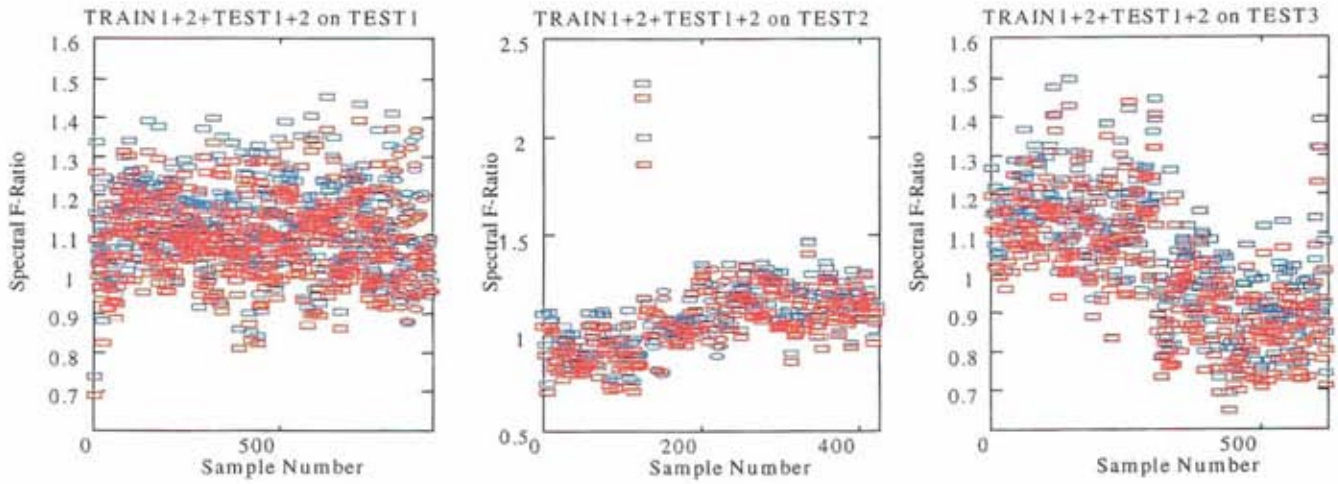

Figure 13 Spectral F-ratios calculated from prediction of TEST data from A1 station using TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 model.

Prediction of data from the different locations using the A1 model is marginal, at best. Most of the spectral F-ratios for the predicted data are higher than was seen with the A1 model predicting the A1 test data, suggesting the data collected at different sites is dissimilar to that collected at A1. If the signals collected at the different sites differed from the A1 data by only a scalar value (i.e., decrease in intensity), it would be expected that the prediction profiles would be similar in shape to those seen when predicting the A1 data. In addition, the resulting spectral F-ratios would be similar to those seen after the prediction of A1 data by an A1 model. As this is not the case, the data collected from the sites other than A1 appears to be inherently different than that collected from A1. These differences could result from environmental parameters local to the specific detector or differences in the detectors (e.g. sensitivity, response)

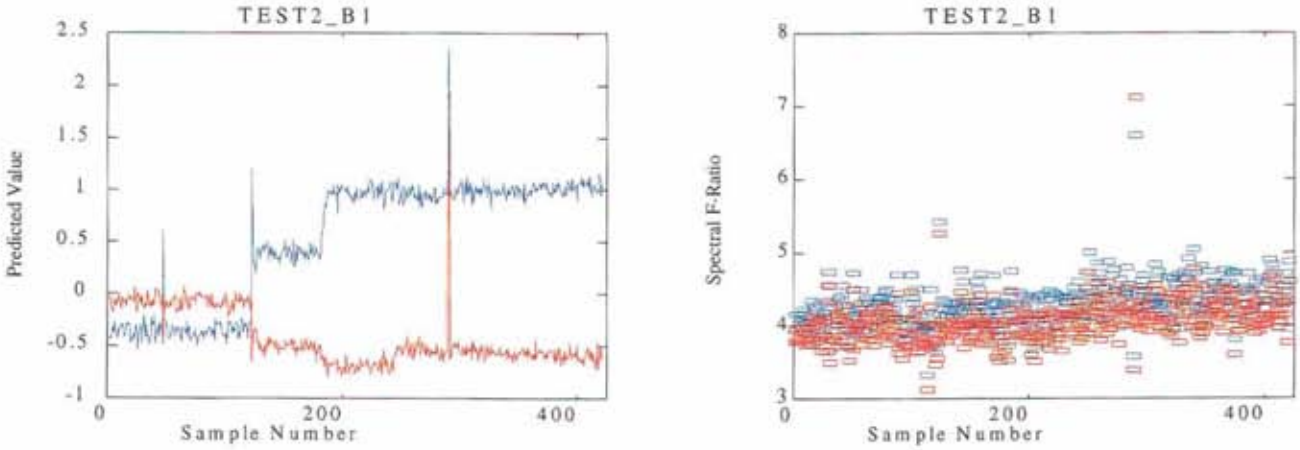

Figure 14 Prediction of TEST2_B1 data using TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 model

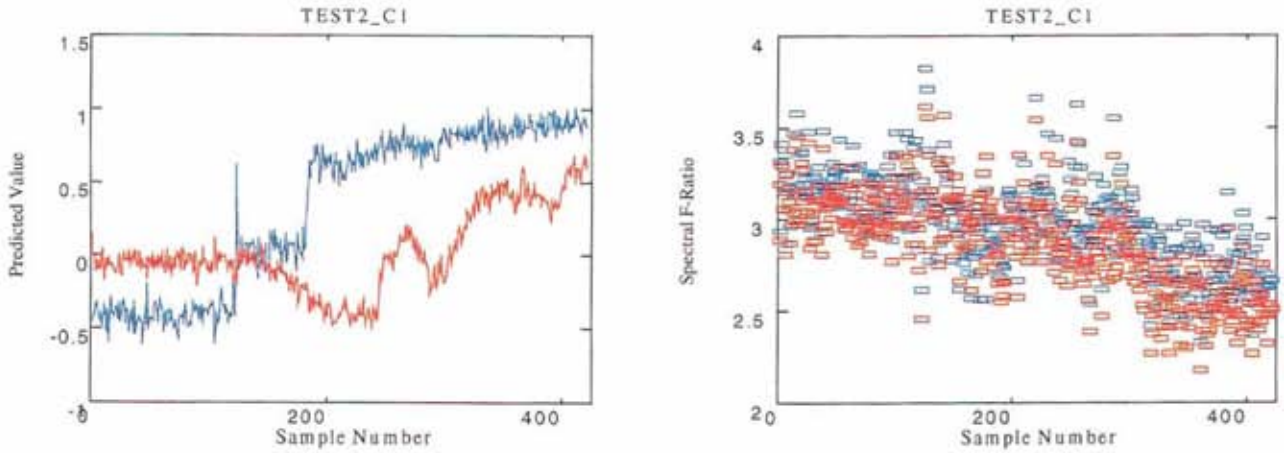

Figure 15 Prediction of TEST2_C1 data using TRAIN1_A1 + TRAIN2_A1 + TEST1_A1 + TEST2_A1 model

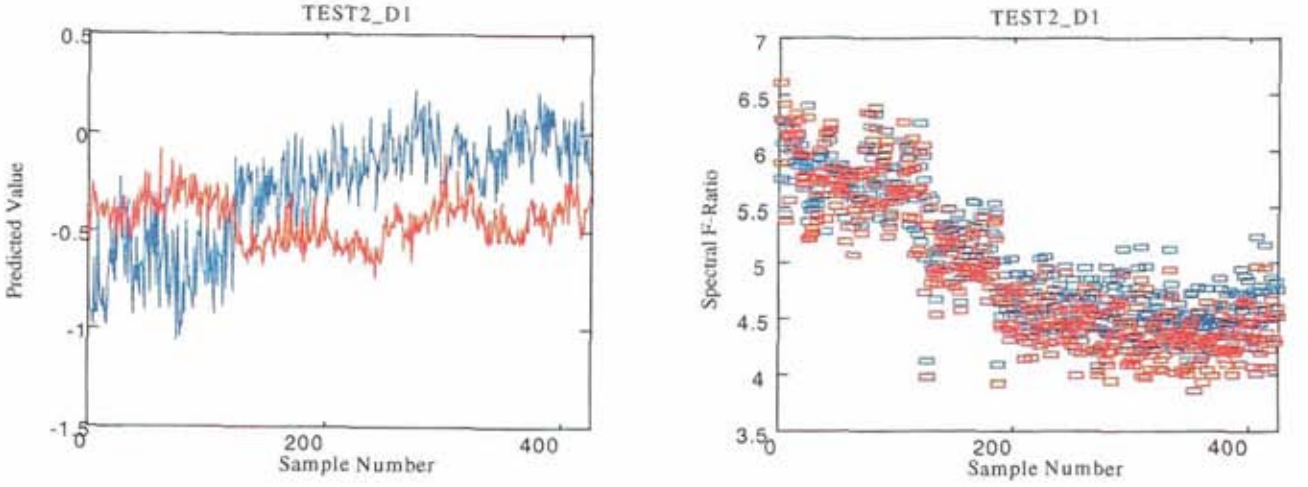

Figure 16 Prediction of TEST2_D1 data using TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 model

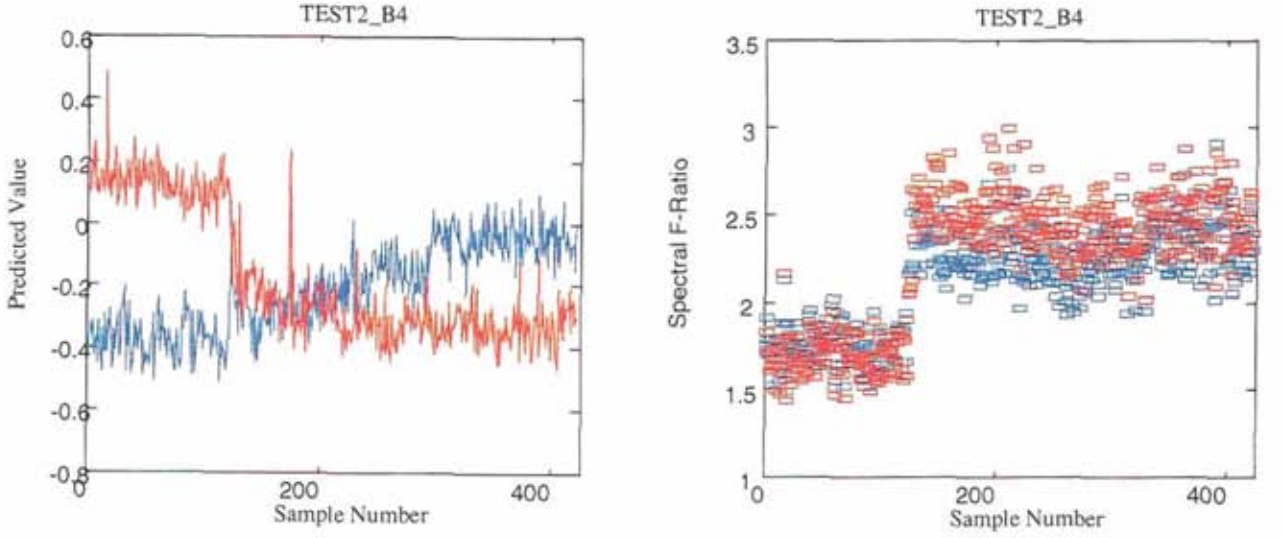

Figure 17 Prediction of TEST2_B4 data using TRAIN1_A1+TRAIN2_A1+TEST1_A1+TEST2_A1 model

Specific Site Models Predicting TEST2 Data

New models were created using B1, C1, D1 and B4 data from TRAIN1, TRAIN2, TEST1 and TEST2 files. Samples used from each set were equivalent to samples used in the A1 model in terms of collection time. These site-specific models were then used to predict TEST2 data from the same site. The quality of the models was noted to be dependent on the location of the sensor, as it was reflected in the R^2 value between the predicted and reference calibration data for each instrument. Table 2 lists the model, the number of factors and the R^2 between the reference values and the

predicted values for the cross validated model. For comparison, the same values are listed for the A1 model as well.

Table 2

Data used for Model	# of Factors, Inst. 1/ # of Factors, Inst. 2	R^2 Instrument 1/R^2 Instrument 2
A1	5/3	0.98/0.99
B1	9/5	0.95/0.97
C1	4/6	0.92/0.91
D1	9/11	0.80/0.85
B4	4/4	0.57/0.53

The model for the B4 data is extremely poor. It is believed that the B4 data have a signal to noise ratio too low to extract good models.

The prediction profiles for the appropriate set of TEST2 data using the models B1, C1 and D1 are shown along with the spectral F-ratios in Figures 18- 20. Prediction profiles of the TEST2_B1 and TEST2_C1 data using the B1 and C1 models respectively, resemble the prediction profiles of the A1 data using A1 models, indicating data from these sensors had adequate signal to noise for calibration and prediction. The prediction profiles for D1 are quite noisy, although some suggestion of a rise and fall of signal is present. Predictions using the B4 model are not shown, due to the lack of a good model.

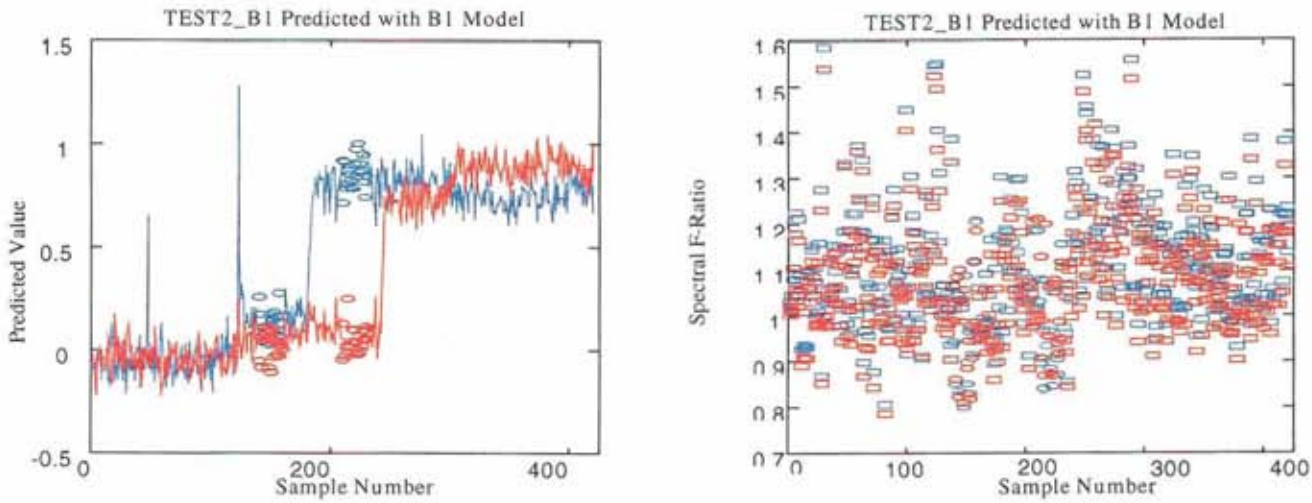

Figure 18 Prediction of the TEST2_B1 data using a TRAIN1_B1 +TRAIN2_B1 + TEST1_B1 + TEST2_B1 model. Circles indicate data points used in the calibration.

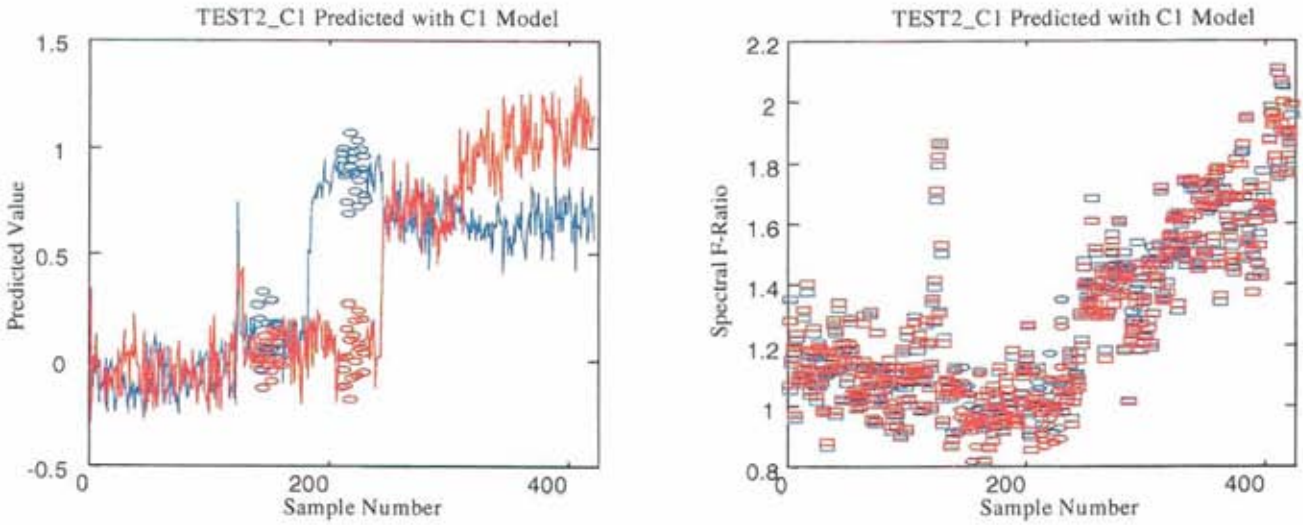

Figure 19 Prediction of the TEST2_C1 data using a TRAIN1_C1+TRAIN2_C1+TEST1_C1+TEST2_C1 model. Circles indicate data points used in the calibration.

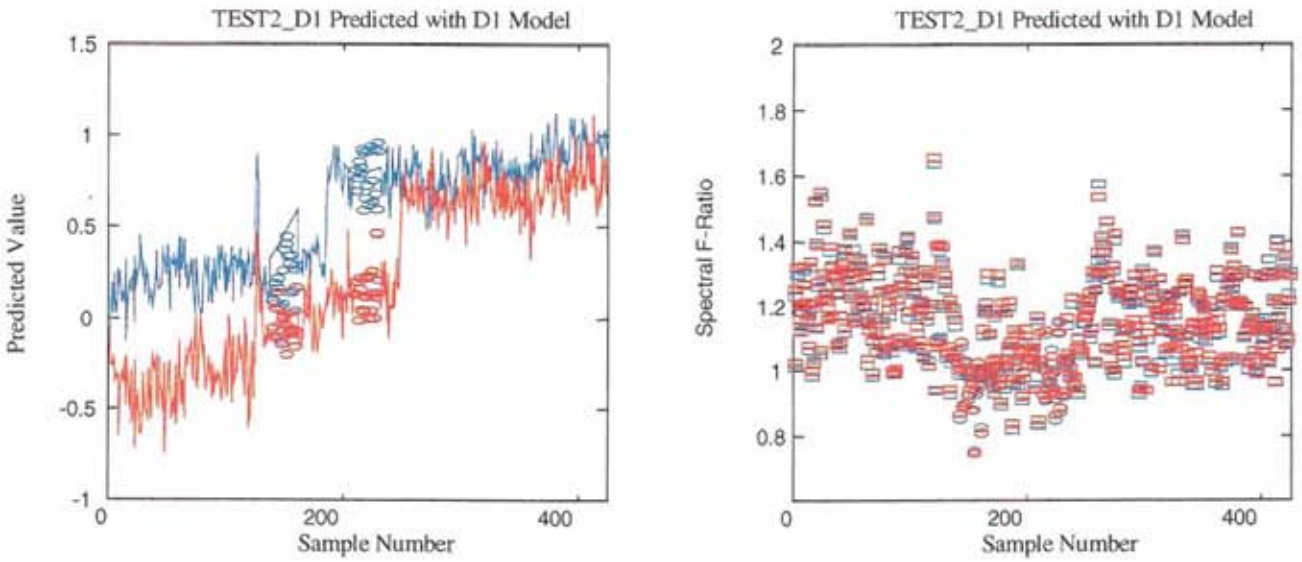

Figure 20 Prediction of the TEST2_D1 data using a TRAIN1_D1 + TRAIN2_D1 + TEST1_D1 + TEST2_D1 model. Circles indicate data points used in the calibration.

Summary/Conclusions

Multivariate data analysis of frequency data obtained from a field test of seismic sensors was performed. The goal of the field test was to evaluate whether seismic signatures could be used to determine when designated pieces of equipment were being turned on or off. Using the multivariate technique Principle Components Regression

(PCR), models were developed for two pieces of equipment using training data collected as the equipment was cycled on and off.

Data was collected from five different sensors: A1, B1, C1, D1 and B4, corresponding to data containing decreasing signal to noise. Models created using TRAIN1_A1 and TRAIN2_A1 data had difficulty in predicting data that contained additional signals over those contained in TRAIN1_A1 and TRAIN2_A1. Spectral F-ratios calculated using the residuals of the predicted data providing insight as to those points in which new or different signals were present. Addition of data comprising the new signals created more robust models as evaluated by the decrease in the number of sample points predicted to be values other than 0 or 1.

A new A1 model was created using additional data from TEST1_A1 and TEST2_A1. This robustified model was used to predict A1-B4 TEST2 data. While the prediction of TEST2_A1 was good, the prediction of B1-B4 TEST2 was poor. The spectral F-ratios for B1-B4 TEST2 data were generally high (>3), suggesting the data obtained from the B1, C1, D1 and B4 sensors contain new or inherently different signals as compared to the data collected from the A1 sensor. Using data from the B1, C1, D1 and B4 sensors to create new models, better prediction results of the B1, C1, and D1 TEST2 data were obtained. Good models could not be created using data from the B4 sensor, suggesting that the data contain signal to noise too low for model building.

This preliminary investigation into the application of PCR to seismic data resulted in prediction data that looks promising. Application of other multivariate methods or variation of PCR may yield further improvements. For instance, PCR Augmentation, the method of adding unknown data to the calibration solely for the definition of the loading vectors, may provide an approach to adding unknown signals to the calibration without having to know the instrument status of the unknown samples. Calibration transfer methods, developed for transferring models from one instrument to another, are a possible technique to be employed for using the A1 models to predict the data from sensors at other locations.

Accounting for the transition status between on to off may also improve calibration and predictions. Once the transition states are included in the calibration,

Partial Least Squares (PLS) methods may offer an advantage over PCR calibration methods.

Finally, prudent selection of calibration samples may also improve calibration and prediction. The models presented above used all the calibration data from TRAIN1 and TRAIN2 and thus were heavily weighted toward data from those sets. Balancing the data from TRAIN1, TRAIN2, TEST1 and TEST2 such that each instrument state was weighted equally could provide improvement in the prediction of those states less heavily weighted in the original model.

Distribution List:

1	MS	0859	Gerard Sleefe
1		0859	Mark Ladd
1		0750	Gregory Elbring
2		0342	Kathleen Alam
1		9018	Central Technical Files, 8940-2
2		0899	Technical Library, 4916
2		0619	Review & Approval Desk, 00111 For DOE/OSTI

References

Geladi, P. and B. Kowalski (1986). "Partial Least Squares Regression: A Tutorial." Analytica Chimica Acta **185**: 1-17.

Haaland, D. M. and E. V. Thomas (1988). "Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information." Analytical Chemistry **60**: 1193-1202.

Strang (1980). Linear Algebra and Its Applications. London, Academic Press.

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. International Symposium on Multivariate Analysis. P. R. Krishnaiah. Dayton, Academic Press: 391-420.

Wold, S., K. Esbensen, et al. (1987). "Principle Component Analysis." Chemometrics and Intell. Lab. Systems **2**: 37-52.