OAK RIDGE
NATIONAL LABORATORY
MANAGED BY UT-BATTELLE
FOR THE DEPARTMENT OF ENERGY

ORNL/TM-2000/5

# A Vector Approach to Regression Analysis and Its Implications to Heavy-Duty Diesel Emissions

**November 2000**

**Prepared by**

**H. T. McAdams**
**AccaMath Services**
**Carrolton, Illinois**

**R. W. Crawford**
**R.W. Crawford Energy Systems**
**Tucson, Arizona**

**G. R. Hadder**
**Oak Ridge National Laboratory**
**Oak Ridge, Tennessee**

UT-BATTELLE

ORNL-27 (4-00)

ORNL/TM-2000/5

# A VECTOR APPROACH TO REGRESSION ANALYSIS AND ITS APPLICATION TO HEAVY-DUTY DIESEL EMISSIONS

**H. T. McAdams**
AccaMath Services
Carrollton, Illinois

**R. W. Crawford**
RWCrawford Energy Systems
Tucson, Arizona

**G. R. Hadder**
Oak Ridge National Laboratory
Oak Ridge, Tennessee

**November 2000**

Prepared for
Office of Energy Efficiency and Renewable Energy
Office of Fossil Energy
Office of Policy
U.S. Department of Energy

Prepared by the
**OAK RIDGE NATIONAL LABORATORY**
Oak Ridge, Tennessee 37831–6472
Managed by
**UT-BATTELLE, LLC**
for the
**U.S. DEPARTMENT OF ENERGY**
under contract DE-AC05-00OR22725

# TABLE OF CONTENTS

LIST OF FIGURES ..... v

LIST OF TABLES ..... v

ACRONYMS AND ABBREVIATIONS ..... vii

ACKNOWLEDGMENTS ..... ix

ABSTRACT ..... xi

EXECUTIVE SUMMARY ..... xiii

1. INTRODUCTION ..... 1

1.1 STATISTICAL PERSPECTIVE ..... 1
1.2 OVERVIEW OF THE VECTOR APPROACH ..... 2
1.3 DIESEL EMISSION CONTROL ..... 7
1.4 ORGANIZATION OF REPORT ..... 7

2. THE VECTOR METHODOLOGY ..... 9

2.1 DESCRIPTION OF DATA BASE ..... 9
2.2 COMPUTER SOFTWARE ..... 11
2.3 THE VECTOR METHODOLOGY ..... 11
2.3.1 Vector Approach to Representing Fuels ..... 11
2.3.2 Use of Eigenfuels in Regression Analysis ..... 18
2.3.3 Application to Diesel Emissions ..... 20
2.4 MODEL INTERPRETATION AND SIMPLIFICATION ..... 26

3. PERSPECTIVE AND EXTENSIONS ..... 33

3.1 PERSPECTIVE ON SUMS OF SQUARES ..... 33
3.1.1 The Many Faces of Sums of Squares ..... 33
3.1.2 Model Sums of Squares as Induced Distributions ..... 38
3.2 IMPLICATIONS FOR THE METHODOLOGY ..... 40
3.2.1 Hypothesis Testing: t versus F ..... 40
3.2.2 Statistical Inference ..... 41
3.3 METHODOLOGICAL EXTENSIONS ..... 43
3.3.1 Comparing Independent Studies ..... 43
3.3.2 Robustness of the Eigenvector Basis ..... 45
3.3.3 Representation of Non-linear Effects ..... 47
3.3.4 Transformation as an Approach to Non-Linearity ..... 48

4. TOWARD AN IMPROVED UNDERSTANDING OF FUELS AND EMISSIONS ..... 51

4.1 IMPROVED THEORY ..... 51
4.2 NEW INSIGHT ON FUELS FORMULATION ..... 52
4.3 MORE AND BETTER ENGINE TESTING DATA ..... 55

5. CONCLUSIONS ..... 57

6. REFERENCES ..... 59

APPENDIX A: Database Development ..... 65

APPENDIX B: MatLab Software ..... 71

APPENDIX C: Theory Underlying the Vector Approach ..... 77

APPENDIX D: Simplification of Vector-based Regression Models ..... 87

APPENDIX E: Tests of Statistical Significance ..... 105

APPENDIX F: Comparing the Results of Independent Studies ..... 117

APPENDIX G: Bootstrap Sampling ..... 135

APPENDIX H: Incorporating Non-Linear Terms in a Vector Model ..... 151

APPENDIX I: Transformations as an Approach to Non-Linear Regression ..... 165

# LIST OF FIGURES

Figure 1.1. Interdependence of Diesel Fuel Properties . . . . . 3
Figure 1.2. Comparative Performance of Regression Models . . . . . 5
Figure 1.3. Significance versus Substantiality of Effects (PM Emissions Model) . . . . . 6
Figure 2.1. Distribution of Eigenfuel Coefficients . . . . . 17
Figure 2.2. Sources of Variance in Diesel Engine Emissions Data . . . . . 21
Figure 2.3. Percentage Contributions to Variance Explanation for Fuels and Emissions . . . . . 24
Figure 3.1. Lattice Representation of SS for Three Variables . . . . . 35
Figure 3.2. Sums of Squares for Subset Models in a Three-Variable Space . . . . . 37
Figure 4.1. Fuels Expressing Eigenvectors 9 through 12 . . . . . 53

# LIST OF TABLES

Table 2.1. The HDD Emissions Database . . . . . 10
Table 2.2. Correlation Matrix for Fuel Properties . . . . . 13
Table 2.3. Eigenvectors of the Test Fuels Data Set . . . . . 13
Table 2.4. Eigenfuel Representation for a Selected Fuel . . . . . 16
Table 2.5. $NO_x$ Regression Using Eigenfuels as Explanatory Variables . . . . . 19
Table 2.6. Sum of Squares for Fuels Adjusted for Engine Effects . . . . . 22
Table 2.7. Summary of Regression Results for $NO_x$ and PM . . . . . 23
Table 2.8. Statistical Significance and SS Partitioning . . . . . 27
Table 2.9. SS Partitioning Based on Simplified $NO_x$ and PM Models . . . . . 29
Table 2.10. Revised Eigenvectors of the Test Fuels Data Set . . . . . 30
Table 2.11. Statistical Significance and SS Partitioning for the Revised $NO_x$ and PM Models in Seven-Variable Space . . . . . 31
Table 3.1. Demonstration Data Set . . . . . 34
Table 3.2. Estimation of Sums of Squares by All Possible Regressions . . . . . 36
Table 3.3. Partitioning of Model SS Among Variables and Eigenvectors . . . . . 38

# ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of Variance |
| CetImprv | Cetane Improvement (number), due to additives |
| CO | Carbon Monoxide |
| COV | Coefficient of Variation |
| DF | Degrees of Freedom |
| DOE | U.S. Department of Energy |
| Eigenfuel | Eigenvectors composed of fuel properties |
| EPA | U.S. Environmental Protection Agency |
| E-Space | Eigenvectors used as explanatory variables |
| FBP | Final Boiling Point |
| GLM | General Linear Model |
| HC | Hydrocarbons |
| HDD | Heavy Duty Diesel |
| HDEWG | Heavy Duty Engine Working Group |
| IBP | Initial Boiling Point |
| ID | Interdependency (equations) |
| MonoArom | Mono-aromatics content |
| MS | Mean Square |
| NatCetane | Natural Cetane (number) |
| $NO_x$ | Nitrogen Oxides |
| OLS | Ordinary Least Squares |
| ORNL | Oak Ridge National Laboratory |
| PCA | Principal Components Analysis |
| PCR | Principal Components Regression |
| PCR+ | Principal Components Regression Plus |
| PM | Particulate Matter |
| PolyArom | Poly-aromatics content |
| P-Space | Fuel properties use as explanatory variables |
| RYM | Refinery Yield Model, maintained by ORNL |
| SAE | Society of Automotive Engineers |
| SS | Sum of Squares |
| svd | Singular Value Decomposition |
| SWRI | Southwest Research Institute |
| T10 | Ten percent evaporation temperature |
| T50 | Fifty percent evaporation temperature |
| T90 | Ninety percent evaporation temperature |

# ACKNOWLEDGMENTS

This research, performed under contract with the Energy Division of Oak Ridge National Laboratory (ORNL), was sponsored by the U.S. Department of Energy (DOE) Offices of Energy Efficiency and Renewable Energy, Fossil Energy, and Policy. ORNL is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. DOE. The authors wish to acknowledge the guidance of Barry McNutt of the U.S. DOE Office of Policy. The authors also wish to acknowledge the assistance of K. G. Duleep of Energy and Environmental Analysis, Inc. (EEA) in providing portions of the data used in this study.

# ABSTRACT

An alternative approach is presented for the regression of response data on predictor variables that are not logically or physically separable. The methodology is demonstrated by its application to a data set of heavy-duty diesel emissions. Because of the covariance of fuel properties, it is found advantageous to redefine the predictor variables as *vectors*, in which the original fuel properties are *components*, rather than as *scalars* each involving only a *single* fuel property. The fuel property vectors are defined in such a way that they are mathematically independent and statistically uncorrelated.

Because the available data set does not allow definitive separation of vehicle and fuel effects, and because test fuels used in several of the studies may be unrealistically contrived to break the association of fuel variables, the data set is not considered adequate for development of a full-fledged emission model. Nevertheless, the data clearly show that only a few basic patterns of fuel-property variation affect emissions and that the number of these patterns is considerably less than the number of variables initially thought to be involved. These basic patterns, referred to as "eigenfuels," may reflect blending practice in accordance with their relative weighting in specific circumstances.

The methodology is believed to be widely applicable in a variety of contexts. It promises an end to the threat of collinearity and the frustration of attempting, often unrealistically, to separate variables that are inseparable.

# EXECUTIVE SUMMARY

Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several predictor variables. In spite of its evident success in many applications, the regression approach can face serious difficulties when the predictor variables are to any appreciable extent covariant. This point was made quite evident in a recently published review, which found that efforts to evaluate the separate effects of fuel variables on the emissions from heavy-duty diesel (HDD) engines were often frustrated by the close association of fuel properties. This report addresses these concerns by offering a new approach to modeling the effects of fuel characteristics on emissions.

The work was motivated by the observation that most HDD engine research was conducted with test fuels that had been "concocted" in the laboratory to vary selected fuel properties in isolation from each other. This approach can eliminate the confounding effect caused by naturally covarying fuel properties, but it departs markedly from the real world, where the reformulation of fuels to reduce emissions will naturally and inevitably lead to changes in a series of interrelated properties. What impact might this method of blending test fuels have on their ability to provide an accurate and reliable basis for assessing the emissions performance of future diesel fuels?

## Development of a New Statistical Methodology

The approach presented here is based on the use of Principal Components Analysis (PCA) to describe fuels in terms of vector quantities called *eigenfuels*. Each eigenfuel represents a unique and mathematically independent characteristic of diesel fuel, and the most important eigenfuels can be related to the refinery and blending processes used in creating the fuels. When applied as predictors for emissions in regression analysis, eigenfuels are found to have many advantages, including:

- Simplification of the analysis, because their mathematical independence eliminates correlations among the variables and the complications introduced by multi-collinearity.
- Economy of representation, because a small number of such vector variables may effectively replace a larger number of the original variables.
- Greater understanding of the patterns of variation that are important to emissions, and how these patterns relate to fuel blending and refinery processes.
- Potentially new insight into the optimal formulation of fuels to reduce emissions, and improved experiment design for the estimation of fuel effects.

The natural covariance of fuel properties – a confounding factor for the original fuel properties – becomes a strength associated with realism and efficiency in the eigenfuel approach.

## Key Findings Regarding Diesel Fuels and Emissions

A database of HDD engine testing was compiled from the literature and used to demonstrate the methodology, recognizing that the existing data are inadequate to answer fully the many questions related to the effect of fuels on emissions. Within this limitation, the study suggests that the eigenfuel approach may lead to a new perspective on the diesel fuel-emissions relationship:

- *Fuel properties are only surrogate variables for underlying causal factors.* Much of the emissions reduction seen in past testing comes from reducing the proportion of high-aromatic cracked stocks in diesel fuels. Because these stocks are low in cetane and high in density, researchers have tended to attribute the emissions reductions to the increase in cetane or reduction in density associated with their removal, rather than to the compositional change.
- *How one varies a fuel property can be the most important factor* in determining the emissions response. A given fuel property can be changed in several ways, and a unit change in that property can produce markedly different effects on emissions depending on how that change is introduced.
- *Past studies may understate the impact of fuels on emissions.* If density is varied in several ways – one of which has a strong effect on emissions and the others not at all – a study will tend to see only the average, diluted effect.

As a result, there is a very real risk that existing testing does not accurately assess the relationship between fuels and emissions and that policy makers may thereby underestimate the potential for fuels reformulation to contribute to emissions control.

## Application of Eigenfuels in Diesel Engine R&D

The eigenfuel approach provides new ways to design test fuels that are far more likely to be representative of future fuels that will be produced in refineries, compared to fuels blended in an effort to vary selected properties independently. The eigenfuel approach can also be used to extract additional insights from the emissions data. Test fuels design could be implemented in at least two ways:

- Develop test fuels to capture the processing and blending variability likely in the production of low sulfur fuels, and then procure the test fuels from several, differently-configured refineries.

- Use the eigenfuel approach to guide test fuel blending in the laboratory, so that the resulting test fuels closely replicate the signature characteristics expected for future low-sulfur diesel fuels.

In either case, the test fuels will express the natural correlations among fuel properties. While these correlations would be confounding factors in conventional analysis, they can be exploited in eigenfuel analysis.

## Recommendations to Improve the Diesel Emissions Database

An improved database is a prime requirement for the future development of a reliable diesel emissions model. The following are recommendations for future testing to correct the limitations of the existing data:

- *More testing of oxygenated fuels* will be required before a complete diesel emissions model can be developed. Few programs to date have evaluated oxygenated fuels, and the available data are too sparse to support an analysis.
- It may be important for new testing to report *a more detailed hydrocarbon speciation*. Existing information is frequently limited to mono- and poly-aromatic content, but it could well be important to know which hydrocarbon species were increased when, for example, aromatics content was reduced.
- An improved database should represent *a substantially larger number of engines and engine characteristics*. The existing database, while representing 280 individual engine tests, is based on only 11 individual engines and cannot support the assessment of engine-related effects.

# 1. INTRODUCTION

Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several predictor variables. In spite of its evident success in many applications, the regression approach can face serious difficulties when the predictor variables are to any appreciable extent covariant. This point is made quite evident in a recent review by Lee *et al.* (1998), in which efforts to evaluate the separate effects of fuel variables on the emissions from heavy-duty diesel (HDD) engines were often frustrated by the close association of fuel properties.

This report is an attempt to address these concerns by offering what is believed to be an ameliorative approach to modeling the effects of fuel characteristics on emissions. The approach is an adaptation of Principal Component Regression (PCR) that has certain advantages over the more commonly encountered method of stepwise regression, which was widely used in the development of the Complex Model for Reformulated Gasoline (U.S. DOE, 1994). A database of HDD engine testing is used to demonstrate the methodology, recognizing that the existing data is inadequate to answer fully the many issues related to the effect of fuels on emissions.

## 1.1 STATISTICAL PERSPECTIVE

The approach demonstrated here is only one of many – including ridge regression, partial least squares, all possible regressions, and PCR – that have been devised to counter regression difficulties, as attested to by the extensive literature on these subjects (Krzanowski *et al.*, 1994; Jackson, 1991; Martens and Naes, 1989). Each has its advantages and disadvantages, and in each a certain degree of art and arbitrariness must be recognized. It is the contention of this report that PCR, because of its seeming difficulty of interpretation and its past criticisms, is used less widely than it might be if better understood and more appropriately applied.

In the statistical literature are numerous evaluations of PCR as practiced under that label (Westerholm and Li, 1994; Hawkins, 1973; Boneh and Mendieta, 1994; Mansfield *et al.*, 1977; Jolliffe, 1972; 1973). Proponents point to the orthogonality of the regressors as a primary advantage. Orthogonality eliminates the aliasing of effects that accompanies non-additivity, and thereby makes statistical inference simple and exact. Moreover, the vector predictors that are the hallmark of PCR, though revealed by mathematical analysis, may have real-world interpretations capable of providing insights as to the physical processes driving the predictive relationship.

However, PCR has been soundly criticized (Hadi and Ling, 1998), and rightfully so, because as practiced up to this point in time, it selects predictors without any reference to

the response variable. Accordingly, it is possible for regressors having major influences on the response variable to be excluded without being given the chance to reveal their predictive importance.

Starting with some of the basic principles of Principal Component Regression (PCR), this report goes far beyond what is currently accepted as the prevailing PCR protocol. It is fitting, therefore, to refer to this extended approach as Principal Component Regression Plus (PCR+). This distinction is essential to prevent the perceived limitations of PCR from being identified with the approach taken here. Indeed, PCR+ is what PCR should have been from its inception. Insistence on selecting predictor variables without regard to their influence on the response is a restriction that never should have been, and it is hard to understand how a flaw this obvious could have survived so long. The simple fact that one can not evaluate predictive capability until there is something to predict should have been self-evident.

It is important that the limitations of PCR, as currently perceived, do not become a stigma that prevents the full potential of PCR+ from being realized in present and future applications. At the outset, PCR+ considers all eigenvectors to be of equal predictive importance, just as Ordinary Least Squares (OLS) would do. It includes all eigenvectors in the model and, by virtue of additivity, partitions the model sum of squares (SS) into separate components indicative of their true relative importance in prediction. The portion of the model SS that is associated with a given eigenvector can be sub-partitioned into parts associated with the components of the eigenvector, namely the original X-space variables. These sub-partitions can then be aggregated to show the relative importance of each of these original variables in the overall model or any subset model. Therefore, the awkwardness and sometime ambiguity of search methods, such as stepwise regression, need no longer be tolerated.

## 1.2 OVERVIEW OF THE VECTOR APPROACH

It is easily demonstrated that predictor variables can be naturally associated in a way that defies their separation. For example, it is evident and unarguable that increasing a fuel component such as olefins implies decreasing one or more other components such as aromatics or paraffins. Other examples abound in the refining world, such as the association of distillation characteristics with chemical composition.

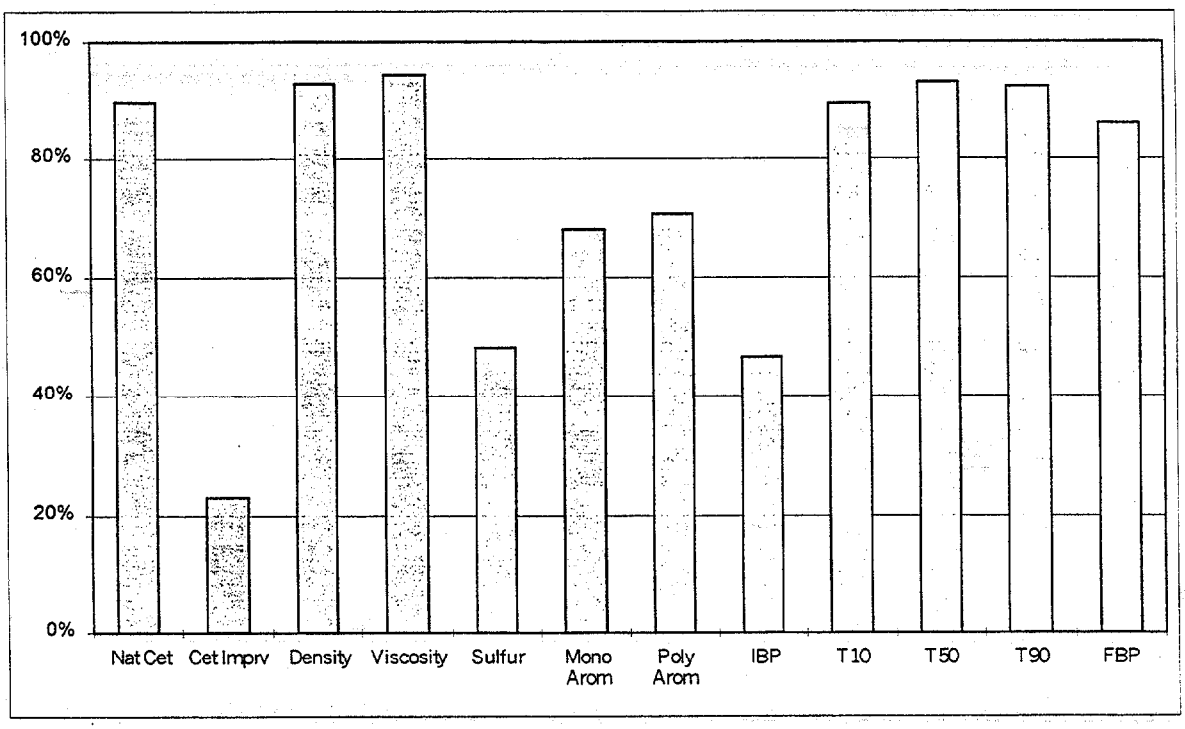
While the association of physical and chemical properties of automotive fuels is well known, the degree of interdependence may be surprising to some. To demonstrate the interdependence, each of twelve variables describing diesel fuel properties was expressed as a function of the other property variables, using conventional multi-linear regression analysis and the database on HDD engine testing developed for this study. In each regression, one of the property variables plays the role of response variable, while its companion variables are used as predictors. The result decomposes the variation of each fuel property into two parts:

1. The portion explained by (or shared with) the other fuel properties, as computed from the regression.

2. The portion independent of the other properties, determined from the residuals between the observed responses and the computed values.

Figure 1.1 shows the extent to which each of the fuel properties depends on all the others, where the interdependence of each variable is measured by the $R^2$ of its regression equation. For many properties – including natural cetane, density, viscosity, and four points on the distillation curve – approximately 90 percent of the variation is shared with the other variables and only about 10 percent is independent. For others – sulfur content, mono- and poly-aromatics content, and IBP – one-half to two-thirds of the variation is shared. Only for cetane improvement is the shared portion relatively small and the independent portion large – a result not surprising inasmuch as cetane enhancers were added to a range of base fuel stocks to create the test fuels.

**Figure 1.1.** Interdependence of Diesel Fuel Properties




These natural associations are not to be confused with apparent associations that arise from inappropriate experiment designs in violation of principles enunciated by R.A. Fisher in his pioneering book *Design of Experiments* (Fisher, 1935). Neither is the vector approach to be confused with applications involving Principal Component Analysis (PCA) and related factor-analytic methods used to understand the interrelation among

descriptive variables, as was pioneered by L. L. Thurstone in his book *The Vectors of Mind* (Thurstone, 1935).

Rather, for variables naturally associated in a physical system, our approach defines new, vector-based predictor variables in such a way that the vector variables are orthogonal. PCA is used to resolve the design matrix into eigenvectors that explain, in the most compact way, the variation among the original variables (here, fuel properties). Then, the eigenvectors are used as candidate predictors of the response variable (here, emissions) in an ordinary least squares (OLS) regression. Because fuels can be specified unambiguously using the eigenvectors, and the eigenvectors are shown to have physical interpretations, we adopt the terminology *eigenfuel* in place of *eigenvector* and treat fuels as mathematical blends of eigenfuels, each of which represents a distinct, mathematically independent characteristic.

Recognizing that any of the eigenfuels may play an important role in prediction, we incorporate, initially, *all* of them in the model. No attempt is made to select a subset of these vectors before performing the regression, as is commonly done in many past applications of PCR. Then, by means developed in this report, we select the most appropriate subset of eigenfuels to retain in the final model. Because of orthogonality, one can partition the model sum of squares (SS) explicitly among vectors and drop from the model those that are deemed unimportant, either because they fail to reach a specified level of significance or because they contribute little to the prediction in terms of magnitude. This approach is similar in many respects to the case studies described by Jeffers (1967).

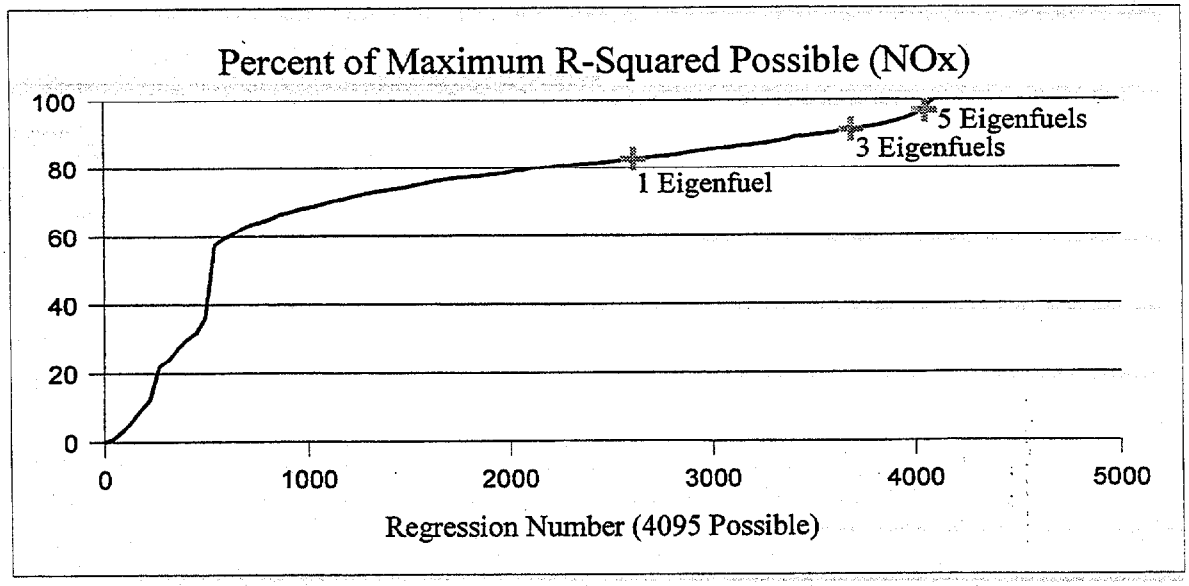
The final step consists of "pruning" the retained eigenfuels of those components (the original fuel variables) that contribute little to prediction. This step is possible because the ability to partition the model SS among eigenfuels implies the ability to partition that SS among their components – namely, the original variables. The method by which the partitioning is realized avoids such commonly used procedures as removing variables one at a time to determine how their removal reduces the model SS.

Knowing the extent of interdependence among the fuel variables, we should not be surprised by the difficulty of selecting an "optimal" set of variables for a regression model. We may believe that natural cetane or density has an important influence on emissions, but either may be *nearly replaced* by a combination of other variables. Stepwise regression, a commonly used technique in model development, searches through a sequence of differing model formulations to find one that is "optimum." In data such as this, there can be many different sets of variables that perform nearly as well as the one set ultimately chosen.

Figure 1.2 brings this point into focus. There are 4,095 different regressions that can be formed from twelve fuel properties, and these form the universe among which stepwise regression searches. It will take all twelve fuel properties to place a model at the very end

of the curve. Forty five different models populate the last 0.02 in $R^2$ and these typically involve 7, 8, or 9 predictor variables.

**Figure 1.2.** Comparative Performance of Regression Models




Models based on a small number of eigenfuels perform well in this context. For example, one eigenfuel explains 83 percent of the fuels-related variance in $NO_x$, three eigenfuels explain 91 percent, and five explain nearly 97 percent. Eigenfuel models for PM emissions perform as well. Given the high degree of interdependence in this data, we prefer to move away from efforts to find the "best" variables and, instead, toward an approach based on decomposing the interdependence into vectors representing the independent, underlying influences.

The transition from scalar to vector predictors brings with it a spate of interpretational and inferential issues. Tests of significance, for example, may need to be viewed in a different light, and it may be appropriate to put more emphasis on the magnitude of an effect rather than its probability of occurring by chance. So firmly embedded in our research culture is the statistical paradigm that most investigators are disinclined to acknowledge the existence of other criteria for judging the worth of a scientific finding. Moreover, the 0.05 level of significance is a fixed icon and tends to be routinely applied, even though the power of the test is strongly dependent on sample size. Further, one tends to accept, without question, that a variable either is or is not "significant," *in toto*.

In the present circumstance, however, one may have to accept the fact that the significance of a fuel variable may depend on its associations, inasmuch as *all* the variables appear as components of *each* of the eigenfuels. This is because rejecting an eigenfuel implies only partial rejection of the original variables comprising it, because the

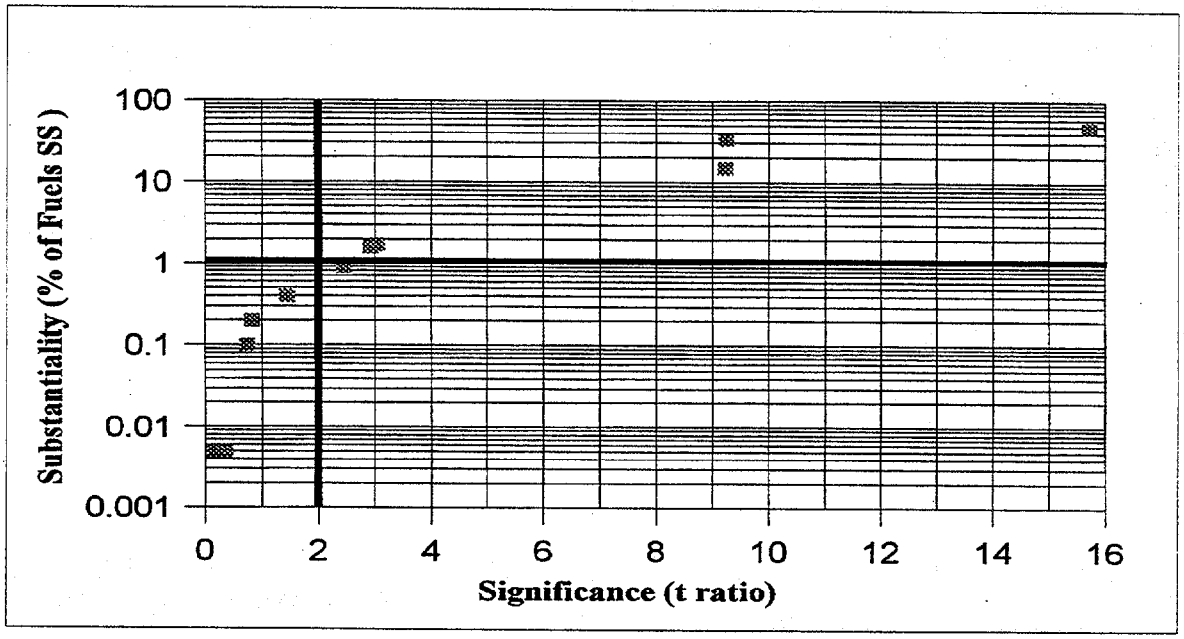
same variables occur in the eigenfuels that are not rejected. In this report we accept the notion of "partial" significance, believing that a variable can be *significant* when found in respectable company and *not significant* in less respectable company.

We also prefer to evaluate effects on the basis of their magnitude in addition to their probability of occurring by chance. We refer to these criteria for choice as substantiality and significance. In the implementation of the substantiality criterion, sample size still plays a role, of course, but in a way that is the dual of its role in a conventional test of significance. For a fixed significance level, such as the classic 0.05, the magnitude of the "least detectable" effect is variable and depends on sample size.

When a fixed magnitude is used as the threshold for acceptance or rejection of an effect, it is the statistical significance level that varies, again depending on sample size. We believe it is essential to balance the two considerations. Even though an effect may be statistically significant, because of large sample size, there is no reason for it to be retained if it makes only a minuscule difference in predictions.

Our approach is illustrated in Figure 1.3. The horizontal axis represents significance, the vertical axis represents substantiality, and the points represent the eigenfuels and where they lie in the plane of the two criteria. We first reject eigenfuels with t-ratios smaller than 1.96, as required for 0.05 significance. We then reject predictors that fail to explain some minimum percentage of the fuels-related SS. In the figure, this threshold is taken as one percent. For particulate emissions, as shown, five eigenfuels satisfy the combined criteria for significance and substantiality.

**Figure 1.3.** Significance versus Substantiality of Effects (PM Emissions Model)



A major advantage of the substantiality criterion is that its critical value does not change with sample size, though the associated significance threshold does. Thus, substantiality serves as a counterbalance to the test of significance, the power of which increases with sample size and often leads us to retain smaller and smaller, and sometimes negligible, effects.

The vector approach developed in this report provides a generally applicable method for identifying an efficient set of vector variables to describe a collection of data. When applied as predictor variables in regression analysis, the vector variables are found to have many advantages, including:

- Economy of representation, because a small number of vector variables may effectively replace a larger number of the original variables.
- Simplification of regression analysis, because properly constructed vector variables (the eigenvectors of the problem) will be mathematically independent and eliminate the complications introduced by multi-collinearity.
- Potentially greater understanding of the patterns of variation present in the data and how these are related to the dependent variable under consideration.

## 1.3 DIESEL EMISSION CONTROL

The background and status of diesel emission research are well documented by the previously cited literature review by Lee *et al.* (1998). Up to this point in time it appears that engine design factors tend to eclipse fuel effects so far as efforts to reduce diesel emissions are concerned. Moreover, as pointed out by the authors, a number of prototype engine technologies are under consideration in order to meet future proposed emission limits in the United States (for 2004) and in the European Union (for 2000 and 2005).

How fuel properties may influence diesel emissions in the future is particularly problematical, especially in those instances where engine or fuel properties play an *enabling* role for the other. Inseparable effects are not necessarily limited to fuels. Some may pertain to engine design features, such as exhaust gas recirculation or catalyst systems, that depend on fuel properties for an enabling effect. Evaluation of such coupled effects may resist conventional statistical means and present yet another opportunity for exploitation of the vector-variable approach.

## 1.4 ORGANIZATION OF REPORT

The report documents research performed in two phases during 1999 and 2000. Portions were previously published in the technical paper series of the Society of Automotive Engineers (SAE) (McAdams *et al.*, 2000). Section 2 presents the concepts of the vector

methodology and demonstrates its application to a database of HDD engine emissions. Section 3 develops a perspective that shows how the vector approach leads to a preferred identification of the explanatory influences in a regression model. It then presents a series of extensions and refinements to the basic methodology to address difficulties that can be expected to arise in practice. Section 4 discusses three areas in which additional work is needed to improve the understanding of diesels fuels and HDD emissions, while Section 5 presents conclusions of the work. Nine appendices are included with this report to amplify on the development of the methodology and document the procedure involved in implementing the approach.

# 2. THE VECTOR METHODOLOGY

This section presents the vector methodology and its application to a database of HDD engine emissions. We begin with a description of the data base and then proceed to a presentation of the methodology.

## 2.1 DESCRIPTION OF DATA BASE

A database representing 280 individual emission tests of HDD engines was compiled from nine publications ( Gonzalez *et al.*, 1993; McCarthy *et al.*, 1992; Schaberg *et al.*, 1997; Sienicki *et al.*, 1990; Spreen *et al.*, 1995; Ullman, 1989; Ullman *et al.*, 1990; 1994; 1995) in the SAE literature where the following criteria were met:

- The Environmental Protection Agency (EPA) transient test cycle was used and either the composite or hot start result was reported. The hot start portion has a $6/7^{ths}$ weight in the composite result.
- At least $NO_x$ and PM emissions were measured, which are the pollutants examined in this study. Eight of the sources measured all four pollutants (HC, CO, $NO_x$, and PM), and one source measured all except CO.
- Emissions testing could be matched to fuels for which the following 12 properties were known: natural cetane, cetane number improvement (resulting from additives), density, viscosity, sulfur content, mono-aromatic content, poly-aromatic content, and five points on the distillation curve.

Table 2.1 lists the variables contained in the database; the field names shown are used to refer to the variables in the tables and figures of this report. Overall, the data represent eleven different engines tested a total of 280 times on 85 different diesel fuels.

Twenty-seven publications were examined in the process of compiling this data base (Akasaka *et al.*, 1997; Cunningham *et al.*, 1990; Daniels *et al.*, 1996; EPA HDEWG Program, 1999; Geiman *et al.*, 1996; Gonzalez *et al.*, 1993; Lange, 1991; Lange *et al.*, 1997; Liotta, 1993; Liotta and Montalvo, 1993; Mann *et al.*, 1998; McCarthy *et al.*, 1992; Nakakita *et al.*, 1998; Nylund *et al.*, 1997; Reynolds, 1993; Rosenthal and Bendinsky, 1993; Schaberg *et al.*, 1997; Schmidt and Gerpen, 1996; Sienicki *et al.*, 1990; Spreen *et al.*, 1995; Star, 1997; Tamanouchi *et al.*, 1997; Tanaka *et al.*, 1996; Ullman, 1989; Ullman *et al.*, 1990; 1994; 1995). Eight publications using the EPA transient test cycle were excluded because one or more of the fuel properties was not reported, most commonly the poly-aromatic content. Ten publications were excluded for reasons related to the emissions data. In one instance, only PM was measured, while European or

**Table 2.1.** The HDD Emissions Database

| Field Name | Units | Description |
|---|---|---|
| Engine ID | text | text description of engine |
| Fuel ID | text | text identifier of fuel |
| Test | number | sequential number |
| Source | text | SAE paper number |
| Cycle | number | 0=EPA Composite<br>1=EPA Hot Start |
| Engine | number | unique engine identifier |
| NRepl | number | number of test replications |
| HC | gm/bhp-hr | hydrocarbon emissions |
| CO | gm/bhp-hr | carbon monoxide emissions |
| NOx | gm/bhp-hr | nitrogen oxide emissions |
| PM | gm/bhp-hr | particulate emissions |
| NatCetane | number | natural cetane |
| CetImprv | number | cetane improvement |
| Density | $gm/cm^3$ | |
| Viscosity | $mm^2/sec$ | at/near 40 degrees C |
| Sulfur | ppm | sulfur content |
| MonoArom | percent | mono-aromatics content |
| PolyArom | percent | poly-aromatics content |
| IBP | Celsius | Initial boiling point |
| T10 | Celsius | 10 percent evaporation point |
| T50 | Celsius | 50 percent evaporation point |
| T90 | Celsius | 90 percent evaporation point |
| FBP | Celsius | Final boiling point |

Japanese test cycles were used in nine other instances. Additional information on the data base may be found in Appendix A.

The 12 fuel properties examined here are a super-set of the fuel properties that have been considered with respect to HDD emissions. They include characteristics such as aromatics content, cetane rating, and sulfur content that a consensus of investigators believes relevant to emissions performance. Additional properties (IBP, T10 and others) are included that could be independent predictors, could be correlated with the consensus variables, or could prove unrelated to emissions. This purposefully casts the net as wide as possible, leaving the identification of the proper subset of predictor variables to the later analysis. However, there is no intent to imply that only these properties could affect engine emissions. For example, fuel oxygen content is likely to affect CO emissions, and perhaps other pollutants, but is not among the selected properties. While some sources tested oxygenated diesel fuels, these were judged to be too few in number to permit including oxygen content in the list.

The eleven HDD engines represented in the data base constitute a very small sample of the engine types present in the on-road vehicle fleet. Nevertheless, they include engines made by the three major manufacturers (Cummins, Detroit Diesel Corporation, and Navistar) and cover a range in model years and horsepower ratings. The engines are generally similar in design, although they are built to varying emissions standards. None are equipped with EGR systems or with catalysts as tested. They can be taken as

reflective of the HDD engines currently on the road, even if the sample is too limited to be considered truly representative.

The large majority of the data base represents individual engine tests, but 41 entries from three sources record the mean values of replicated tests. For this work, we have duplicated the mean values so that each is represented as many times as the tests were replicated. This approach, which gives a total of 280 emission tests, maintains an equal weighting among individual tests and improves the estimation of the total variance. However, the total variance is understated to an unknown extent because the data omit the variation of the (unknown) individual test results around the (reported) mean values.

## 2.2 COMPUTER SOFTWARE

The analysis was conducted using MatLab, a commercially available software package designed for matrix processing (MatLab, 2000). MatLab offers a built-in function *svd*, which extracts the eigenvalues and eigenvectors of a matrix using the singular value decomposition procedure. Other statistical procedures such as computation of correlation matrices and multivariate regression analysis are available as built-in functions or can be easily written using matrix notation. The methodology demonstrated here can be implemented in any computational environment that provides for the calculation of matrix eigenvalues and eigenvectors. Appendix B contains a listing of the major MatLab procedures created in implementing the methodology.

## 2.3 THE VECTOR METHODOLOGY

We first demonstrate how PCA can be used to resolve the matrix of fuels into a representation based on eigenvectors and then demonstrate the use of eigenvectors in regression analysis. Appendix C develops the statistical theory that underlies the methodology presented here.

### 2.3.1 Vector Approach to Representing Fuels

Consider the subset of data describing the fuels used in emissions testing. This test fuels data set consists of the 12 selected properties measured for the 85 different fuels used in the engine testing and replicated a varying number of times corresponding to the number of emissions tests in which each was used. The data set can be viewed as a matrix X of dimension 280 rows (engine tests) by 12 columns (fuel properties) that contains all of the fuels-related information available as predictors for emissions.

Summary statistics for the 280 x 12 data set consist of a mean vector and variance-covariance matrix for the 12 variables. The mean vector is a row vector consisting of 12 components presenting, respectively, the means for the 12 fuel properties: natural cetane, cetane improvement, ... FBP. The 12 x 12 variance-covariance matrix displays the

variance of the 12 fuel properties along its main diagonal. The off-diagonal elements represent the covariances of pairs of fuel variables.

Prior to further analysis, each column (fuel variable) is standardized to mean zero, variance one. All subsequent analysis will be based on these standardized variables, each of which measures fuel properties in units of standard deviations from the mean. Much of the thrust of the analysis will be aimed at reducing the variance-covariance matrix to diagonal form – that is, zeros everywhere except on the main diagonal.

The starting point for this is the correlation matrix, shown in Table 2.2. Correlations greater than 0.50 in absolute magnitude (an arbitrary threshold) have been highlighted to emphasize the pairwise interdependencies of the physical properties. For example, and not surprisingly, the five points on the distillation curve are highly correlated with each other and with viscosity. Other correlations reflect known relationships encountered in fuel blending. Decreased natural cetane is correlated with increased density and aromatic content as would be expected in a fuel blend where the proportion of high-aromatic cracked stocks, which have low cetane and high density, has been increased.

That fuel properties are correlated, sometimes to a large degree, implies that there are fewer *independent* variables than the number of physical properties measured. To demonstrate this fact, a singular value decomposition analysis was performed to extract the 12 eigenvalues and eigenvectors from the correlation matrix. The eigenvectors are defined in the computational procedure in a manner that partitions the total variance into orthogonal components, where the eigenvalues are the variances associated with the corresponding eigenvectors. In this context, orthogonality means that the eigenvectors are linearly independent of each other and, as a result of their definition, the correlation between any two eigenvectors over the data set is exactly zero.

Table 2.3 presents the twelve eigenvalues and eigenvectors of the test fuels data set. Each eigenvector is a linear combination of the original 12 fuel properties. For example, the first eigenvector is described by the coefficients or weights (0.061, 0.034, 0.285, ... 0.365) applied to the fuel properties (natural cetane, cetane improvement, density, ... FBP). The largest coefficients have been highlighted to emphasize the most important fuel property components.

The variance among fuels, indicated by the eigenvalues, is highly concentrated in the first few eigenvectors. The first eigenvector accounts for nearly 40 percent of the total variation among the fuels, the first six together account for more than 90 percent, and the first nine for essentially all (nearly 99 percent). Thus, while the data set contains 12 distinct variables, its total variance is concentrated in a much smaller number of orthogonal patterns that are described by the eigenvectors with the largest eigenvalues.

It is often desirable to develop a conceptual interpretation of the eigenvectors to aid the analyst's understanding, although this may not be completely possible in complex systems. Physical systems (of any kind) are normally created from more basic building

**Table 2.2.** Correlation Matrix for Fuel Properties

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 1 | 1.000 | | | | | | | | | | | |
| CetImprv | 2 | -0.233 | 1.000 | | | | | | | | | | |
| Density | 3 | -0.613 | 0.105 | 1.000 | | | | | | | | | |
| Viscosity | 4 | 0.219 | 0.051 | 0.460 | 1.000 | | | | | | | | |
| Sulfur | 5 | -0.022 | -0.220 | 0.202 | -0.054 | 1.000 | | | | | | | |
| MonoArom | 6 | -0.633 | 0.247 | 0.667 | 0.121 | -0.030 | 1.000 | | | | | | |
| PolyArom | 7 | -0.393 | -0.040 | 0.523 | -0.084 | 0.511 | 0.298 | 1.000 | | | | | |
| IBP | 8 | 0.097 | -0.058 | 0.292 | 0.514 | -0.063 | 0.074 | -0.029 | 1.000 | | | | |
| T10 | 9 | 0.225 | -0.098 | 0.444 | 0.900 | 0.016 | 0.071 | 0.006 | 0.622 | 1.000 | | | |
| T50 | 10 | 0.275 | -0.002 | 0.497 | 0.889 | 0.014 | 0.144 | 0.097 | 0.382 | 0.792 | 1.000 | | |
| T90 | 11 | 0.295 | 0.114 | 0.307 | 0.692 | -0.038 | 0.226 | 0.144 | 0.237 | 0.523 | 0.775 | 1.000 | |
| FBP | 12 | 0.211 | 0.168 | 0.318 | 0.607 | -0.096 | 0.224 | 0.118 | 0.278 | 0.445 | 0.633 | 0.897 | 1.000 |

**Table 2.3.** Eigenvectors of the Test Fuels Data Set

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.061 | -0.556 | 0.163 | -0.220 | 0.071 | -0.068 | 0.138 | -0.458 | 0.104 | -0.456 | 0.045 | 0.391 |
| CetImprv | 0.034 | 0.143 | -0.549 | -0.212 | 0.782 | 0.061 | -0.024 | -0.106 | 0.001 | -0.054 | -0.048 | 0.004 |
| Density | 0.285 | 0.449 | 0.049 | 0.168 | -0.047 | 0.163 | -0.109 | 0.237 | 0.343 | -0.418 | -0.272 | 0.473 |
| Viscosity | 0.432 | -0.120 | -0.017 | 0.142 | 0.084 | 0.308 | 0.004 | 0.150 | -0.156 | 0.292 | 0.638 | 0.371 |
| Sulfur | 0.002 | 0.180 | 0.636 | -0.202 | 0.393 | 0.118 | 0.556 | 0.175 | 0.013 | 0.058 | -0.024 | -0.100 |
| MonoArom | 0.146 | 0.464 | -0.256 | 0.040 | -0.305 | -0.028 | 0.548 | -0.500 | -0.134 | -0.078 | 0.164 | -0.037 |
| PolyArom | 0.076 | 0.418 | 0.385 | -0.272 | 0.092 | -0.289 | -0.552 | -0.343 | -0.171 | 0.030 | 0.225 | 0.049 |
| IBP | 0.262 | -0.072 | 0.052 | 0.537 | 0.248 | -0.697 | 0.118 | -0.033 | 0.234 | 0.128 | 0.015 | -0.024 |
| T10 | 0.399 | -0.113 | 0.128 | 0.316 | 0.109 | 0.192 | -0.093 | -0.123 | -0.642 | -0.186 | -0.392 | -0.198 |
| T50 | 0.431 | -0.083 | 0.064 | -0.053 | -0.028 | 0.319 | -0.149 | -0.240 | 0.555 | 0.010 | 0.065 | -0.552 |
| T90 | 0.392 | -0.074 | -0.075 | -0.428 | -0.158 | -0.134 | 0.071 | -0.037 | 0.021 | 0.551 | -0.486 | 0.252 |
| FBP | 0.365 | -0.048 | -0.147 | -0.409 | -0.141 | -0.359 | 0.069 | 0.476 | -0.157 | -0.406 | 0.205 | -0.254 |
| Eigenvalues | 4.531 | 2.591 | 1.549 | 1.181 | 0.681 | 0.564 | 0.390 | 0.230 | 0.140 | 0.083 | 0.036 | 0.026 |
| Pct Variance | 37.75 | 21.59 | 12.90 | 9.83 | 5.67 | 4.69 | 3.24 | 1.92 | 1.17 | 0.68 | 0.29 | 0.21 |
| Cumulative Pct | 37.75 | 59.34 | 72.25 | 82.09 | 87.76 | 92.46 | 95.70 | 97.62 | 98.79 | 99.48 | 99.78 | 100.0 |

blocks according to a set of rules that reflect a natural structure. If these building blocks are fully described by the chosen set of variables, one hopes to find an expression of this structure in the eigenvectors. In the context of diesel fuels, the underlying structure (and therefore the eigenvectors) should reflect the properties of the refinery processes and blending stocks used to create these fuels.

The following discussion interprets the first four eigenvectors in terms of the associations among fuel properties. Where possible, we have suggested identifications of the eigenvectors with known refinery or blending processes. These largest eigenvectors, as identified by the proportion of the total variance shown in parentheses, are likely to represent generalized characteristics of fuels and therefore to be most amenable to variation in reformulating diesel fuels.

***Primary viscosity/density characteristic (38 percent).*** A direct relationship among viscosity, distillation temperatures, and to a lesser extent density. This characteristic is associated with the largest eigenvalue, meaning that the test fuels vary most among themselves with respect to this characteristic. More viscous compounds found in diesel fuels have higher boiling points, and predictive equations show that viscosity is directly related to the square root of density (Thomas, 1946). Diesel blend stocks exhibit a similar relationship among viscosity, distillation temperatures, and density as demonstrated independently by correlation analysis using the data base of blend stocks in the Refinery Yield Model (RYM) maintained by Oak Ridge National Laboratory.

***Primary aromatics characteristic (22 percent).*** An increase in aromatics content (both mono- and poly-aromatic) is associated with higher density and a decrease in natural cetane. This characteristic reflects a known property of the high-aromatic cracked stocks that are used in blending diesel fuels. These stocks have higher densities and their aromatics content is known to delay ignition and thereby decrease cetane rating.

***Primary sulfur/quality characteristic (13 percent).*** This appears to represent sulfur content and its related impact on the boost from cetane improvers, which declines as the quality of diesel fuel declines. Information from the Ethyl Corporation (Ethyl Corporation, 1995) shows that cetane boost is reduced with lower clear cetane, as for fuels with higher sulfur and poly-aromatics. This characteristic might be explained by the presence of high sulfur-content dibenzothiophenes, which can be present in light cycle oils produced by fluid catalytic cracking.

***Primary blend balancing characteristic (10 percent).*** Fuels with increased temperatures at the low end of the curve (IBP and T10) tend to be associated with decreased temperatures at the upper end (T90 and FBP). This slope characteristic for the distillation curve complements the height characteristic found in eigenvector 1 and may be related to meeting blending specifications. For example, flash point might be satisfied by using heavier blend stocks at the low end of the distillation temperatures, while lighter blend stocks are used on the high end to meet the pour point requirement.

When the variance associated with individual eigenvectors falls to relatively small percentages, regression coefficients for these eigenvectors are subject to wide error bounds. Moreover, such eigenvectors may tend to reflect factors specific to the blending of test fuels in individual sources, rather than characteristics found in a range of fuels. For example, the smallest eigenvectors could represent specific blend stocks used in one or more sources to vary fuel properties for test purposes or to control one or more fuel properties to fixed values, once another property had been varied for experimental purposes. For this reason, we do not attempt to offer physical interpretations for the smaller eigenvectors. Overall, more work is needed to understand the eigenvector characteristics of diesel fuels, particularly as those characteristics may differ between commercially available fuels and test fuels created for use in the laboratory.

Because the eigenvectors form an orthogonal basis, they can be used to re-express the original matrix in orthogonal terms. This process is closely analogous to Fourier transform analysis, in which a time-varying signal is decomposed into individual frequencies and then re-expressed as a weighted sum over frequencies. In Fourier analysis, the continuum of harmonic frequencies from $\omega = 0$ to $\infty$ forms an orthogonal basis from which any time-varying signal can be constructed. In the vector approach defined here, the basis vectors are developed from the experimental data in a manner expressly defined to be orthogonal. Data analysts may perceive the eigenvector approach to be similar to the use of orthogonal polynomials in a regression equation consisting of successive powers of the predictor variable (Fisher and Yates, 1948).

An experimental design matrix $X_{(m \times n)}$ of m rows and n variables can be represented in eigenvector terms as the linear combination $A_{(m \times k)} * V^{T}_{(n \times k)}$, where $A_{(m \times k)}$ is a matrix of coefficients for the k eigenvectors and $V_{(n \times k)}$ is a matrix in which the eigenvectors k, composed of n components each, form the columns. The coefficients $A_{(m \times k)}$ are calculated from the relationship $A_{(m \times k)} = X_{(m \times n)} * V_{(n \times k)}$. In algebraic form, any row m of the X matrix can be expressed as a linear combination of coefficients $a_m(k)$ and eigenvectors $v_j(k)$:

$$X_m(j) = a_m(1)*v_j(1) + ... + a_m(12)*v_j(12) \qquad (1)$$

where: $X_m(j)$ = value of the $j^{th}$ variable (fuel property) for the $m^{th}$ fuel; $a_m(k)$ = coefficient of eigenvector k in the $m^{th}$ fuel ; and $v_j(k)$ = component weight for the $j_{th}$ variable in eigenvector k.

The example in Table 2.4 may help to make these relationships more understandable. Here, the observed values[1] have been taken from a selected observation in the X matrix. Below this, the calculation given by Eq. 1 is shown to exactly reproduce the original

[1] These are the physical properties in standardized form where mean = 0 and variance = 1.

**Table 2.4.** Eigenfuel Representation for a Selected Fuel

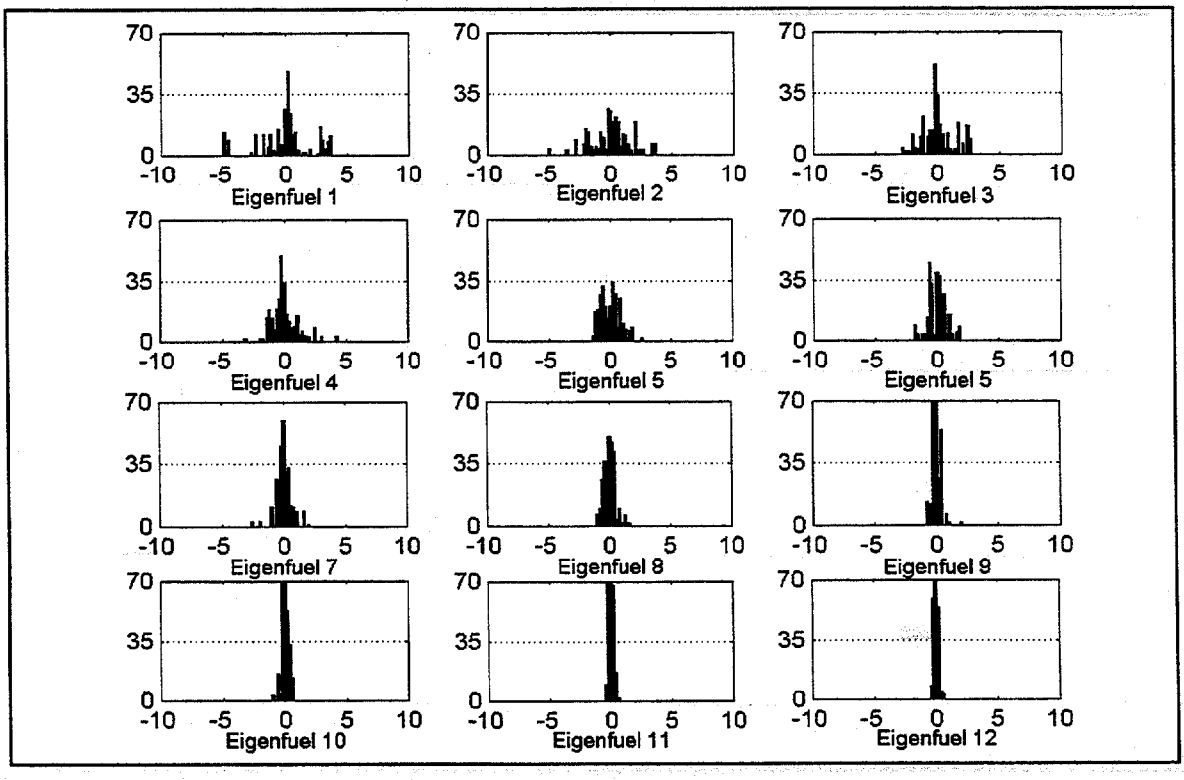
| Fuel Property | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Observed Values | -0.305 | -0.535 | 0.375 | 0.025 | -0.341 | 0.042 | 0.005 | -0.083 | -0.123 | 0.331 | 0.154 | -0.268 |
| Calculated Values | -0.305 | -0.535 | 0.375 | 0.025 | -0.341 | 0.042 | 0.005 | -0.083 | -0.123 | 0.331 | 0.154 | -0.268 |
| k | Coefficient | Eigenvector Components | | | | | | | | | | |
| 1 | 0.121 | 0.061 | 0.034 | 0.285 | 0.432 | 0.002 | 0.146 | 0.076 | 0.262 | 0.399 | 0.431 | 0.392 | 0.365 |
| 2 | 0.213 | -0.556 | 0.143 | 0.449 | -0.120 | 0.180 | 0.464 | 0.418 | -0.072 | -0.113 | -0.083 | -0.074 | -0.048 |
| 3 | 0.066 | 0.163 | -0.549 | 0.049 | -0.017 | 0.636 | -0.256 | 0.385 | 0.052 | 0.128 | 0.064 | -0.075 | -0.147 |
| 4 | 0.259 | -0.220 | -0.212 | 0.168 | 0.142 | -0.202 | 0.040 | -0.272 | 0.537 | 0.316 | -0.053 | -0.428 | -0.409 |
| 5 | -0.632 | 0.071 | 0.782 | -0.047 | 0.084 | 0.393 | -0.305 | 0.092 | 0.248 | 0.109 | -0.028 | -0.158 | -0.141 |
| 6 | 0.229 | -0.068 | 0.061 | 0.163 | 0.308 | 0.118 | -0.028 | -0.289 | -0.697 | 0.192 | 0.319 | -0.134 | -0.359 |
| 7 | -0.294 | 0.138 | -0.024 | -0.109 | 0.004 | 0.556 | 0.548 | -0.552 | 0.118 | -0.093 | -0.149 | 0.071 | 0.069 |
| 8 | 0.011 | -0.458 | -0.106 | 0.237 | 0.150 | 0.175 | -0.500 | -0.343 | -0.033 | -0.123 | -0.240 | -0.037 | 0.476 |
| 9 | 0.371 | 0.104 | 0.001 | 0.343 | -0.156 | 0.013 | -0.134 | -0.171 | 0.234 | -0.642 | 0.555 | 0.021 | -0.157 |
| 10 | 0.205 | -0.456 | -0.054 | -0.418 | 0.292 | 0.058 | -0.078 | 0.030 | 0.128 | -0.186 | 0.010 | 0.551 | -0.406 |
| 11 | -0.119 | 0.045 | -0.048 | -0.272 | 0.638 | -0.024 | 0.164 | 0.225 | 0.015 | -0.392 | 0.065 | -0.486 | 0.205 |
| 12 | 0.048 | 0.391 | 0.004 | 0.473 | 0.371 | -0.100 | -0.037 | 0.049 | -0.024 | -0.198 | -0.552 | 0.252 | -0.254 |

observation. The values are calculated, for any fuel property j, as the product of the coefficient $a_k$ times the coefficient for the $j^{th}$ component of eigenvector k, summed over all eigenvectors k = 1, 2, . . . , 12. The eigenvectors k are placed in row form in this table, while the fuel properties j form the columns.

Thus, we can express any fuel as the vector of coefficients $a_m(k) = (a_1, a_2, ..., a_{12})$ corresponding to the 12 eigenvectors, instead of describing the fuel by its physical properties. Because these coefficients and their associated eigenvectors specify a fuel unambiguously, and the eigenvectors have been shown to have physical interpretations, we adopt the terminology *eigenfuel* in place of *eigenvector* and think of these eigenfuels as hypothetical components of real-world fuel blends. We then treat fuels as *mathematical blends* of eigenfuels, each of which represents a distinct, mathematically independent characteristic. The coefficients $a_m(k)$ become measures of how fuel m is composed of the eigenfuels k.

Once a data set is translated into this representation, the eigenfuel coefficients are distributed with mean zero and variance equal to the corresponding eigenvalue. Figure 2.1 shows histograms of the coefficients $a_m(k)$ for the data set used here. The coefficient distributions are broad (have large variance) for the first several eigenfuels, consistent with their large eigenvalues. The distributions narrow as one moves through the series, until they approach a peak clustered around zero by the end. Thus, the fuels vary most

**Figure 2.1.** Distribution of Eigenfuel Coefficients



widely with respect to the characteristics expressed by the first several eigenfuels and differ only to a very minor extent with respect to those represented by the later eigenfuels.

A fundamental property of eigenfuels is that they form an orthogonal set and their coefficients are uncorrelated – i.e., the off-diagonal elements of the correlation matrix for the coefficients $A_{(m \times k)}$ are zero. That the eigenfuel coefficients are mathematically independent and uncorrelated will prove very important when they are used as predictor variables.

### 2.3.2 Use of Eigenfuels in Regression Analysis

Having reviewed the properties of eigenfuels, we now turn to their use as predictor variables in regression analysis. The empirical relationship between engine emissions and fuel properties is usually determined through regressing an emissions variable Y against one or more fuel property variables $X_i$ in a form similar to:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n \qquad (2)$$

where the coefficients $b_i$ are determined by the regression, and we consider the variables $X_i$ to be scalar quantities. In the vector approach developed here, the regression model will be of comparable form:

$$Y = b_0 + b_1 A_1 + b_2 A_2 + \ldots + b_n A_n \qquad (3)$$

where the new variables $A_i$ are the coefficients of eigenfuel i in the vector representation. Consistent with other work, the dependent variable Y is taken to be the natural logarithm of emissions. (See Appendix I for further discussion of the use of variable transformations in regression analysis.)

When used as predictor variables in regression analysis, eigenfuels have two important properties resulting from their mathematical independence, as demonstrated in Table 2.5. Here, $NO_x$ emissions have been regressed against each eigenfuel k individually using equations of the form:

$$\ln(NO_x) = a_0 + a_1 A_k \quad \text{for } k = 1, 2, \ldots 12 \qquad (4)$$

The regression sum of squares and coefficient values are then tabulated against the results of a regression in which all 12 eigenfuels are present simultaneously (see the rightmost column of the table). The regression sum of squares summed across the twelve individual regressions equals the sum of squares in the regression containing all 12 eigenfuels. Further, the intercept and eigenfuel coefficients for the individual regressions are identical to those estimated in the regression containing all eigenfuels. Thus, the regression sums of squares are *additive* and the coefficient values are *invariant* with respect to the selection of eigenfuels for inclusion in the regression. There is, in fact, a unique partitioning of the variance in the dependent variable into the components

**Table 2.5.** $NO_x$ Regressions Using Eigenfuels as Explanatory Variables

| | Eigenfuels included in Regression | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | ALL |
| **Regression Sum of Squares** | .0236 | .7805 | .0010 | .0506 | .1034 | .0007 | .0148 | .0000 | .0547 | .1229 | .0113 | .0000 | 1.1633 |
| **Cumulative SS** | .0236 | .8041 | .8051 | .8557 | .9591 | .9597 | .9745 | .9745 | 1.0292 | 1.1521 | 1.1633 | 1.1633 | |
| | | | | | | | | | | | | | |
| **Intercept** | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 | 1.5372 |
| **Eigenfuel 1** | .0043 | | | | | | | | | | | | .0043 |
| **Eigenfuel 2** | | .0329 | | | | | | | | | | | .0329 |
| **Eigenfuel 3** | | | .0016 | | | | | | | | | | .0016 |
| **Eigenfuel 4** | | | | .0124 | | | | | | | | | .0124 |
| **Eigenfuel 5** | | | | | -.0233 | | | | | | | | -.0233 |
| **Eigenfuel 6** | | | | | | .0021 | | | | | | | .0021 |
| **Eigenfuel 7** | | | | | | | .0117 | | | | | | .0117 |
| **Eigenfuel 8** | | | | | | | | .0002 | | | | | .0002 |
| **Eigenfuel 9** | | | | | | | | | -.0374 | | | | -.0374 |
| **Eigenfuel 10** | | | | | | | | | | -.0730 | | | -.0730 |
| **Eigenfuel 11** | | | | | | | | | | | -.0335 | | -.0335 |
| **Eigenfuel 12** | | | | | | | | | | | | .0012 | .0012 |

identified with the eigenfuels. It can be shown (see Appendix C) that the contribution of eigenfuel k to the regression sum of squares and $R^2$ statistic is proportional to the product of the eigenvalue $\lambda_k$ and the square of the regression coefficient $b_k$.

This outcome contrasts with the usual result in regression analysis when the predictor variables are correlated with each other. The multi-collinearity existing in such circumstances means that parameter estimates change when predictors are added to or removed from the regression. In addition, combining individual variables to create a pooled regression does not increase the regression sum of squares and $R^2$ statistic to the extent that might be expected. Working through the "fog" created by multi-collinear variables is part of the art of regression analysis and is one of the reasons why independent analysts can reach differing conclusions from the same data. The use of the linearly independent, vector variables eliminates this fog.

The table suggests insights that will be developed in the next section. Not surprisingly, there is a difference between the *importance* of an eigenfuel in describing the variation among fuels and the *strength* of its relationship to $NO_x$ emissions or another dependent variable. Eigenfuels 1 through 12 are defined in decreasing order of variance among the fuels, so that eigenfuel 1 accounts for 38 percent of the fuel variance, followed by eigenfuel 2 at 22 percent, and eigenfuel 10 at only 0.7 percent. However, the regression results indicate that eigenfuel 10 has the strongest relationship to $NO_x$ emissions, as measured by its coefficient, followed by eigenfuels 11 and 2. We also see that some of the eigenfuels (numbers 1, 3, 6, 8, and 12) have very weak relationships to $NO_x$ and are likely candidates to drop from the analysis. However, before attempting to draw conclusions from such results, we must first consider and control for other factors, such as engine characteristics, that contribute to the variance in emissions.

### 2.3.3 Application to Diesel Emissions

In this section we show the application of the vector approach to diesel engine emissions and obtain a first look at its implications. There are many factors beyond fuel composition that contribute to the variance in engine emissions, including differences among engines, test cycles, and the sources from which the data are drawn. The intent is to extract these fixed effects and then re-compute the regression equation involving the eigenfuels. This will be done for both $NO_x$ and PM emissions.

Figure 2.2 suggests the sources of variation to be found in the data base of diesel emissions data. The tested engines are taken to be generally reflective of the population of HDD engines currently on the road. Nine different publications reported tests for 11 individual engines, representing 11 different engine designs, on 85 different fuels using one of two different EPA test cycles. Most, but not all, engine tests were replicated at least once.

**Figure 2.2.** Sources of Variance in Diesel Engine Emissions Data

**Sources – nine different publications**

**Engine Lines – 11 different engine lines**

**Individual Engines – 11 individual engines tested**

**Fuels – 85 unique fuels**

**Test Cycles – 2 different EPA cycles (composite and hot start)**

**Test Replications – test-to-test variation**

We can hypothesize a series of terms in the overall emissions model that represent, for example, the average emissions level $E_0$ of the existing fleet, the variation of average emissions $E_i$ for engine line i around $E_0$, and the variation of the average emissions $E_{ij}$ for individual engine j around its engine line average $E_i$. Other terms in the model would include an effect $S_k$ for differences among the sources and an effect $T_l$ for the different average emissions levels of the two EPA test cycles. This series of terms would be in addition to the effects of fuels on emissions, which are of primary interest here.

We are unable, however, to estimate an overall emissions model at present because of size and coverage limitations of the data base that make it impossible to separate the effects of engine designs, individual engines, test cycles, and sources. Each engine design is represented by only a single specimen and each individual engine has been tested using only one of the two test cycles. For purposes of this exploratory study, we have incorporated a single fixed effect for individual engines in the regression models. This engine effect represents an undifferentiated, composite effect due to engine designs, individual engines, sources, and test cycles.

Thus, we use regression equations of the form:

$$\ln(E_{i,j}) = b_o + \Sigma\, b_i * \delta_i + \Sigma\, b_{j,k} * A_{j,k} \qquad (5)$$

where the dependent variable is the natural logarithm of emissions for engine i tested on fuel j, $\Sigma\, b_i * \delta_i$ represents a dummy variable formulation for the variation in mean emission levels among individual engines i = 2, ..., 11 and $\Sigma\, b_{j,k} * A_{j,k}$ represents the emission effects of fuel j expressed in terms of the 12 eigenfuel coefficients k.

While the eigenfuels are defined to be mathematically independent of each other, correlations exist between the eigenfuels and the engine dummy variables. There is no unique partitioning of the variance between fuel and engine effects, as a result, and fuel effects should be computed within engines and the separate estimates pooled. Otherwise, differences between emission levels for the various engines can modify the fuel-effect

estimates. This separation of fuel and engine effects is achieved by computing the model sum of squares with both engine and fuel variables included and then with only engine variables included. The total model sum of squares with fuel and engine variables, less the model sum of squares when the model is constrained to engine effects only, is the sum of squares attributed to fuels. This assigns to fuels only the sum of squares reduction that can be uniquely associated with fuels and is referred to as "fuels adjusted for engine effects."

As shown in Table 2.6, the combination of engine effects (representing the composite of engine designs, individual engines, sources, and test cycles) and the fuel effects explain 91.1 and 98.6 percent of the sum of squares for $NO_x$ and PM, respectively. This suggests that the variability of test-to-test replication (for a given engine and fuel) is relatively small compared to the differences among engines and fuels within this data base. The engine effects explain 45.5 percent of the sum of squares for $NO_x$ and 95.4 percent for PM, while the fuel effects represented by the eigenfuel terms explain 45.6 percent and 3.2 percent respectively. The importance of engine effects for PM emissions, while real, is greatly increased by one, older engine whose PM emissions are much above the others.

**Table 2.6.** Sum of Squares for Fuels Adjusted for Engine Effects

| | $NO_x$ EMISSIONS | | | |
|---|---|---|---|---|
| **Source of Variation** | **SS** | **DF** | **MS** | **$R^2$** |
| Regression SS | 1.6334 | 22 | 0.0742 | 0.911 |
| Engine SS (Unadjusted) | 0.8156 | 10 | 0.0816 | 0.455 |
| Fuel SS (Adjusted for Engines) | 0.8178 | 12 | 0.0818 | 0.456 |
| Error SS | 0.1602 | 257 | 0.0006 | 0.089 |
| Total SS | 1.7936 | 279 | 0.0064 | 1.000 |
| | **PM EMISSIONS** | | | |
| **Source of Variation** | **SS** | **DF** | **MS** | **$R^2$** |
| Regression SS | 104.2 | 22 | 4.736 | 0.986 |
| Engine SS (Unadjusted) | 100.8 | 10 | 10.080 | 0.954 |
| Fuel SS (Adjusted for Engines) | 3.4 | 12 | 0.283 | 0.032 |
| Error SS | 1.5 | 257 | 0.006 | 0.014 |
| Total SS | 105.7 | 279 | 0.379 | 1.000 |

It is well known that engine design factors have important effects on emissions, and it is all the more to be expected when, as in present circumstances, the vehicles were designed to varying certifications standards. Further, engine and fuel effects are correlated in this data set – to a substantial extent for PM – so that we are not able to clearly separate their contributions at present. We take these preliminary results to suggest that fuels may have substantial effects on engine emissions, but further work with additional data is clearly needed to resolve the competing importance of engines and fuels. This report focuses primarily on the relative contributions made by the eigenfuels to the portion of the total emissions variation that can be attributed to fuels.

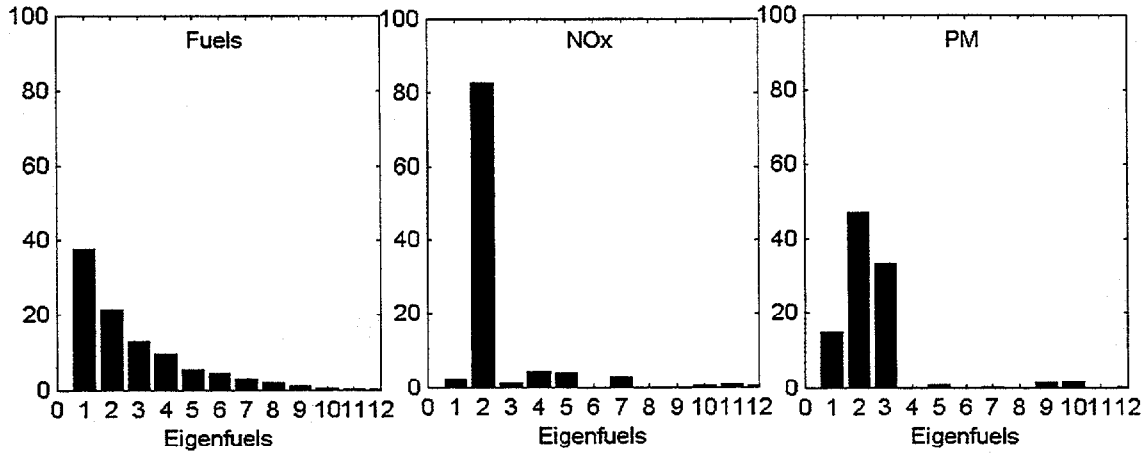
**Table 2.7.** Summary of Regression Results for $NO_x$ and PM

| | Ln($NO_x$) Emissions | | Ln(PM) Emissions | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **t ratio** | **Estimate** | **t ratio** |
| Engines | | | | |
| Intercept | 1.5229 | 248.90 | -0.7683 | 40.15 |
| Engine 2 | 0.0217 | 3.19 | -0.6084 | 28.62 |
| Engine 3 | -0.0308 | 4.63 | -0.5929 | 28.53 |
| Engine 4 | 0.0293 | 3.57 | -0.9771 | 38.03 |
| Engine 5 | 0.0643 | 5.84 | -0.7530 | 21.87 |
| Engine 6 | 0.0339 | 3.98 | -1.7223 | 64.68 |
| Engine 7 | -0.1082 | 7.82 | -0.9844 | 22.75 |
| Engine 8 | 0.0670 | 6.79 | -1.4494 | 47.01 |
| Engine 9 | 0.0413 | 4.07 | -1.6322 | 51.49 |
| Engine 10 | -0.1323 | 11.59 | -1.6107 | 45.11 |
| Engine 11 | 0.0351 | 3.18 | -0.8002 | 23.17 |
| Fuels | | | | |
| Eigenfuel 1 | 0.0043 | 5.30 | 0.0233 | 9.23 |
| Eigenfuel 2 | 0.0344 | 30.78 | 0.0549 | 15.72 |
| Eigenfuel 3 | 0.0051 | 2.46 | 0.0595 | 9.24 |
| Eigenfuel 4 | 0.0120 | 7.09 | -0.0014 | 0.26 |
| Eigenfuel 5 | -0.0149 | 7.62 | 0.0150 | 2.45 |
| Eigenfuel 6 | 0.0027 | 0.98 | 0.0013 | 0.16 |
| Eigenfuel 7 | 0.0173 | 5.87 | 0.0134 | 1.45 |
| Eigenfuel 8 | -0.0031 | 0.66 | 0.0050 | 0.35 |
| Eigenfuel 9 | -0.0057 | 1.22 | -0.0430 | 2.93 |
| Eigenfuel 10 | -0.0156 | 2.59 | -0.0575 | 3.05 |
| Eigenfuel 11 | 0.0294 | 2.85 | 0.0239 | 0.74 |
| Eigenfuel 12 | 0.0325 | 2.62 | -0.0318 | 0.82 |

Table 2.7 summarizes the $NO_x$ and PM regressions. Inspection of the table reveals that all of the engine effects are statistically significant at the 0.05 level (t value exceeding 1.96). Among the fuel effects, all are significant at the 0.05 level except eigenfuels 6, 8, and 9 for $NO_x$ and eigenfuels 4, 6, 7, 8, 11, and 12 for PM. Thus, many fuel effects might be retained in the model if the selection were based solely on statistical significance. However, as we argue, the predictive capability of an effect should temper its selection or rejection in concert with its significance.

The "predictive capability" statistic, computed by normalizing the quantity $\lambda_k * b_k$ to a value of one, identifies the relative contribution of each eigenfuel to the predictive power contributed by all fuels-related information. Using this statistic, Figure 2.3 shows that only one eigenfuel (number 2) for $NO_x$ and three eigenfuels (numbers 1, 2, and 3) for PM account for nearly all of the predictive power that can be ascribed to fuels. This figure also demonstrates, by comparison to the variance explanation for fuels, why all eigenfuels should initially be retained in the regression and considered for deletion only after the relationship to the response variable has been determined. Overall, these results mean that, regardless of the statistical significance of other coefficients, the regression models could be reduced to include at most a few eigenfuel terms (in addition to the engine effects) without a significant reduction in the ability to capture the impact of fuels.

**Figure 2.3.** Percentage Contributions to Variance Explanation for Fuels and Emissions




Let us now briefly examine the substantive meaning of the regression results. For $NO_x$, only eigenfuel number 2 has substantial predictive power. This eigenfuel was previously described as representing a primary aromatics characteristic, in which an increase in aromatics content (both mono-aromatics and poly-aromatics) was associated with increased density and decreased natural cetane. From a refinery perspective, this was identified as representing high-aromatic cracked stocks.

Exponentiating the individual terms of the regression equation, the results indicate that $NO_x$ emissions are decreased by a factor of exp(0.0344) - 1 = 3.5 percent for each unit reduction in this fuel characteristic. Because the variance associated with the second eigenfuel is 2.591, a unit reduction corresponds to 1.000/sqrt(2.591) = 0.62 standard deviations. Therefore, a one standard deviation reduction corresponds to reducing $NO_x$ emissions by 3.5/0.62 = 5.6 percent. A reduction by one standard deviation corresponds to approximately one-third of the total change that is possible and is used here as a rule-of-thumb measure of what might be possible to achieve in practice.

This eigenfuel expresses three individual fuel properties that are widely believed to influence $NO_x$ emissions – aromatics content, natural cetane, and density. However, it represents a single mode of variation involving simultaneous changes in the three variables. Thus, *it may be more correct to speak of reducing $NO_x$ emissions by decreasing the content of high-aromatic cracked stocks*, rather than by varying any of the three properties independently.

The results are somewhat more complicated for PM, since three of the eigenfuels – numbers 2, 3, and 1 in order – are found to contribute substantially to the fuels-related predictive power. The most important, eigenfuel number 2, was identified with the content of high-aromatic cracked stocks. Using the calculation shown above, a one standard deviation reduction corresponds to a (exp(0.0549)-1)*sqrt(2.591) = 9.1 percent reduction in PM. Eigenvector 3, involving a primary association between sulfur content and cetane boost, corresponds to a 7.6 percent PM reduction for each standard deviation change. Eigenvector 1, involving a primary viscosity and distillation curve characteristic, corresponds to a 5.0 percent reduction per standard deviation. The effects are additive, since the eigenfuels are independent, and would reach a total of 21.7 percent if a one standard deviation reduction were made in all three. All other eigenfuels make negligible contributions, whether they are found to be statistically significant or not.

As indicated in the review by Lee, Pedley, and Hobbs, there is only a weak consensus on how fuel properties affect PM, except that reducing sulfur content is generally accepted to reduce emissions. Eigenvector 3 appears to express this consensus relationship. The properties involved in the other eigenfuels (1 and 2) include aromatics content, natural cetane, viscosity, the distillation curve, and to a lesser extent density. Density and poly-aromatics content are thought to have a small effect on PM emissions in some engine groups, while there is no consensus on whether cetane, mono-aromatic content, viscosity, or distillation curve parameters are important. We can not resolve these points of potential difference based on the current data base and analysis. However, the results presented here suggest there is more than one way in which PM emissions can be reduced.

Note also that several of the smaller eigenfuels (numbers 11 and 12 for $NO_x$ and 9 and 10 for PM) are indicated to have strong relationships to emissions. It is a natural statistical consequence that eigenvectors associated with small eigenvalues have regression coefficients that are subject to very wide error bounds. However, these "statistically challenged" eigenfuels *could* point toward unexploited modes of reducing $NO_x$ and PM emissions, even though the current state of the analysis and the size and scope of the engine emissions data base are inadequate to draw conclusions on their meaning or potential importance. These findings could also result from correlations to factors that have been inadequately controlled in the regressions. If the emissions relationships are real, it might still prove impractical, if not impossible, to blend fuels that vary significantly in these characteristics because of considerations of cost, safety, driveability, or other reasons.

## 2.4 MODEL INTERPRETATION AND SIMPLIFICATION

Investigators accustomed to scalar predictor variables may perceive vector variables to be needlessly complex and may find it difficult to interpret the response variable in terms of eigenvectors. In the psychological and social sciences, this objection may be particularly *apropos* because of the lack of underlying theory as a logical assist. In the physical sciences, however, interdependent variables may often be readily recognized. Certainly, in the present instance, blending experience was a valuable interpretational aid.

Nevertheless, a model for predicting emissions in terms of fuel characteristics can not be considered complete until all means of simplification have been explored. Therefore, simplification procedures have been an important concern from the beginning. Though well aware of rotation procedures (varimax, quartimax and equimax), our experience with these procedures failed to provide any significant insights in the context of this problem. Instead, with regard to emissions response, which is our main concern, we elected to pursue a direction aimed at better understanding the relation between eigenvectors and their fuel-property components.

A very useful outcome is a scheme for re-expressing the SS partitioning among eigenvectors as a partitioning among the underlying fuel variables. The conversion is made possible, of course, by virtue of the fact that each eigenvector can be expressed as a linear combination of the original variables. As shown in Appendix D, the SS attributed to each eigenvector can be segregated into the contributions made by each of the original variables comprising the eigenvector, and these individual contributions can be summed over the eigenvectors to yield the SS attributable to each variable. Since it corresponds directly to the eigenvector partitioning, the derived partitioning among variables is in a sense unique. Certainly it has a claim to distinction from the multiple partitionings that can arise in stepwise regression or when all possible subset models are considered.

The SS partitionings for $NO_x$ and PM are shown in Table 2.8, both by eigenvectors and by the original fuel variables. The difference in partitioning for the two pollutants is evident and again emphasizes the futility of attempting to select variables independently of the role they play as predictors. We show in the following a means for simplifying the emission models through a combination of statistical tests of significance and evaluation of the magnitude of effects.

Eigenfuels can be rejected, as is usually done in OLS, by dropping those that fail to meet a specific level of significance. At the 0.05 level, those eigenfuels tagged * in the table would be removed. However, it is to be noted that several other eigenfuels, tagged ** in the table, each contribute less than 1 percent to the model SS. On the assumption that elimination of these eigenfuels would have little influence on the predictive capability of the model, they can be considered for removal, because it is irrelevant whether they are statistically significant or not.

**Table 2.8.** Statistical Significance and SS Partitioning

| | log( NOx ) Emissions | | log( PM ) Emissions | |
|---|---|---|---|---|
| **Eigenvectors** | t ratio | Model SS | t ratio | Model SS |
| **1** | 5.30 | 2.24 | 9.23 | 14.8 |
| **2** | 30.78 | 82.49 | 15.72 | 47.1 |
| **3** | 2.46 | 1.07 | 9.24 | 33.1 |
| **4** | 7.09 | 4.56 | 0.26* | 0.0** |
| **5** | 7.62 | 4.08 | 2.45 | 0.9** |
| **6** | 0.98* | 0.11** | 0.16* | 0.0** |
| **7** | 5.87 | 3.15 | 1.45 | 0.4** |
| **8** | 0.66* | 0.06** | 0.35* | 0.0** |
| **9** | 1.22* | 0.12** | 2.93 | 1.6 |
| **10** | 2.59 | 0.54** | 3.05 | 1.7 |
| **11** | 2.85 | 0.84** | 0.74* | 0.1** |
| **12** | 2.62 | 0.73** | 0.82 | 0.2** |
| | | | | |
| **NatCetane** | | 19.28 | | 4.98 |
| **CetImprv** | | 1.55 | | 1.55 |
| **Density** | | 26.55 | | 17.72 |
| **Viscosity** | | 0.24** | | 0.59** |
| **Sulfur** | | 3.60 | | 31.87 |
| **MonoArom** | | 38.14 | | 6.61 |
| **PolyArom** | | 8.02 | | 27.94 |
| **IBP** | | 0.06** | | 0.32 |
| **T10** | | 0.01** | | 6.13 |
| **T50** | | 0.66** | | 0.74** |
| **T90** | | 1.72 | | 0.25** |
| **FBP** | | 0.18** | | 1.30 |

The disposition of eigenfuels on the basis of statistical significance is, of course, dependent on the significance level adopted. Similarly, the cutpoint for practical significance is arbitrary. It is here that art and judgement, tempered by experience, enter just as it does in any other scheme for selecting variables. It is not the intent of this report to resolve these issues, but rather to illustrate a methodology for their resolution. We suggest, therefore, a possible scenario for model simplification:

- Reject those eigenvectors that have a t-ratio smaller than 1.96, corresponding to 0.05 significance for large samples.
- Renormalize the percent SS and transform the eigenvector contributions into the contributions to SS by individual variables.

- Prune those components that contribute less that 1 percent to the model SS; we call this criterion a "substantiality threshold."

In this way we have retained predictive components that are both statistically and practically significant, given the criteria used for choice. It removes one objection sometimes raised in connection with PCR, namely that the model retains all variables and therefore effects no simplification. Further refinement can be had by recomputing the eigenvectors in the reduced-variable space. The simplified internal structure resulting from the pruning of some variables makes interpretation of the eigenvectors easier.

In addition, the scenario has isolated those fuel variables that play important roles in prediction. The model can be restated, if desired, in terms of the fuel property variables, rather than eigenfuels, in which case the relative contributions can be appraised with the knowledge that the partitioned SS originate in the"clean" eigenvector environment as opposed to the multi-collinear environment of stepwise or all-subset regression.

Let us now apply the simplification criteria to the data set and follow the pruning process through to its completion in a reduced-variable space. Dropping those eigenvectors tagged * or ** in the prior table, we have the following surviving eigenvectors that are statistically significant at the 0.05 level and contribute at least 1.0 percent to the model SS:

- For $NO_x$, the six eigenvectors numbered 1, 2, 3, 4, 5, and 7
- For PM, the five eigenvectors numbered 1, 2, 3, 9, and 10.

The 6 eigenvectors remaining in the $NO_x$ model account for 97.6 percent of the total SS that can be attributed to fuels, while reducing the dimensionality of the analysis space by half – i.e., from 12 to 6 independent regressors. The 5 eigenvectors remaining for PM account for 98.3 percent of the SS and reduce the dimensionality by more than one-half.

Let us now see what the simplified $NO_x$ and PM models imply for the SS partitioning by the original variables. We will apply the criterion that variables contributing less than 1.0 percent of the total SS, tagged ** in Table 2.9, can be dropped from further consideration. Seven variables would be dropped for $NO_x$, including viscosity and all five points on the distillation curve. Five variables would be dropped for PM, including viscosity and four of the points on the distillation curve. One could retain separate variable slates for $NO_x$ and PM in model development, but for the purposes of demonstrating the approach we will retain the seven variables common to both, specifically: natural cetane, cetane improvement, density, sulfur content, mono- and poly-aromatics content, and T10. These seven account for 98.7 percent of the SS in the 6-eigenfuel $NO_x$ model and 98.0 percent of the SS for the 5-eigenfuel PM model.

Note that the pruning criteria must be chosen with consideration for their cumulative effect. The criteria illustrated here retained variables accounting for 97.6 of the total

fuels-related SS in the simplified $NO_x$ model and 98.7 percent of this amount in the reduced variable slate. Accounting for the SS "lost" in reducing the models to six and five eigenfuels for $NO_x$ and PM, respectively, the reduced variable slate retains 96.3 percent of the total fuels-related SS for $NO_x$ and a corresponding 97.0 percent for PM. Thus, we see that the dimensionality of the problem, in terms of individual fuel properties, can be reduced from 12 to 7 with only minor loss of information related to the fuels effect.

**Table 2.9.** SS Partitioning Based on Simplified $NO_x$ and PM Models

| | NOx Model SS | | PM Model SS | |
|---|---|---|---|---|
| | All Vectors | 6 Vectors | All Vectors | 5 Vectors |
| **NatCetane** | 19.28 | 27.22 | 4.98 | 4.84 |
| **CetImprv** | 1.55 | 1.70 | 1.55 | 3.96 |
| **Density** | 26.55 | 23.74 | 17.72 | 21.25 |
| **Viscosity** | 0.24 | 0.10** | 0.59 | 0.33** |
| **Sulfur** | 3.60 | 4.33 | 31.87 | 23.54 |
| **MonoArom** | 38.14 | 34.96 | 6.61 | 6.56 |
| **PolyArom** | 8.02 | 6.71 | 27.94 | 31.33 |
| **IBP** | 0.06 | 0.42** | 0.32 | 0.13** |
| **T10** | 0.01 | 0.00** | 6.13 | 6.54 |
| **T50** | 0.66 | 0.13** | 0.74 | 0.58** |
| **T90** | 1.72 | 0.41** | 0.25 | 0.03** |
| **FBP** | 0.18 | 0.27** | 1.30 | 0.91** |

Even greater simplification could be achieved by raising the substantiality threshold to prune eigenvectors and fuel property variables that contribute between 1 and 2 percent to the total, at the expense of reducing the proportion of the fuels information that is retained in the analysis. There are as many ways to select the simplification and pruning criteria as there are ways to view what is important to a problem. In many contexts the priority will be to retain almost all of the explanatory power while achieving some benefit from the simplification. Highly simplified models may be more important in other instances and will argue for retaining a smaller portion of the total.

Having eliminated fuel property variables that are superfluous to the problem, let us now re-compute the eigenvectors. Table 2.10 summarizes the results of the principal components analysis for the reduced variable space; the largest components in each eigenfuel have been highlighted for comparison to the previous Table 2.3. Important points of comparison are:

- The first revised eigenfuel corresponds to original eigenvector 2, identified with high-aromatic cracked stocks. This eigenfuel has natural cetane, density, and aromatics content as its most important components.
- The second revised eigenfuel corresponds to original eigenfuel 3, identified with a fuel quality feature, with cetane improvement, sulfur content, and poly-aromatics content as its most important components.
- The third revised eigenfuel is one not seen before. It loads predominantly on T10 and accounts for 17.6 percent of the fuels variance. With elimination of the other temperature points, this appears to be a representation of the overall distillation curve in a highly simplified way.
- The fourth revised eigenfuel corresponds (with a sign inversion) to original eigenfuel 5, which was not interpreted in refinery terms.
- The fifth revised eigenfuel corresponds to original eigenfuel 7, which also was not given an interpretation.
- The sixth and seventh revised eigenfuels do not correspond to any of the original eigenfuels.

**Table 2.10.** Revised Eigenvectors of the Test Fuels Data Set

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| NatCetane | **0.484** | -0.261 | 0.248 | -0.280 | 0.050 | **0.598** | **0.444** |
| CetImprv | -0.134 | **0.538** | -0.018 | **-0.831** | 0.005 | -0.036 | -0.003 |
| Density | **-0.547** | -0.079 | 0.298 | 0.038 | -0.037 | -0.242 | **0.737** |
| Sulfur | -0.171 | **-0.608** | -0.307 | **-0.348** | **0.596** | -0.159 | -0.072 |
| MonoArom | **-0.488** | 0.255 | 0.098 | 0.218 | **0.451** | **0.642** | -0.152 |
| PolyArom | **-0.415** | **-0.365** | -0.274 | -0.183 | **-0.659** | **0.363** | -0.141 |
| T10 | -0.095 | -0.261 | **0.819** | -0.165 | -0.053 | -0.109 | **-0.459** |
| Eigenvalues | 2.698 | 1.538 | 1.233 | 0.731 | 0.394 | 0.309 | 0.097 |
| Pct Variance | 38.545 | 21.973 | 17.608 | 10.449 | 5.625 | 4.410 | 1.390 |
| Cumulative Pct | 38.545 | 60.518 | 78.126 | 88.575 | 94.200 | 98.610 | 100.000 |

The eigenfuels that were found to be important features of diesel fuels in the original 12-variable space are retained in some form in the revised eigenvector slate. However, the first original eigenfuel has disappeared from the revised set. It was interpreted as

representing a generalized fuel weight feature with density, viscosity, and the overall height of the distillation curve as its major features. Of these variables, only density and T10 have been retained, and there is no way to observe the variation of density in company with viscosity and the entire distillation curve in the revised data set. This variation remains in the data, but it will be distributed among the revised eigenvectors.

Revised regression models of the form given in Eq. 5, but including the seven revised eigenfuels with engine effects, were estimated for $NO_x$ and PM emissions. Table 2.11 summarizes the results for statistical significance and SS partitioning for the fuels-related effects in the models. As before, we find that one eigenfuel (the first revised eigenfuel representing high-aromatic cracked stocks) is the dominant explanatory factor for $NO_x$ emissions and the largest of several for PM.

**Table 2.11.** Statistical Significance and SS Partitioning for the Revised $NO_x$ and PM Models in Seven-Variable Space

| | log( $NO_x$ ) Emissions | | log( PM ) Emissions | |
|---|---|---|---|---|
| **Revised Eigenvector** | t ratio | Model SS (percent) | t ratio | Model SS (percent) |
| **1** | 30.94 | 87.95 | 19.36 | 61.64 |
| **2** | 2.25 | 0.78** | 11.46 | 36.31 |
| **3** | 3.18 | 0.99** | 0.63* | 0.07** |
| **4** | 9.34 | 7.17 | 2.73 | 1.10 |
| **5** | 5.28 | 3.09 | 1.89* | 0.71** |
| **6** | 0.31* | 0.01** | 0.48* | 0.04** |
| **7** | 0.14* | 0.00** | 0.94* | 0.14** |
| | | | | |
| **NatCetane** | | 27.94 | | 4.97 |
| **CetImprv** | | 2.00 | | 1.70 |
| **Density** | | 24.33 | | 19.33 |
| **Sulfur** | | 6.66 | | 36.10 |
| **MonoArom** | | 31.13 | | 5.55 |
| **PolyArom** | | 7.90 | | 26.51 |
| **T10** | | 0.04** | | 5.84 |

For $NO_x$, the one dominant eigenvector accounts for 88.0 percent of the fuels related SS, while its counterpart in the original slate of 12 accounted for 82.5 percent. Reducing this fuel feature by one standard deviation would reduce $NO_x$ emissions by 5.6 percent. The 12-eigenvector analysis (see Section 2.3.3) estimates the same one standard deviation reduction of the corresponding eigenvector also to reduce $NO_x$ by 5.6 percent. The fourth

revised eigenvector accounts for an additional 7.2 percent of the fuels-related SS and would reduce $NO_x$ by an additional 1.6 percent.

For PM, we have two revised eigenfuels (numbers 1 and 2) that account for almost all of the fuels effect, compared to three (eigenfuels 3, 2, and 1) in the original analysis. These two correspond to the properties described in the original analysis by eigenfuels 2 and 3. As noted, the original first eigenfuel (fuel weight) has disappeared from the revised set, and its impact on PM emissions is distributed to other revised eigenfuels.
Reducing revised eigenfuel 1 by one standard deviation is estimated to reduce PM emissions by 10.4 percent, while reducing revised eigenfuel 2 by one standard deviation gives a PM reduction of 8.0 percent. Our original analysis obtained estimates of 9.1 percent and 7.6 percent reductions for the corresponding eigenfuels.

Thus, we have two-eigenfuel models for both $NO_x$ and PM that capture about 95 percent of the fuels-related SS for emissions. A 7.2 percent reduction in $NO_x$ could be achieved if revised eigenfuels 1 and 4 were each reduced by one standard deviation, and an 18.4 percent reduction in PM could be achieved if revised eigenfuels 1 and 2 were each similarly reduced. Overall, the regression analysis and the SS partitioning for both eigenfuels and fuel properties tell the same story seen in the original analysis:

- $NO_x$ is influenced primarily by natural cetane, density, mono- and poly-aromatics content.
- PM is influence primarily by density, sulfur content, and poly-aromatics.

Predictions made by the simplified model are, for all practical purposes, the same as in the original analysis. It seems that 7 individual fuel properties suffice, and that $NO_x$ and PM emissions can each be modeled with relationships involving 2 eigenfuels.

# 3. PERSPECTIVE AND EXTENSIONS

The development of the vector emissions model is, thus far, admittedly simplistic and does not cover many of the difficulties that can be expected to arise in practice. We are well aware that the data base used in this demonstration has many shortcomings. It is a pooling of separate studies, each of which was performed with a specific objective in mind. Vehicle sampling was inadequate and did not allow estimation of effects attributable to specific engine characteristics. We urge caution in interpreting results and stress again that the major purpose is to demonstrate methodology, rather than to draw conclusions regarding the fuel/emissions relationship.

In the following sections, we first develop a perspective on the meaning and interpretation of sums of squares to emphasize how the orthogonalization achieved in the vector approach leads to a unique and preferred identification of the explanatory influences in a regression model. We then discuss implications of this perspective for how the vector approach should be applied in future work. Finally, we address a series of methodological extensions that may be needed to meet the challenges of real-world data. The discussion within this section summarizes key points and findings and, in many instances, leaves the more-detailed development and explanation to the referenced appendices.

## 3.1 PERSPECTIVE ON SUMS OF SQUARES

### 3.1.1 The Many Faces of Sums of Squares

Consider the general case to which the problem of HDD engine emissions belongs – i.e., a regression model involving N predictor variables and the problem of selecting the ideal subset of these variables to include in the regression equation. Each of the N variables can be included or excluded independently of all the others. Consequently, there are $2^N$ possible subsets, each being a possible choice of predictor variables to include in the model. The number of included variables can range from zero (the empty subset) to N (the universal subset).

If the design matrix is not columnwise orthogonal, it is not possible to compute directly the fraction of the model sum of squares (SS) attributable to a given predictor. Customarily, this fraction is estimated by computing the model SS with all N predictors and then with all predictors except the one for which the SS is to be estimated. Then, as seems reasonable, the *difference* between the two sums of squares is taken as the contribution attributed to the excluded variable. This estimates the SS contribution *at the*

*margin*, given the other variables present, but does not determine an absolute or non-conditional SS contribution.

This procedure can be broadly generalized into an "all possible regressions" approach. For any one of the predictor variables, exactly one-half of all the $2^N$ subsets contain that variable, while the other half do not. Therefore, it is possible to compute $2^{N-1}$ estimates of the SS attributable to any given predictor. For 12 such variables, there are $2^{N-1} = 2048$ possible estimates of the SS that conceivably might be attributed to a given predictor variable. It is little wonder, then, that different researchers often come to different conclusions. On the other hand, if the design matrix is columnwise orthogonal, all of those $2^{N-1}$ estimates are identical. It would seem that no further argument should be necessary to justify orthogonalization as the method of choice for estimating the relative impact of predictor variables on response.

### 3.1.1.1 A Three-Variable Example

With modern computers it is quite feasible to compute all possible regression estimates for every one of the N predictor variables. To enumerate all such estimates is hardly practical except for small N, but the estimates for large N can be summarized statistically and expressed as tables or histograms. A small, albeit trivial, example containing only three variables will provide further insight into the many faces of sums of squares in an environment of correlated variables.

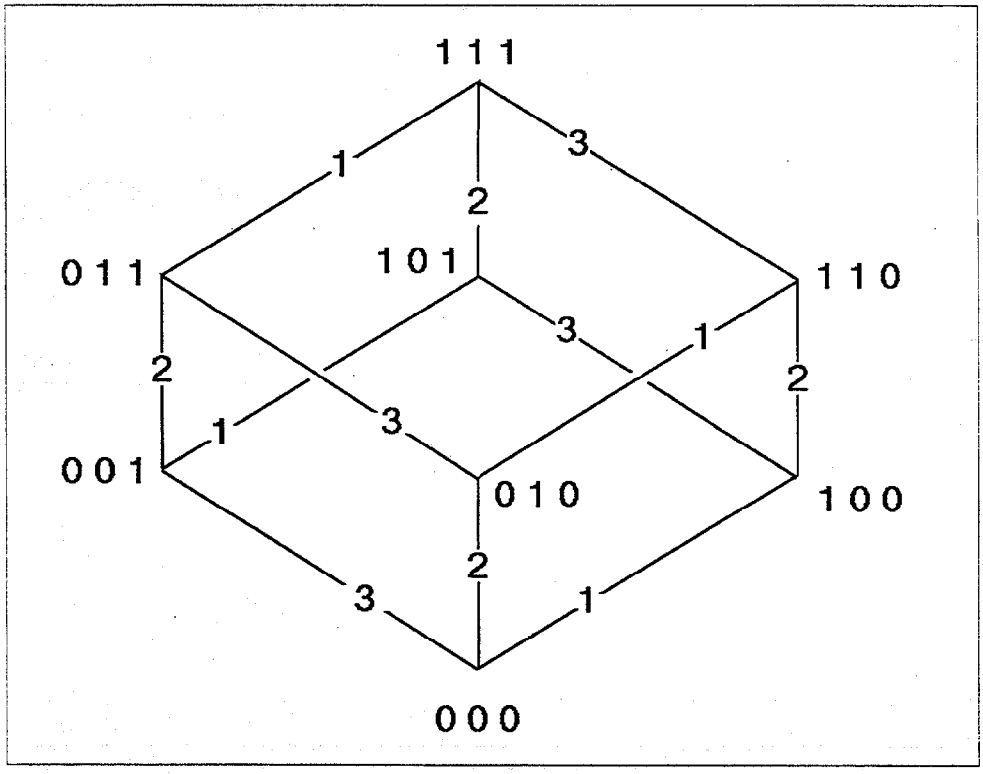
**Table 3.1.** Demonstration Data Set

| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| 7 | 4 | 3 | 11.0 |
| 4 | 1 | 8 | 3.2 |
| 6 | 3 | 5 | 5.1 |
| 8 | 6 | 1 | 19.1 |
| 8 | 5 | 7 | 9.5 |
| 7 | 2 | 9 | 5.6 |
| 5 | 3 | 3 | 5.8 |
| 9 | 5 | 8 | 11.7 |
| 7 | 4 | 5 | 8.0 |
| 8 | 2 | 2 | 14.2 |

The demonstration data set is shown in Table 3.1. Let us use the convention that, if a variable is included in a submodel, its inclusion is denoted by 1. If the variable is excluded, its exclusion is denoted by 0. Thus [1 1 1] means that all three variables are included in the model, while [0 1 1] denotes that variable $x_1$ is excluded, while variables $x_2$ and $x_3$ are included. The possible subsets of the model are displayed in Figure 3.1 as a partially ordered system called a *lattice*.

If the figure is viewed as a cube, corners of the cube represent SS estimates for the eight subset regressions. Edges of the cube along the three principal directions

**Figure 3.1.** Lattice Representation of SS for Three Variables




correspond to the SS estimates developed from the differences between paired regressions, in each case defined to include and exclude only one of the three variables. For example, the difference between the model SS computed for the corner [1 1 1] and for [0 1 1] provides an estimate of the SS for variable $x_1$. Similarly one can compute estimates for $x_2$ and $x_3$. There are four such estimates for each variable and twelve different SS estimates in this system.

Table 3.2 summarizes the model sums of squares and $R^2$ for these three variables as estimated by difference using the ANOVA result for each of the possible regression equations. Just how widely the estimates vary can be appreciated by the bar graph of Figure 3.2. How can this variability be explained, and how can the various estimates be mapped, if possible, into a *single, unique estimate* for each of the three variables?

A reasonable answer to this question is provided by considering the estimates as elements of a partially-ordered system. Beginning with the vertex at the top of the lattice, the point [1 1 1] signifies that all three variables are included in the model. Any one of those variables can be dropped out, as indicated by the three downward branches emanating from the top vertex. These first three estimates represent the effects of $x_1$, $x_2$, and $x_3$ *when all three variables are included in the model.*

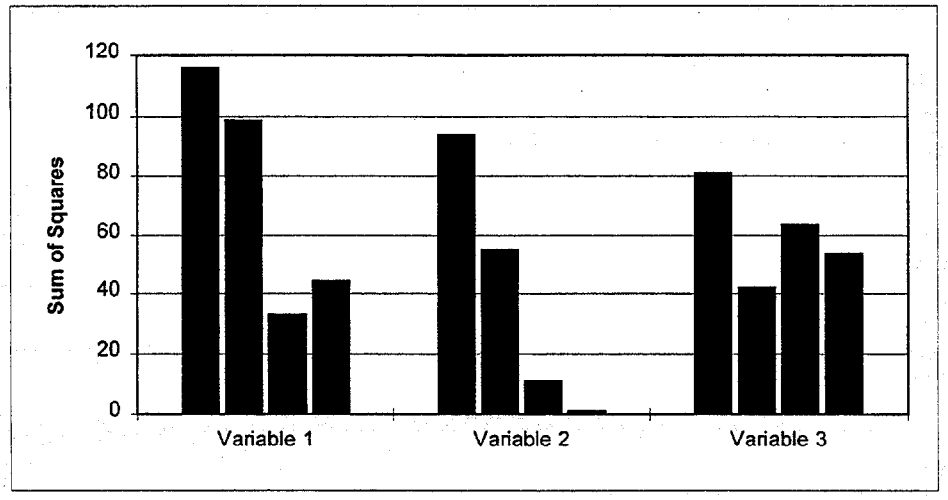
**Table 3.2.** Estimation of Sums of Squares by All Possible Regressions

| Inclusion Set | | | | Exclusion Set | | | | SS | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Variable $X_3$ | | | | | | | | | |
| 0 | 0 | 1 | | 0 | 0 | 0 | | 81.0814 | 0.3839 |
| 0 | 1 | 1 | | 0 | 1 | 0 | | 42.1636 | 0.1996 |
| 1 | 0 | 1 | | 1 | 0 | 0 | | 63.2768 | 0.2996 |
| 1 | 1 | 1 | | 1 | 1 | 0 | | 53.1949 | 0.2519 |
| Variable $X_2$ | | | | | | | | | |
| 0 | 1 | 0 | | 0 | 0 | 0 | | 93.6380 | 0.4433 |
| 0 | 1 | 1 | | 0 | 0 | 1 | | 54.7182 | 0.2591 |
| 1 | 1 | 0 | | 1 | 0 | 0 | | 11.1022 | 0.0526 |
| 1 | 1 | 1 | | 1 | 0 | 1 | | 1.0203 | 0.0048 |
| Variable $X_1$ | | | | | | | | | |
| 1 | 0 | 0 | | 0 | 0 | 0 | | 115.9143 | 0.5488 |
| 1 | 0 | 1 | | 0 | 0 | 1 | | 98.1097 | 0.4645 |
| 1 | 1 | 0 | | 0 | 1 | 0 | | 33.3804 | 0.1580 |
| 1 | 1 | 1 | | 0 | 1 | 1 | | 44.4118 | 0.2103 |

At the next level of the lattice, SS estimates can be computed for the three variables, but in these cases representing the contributions each variable makes *when only it and one other variable are included in the model* – i.e., given that one of the other variables had already been eliminated from the model by some means. Finally, at the lowest level of the lattice, the difference approach estimates the SS contributions for each variable *when that variable is the only variable in the model.*

The conclusion to be drawn here is that, unless all variables are mutually orthogonal, the apparent impact of any one of the predictors is determined to a considerable degree by the environment in which it is situated. Observe, too, that these various estimates correspond to steps in a stepwise regression approach, because any intermediate model in stepwise regression corresponds to one of the $2^N$ submodels among "all possible regressions." Also, it should be evident that there is a (structured) multiplicity of paths by which any stepwise model can be reached.

**Figure 3.2.** SS for Subset Models in Three-Variable System




### 3.1.1.2 Relationship to the Vector Approach

It is our contention that, in an N-dimensional regression analysis, any of the $2^{N-1}$ estimates is a legitimate choice for the effect of a selected variable and that the method of dropping the variable from the full model should not be taken as the only way to estimate its SS contribution. In fact, the only unique estimate is obtained by orthogonalization, as achieved in the vector approach.

In the vector approach, one shifts from P-Space, in which regression coefficients are computed for the correlated fuel property variables, to E-Space, in which regression coefficients are computed for uncorrelated eigenvectors. Since each of these eigenvectors has the original variables as components, one can combine the independent E-Space estimates into an aggregated P-Space estimate. This estimate represents the SS contribution of the original variables resulting from *the very special orthogonal environment that isolates the independent components of the SS.* The SS partitioning developed in the vector approach is distinct – i.e., it is not any one of the difference-based estimates – and is to be preferred over all others because of the orthogonal environment in which it originates.

The interpretation of this particular partitioning of the SS can be better appreciated by considering the contributions of the P-Space variables in two stages. If $p_1$, $p_2$, and $p_3$ denote, respectively, the percentage contributions of eigenvectors 1, 2, and 3 to the model SS, then one can break down each eigenvector's individual contribution into the contributions made by its components $x_1$, $x_2$, and $x_3$. These percentage contributions can be denoted, respectively, as $p_{11}$, $p_{12}$, $p_{13}$, $p_{21}$, $p_{22}$, $p_{23}$, $p_{31}$, $p_{32}$, and $p_{33}$. This micro-partitioning of the model SS is shown in Table 3.3.

**Table 3.3.** Partitioning of Model SS Among Variables and Eigenvectors

| | Eigenvector 1 | Eigenvector 2 | Eigenvector 3 | Sum |
|---|---|---|---|---|
| Variable 1 | 35.7617 | 0.8980 | 3.1525 | 39.8122 |
| Variable 2 | 40.8741 | 0.0573 | 3.7231 | 44.6545 |
| Variable 3 | 10.1287 | 5.1132 | 0.2915 | 15.5334 |
| Sum | 86.7646 | 6.0684 | 7.1670 | 100.0000 |

As would be expected, the original variables exert their influence primarily by way of eigenvector 1, because that eigenvector as a whole accounts for 86.8 percent of the model SS. However, it is to be noted that the components of that eigenvector account, respectively, for 35.8, 40.9, and 10.1 percent of its contribution. When the minor contributions from the other two eigenvectors are added in, the contributions of variables 1, 2, and 3 are, respectively, 39.8, 44.7, and 15.5 percent.

It is to be noted that when variables are removed to make subset models, the percentage contributions to the model SS changes, just as it did when models were formed from the original variables, but in a much simpler way. Furthermore, if model simplification is performed by first removing nonsignificant and/or nonsubstantial eigenvectors, the variable removal process approaches a fully-ordered, rather than a partially-ordered, system.

The major aspects of the simplification are fairly obvious. Whereas in P-space regression there are $2^{N-1}$ ways to compute the sum of squares contribution for each variable, there is only $2^0 = 1$ way in E-space. Thus, we reduce the "entropy" or uncertainty by a factor of $2^{N-1}$. It is hard to visualize the branching system for 12 variables, which would have thousands of vertices and branches. We need not do so when the system is examined in E-Space. In short, the eigenvector-based partitioning of the sums of squares makes possible a step-by-step simplification process while avoiding the pitfalls of stepwise algorithms.

### 3.1.2 Model Sums of Squares as Induced Distributions

It is evident that what one computes as contributions to the model sum of squares is determined, in large degree, by how the predictor space is sampled. Indeed, there is no such thing as a SS partitioning without reference to a sampling scheme. Consider the case of a predictive equation where we can consider the coefficients to be fixed, immutable, and not subject to chance, as might be derived from an applicable theory. How much each coefficient contributes to the model SS depends on how each of the arguments of the equation are sampled to produce a set of data.

As computed by the ANOVA table in a typical regression program, all sums of squares are based on sampling *the design matrix* or some subset thereof. In fact, the residuals pertain only to the points actually sampled in the experiment. They can tell us nothing about differences between actual and computed responses at points not in the design. Indeed, it is possible to fit an unlimited number of equations to the data in such a way that they yield exactly the same error sum of squares, and even more, have exactly the same point by point residuals.

This fact raises an interesting question. What if the regression equation is evaluated at points *not included* in the data set? Then, how would the relative importance of the predictor variables appear? Certainly they would not necessarily duplicate what was found from ANOVA in the design sample space. For example, it is entirely conceivable that all predictor variables, except one, are held constant and the one is varied over some range of its allowable values. The partitioning in this case would assign 100 percent of the SS to that single variable. Thus, we see clearly that the partitioning assigned by the regression ANOVA does not necessarily define the SS partitioning to be found from the application of the regression equation to the real world.

As a mathematical construct, SS partitioning is analogous to defining an induced distribution as a distribution of a *function* of a random variable. For example, let f(x) be N(0,1) and let g(x) = sin(x), $0 <= x <= 2\pi$. Then, f( g(x) ) is a distribution induced by a normal distribution operating on a sine function. There is nothing "random" about the sine function, but it can be sampled in accordance with a defined probability distribution to produce a random outcome. Most importantly, what comes out depends on what is put in for the sampling distribution – it can be continuous or discrete, infinite or finite, even just a set of numbers like the set of "treatments" in a design matrix.

Random balance sampling is an example of the induced distribution concept that was found useful in the development of the Complex Model for Reformulated Gasoline (U.S. DOE, 1994). In this example, the predictor variables were often correlated, so that it was difficult to isolate the predictor variables that contributed most or least to the response. Yet, in the interest of simplification, it was desirable to eliminate from the model any variable that contributed minimally to prediction capability.

Again, orthogonality seemed essential to making appropriate choices as to what terms to include in the regression equation and what terms to exclude. Random sampling from a multivariate rectangular distribution was found to provide a means for making these choices in the Complex Model. It was postulated that, in applications of the model, values of the predictor variables between certain limits were equally likely to be used in the formulation of fuel blends. Random samples from a multivariate uniform distribution can be shown to be uncorrelated, so that the relative contributions to the model SS from predictor variables are essentially additive, and their relative importance can be readily assessed.

More generally, it is our belief that attention should be directed not only to the data set, but also to the probable set of "treatments" likely to be used in real-world applications of the model. This will have influence on the preferred approach to statistical inference and on the range of variation in fuel properties that is to be considered in the development of an emissions model.

## 3.2 IMPLICATIONS FOR THE METHODOLOGY

### 3.2.1 Hypothesis Testing: t versus F

The foregoing discussion argues that the PCR method of partitioning the model SS among the eigenvectors and, by extension, among the variables that are components of the eigenvectors, yields a result that is preferred as an indication of the explanatory power of the individual regressors. This partitioning is unique and differs from the result of the other methods used to attribute SS to variables in a non-orthogonal environment. As a result, we must reconsider the methods used for hypothesis testing in the eigenvector approach because the SS estimates at issue play an integral role in the commonly used tests.

In conventional regression analysis, tests of significance are typically of two types:

- t-test for the significance of individual regression coefficients
- F-test for the significance of the model as a whole.

The t-tests involve the ratio of a selected coefficient to its standard error. The F-test involves the ratio of the model mean square to the error mean square. However, it should be realized that the F-test is applicable to the ratio of any two random variables distributed as Chi-square and can be used to test the significance of a term in a regression model if, as is true when the regressors are orthogonal to each other, one can compute a corresponding mean square for that term.

Appendix E develops the use of the F-test for this purpose and demonstrates it on a simplified version of the emissions model developed in this report. As shown there, the t-test and F-test yield identical results in the orthogonal case. The observed value of F is exactly equal to the square of the t-ratio for F-tests involving only a single degree of freedom in the numerator (corresponding to the one degree of freedom associated with each of the eigenvector regressors). The t- and F-tests lead, however, to discordant results in the non-orthogonal case.

When a response variable is regressed against the original, correlated P-Space variables, it is possible for a variable to fail the t-test of significance, even though, on the basis of SS partitioning in E-Space, that variable accounts for a substantial part of the model SS and can easily pass the F-test. The E-Space partitioning is accomplished by first partitioning

the SS among the eigenvectors and then translating that partitioning into the equivalent partitioning among the property variables.

The apparent discrepancy between the two statistical tests arises from aliasing effects that pervade the P-Space variables but do not affect the eigenvectors. For many of the original predictor variables, it is possible to express as much as 90 percent of its variability in terms of other P-Space variables. Thus, a t-test of that variable in P-Space is not really a test of that variable *per se*, but is rather a test of a conglomerate of confounded effects parading under the variable name.

In addition to being distorted, the results of that test are applicable *only* in the company of whatever variables are included in the model. The indicated significance of the subject variable may change markedly if one or more of the other predictor variables is deleted from the model. Although not usually acknowledged, this flaw in the t-test stems directly from the non-additivity of components of the model SS. Means for circumventing this deficiency of the t-test are considered in Appendix E.

We recommend the following approach to hypothesis testing:

- The conventional t-test can be used to test the significance of regression coefficients for eigenvectors, since t- and F-tests are identical in the orthogonal case.
- The F-test should be used in significance tests involving the original variables, as will be encountered at the stage of the simplification process where decisions are made to include or exclude variables from the analysis.

It may also be necessary to use the t-test as a default procedure when the eigenvector terms are mixed with other effects to which they are not orthogonal, an example being fuel and engine effects. One can test both the eigenvectors and engine effects by means of a t-test as a default procedure, but it should be understood that confounding of fuel and engine effects may result if the correlations among them are large. Shortcomings of this procedure, and a possible alternative, are presented in Appendix E.

### 3.2.2 Statistical Inference

The 0.05 level of significance for testing a null hypothesis goes essentially unchallenged in a world where risk, and the consequences of risk, are anything but constant. This icon is comforting because it assures us that we will erroneously reject the null hypothesis (a Type I error) only one time in twenty, and that seems like very good odds. On the other hand, if the effect being rejected actually is real, and not an artifact of sampling, we will be wrong 19 times out of 20 (a Type II error), and that seems like very poor odds. There is a continuum of tradeoffs that can be made in arriving at an optimum policy for managing Type I versus Type II risks, although such methods for risk management are all-too-seldom considered in the actual practice of data analysis.

But what impact does this have on our ability to predict real-world changes in response? Much depends on sample size. If a regression is based on several hundred observations and if the error standard deviation is sufficiently small, it may be possible to declare an effect statistically significant even though its *magnitude*, so far as accuracy of prediction is concerned, may be negligibly small. On the other hand, if sample size is small, we may erroneously accept the null hypothesis unless the effect being tested is, in fact, fairly large. The bottom line is that, as sample size varies from one investigation to another, so will the magnitude of what we can declare to be statistically significant under a fixed probability of Type I error.

For large sample sizes, the outcome of retaining, as statistically significant, effects that have little effect on the response was seen clearly in early stages of the Complex Model for RFG, and the potential for this outcome can also be seen in the smaller, but statistically significant, eigenfuel effects found in Table 2.7. The simplification procedures demonstrated in Section 2 introduced a *substantiality* test to limit unnecessary complexity by requiring that an effect must explain at least a minimum portion of the variation in the dependent variable to be retained in the model.

What one computes as contributions to the model sum of squares is determined, in large degree, by how the predictor space is sampled in the data at hand. If we have reason to believe the existing data appropriately samples the variation in fuel properties that will be encountered in applying the regression model to real-world problems, then we risk little in simplifying the model using substantiality tests based on the sample data. On the other hand, if the sampling distribution induced in our data is dissimilar to that of the real world, effects excluded because they contribute little to the model sum of square could represent heretofore unrecognized and unexploited means of reducing emissions.

Experience tells us that it is often difficult to be confident of the representativeness of one's data. In regard to diesel fuels and emissions, the existing data set covers the range of variation in fuel properties that many researchers consider likely to be encountered in reformulating diesel fuels. Yet, it is unlikely that the data set considers all methods that could be used to reformulate diesel fuel, and it may effectively weight (in terms of numbers) the various reformulation methods differently than actual future practice in refinery operations will show. Thus, it is difficult to judge whether the sampling distribution induced in the data set is representative of future diesel fuels.

An alternative approach to tests of both significance and substantiality is to fix at the outset of a study the magnitude of the effect (i.e., regression coefficient) that one would be willing to ignore. As sample size varies from problem to problem, the Type I error probability associated with this magnitude would vary. If the effect is negligible in terms of its impact on the predicted response, the associated level of uncertainty would be irrelevant. If the effect is large enough to be important, but its associated level of uncertainty is unacceptably high, additional testing to reduce the uncertainty would be a more appropriate response than merely excluding the effect from the model. Thus, we

would prefer to establish at the outset of an investigation the smallest effect that is meaningful to our purpose and drop any that are found to be smaller than this threshold.

## 3.3 METHODOLOGICAL EXTENSIONS

### 3.3.1 Comparing Independent Studies

The scientific method requires that one evaluate research questions based on the preponderance of the evidence in repeated, independent studies. One study is inadequate, independent studies frequently disagree and, therefore, one must search for a consensus. When using conventional regression analysis, we can apply well known procedures to determine whether the differences between two studies are statistically significant – typically by testing whether the difference in regression coefficients for, say, the impact of sulfur on PM emissions, is sufficiently large that it is unlikely to arise solely by chance.

How does one make similar comparisons when using the eigenfuel methodology developed here? After all, two researchers using independent data sets will estimate eigenfuel slates that differ from each other to at least some extent. Since these are the independent variables, can one even formulate a proper comparison between the regression coefficient estimated for eigenvector 1 in the first data set and a different eigenvector 1 in the second data set? As will be shown, comparisons can properly and readily be made by transforming the results to a common basis, and then applying appropriate tests of statistical significance.

The need to transform to a common basis may appear to be a limitation or complication of the eigenfuel methodology. In conventional regression analysis, the independent variables are fixed, and we can compare, for example, a sulfur effect in one study with a sulfur effect in another. Yet, we should remember that a variable used in a regression model will incorporate the influence of other variables, not included in the model, with which it is correlated. The extent of the correlations will differ in another data set and study, as will the variables included in the regression models. A variable such as sulfur content will actually stand for a unique set of influences in each independent data set and study, and the comparison between two studies will be confounded by these differences. Thus, the comparison problem is merely more apparent in the eigenfuel methodology and less often acknowledged in conventional regression analysis.

We have several choices in seeking a common basis for comparison. First, researchers could conduct the comparison in P-space by transforming their regression coefficients and sums-of-squares partitioning into comparable results for the fuel properties. This comparison could always be made if researchers were to adopt the convention of publishing the P-space transformations of their results. It would be an easy and familiar basis for comparison, but it would also lose any information in the studies that was linked to the eigenvectors, but not to their individual components.

To retain this information, one could transform the results of study A into the eigenvector space of study B (or vice versa), or one could transform both studies into the eigenvector space of a third, reference study. We prefer and will adopt the latter approach and use the eigenvectors of the full data set as a reference space for this purpose. However, the reference space could as well be the eigenvectors of some other study.

Appendix F demonstrates the methods for comparing studies by dividing the existing data base into two subsets corresponding to hypothetical studies A and B. The subsets were defined in a systematic manner in an effort to maximize their differences and then subjected to the following process:

- Emissions studies were conducted independently for each subset using the vector methodology developed in this report. The eigenvector structures for subsets A and B were found to have similarities, but also to show obvious differences. For purposes of the demonstration, all eigenfuel effects were retained in the emission models, although researchers would be more likely to compare simplified models.
- As the first step in the comparison, the eigenvector slates for the subsets are transformed to the vector basis of the reference space (here, representing the full data set), where they are seen to share much in common, particularly for the eigenfuel characteristics with the largest eigenvalues.
- The coefficients for $NO_x$ and PM regression equations developed from each subset are then transformed to the equivalent regression coefficients and standard errors applicable to the reference vectors. The transformed coefficients, and standard errors, are identical to those obtained by using the reference eigenvectors directly in a regression conducted for the subset. When, as true here, the regression model mixes orthogonal eigenvector terms with dummy variables for engine effects, the regression coefficients are still exact, but the standard errors approximate only those that would be obtained directly. They are thought to be adequate for purposes of comparison.
- Finally, the vector-based emission regressions for each subset are transformed to P-Space where the predicted impact of the original variables can be compared directly.

Although this report is not primarily intended to assess the fuels/emissions relationship, some of the results of the subset comparison are instructive:

- Both subsets agree that reference vector number 2, representing high-aromatic cracked stocks in diesel fuel blends, is the dominant effect on both $NO_x$ and PM emissions, accounting for approximately 80 percent of the fuels-related SS for $NO_x$ and 45 percent for PM.

- The subsets disagree on the magnitude of the $NO_x$ effect for reference vector 2, with subset B estimating an effect approximately twice as large. Where comparisons are meaningful, subset B estimates a larger impact of fuels on $NO_x$ emissions overall.
- In both subsets, a second reference vector plays a major role in PM emissions, although they disagree on its identity. Reference vector 3 is next in importance in subset A, while vector 1 is next in subset B.

Thus, the results of the studies have much in common when compared on the common basis formed by the reference eigenvectors, and the most important differences between the subsets are clearly displayed.

As seen in some detail in Appendix F, the similarities are fewer and the differences both greater and more confusing when the studies are compared on the basis of the original variables. On this basis we find, for example, that mono-aromatics content, followed by poly-aromatics content and T90, appear to have the largest effects on $NO_x$ emissions in subset A, while it is density, followed by natural cetane and T90, in subset B. And where the same variables are found important, the subsets often differ on the algebraic sign (directional impact) of the effects. There is also substantial discord between the subsets regarding the effect of fuel properties on PM emissions.

We are encouraged to find a general, though not complete, degree of accord for the subsets when viewed in E-space, while we are not surprised to discover substantially more discord and confusion for the comparisons made in P-space. A recurring theme in the literature review by Lee, Pedley, and Hobbs (Lee *et al.*, 1998) and in recent conference presentations (CEC/SAE, 2000) is that the existing studies frequently disagree, a consensus is hard to find, and the reasons for the many differences are not readily apparent.

As we have argued throughout this report, studies conducted in a non-orthogonal environment are subject to confounding influences – specifically, that the apparent significance and importance of a variable will depend on its correlations to other variables, both those present in the model and those excluded, and that the meaning of a variable can change from one data set to another as these correlations change. We believe that these confounding influences are responsible, at least in part, for the degree of confusion existing on this topic and that the vector methodology offers a means to clear some of this fog.

### 3.3.2 Robustness of the Eigenvector Basis

Once an orthogonal basis of eigenfuels has been developed, how "robust" is that basis when applied to a different set of data? The question is not appreciably different from that faced by a conventional regression model when it is applied in a context different from that in which it was developed. If eigenvectors are found to differ greatly from one

data set to another, they may have local use, but will be limited in their global applicability. On the other hand, if eigenvectors prove to display recurring, common features across data sets, we may be more confident that they are robust in describing fuels. The preceding comparison of the data subsets begins to touch on this issue.

Validation sampling, in which the data set is divided into two or more parts and results compared, is an acknowledged approach to this problem (Wold, 1978). We consider bootstrap sampling (Efron and Tibshirani, 1998), in which samples are drawn with replacement from the data set, to be an extension of the validation idea, and one that does not suffer from reduction in sample size. To some, however, validation sampling in either form may not seem like a test at all, because all samples are drawn from the same source. More appropriate would be a procedure in which samples are drawn from different but similar sources with a view to finding the common aspects of response. This approach sometimes goes under the name "system identification" and has been explored by one of the authors in an earlier and different context (McAdams, 1972), although its requirements for multiple sources may limit its applicability to the diesel emissions problem at this point in time.

Bootstrap sampling was explored in this study as a technique that could shed light on the local versus global applicability of the eigenvector slate developed from the data set. Bootstrap sampling has been described as an empirical approach to developing confidence intervals or other measures of accuracy for statistical estimates when applicable theory is either missing or difficult to apply. It treats a sample data set as a pseudo-population from which new samples can be drawn and then analyzed. Assuming the sample data set is representative of a larger population of interest, each simulated sample that is drawn from it (with replacement) is a possible outcome of sampling directly from the larger population. We may learn much about the empirical distribution and variability of statistical measures, even when we are unable to ascribe a theoretical treatment, by analyzing each sample and compiling the results across the samples.

As just one example, we realize that the coefficient values for the component variables forming the eigenvectors will vary from one sample to another, even if all of the samples share a common eigenstructure. Where the coefficients are sufficiently small in a particular vector, we should not be surprised to find that the algebraic signs vary from plus to minus in different samples. How does one form the confidence intervals for these coefficients, so that one can identify terms in a vector that can not reliably be distinguished from zero? The bootstrap sampling procedure provides a method of directly simulating this variation leading to the determination of confidence intervals for the eigenvector coefficients.

Appendix G explores the bootstrap technique as a method for understanding the behavior of diesel fuel eigenvectors, including the determination of mean and median values and confidence intervals for the eigenvector coefficients. The important questions considered in the appendix are whether the eigenvectors tend to repeat themselves from one sample

to the next and, if so, whether they tend to be found in the same eigenvalue-determined order in each sample. The result of the bootstrap experiment shows a substantial degree of stability in vector composition and order across the samples and supports the belief that the reference eigenvectors represent generally applicable and repeatable features of diesel fuels.

### 3.3.3 Representation of Non-linear Effects

Work done to date has been based on the assumption that all vector predictor variables exert a linear effect on the response variable. Experience tells us that the effect of a predictor variable may be linear over much of its range but may show curvature for extreme values (e.g., saturation effects). In other instances, theory may suggest that the effect of the predictor may be non-linear over its entire range. The methodology for a vector approach to regression can not be considered complete unless it can accommodate such non-linear effects.

We consider it straightforward to add non-linear effects to the model simply by incorporating basis vectors exhibiting the desired non-linear characteristics. Appendix H develops our preferred approach in some detail. In brief, the linear terms $x_i$ constituting the original variables may be augmented by squared terms $x_i^2$, interactive terms $x_i\, x_j$, or more generalized non-linear functional forms $f(x_i)$ in forming the X matrix. Moreover, there appears to be no difficulty in retaining the orthogonality of all vectors in the model, provided that the non-linear terms are introduced to the X matrix before its standardization to mean 0 and standard deviation 1.

Interpreting the role played by non-linear elements incorporated in *each* of the basis vectors may seem to pose a serious problem to the data analyst. Because of the correlation of the non-linear variable with one or more of the *linear* variables, the non-linear term may be significant and substantial in *some* eigenvectors but not in others, just as in a purely linear model. The effect is most likely to be induced by correlation between the non-linear function and the argument or arguments of that function. However, it is entirely possible for the non-linear component to be associated with variables *other* than its argument, in which case there would be aliasing that remains unsuspected without eigenvector decomposition.

A measure of the degree to which the non-linear variable is associated with the eigen-vectors of which it is a component can be had by examining the correlation between the non-linear term and each of the eigenvectors. This measure should help identify the eigenvector or eigenvectors by which the non-linear component shows its effect. That, combined with physical reasoning, should provide insight into the mechanism(s) giving rise to the non-linear effect.

Interpretation of eigenvectors containing one or more non-linear elements also presents a challenge to the data analyst. On the other hand, it forces him or her to recognize that the effect of the non-linear element, either by chance or by design, may be associated with

other variables in a way not anticipated and not evident without eigenvector resolution. Further, it may encourage the analyst to perform a thoughtful examination of variables during the model simplification process.

The considerations presented above apply equally to interaction effects, expressed as a weighted *product* of two variables. Moreover, quadratic and product terms do not exhaust the variety of non-linear elements that might be included in a regression model. Extremely complex terms are known to have been included in a model (Hewlett, 1963), though their *raison d'etre* may not be apparent. Our view is that a regression model should be as simple as possible and that non-linear effects may be better modeled by more conventional approaches such as transformation of variables.

### 3.3.4 Transformations as an Approach to Non-Linearity

One of the assumptions of least squares regression is that errors are normally distributed and homoscedastic. Since regression deals with estimating the average response to the predictor variables, the errors are likely to be approximately normally distributed as a consequence of the very powerful Central Limit Theorem. Homoscedasticity, though, is another matter. If errors get larger or smaller as we go from lower to higher emissions, we are likely to get biased estimations if we assume that the error variance is constant over the whole range of emissions.

Usually, we make transformations for one or more of several reasons:

1. To linearize or otherwise simplify the form of the regression equation
2. To "stabilize" the error variance – that is, to change a heteroscedastic error distribution to one that is at least approximately homoscedastic
3. To produce a result that is more consistent with real-world experience.

Linearization is a convenient technique for converting a non-linear equation to one of linear form, so much so that one is tempted to make that conversion without regard for its consequences to points 2 and 3. Similarly, error stabilization may be indulged in quite legitimately but could distort the regression curve, surface, or hyper-surface to such an extent that point 3 is violated.

Finally, whether a result is consistent with real-world experience is often a judgement call. However, there may be good reason to believe that the effect of an incremental change in a predictor on the response is not constant, but is proportional to the value of the response at which the increment is applied. Again, a transformation made to implement that aspect of experience may not necessarily be consistent with points 1 and 2.

In short, in making a transformation, we often deal with conflicting requirements and are lucky if we do not get an effect that we do not want, along with the one that we do. For example, if our error distribution really is homoscedastic and we make a log transformation for purposes of linearization, we may produce a heteroscedastic distribution that works to our disadvantage, so far as goodness of fit is concerned, to yield biased predictions.

Appendix I explores in more detail the issues surrounding transformations as an approach to non-linear regression. The linear transformations of mean removal and rescaling are considered, as are the common non-linear log and log-log transformations. In each case the differential effect and error weighting is identified. The appendix also presents examples of the consequences for goodness of fit that violation of the distributional assumptions underlying regression analysis can entail.

The emission models developed in this report employ a rescaling transformation of the independent variables – in which the mean value is removed from each fuel property variable and the range of variation is equalized across variables – that produces so-called "standardized scores." A log transformation of the dependent (emissions) variable is employed, consistent with widespread practice among researchers, based on the tendency for emission test variation to be proportional to mean emission level. In such instances, the log transformation has the effect of stabilizing the variance, but it also implies that a given change in a fuel property is expected to exert a larger effect on higher emitting engines, compared to lower emitting engines. This, also, is consistent with the expectations of some researchers (Lee *et al.*, 1998) that the effect of fuels is larger in older engines certified to higher emission standards.

The advantages and limitations of transformations are as relevant to vector-based regression as they are to conventional regression analysis. Every effort should be made to assure that the assumptions regarding error distribution and model correctness are sound before proceeding with the modified PCR approach. Any errors of judgement made at this stage will be propagated to E-Space and may vitiate advantages that might be gained by the eigenvector approach.

# 4. TOWARD AN IMPROVED UNDERSTANDING OF FUELS AND EMISSIONS

The research conducted in this study has identified at least three areas in which additional work is needed to move government and industry toward an improved understanding of diesel fuels and the interface with engine emissions. This section identifies these key needs.

## 4.1 IMPROVED THEORY

The state of uncertainty attending the existing test results makes clear that additional HDD emissions data is needed if analysts are to reach a consensus on the emissions benefit of fuel reformulation. Paralleling this prerequisite is the need for an improved theoretical understanding of *how* fuels affect emissions. Gains in this area can inform the measurement of relevant fuel properties and guide the development and interpretation of appropriate regression models.

As an example, the empirical results of this study strongly suggest that reduction of the amount of high-aromatic cracked blend stocks used in diesel fuels is the primary means of reducing $NO_x$ emissions. This effect also appears to be the reason for the correlations found between emissions and the variables natural cetane and density, since the high-aromatic cracked blend stocks are characteristically low in cetane and high in density. Recognizing cetane and density as surrogate measures for high-aromatic blend stocks is a step forward in our understanding – in particular, it becomes clear that other methods of increasing cetane or reducing density may not necessarily reduce emissions in the same way.

That reducing high-aromatic blend stocks reduces $NO_x$ emissions allows us to infer that the presence of the blend stocks must modify the combustion process, since $NO_x$ is formed as a direct byproduct of combustion. It is also likely that these blend stocks are just one example of a larger set of compounds whose presence in diesel fuels has similar combustion effects. Connecting these empirical points by elements of an applicable theory could lead to a clear-cut road map for reducing $NO_x$.

A recent paper (Matheaus *et al.*, 2000) by members of the EPA Heavy-Duty Engine Working Group provides tantalizing indications of the outline of such a theory, at least in regard to $NO_x$ emissions. This reports on test results from an experimental program to investigate the effects of cetane, density, and mono- and poly-aromatics content on emissions from a prototype HDD engine developed to meet the 2004 $NO_x$ emission standard. The paper shows that reductions in density and aromatics content are

associated with reduced $NO_x$ emissions, while changes in density appear to have no $NO_x$ effect. The authors note that "... the combustion processes were dominated by diffusion burning, especially at the higher load conditions, which contribute very heavily to the weighted $NO_x$ emissions." The implication of this finding is that "... the flame temperature history during combustion is basically the stoichiometric adiabatic flame temperature. The flame temperature is controlled by the local thermodynamic conditions in the combustion chamber and the fuel properties of heat of combustion and the hydrogen and carbon contents."

The importance of this finding was illustrated by a graph of $NO_x$ emissions versus the single variable of fuel hydrogen content in percent. There is very little scatter of the emissions data around the hydrogen-content regression line, compared to substantial scatter around every other individual predictor variable. In the discussion period following this paper at a recent conference (CEC/SAE, 2000), the authors elaborated that the calculated adiabatic flame temperature of a fuel could be a nearly-ideal predictor for $NO_x$ emissions. However, the publicly available procedures for measuring the fuel carbon content (one of the variables in the computation) are insufficiently accurate to permit its practical use for this purpose.

Nevertheless, the theoretical insight and empirical evidence offered in the cited work suggest that future testing should examine hydrogen content and/or other surrogate measures for adiabatic flame temperature as a predictor for $NO_x$. Similar advances in understanding could benefit the analysis of PM emissions.

## 4.2 NEW INSIGHT ON FUELS FORMULATION

The discussion of the previous section painting a perspective on sums of squares showed that an experimental data set presents a distribution of sums of squares that is, to a large extent, induced by the sampling scheme underlying the experimental design. If the experimental design were to differ, we would partition the variation among fuels in a different manner and, even if the unit relationship to emissions is unchanged, would also partition the model (emissions) sums of squares differently. Thus, while we must work with the data at hand, we must also be cautious of the potential that our data excludes, to a greater or lesser degree, the influences that will be encountered in real-world applications of the model.

The vector regression models developed in Section 2.3 indicated that the smaller eigen-vectors (numbers 9 through 12) may be strongly related to emissions even though the variance associated with these eigenvectors is so small that they make only a negligible contribution to the emissions variance *in this data set*. These findings could result from correlations to factors that have been inadequately controlled in these regressions, but it is also possible they point toward unexploited modes of reducing $NO_x$ and PM emissions. Therefore, an effort was made to identify the fuels that express these eigenvector characteristics most strongly.

The key results of the examination are summarized in Figure 4.1. A total of only 14 different fuels (but accounting for more engine tests) were found to express one or more of these characteristics strongly. These 14 fuels, out of the 85 distinct fuels in the data set, are found in three test sources. Eight of the fuels were used in tests conducted at Southwest Research Institute (SWRI) as part of the Coordinating Research Council (CRC) VE test series and one unrelated program. Five fuels were used in testing by Sienicki (Sienicki *et al.*, 1990) and one fuel by Gonzalez (Gonzalez *et al.*, 1993). Factors found to be common to fuels 1 and 8-11 may aid the understanding of eigenfuels 9 and 10, while factors shared by fuels 2, 5-6, and 12-14 may assist with eigenvectors 11 and 12.

**Figure 4.1.** Fuels Expressing Eigenvectors 9 through 12

| | | | | Eigenfuel Number | | | |
|---|---|---|---|---|---|---|---|
| | | | | 9 | 10 | 11 | 12 |
| **Contribution to Variance Among Fuels** | | | | 1.17% | 0.68% | 0.29% | 0.21% |
| **Regression Coefficients** | | | | | | | |
| **NOx** | | | | ns | -0.0156 | 0.0294 | 0.0325 |
| **PM** | | | | -0.0430 | -0.0575 | ns | ns |
| **Fuels Loading Highly on the Eigenvectors** | | | | | | | |
| **Percent of Eigenfuel Variance** | | | | 56% | 72% | 54% | 67% |
| **Fuel Description** | | | | | | | |
| | Source | Program | Fuel Identification | Coefficient in Eigenfuel Representation | | | |
| 1 | 941020 | SWRI/CRC VE-10 | Fuel E | -4.00 | | | |
| 2 | 902171 | SWRI/CRC VE-1 | Fuel X1 | | | -2.74 | -3.29 |
| 3 | 922267 | Sienicki | Fuels G, H | -1.88 | -2.87 | -1.81 | -1.84 |
| 4 | 892072 | SWRI/CRC VE-1 | Fuel 1 #685 | -1.74 | -2.48 | -1.86 | |
| 5 | 972898 | SWRI/SSPD | Fuel C: CARB Diesel | | | -1.67 | |
| 6 | 892072 | SWRI/CRC VE-1 | Fuels 0, 2 | | | -1.52 | |
| 7 | 922267 | Sienicki | Fuel D | | | | 1.80 |
| 8 | 932731 | Gonzalez | Fuel H | | 1.52 | | |
| 9 | 892072 | CRC VE-1 | Fuel 4 #688 | 1.68 | 1.85 | | |
| 10 | 941020 | CRC VE-10 | Fuel H | 1.83 | | | |
| 11 | 902171 | CRC VE-1 | Fuel X2 #952 | 2.09 | | | |
| 12 | 902172 | Sienicki | Fuel 5 | | | | 1.91 |
| 13 | 902172 | Sienicki | Fuels 2, 2A, 2B, 2S | | | 1.54 | 2.18 |
| 14 | 902172 | Sienicki | Fuels 4, 4B | | | | 2.18 |
| | | | | Increase to reduce NOx and PM | | Decrease to reduce NOx | |

There is not enough information on these fuels to identify the constituents that make them unique, nor is that task within the scope of this work. Additional work is needed to understand more clearly the following basic trends observed in these fuels:

- Fuels low in eigenfuel 9 tend to have very low viscosity and low temperatures at the beginning of the distillation curve (IBP and T10), while those high in eigenfuel 9 have correspondingly high viscosities and high initial temperatures. Increasing this feature appears to reduce PM emissions in this instance, although it is more generally true that "heavier" fuels tend to have higher PM emissions.
- Fuels low in eigenfuel 10 tend to have very low viscosity and low temperatures across the distillation curve, while fuels high in the eigenfuel display the opposite trends. Increasing the feature also appears to reduce $NO_x$ and PM emissions.
- Fuels low in eigenfuel 11 show a tendency to have low viscosity and low distillation temperatures, although SWRI/CRC's Fuel X1 has average viscosity and distillation characteristics. Increasing this feature appears to increase $NO_x$ emissions.
- Fuels low in eigenfuel 12 show, again, a tendency toward lower viscosity and distillation temperatures, but SWRI/CRC's Fuel X1 counters the trend. Increasing this feature appears, again, to increase $NO_x$ emissions.

The evidence outlined above offers tantalizing hints that as-yet unrecognized properties or components of fuels may have a large impact on emissions. However, we can not rule out the possibility that the emission results are artifacts caused by correlations to factors that have been inadequately controlled in the regressions or by the wide error bounds associated with small eigenvalues. Future work should make an effort to understand these results.

More importantly, perhaps, these eigenfuels again emphasize that emission changes can not be ascribed uniquely to variations in selected properties alone. Emissions may be increased when varying properties (here, viscosity and distillation temperatures) in one way, while similar variations occurring in another manner may have the reverse effect. Eigenfuels 9 and 10 may represent blending components that should be emphasized in fuels reformulation in order to reduce $NO_x$ and PM emissions, while eigenfuels 11 and 12 may represent ones to be avoided.

The identification of test fuels associated with small eigenvalues, but apparently substantial emissions effects, poses something of a dilemma to the data analyst. In particular, it again brings into focus the question of the relative emphasis that should be placed on the indicated size of an effect relative to its indicated statistical significance.

Small eigenvalues are the result of limited variation among the coefficients of the associated eigenvectors when the test fuels are resolved into their equivalent eigenfuel representation. The standard error of a regression coefficient, computed for the purpose of testing the null hypothesis that the estimated effect arises solely by chance, is inversely proportional to the variance of the independent variable.

It is clear, therefore, that under these conditions the chance of finding an effect to be statistically significant is considerably reduced. Therefore, one should not be too quick to exclude summarily what appears to be a substantial effect simply because it fails to pass a test of significance based on an arbitrarily selected significance level. Rather than to ignore such effects and risk committing an error of omission, it would be prudent to re-sample fuels for supplementary testing in such a way as to increase the range of variation of the eigenfuels of interest.

## 4.3 MORE AND BETTER ENGINE TESTING DATA

As has been made clear in the course of this work, an improved data base is a prime requirement for the future development of a reliable diesel emission model. The following are our recommendations for future testing to correct the most important limitations of the existing data base:

- The data base omits at least one fuel property – oxygen content – that is likely to affect emissions. Few test programs to date have evaluated oxygenated fuels, and more testing of such fuels will be required before a complete diesel emissions model can be developed.
- The data base lacks information on hydrocarbon composition beyond mono- and poly-aromatic content, as do most existing testing programs. It may be important for new testing to report a more detailed hydrocarbon speciation because, when a fuel is changed by the substitution of one constituent for another, it is not possible to attribute an emissions change uniquely to the one constituent (or the other). While such an effort could open a Pandora's box if carried too far, it could very well be important to know whether it was hydrocarbon species 1 or 2 that was substituted when, for example, aromatics content was reduced.
- The data base represents too few engine types and individual engines to be taken as representative of the on-road HDD fleet or to permit the assessment of engine-related effects. Although the total of 280 emissions tests is relatively large, the data are based on only 11 individual engines. An improved data base should represent a substantially larger number of engines sampled in a representative manner from the cells created by the intersections of model year, emission certification standard, manufacturer, and engine design. It would be desirable that two or more specimens be included for each major engine design to permit the estimation of design-specific effects, and that all testing use the one test cycle on which certification decisions for fuels and engines will be based.

The properties of test fuels used in future testing should be varied over the widest practical range and should include any fuel property indicated to affect emissions in a

substantive way. Also, the fuel properties should be varied in accordance with their *cooperative* effects, as indicated in the eigenfuels, rather than as separate variables. Just how the test fuels should be blended is an important topic.

One method to achieve such sampling might be to vary a given fuel property over its desired range by blending commercially feasible blend stocks and relying on the natural association of fuel properties to produce the desired cooperative effects. As an extension of this approach, one might attempt to blend fuels that meet multiple property specifications simultaneously. This latter approach has been the one frequently followed in past testing where, for example, mono- and poly-aromatics content might be varied subject to controlling fuel density and viscosity to predetermined values.

It is here, though, that blending in terms of eigenfuels shows its major advantage by providing an explicit mathematical scheme for producing blends having the desired properties. It is necessary only to resolve a desired eigenfuel into a corresponding set of available blend stocks. To match an eigenfuel exactly would require up to 12 blend stocks. Note, however, that in eigenfuel 2, which accounts for 82 percent of the $NO_x$ regression SS, only 4 of its components – natural cetane, density, mono-aromatic and poly-aromatic content – make major contributions. Therefore, no more than four blend stocks should be sufficient to approximate eigenfuel 2. Though the mechanics of the solution mimic the conventional, it employs a vastly more effective strategy for formulating optimum test fuels.

In varying engine characteristics, the primary requirement is to have a sampling of vehicles that covers all of the design factors that might influence emissions. It may turn out that some of these design factors have relatively little effect on emissions. Here, again, a fixed magnitude of the engine effect should be used as the criterion for including or excluding that factor, rather than an arbitrary test of significance. Since the engine effects are most likely not orthogonal, it might be appropriate to induce orthogonality by a method such as random balance assessment (McAdams, 1995).

Engine effects and fuel effects may interact. If so, the computation of "fuel effects adjusted for engine effects" is inadequate, because the adjustment corrects only for the difference in the mean level of emissions among engines, it being assumed that the incremental effect of a fuel change is constant for all engine classes. It is quite possible, however, that the effect of an incremental change in a fuel property is greater for one class of vehicle than another, especially if the fuel change is designed to play an "enabling" role for a vehicle design change. Resolving such issues would require more extensive testing and a model of greater complexity.

# 5. CONCLUSIONS

Though demonstrated in the context of emissions from HDD engines, the vector approach to regression analysis is believed to be applicable to a wide range of problems. Indeed, it should be considered in any circumstance where variables are inextricably covariant. It may be the favored approach whenever the covariance of predictor variables can not reasonably be "broken," but it does not preclude the use of other multivariate methodologies or of scalar predictor variables when the predictor variables can be varied independently of each other in circumstances such as balanced experimental designs.

When applied to the fuel/emissions relationship, the findings of this report illustrate the range of benefits that the vector approach offers:

- Simplification of the regression analysis as a result of the desirable mathematical properties of vector variables – their independence and absence of correlations, and their economy of representation.
- Greater understanding of the patterns of variation that are important to emissions reduction, in this instance, and how these patterns relate to fuel blending and refinery processes.
- Potentially new insight into the optimal formulation of fuels to reduce emissions.
- Improved experiment design for more efficient estimation of fuel effects.

Disadvantages of the methodology that may be perceived by some are:

- The ineffectiveness of selecting predictor variables by means of PCA as noted in the cautionary notes by Hadi and Ling (1994).
- Potential difficulties in interpreting the effects of the predictor vectors on the dependent variable.

We have shown that variable selection can not, and should not, be attempted without regard to the response variable. This fact is made clear by the difference between $NO_x$ and PM response, even though the PCA analysis of the design space is the same for both.

With regard to interpretation, it may appear that the multi collinearity "fog" disposed of by orthogonalization is merely replaced by a different kind of fog arising from the difficulties of interpreting eigenvectors. We believe that viewing response in terms of eigenvectors is not so much *difficult* as it is *unconventional* and that it affords an improved basis for understanding the factors that actually drive the response. Finally, we

offer a means for alternatively viewing response in terms of the original variables, but within the context of an orthogonal, eigenvector solution.

It is, perhaps, most important that the vector approach to analysis does not require or benefit from the attempt to "break" naturally-occurring associations among fuel properties in the blending of test fuels. Without the need to artificially separate these associations, a wider range of real-world diesel fuels, representing current and future refinery configurations and processes, could be used in engine testing, thereby avoiding possibly unrealistic or unrepresentative emission results. These benefits imply an increased accuracy in assessing emissions benefits and an improved basis for measuring cost effectiveness.

# 6. REFERENCES


1. Akasaka, Y., T. Suzuki and Y. Sakurai. (1997). *Exhaust Emissions of a DI Diesel Engine Fueled with Blends of Biodiesel and Low Sulfur Fuel*, SAE 972998.

2. Boneh, S. and G.R. Mendieta. (1994). *Variable Selection in Regression Models Using Principal Components*. Communications in Statistics – Theory and Methods 23, pp. 197-213.

3. CEC/SAE International Spring Fuels and Lubricants Meeting. (2000). Paris, France. Sessions on Diesel Performance and Additives, June 19-22, 2000.

4. Cunningham, L.J., T.J. Henly and A.M. Kulinowski. (1990). *The Effects of Diesel Ignition Improvers in Low-Sulfur Fuels on Heavy-Duty Diesel Emissions*," SAE 902173.

5. Daniels, T.L., R.L. McCormick, M.S. Graboski, P.N. Carlson, V. Rao and G.W. Rice. (1996). *The Effect of Diesel Sulfur Content and Oxidation Catalysts on Transient Emissions at High Altitude from a 1995 Detroit Diesel Series 50 Urban Bus Engine*, SAE 961974.

6. Efron, B. and R.J. Yates. (1998). *An Introduction to the Bootstrap*, Chapman & Hall/CRC

7. EPA HDEWG Program: Phase II, Briefing for Meeting of the Mobile Sources Technical Review Subcommittee. (1999). Clean Air Act Advisory Committee, Washington, DC, January 13.

8. Ethyl Corporation. (1995). HITEC Performance Chemicals. *A Distillate Fuel Response to "Diesel Ignition Improver."*

9. Fisher, R. A. (1935). *The Design of Experiments*, Edinburgh: Oliver and Boyd.

10. Fisher, R. A. and F. Yates. (1948). *Statistical Tables for Biological, Agricultural, and Medical Research*, 3rd Edition, Edinburgh, Oliver and Boyd, Ltd.

11. Geiman, R.A., P.B. Cullen, P.R. Chant, P.N. Carlson and V. Rao. (1996). *Emission Effects of Shell Low $NO_x$ Fuel on a 1990 Model Year Heavy Duty Diesel Engine*, SAE 961973.

12. Gonzalez, D., G.R. Rodriguez, R. Galiasso and E. Rodriguez. (1993). *A Low Emission Diesel Fuel: Hydrocracking Production, Characterization, and Engine Evaluations*, SAE 932731.

13. Hadi, A.S. and R.F. Ling. (1998). *Some Cautionary Notes on the Use of Principal Components Regression.* The American Statistician, Vol. 52, No. 1. February.

14. Hawkins, D. (1973). *On the Investigation of Alternative Regressions by Principal Component Analysis.* Applied Statistics 22, pp. 275-286.

15. Hewlett, R.F. (1963). *Computer Methods in Evaluation, Development, and Operations in an Ore Deposit.* American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc., Dallas, TX, February. Preprint No. 63151.

16. Jackson, J.E. (1991). *A User's Guide to Principal Components.* John Wiley & Sons.

17. Jeffers, J.N. (1967). *Two Case Studies in the Application of Principal Component Analysis.* Applied Statistics, Vol 16, pp. 225-236.

18. Jolliffe, I.T. (1973) *Discarding Variables in a Principal Component Analysis. II: Real Data.* Applied Statistics 22, pp. 21-31.

19. Jolliffe, I.T. (1972). *Discarding Variables in a Principal Component Analysis. I: Artificial Data.* Applied Statistics 21, pp. 160-173.

20. Krzanowski, W.J., and F.H.C. Marriott. (1994). *Multivariate Analysis.* Kendall's Library of Statistics. Halstead Press.

21. Lange, W.W. (1991). *The Effect of Fuel Properties on Particulates Emissions in Heavy-Duty Truck Engines Under Transient Operating Conditions*, SAE 912425.

22. Lange, W.W., J.A. Cooke, P. Gadd, H.J. Zurner, H. Schogel and K. Richter. (1997). *Influence of Fuel Properties on Exhaust Emissions from Advanced Heavy-duty Engines Considering the Effect of Natural and Additive Enhanced Cetane Number*, SAE 972894.

23. Lee, R., J. Pedley and C. Hobbs. (1998). *Fuel Quality Impact on Heavy Duty Diesel Emissions - A Literature Review*, SAE 982649.

24. Liotta Jr., F.J. (1993). *A Peroxide Based Cetane Improvement Additive with Favorable Fuel Blending Properties*, SAE 932767.

25. Liotta Jr., F.J. and D.M. Montalvo. (1993). *The Effect of Oxygenated Fuels on Emissions from a Modern Heavy Duty Diesel Engine*, SAE 932734.

26. Mann, N., F. Kvinge and G. Wilson. (1998). *Diesel Fuel Effects on Emissions: Towards a Better Understanding*, SAE 982486.

27. Mansfield, E.R., J.T. Webster and R.F. Gunst. (1977). *An Analytical Variable Selection Technique for Principal Component Regression.* Applied Statistics 26, pp. 34-40.

28. Martens, H. and T. Naes. (1989). *Multivariate Calibration.* John Wiley & Sons.

29. Matheaus, A. C., G. Neeley, T. Ryan, R. Sobotowski, J. Wall, C. Hobbs, G. Passavant, T. Bond. (2000). *EPA HDEWG Program-Engine Test Results*, SAE 2000-01-1858.

30. *MatLab*, The MathWorks, Inc. (2000). 24 Prime Park Way, Natick, MA 01760-1500. Website http://www.mathworks.com.

31. McAdams, H.T. (1995). *A Random Balance Procedure for Simplifying a Complex Model.* American Statistical Association, 1995 Proceedings of the Section on Statistics and the Environment, Alexandria, VA.

32. McAdams, H.T. (1972). *A Factor Analytic Approach to the Identification of Manufacturing Systems.* Proceedings of CIRP Seminars on Manufacturing Systems, Vol I, No. 2. pp. 79-97.

33. McAdams, H.T., R.W. Crawford and G.R. Hadder. (2000). *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*, SAE 2000-01-1961.

34. McCarthy, C.I., W.J. Slodowske, E.J. Sienicki and R.E. Jass. (1992). *Diesel Fuel Property Effects on Exhaust Emissions from a Heavy Duty Diesel Engine that Meets 1994 Emissions Requirements*, SAE 922267.

35. Nakakita, K., S. Takasu, H. Ban, T. Ogawa, H. Naruse, Y. Tsukasaki and L. Yeh. (1998). *Effect of Hydrocarbon Molecular Structure on Diesel Exhaust Emissions Part 1: Comparison of Combustion and Exhaust Emission Characteristics Among Representative Diesel Fuels*, SAE 982494.

36. Nylund, O., P. Aakko, S. Mikkonen and A. Niemi. (1997). *Effects of Physical and Chemical Properties of Diesel Fuel on $NO_x$ Emissions of Heavy Duty Diesel Engines*, SAE 972997.

37. Reynolds, E.G. (1993). *The Effect of Fuel Processes on Heavy Duty Automotive Diesel Engine Emissions*, SAE 932350.

38. Rosenthal, M.L. and T. Bendinsky. (1993). *The Effects of Fuel Properties and Chemistry on Emissions and Heat Release of Low Emission Heavy Duty Diesel Engines*, SAE 932800.

39. Schaberg, P.W., I.S. Myburgh, J.J. Botha, P.N. Roets, C.L. Viljoen, L.P. Dancuart and M.E. Starr. (1997). *Diesel Exhaust Emissions Using Sasol Slurry Phase Distillate Process Fuel*, SAE 972898.

40. Schmidt, K. and J. Van Gerpen. (1996). *The Effect of Biodiesel Fuel Composition on Diesel Combustion and Emissions*, SAE 961086.

41. Sienicki, E.J., R.E. Jass and W.J. Slodowske. (1990). *Diesel Fuel Aromatic and Cetane Number Effects on Combustion and Emissions from a Prototype 1991 Diesel Engine*, SAE 902172.

42. Spreen, K.B., T.L. Ullman and R.L. Mason. (1995). *Effects of Cetane Number, Aromatics and Oxygenates on Emissions from a 1994 Heavy Duty Diesel Engine with Exhaust Catalyst*, SAE 950250.

43. Starr, M.E. (1997). *Influence on Transient Emissions at Various Injection Timings, using Cetane Improvers, Bio-diesel and Low Aromatics Fuels*, SAE 972904.

44. Tamanouchi, M., H. Morihisa, S. Yamada, J. Iida, T. Sasaki and H. Sue. (1997). *Effects of Fuel Properties on Exhaust Emissions for Diesel Engines with and without Oxidation Catalyst and High Pressure Injection*, SAE 970758.

45. Tanaka, S., M. Morinaga, H. Yoshida, H. Takizawa, K. Sanse and H. Ikebe. (1996). *Effects of Fuel Properties on Exhaust Emissions from DI Diesel Engines*, SAE 962114.

46. Thomas, J. (1946). *J. Chem. Soc.*, Part II, pp. 573-579.

47. Thurstone, L. L. (1935). *The Vectors of Mind*, Chicago: U. of Chicago Press.

48. Ullman, T.L. (1989). *Investigation of the Effects of Composition and Injection and Combustion System Type on Heavy-Duty Diesel Exhaust Emissions*, SAE 892072.

49. Ullman, T.L., K.B. Spreen and R.L. Mason. (1995). *Effects of Cetane Number on Emissions from a Prototype 1998 Heavy Duty Diesel Engine*, SAE 950251.

50. Ullman, T.L., K.B. Spreen and R.L. Mason. (1994). *Effects of Cetane Number, Cetane Improver, Aromatics, and Oxygenates on 1994 Heavy Duty Engine Emissions*, SAE 941020.

51. Ullman, T.L., R.L. Mason and D.A. Montalvo. (1990). *Effects of Aromatics, Cetane Number, and Cetane Improver on Emissions from a 1991 Prototype Heavy Duty Diesel Engine*, SAE 902171.

52. U.S. Department of Energy. (1994). *Estimating the Costs and Effects of Reformulated Gasoline*. DOE/PO-0030. December.

53. Westerholm, R. and H. Li. (1994). *A Multivariate Statistical Analysis of Fuel-Related Polycyclic Aromatic Hydrocarbon Emissions from Heavy-Duty Diesel Vehicles*. Environ. Sci. Technol. 28, pp. 965-972.

54. Wold, S. (1978). *Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Models*. Technometrics 20, No. 4. November.

# APPENDIX A: DATABASE DEVELOPMENT

This appendix describes the development of a database on Heavy Duty Diesel (HDD) engine emissions and diesel fuels to support this work. Data from 27 technical publications of the Society of Automotive Engineers (SAE) were reviewed in preparing the database, in addition to information available through the U.S. EPA. This effort included all sources known and available at the time the work was performed, although it was not necessarily a comprehensive literature search. It does not include additional data released since that time.

The first section summarizes the final database developed as a result of this effort. The second section describes briefly the engines covered in this database. The third section discusses additional Data sets that may be used in portions of the analysis related exclusively to fuel properties. Brief comments on data preparation are given in the concluding section.

## A.1 SUMMARY OF THE DATABASE

The final database consists of 198 entries representing 280 individual HDD emissions tests and a matching slate of 12 properties for the 85 different diesel fuels on which the engines were operated. Test results were selected from nine SAE publications where the following criteria were met:

- Uses the EPA transient test cycle and reports either the composite result or the hot start portion. The hot start portion has a $6/7^{ths}$ weight in the composite test.
- Measures at least NOx and PM. Eight of the sources measured all four pollutants (HC, CO, NOx, and PM), and 1 source omitted only CO.
- Emissions testing could be matched to fuels for which the following 13 properties were known: natural cetane, cetane number improvement (resulting from additives), API gravity, density, viscosity, sulfur content, mono aromatic content, poly aromatic content, and five points on the distillation curve (IBP, T10, T50, T90, FBP).

The 13 fuel properties were selected to be a super-set of properties that researchers in the field have examined for their effect on HDD emissions. As such, they include the properties which a consensus of researchers believes to be relevant to emissions performance with additional properties that may be highly correlated with other properties

**Table A.1: Comparison of Fuel Properties to the Revelant Ranges**

| Fuel Parameter | Relevant Range (Min - Max) | Range in Database (Min - Median - Max) |
|---|---|---|
| **Aromatics (percent)** | 10 - 40 | 2.7 - 26.5 - 47.5 |
| **Mono Aromatics** | – | 0.5 - 18.0 - 31.8 |
| **Poly Aromatics** | – | 0 - 6.4 - 26.7 |
| **Cetane Number (total)** | 35 - 60 | 35 - 51 - 74 |
| **Natural Cetane** | – | 35 - 44 - 74 |
| **Sulfur Content (ppm)** | 5 - 500 | 0 - 384 - 3360 |
| **T50 (Celsius)** | 230 - 295 | 211 - 264 - 292 |
| **T90 (Celsius)** | 275 - 330 | 274 - 318 - 337 |
| **Viscosity ($mm^2/sec$)** | 1.5 - 3.5 | 1.58 - 2.68 - 3.87 |
| **Density ($gm/cm^3$)** | – | .78 - .84 - .88 |

or prove unrelated to emissions. This purposefully casts the net as wide as possible for variable selection, leaving to the analysis the identification of the proper set of explanatory variables. Table A.1 demonstrates that these 85 fuels cover the range of fuel properties thought[1] to be of interest to the analysis of diesel emissions.

This selection should not be interpreted to mean that only these properties have an effect on engine emissions. For example, fuel oxygen content is likely to have an impact on CO emissions and perhaps other pollutants, but is not covered in the list of fuel properties. While some papers reviewed in this work tested diesel fuels containing oxygen, these were judged to be too few in number to permit including oxygen content in our list.

Eight publications that used the EPA transient test cycle were excluded because one or more of the fuel properties were not available. The most common reason was that the poly-aromatic content was not reported. Ten SAE publications were excluded for reasons related to the emissions data. In one instance, only PM was measured. Non-EPA test cycles were used in the nine other instances; most commonly these were European or Japanese test cycles.

Unfortunately, the entire body of data collected by the EPA Heavy Duty Engine Work Group (HDEWG) in Phase II testing was excluded from the database because of its test

[1] Personal fax communication from Mr. Barry McNutt dated 8-30-95.

cycle. This is an on-going cooperative program between government and industry which is creating a substantial database on the HDD emission response to fuel properties. However, the engine used in this testing has a prototype EGR system which is not compatible with the EPA transient test cycle. The program therefore uses a steady state procedure, exclusively covering eight modes based on engine load and speed. NOx emissions are not measured because the steady-state values correlate poorly with NOx on the EPA cycle.

Overall, the database represents 11 different engines tested a total of 280 times on 85 different diesel fuels. The large majority of database entries represent individual engine tests, but 41 entries from three sources record the mean values of replicated tests. The number of test replications is coded as a field in the database and should be used as a weighting factor in statistical analysis. Three papers clearly contributed mean values but did not clearly indicate the number of test replications. We have assumed a replication factor of three based on the replications recorded in other sources that were ultimately excluded from the database.

## A.2 ENGINE COVERAGE

The eleven HDD engines included in the final database represent a very small sample of the engine types present in the on-road vehicle fleet. Nevertheless, they include engines made by three major manufacturers (Cummins, Detroit Diesel Corporation, and Navistar) and cover a range in model years (1987 through 1998) and horsepower ratings. The engines are thought to be generally similar in design, although they are built to varying emissions standards. None are equipped with EGR systems or with catalysts. They can be taken as reflective of the HDD engines currently on the road, even if the sample is too limited to be considered truly representative. Basic engine characteristics are summarized in Table 2.2.

This database will be used to develop statistical parameters that describe for the existing vehicle fleet the relationship between engine emissions and fuel properties and the associated variance. Some studies have concluded that this relationship varies with the design emissions level, such that lower-emitting engines demonstrate a smaller sensitivity to changes in fuel properties. Whether and how to segregate the engines into groups or strata may be a question faced in future analyses.

Energy and Environmental Analysis, Inc. (EEA) was asked to propose an *a priori* method of stratification based on engineering judgement. Its recommendation was to group engines based on the value of the expression NOx + 10*PM, which gives the empirical tradeoff between NOx and PM that is faced in engine design. Once individual calibrations are accounted using the expression, the resulting value reflects an approximate emissions control stringency level. EEA recommended three stratifications:

the one model year 1987 engine with a NOx + 10*PM value of 9.7; a second stratum ranging from 6.6 to 7.5; and a third stratum ranging from 4.9 to 6.3.

**Table A.2. Summary of Engine Characteristics**

| SAE Source | Model Year | Engine Number | Engine ID | HP | NOx + 10PM Value | Stratum |
|---|---|---|---|---|---|---|
| 892072 | 87 | 1 | Cummins NTCC 4000 | 400 | 9.7 | 1 |
| 892072 | 88 | 2 | DDC Series 60 | 315 | 7.5 | 2 |
| 892072 | 89 | 3 | Navistar 7.3L | 185 | 7.5 | 2 |
| 902171 | 91 | 4 | DDC Series 60 (prototype) | 330 | 6.6 | 2 |
| 902172 | 91 | 5 | Navistar DTA-466 (prototype) | 230 | 7.1 | 2 |
| 972898 | 91 | 21 | DDC Series 60 | 350 | 6.3 | 3 |
| 922267 | 93 | 8 | Navistar DTA-466 | 210 | 5.6 | 3 |
| 932731 | ?? | 9 | unknown | 153 | 5.4 | 3 |
| 941020 | 94 | 13 | DDC Series 60 (prototype) | 320 | 5.9 | 3 |
| 950250 | 94 | 14 | Navistar DTA-466 (prototype) | 199 | 5.5 | 3 |
| 950251 | 98 | 15 | DDC Series 60 (prototype) | 325 | 4.9 | 3 |

## A.3 ADDITIONAL DATA SETS CHARACTERIZING DIESEL FUELS

Additional information exists on diesel fuel properties that is available for use in the analysis. First, a data set of 110 diesel fuels exists for which the 12 selected fuel properties are known. The 85 fuels found in the final database are simply those for which qualifying emissions test results were available.

Two additional and smaller samples of diesel fuels may be used in the study for purposes of demonstrating the basic properties of eigenvectors applied to fuel characteristics. One data set consists of 30 fuels created for SAE paper 962114 *Effects of Fuel Properties on Exhaust Emissions from DI Diesel Engines*. This paper reports on a test program at the Japanese COSMO Research Institute for which a significant effort was made to create test fuels whose properties could be varied independently. The fuels can be viewed as one extreme of fuel blending that is achievable in a laboratory setting, but is not likely to be representative or achievable in the refinery. This source was excluded from the database

and from the 110-fuel data set because it uses a Japanese test cycle and only 11 of the 12 selected properties were known.

A second sample of diesel fuels is the group of 18 fuels blended for the test program conducted for the Heavy Duty Engine Work Group. These fuels were created for an experimental design in which density, cetane number, mono-aromatics content, and poly-aromatics content were varied independently. These fuels were blended solely from blend stocks found in refinery streams, so that they are at least technically realizable in a refinery. They provide a contrast to the laboratory fuels fabricated in SAE 962114. These fuels are included in the 110-fuels data set, although the emissions results were excluded from the final database because the testing did not use the EPA transient cycle.

## A.4 DATA PREPARATION ISSUES

A range of problems was encountered in compiling the database, particularly in regard to information on fuel properties. Similar problems will be encountered by any researcher attempting to enlarge this database.

The individual sources on fuel properties vary widely in terms of the test methods employed, properties reported, and the units used to report data. For this study, we attempted to reconcile the sources into a common format, but have made compromises that may not be appropriate for later studies. These compromises were thought to be appropriate because of the exploratory nature of this work and the limited data available to it.

In reconciling the fuels data, we attempted to retain as much information on fuels as was reported in the original sources. The initial fuels data set grew to include more than 30 different characteristics to accommodate the wide range of values reported. The process of reconciliation involved the following tasks:

- Correcting units of measure to a common basis. For example, sulfur content was reported as both parts per million (ppm) and as the weight percent, but were standardized to the ppm basis used in the database. Another such example is conversion of the distillation curve in degrees Fahrenheit to the degrees Celsius used in the database.

- Estimating characteristics not reported in the original source. For example, API gravity and density are so highly correlated that one was estimated from the other when both were not reported, so that we did not lose the data. Only density has been used in the analysis due to the high degree of correlation.

- Combining similar measures – e.g., viscosity values at varying temperatures. Most sources reported viscosity at two temperatures – ambient temperature and

engine operating temperature. Ambient temperature values were compiled in the database, although these may represent 30 degrees Celsius, 40 degrees Celsius, or 100 degrees Fahrenheit. This is one example of the compromises made.

- Selecting among competing measures – e.g., fuel aromatics content stated as either volume percent or weight percent. In this instance, the weight percent values were the preferred measure. However, volume percent values were accepted where no other information on aromatics content was available. This is a second example of compromise.

- We imputed a replication factor of 3 for the three sources that clearly reported mean values but did not state the number of repeated tests. This assumed value was based on the replication factors found in other sources that were ultimately excluded from the database.

Finally, fuel properties were stated in terms that eliminated additive relationships among the final variables. If $Y = X + A$, then X and Y will be correlated even if X and A are independent. Fuel cetane ratings were stated as the natural cetane and the cetane number improvement, which are independent, rather than natural cetane and total cetane, which are correlated. Aromatics content was similarly stated as mono-aromomatic content and poly-aromatic content.

In general, no effort was required to reconcile or standardize the emissions test results. All results were stated in gm/bhp-hr when the EPA transient test cycle was used. It is conceivable that some publications would express EPA test cycle results in metric units and therefore require conversion, but this was not encountered in the sources used.

# APPENDIX B: MATLAB SOFTWARE

The following figures display the primary computer programs used with the MatLab matrix processing package to conduct the analysis shown in this paper. These consist of a main analysis script plus two supporting functions that standardize the data and perform the regression analysis. The main processing script performs the following functions:

- Load the emissions/fuel data base and prepare variables needed for the analysis
- Standardize the fuel variables to mean 0 and variance 1 using the supporting function *standardform*
- Form the covariance matrix of the standardized fuels data and extract the eigenvalues (variable latent) and eigenvectors (variable fuel_pc) using the built-in function *svd*
- Re-express the fuels data in eigenvector terms
- Set up and perform multiple linear regressions using the supporting function *reggaer*.

The standardization function *standardform* converts a matrix *x* into a standardized matrix *y* of mean 0 and variance 1. The means values and standard deviations of the column vectors constituting matrix x are returned as outputs *xmean* and *xstd*. An optional weighting vector *wtvec* may be supplied as an input, but this capability was not used in this study.

The regression routine *reggaer* performs a generalized multivariate regression based on minimizing least squares. Its inputs are: the matrix *x* containing the independent variables; the column vector *y* containing the dependent variable; and the variable *intercept*, which indicates whether an intercept term is to be fitted (value of 1) or suppressed (all other values). Outputs of the regression routine are:

- Matrix *tpart*, an n x 3 matrix containing the n parameter estimates, the standard errors of the estimates, and the t statistic
- Matrix *apart*, a 4 x 3 matrix giving the analysis of variance outputs
- Scalar *f*, giving the F statistic for the overall model
- Scalar *rr*, giving the model $R^2$.

**Figure B.1: Main Analysis Script (Part 1)**

```
% This m-file sets up a workspace for the emissions test database
% and conducts a regression of NOx and PM against engine effect and
% eigen fuel properties.
% Uses the eigenfuels for the 198 fuel dataset as explanatory
variables.
% After replication to account for weighting, data set has 280
entries.
% API Gravity dropped from list of variables.
% load emissions data
[xdata, varnames, casenames] = tblread('HCNPFuel.csv','comma');
% drop API gravity from the variable list
xdata(:,12)=[];
varnames(12,:)=[];
% expand data set to account for replication factor
v=xdata(:,5)==3;
part1=xdata( v,:);
part2=xdata(~v,:);
xdata=[part1;part1;part1;part2];
xdata(:,5)=1;
clear v part1 part2 part3
% set up variables for emissions and fuels
[obs,vars]=size(xdata);
emissions=xdata(:,1:9);
testfuels=xdata(:,10:vars);
fuelnames=varnames(10:vars,:);
[obs,vars]=size(testfuels)
HC = log(emissions(:,6));
CO = log(emissions(:,7));
NOx= log(emissions(:,8));
PM = log(emissions(:,9));
NRepl=ones(obs,1);
weights = diag( sqrt(NRepl) );
% standardize variables to mean 0 and variance 1
xvariables=testfuels;
[xvariables,testfuels_mean,testfuels_std] =
standardform(testfuels,NRepl);
% form covariance matrix and perform svd decomposition
xnormcov=cov(xvariables);
[u,latent,fuel_pc]=svd(xnormcov);
latent=diag(latent);
variance=sum(latent);
explained=100*latent/variance;
fuel_pc
latent
variance
% Re-express x variables as PCA coefficients
PCA_coeffs = xvariables*fuel_pc;
   oldxvar = xvariables;
xvariables = PCA_coeffs;
```

**Figure B.1: Main Analysis Script (continued)**

```
% form & print correlation matrix
corrdata = [ NOx, PM, xvariables ];
x_corr = corrcoef(corrdata)

% Set up dummy variables for individual engines
engines=xdata(:,3);
engvals=unique(engines);
[engs,null]=size(engvals);
enginedummy=zeros(obs,0);
for i=2:engs
   enginedummy=[enginedummy engines==engvals(i)];
end;

% Begin regression section
xmatrix=[enginedummy xvariables];
[t_NOx, a_NOx, f_NOx, r_NOx] = reggaer(xmatrix, NOx);
t_NOx
a_NOx
f_NOx
r_NOx

xmatrix=[enginedummy xvariables];
[t_PM, a_PM, f_PM, r_PM] = reggaer(xmatrix, PM);
t_PM
a_PM
f_PM
r_PM
```

**Figure B.2: Standardization Function *standardform***

```
function [y,xmean,xstd]=standardform(x,wtvec)
% standardform: convert matrix x to standard form of mean 0 and variance 1
% where r is the number of rows, and
%       c is the number of columns
[r,c]=size(x);

if (nargin <= 1)
   weights=repmat(1,r,c);
else
   weights=repmat(wtvec,1,c)*(r/sum(wtvec));
end

xmean=mean(weights.*x);
xstd=sqrt((sum(weights.*((x-kron(xmean,ones(r,1))).^2)))/(r-1));
```

**Figure B.3: Regression Routine *reggaer***

```
% This m-file computes the regression equation - i.e., intercept
% and slopes, and computes t for testing the significance of the
% slopes of the regression line against the null hypothesis
% that each slope = 0.
%
% It also provides an analysis of variance, an f-ratio and r square.
% The ANOVA gives sum of squares, degrees of freedom
% and mean squares for the mean, regression and residuals.

function [tpart,apart,f,rr]=reggaer(x,y,intercept)

[u,v]=size(x);

if (nargin <= 2)
   intercept = 1;
end

if (intercept == 1)
   x=[ones(u,1) x];
end

tt=x'*x;
q=inv(tt);
b=q*x'*y;
%b is the vector of regression coefficients: intercept and slope
         z=y-x*b;
            [h,k]=size(b);
      t1=z'*z;
%t1 is the residual sum of squares
      t2=u-h;
%t2 is the residual degrees of freedom
      t3=t1/t2;
%t3 is the residual mean squares
         z=z/(u-h);
%z is the error variance for the regression line
         varb=t3*q;
%varb is the variance of the regression coefficients.
stdb=sqrt(diag(varb));
t=abs(b./stdb);
s1=y'*y;
```

**Figure B.3: Regression Routine reggaer (continued)**

```
%s1 is the total sum of squares
[q,r]=size(y);
    s2=q;
%s2 is the total number of degrees of freedom
    s3=s1/s2;
%s3 is the total mean squares
        q1=sum(y)*sum(y)/q;
%q1 is the sum of squares for the mean
        q2=r;
%q2 us the degrees of freedom for the mean
       q3=q1/q2;
%q3 is the mean squares for the mean
   w1=b'*x'*y-q1;
%w1 is the regression sum of squares
   w2=h-q2;
%w2 is the regression degrees of freedom
   w3=w1/w2;
%w3 is the regression mean squares
         f=w3/t3;
sumsq=[q1 w1 t1 s1]';
df=[q2 w2 t2 s2]';
meansq=[q3 w3 t3 s3]';
b=b';
stdb=stdb';
t=t'\
sumsq=sumsq';
df=df';
meansq=meansq';
tpart=[b' stdb' t'];
apart=[sumsq' df' meansq'];
f=f;
rr=w1/(w1+t1);
```

# APPENDIX C: THEORY UNDERLYING THE VECTOR APPROACH

Multivariate regression is based on the General Linear Model as found in most statistical analysis software. In this Appendix, it will be shown how this model is converted to a form in which the predictor variables are vector quantities rather than scalars.

## C.1 CONVENTIONAL FORM OF THE GENERAL LINEAR MODEL

We assume a linear model of the form

$$y = b_0 f_0 (x_1, x_2, \ldots x_k) + b_1 f_1 (x_1, x_2, \ldots x_k) + \ldots + b_n f_n (x_1, x_2, \ldots x_k) \qquad (1)$$

where y is the response variable and there are k predictor variables $x_1, x_2, \ldots x_k$. These variables may be inserted into the model as arguments in the n functions $f_1, f_2, \ldots f_n$, where $n >= k$, and the functions can be of arbitrary form. The n-dimensional space in which the terms of (1) are embedded is often referred to as the X-space, and the space containing all responses may be referred to as the Y-space.

Typical is the so-called *second order model*, in which the first and second powers of each of the predictor variables are included, as well as all *first-order interaction terms*, as represented by the products of all pairs of variables. These functions are said to provide a *basis* for the model – that is, the terms constitute a set of *basis functions*, linear combinations of which are capable of representing any response y in the problem context. The coefficients of these basis functions are usually computed by the method of least squares.

Conventionally, the strategy at this point is to test the coefficient of each basis function for *statistical significance* and eliminate any terms that do not reach the assigned significance level. Because of correlation that often accompanies the basis functions, it is usually necessary to re-compute the regression coefficients, and it is to this end that *stepwise regression* may be invoked.

Many problems are encountered in applying the conventional approach, and these are well documented in the statistical literature. One problem often fails to receive adequate attention, however, and there is evidence that many analysts, though aware of it, may not fully appreciate its implications.

The truth is that the model, as represented by (1), promises more than it can actually deliver. In particular, the equation seems to *imply* that the predictor variables, the basis

functions and the response variable are *continuous* and that their domain of definition is the real-number continuum. In reality, their domain is much more restricted; indeed, the response is known *only* at the finite number of locations in X-space at which samples were taken in the collection of the data. Responses imputed at points other than the sample points constitute an *extension* of the domain – that is, they extend what we know at a finite number of points to the entire real-number continuum.

There is no justification for such an extension other than intuition – the same intuition that teaches us that it is reasonable to connect data points in an x,y-plot with a continuous line or curve. In a more penetrating look at the problem, it must be recognized that there are infinitely many *possible* extensions to a set of data and that goodness-of-fit criteria, such as the residual sum of squares or the coefficient of determination ($R^2$) apply *only* to the *sample* points and may tell us little about the space between these points. Indeed, there can be infinitely many extensions that give *exactly* the *same* sum of squares and even the same residuals on a point-by-point basis.

Let us confine our attention, then, to just the points in X-space at which we have data. If we retain the notation of (1), we must think of the basis functions as being defined at *only* those points. To be exact, the notation of (1) would have to be modified by inclusion of appropriate *impulse functions* defined as zero everywhere except at the points in X-space at which the response is known. That being the case, it is advantageous and more appropriate to express the basis functions as *basis vectors* having a finite number of components and to view the set of responses as a *response vector* having only as many components as there are observations in the data.

## C.2 AN ALTERNATIVE, VECTOR-BASED MODEL

By means of certain theorems of matrix algebra, a procedure will be developed to transform the original data vectors into an orthogonal basis that eliminates many of the difficulties, such as collinearity, that accompany the conventional linear-model format. It will be seen that our approach to this problem follows a quite different strategy. Rather than pre-ordaining the form of the response space, it seeks to formulate a set of basis functions *derived completely from the data.*

The vectors arise from the data as follows. There are m observations, each consisting of an observed response and a vector description of the point in X-space at which the observation was made. These give rise to a column vector $\underline{y}$ (m x 1 matrix) of responses and a X-matrix consisting of m rows and n columns, where m is the number of observations and n is the number of basis elements. The two are related by the matrix equation

$$\underline{y} = X\,\underline{b} \qquad (2)$$

where $\underline{b}$ is a column vector having n coefficients, one for each of the basis vectors. The elements of $\underline{b}$ are computed by least squares so as to minimize the sum of squares of the residuals.

There is another set of coefficients or weighting factors, however, that fail to be recognized as such when the regression equation is written in the functional form (1). The need for these coefficients becomes quite evident when one attempts to express a generic term $b_i f_i (x_1, x_2, \ldots x_k)$ in its corresponding vector form. It turns out that this generic term, when evaluated at one of the observed data points in X-space, is actually the *multiplier*, or *coefficient*, for a basis vector in which all components are zero except the one corresponding to the term of (1) under consideration.

The point of departure for what we have termed vector variables, therefore, is simply a *generalization* of the conventional approach. Let the components of the general basis vector be unrestricted rather than requiring that all components except one be *zero*. Then construct the basis vectors in such a way that they are uncorrelated and that their associated coefficients are perforce uncorrelated also.

At this point it should be evident that there are two *levels* of coefficients. The first level deals only with the X-space – that is, the space of predictor variables. The second deals with the *regression* coefficients and needs to take into account both the predictor and response variables. These two aspects of the problem will be discussed in the following two sections.

### C.2.1 Reformulating the X-space

We seek to transform a non-orthogonal set of basis vectors, each containing only one non-zero element, into a set of basis functions that are orthogonal and may exhibit *all* non-zero components. For this purpose, we invoke the Principal Axes Theorem of matrix algebra:

> Any real symmetric matrix M is simultaneously similar and congruent to a diagonal matrix D. That is, there exists an orthogonal matrix V such that $D = V^{-1} M V$, the condition for similarity and $D = V' M V$, the condition for congruency

It is to be recalled, also, that a matrix V is orthogonal if and only if $V^{-1} = V'$ and that any matrix M, when multiplied by its transpose M' generates a symmetric matrix.

For purposes of further discussion, and without loss of generality, we assume that the X-matrix has been transformed into standard form, each column having zero mean and unit standard deviation. The correlation matrix of X can therefore be defined as

$$R = (1/N)\, X'X \qquad (3)$$

which is real and symmetric and which therefore satisfies the conditions of the Principal Axes Theorem – that is:

$$D = V' R V \tag{4}$$

This equation is satisfied if V is the matrix of eigenvectors of R and D is a diagonal matrix with the eigenvalues (or latent roots) of R displayed along the diagonal.

It is clear that the vectors of X can be expressed as a linear combination of the vectors of V:

$$X = A V' \tag{5}$$

where A is a matrix of weighting factors or coefficients. The equation can be solved for A by post-multiplying by V:

$$A = X V \tag{6}$$

Just as it was the coefficients of the original basis vectors that constituted the original, non-orthogonal X-matrix in regression, so it is the coefficients of the reformulated, orthogonal basis vectors that constitute the revised X-matrix. However, it remains to be shown that this revised matrix is column-wise orthogonal.

Recall that the correlation matrix R is defined as $R = (1/N) X'X$ and that $X = A V'$. Therefore, by substitution of (5) in (3):

$$R = (1/N) (A V')' (A V') = (1/N) V A'A V' \tag{7}$$

Premultiplying by V' and postmultiplying by V, we have:

$$V' R V = (1/N) V' V A' A V' V = (1/N) A' A \tag{8}$$

But, it is known by factorization of R, that:

$$V D V' = R \tag{9}$$

Therefore, by substitution of (9) for R in (8):

$$V' V D V' V = (1/N) A' A \tag{10}$$

or

$$(1/N) A' A = D \tag{11}$$

or

$$A' A = N D \tag{12}$$

where N is sample size and D is the matrix of eigenvalues of R.

### C.2.2 Reformulating the Y-Space

In the previous section, our concern was with providing an orthogonal basis for the predictor variables, primarily with the purpose of eliminating correlation among the predictor variables by substituting orthogonal vectors consisting of weighted combinations of the predictor variables. It was also shown, however, that the coefficient matrix A derived from that exercise presents a revised regression X-matrix that is column-wise orthogonal, a fact that makes for easy computation of the regression coefficients and for easy interpretation of the contribution made by each basis vector to the prediction of the response.

Recall equation (2):

$$\underline{y} = X \underline{b}$$

and for X substitute the matrix A derived above.

$$\underline{y} = A \underline{b} \tag{13}$$

The solution for b is given by

$$\underline{b} = (A'A)^{-1} A' \underline{y} \tag{14}$$

Since A'A is diagonal, the terms in the regression equation are independent and the sums of squares associated with each are additive. Removing one or more of the basis vectors from the regression equation does not require re-computation of the other regression coefficients. In the case of a non-orthogonal X-matrix, removal of one or more predictors would require re-computation of the regression coefficients, and it would be seen that sums of squares associated with individual terms are not additive. Further complications and uncertainties can cause iteration of the process, often by means of stepwise regression, a procedure that does not enjoy uniform acceptance by professional statisticians.

The complications noted above can be avoided only if the matrix of the normal equations, generically denoted as the X'X matrix, is diagonal, and that can occur only if the columns of X are orthogonal – i.e., display no correlation. Principles of experiment design can be invoked to provide a set of treatments exhibiting this property. Whether these treatments can be realized, however, is another matter. In many instances they can not, because the

variables may be physically related and impossible to vary independently of each other. It is to this circumstance that the theory set forth in this Appendix is addressed.

## C.3 PARTITIONING THE SUMS OF SQUARES

A useful outcome of regression analysis is its ability to partition the sums of squares into two parts: that attributable to the model and that attributable to error. In the vector-based model, it will be seen that further partitioning can be done to determine the relative contributions of the eigenfuel vectors to the model sum of squares.

Assume m samples, n basis vectors each with k components, and the linear model

$$\underline{y} = X\,\underline{b} + \underline{e} \tag{15}$$

where X is the m x n matrix of least-squares normal equations
$\underline{b}$ is the n x 1 vector of regression coefficients
$\underline{e}$ is the m x 1 vector of residuals, and
$\underline{y}$ is the m x 1 vector of responses.

The normal equations can be written in matrix notation as

$$X'X\,\underline{b} = X'\,\underline{y} \tag{16}$$

and the least-squares solution for $\underline{b}$ is

$$\underline{b} = (X'X)^{-1}\,X'\underline{y} \tag{17}$$

Note, however, that $\underline{b}$ is only an estimate and that $X\,\underline{b}$ does not reproduce the response vector $\underline{y}$. Rather, it produces an estimated or calculated vector of responses $\underline{y}_c$. The *residuals*, therefore, are defined by the error vector

$$\underline{e} = \underline{y} - \underline{y}_c \tag{18}$$

and the sum of squares of the residuals is

$$\underline{e}'\underline{e} = (\underline{y} - \underline{y}_c)'\,(\underline{y} - \underline{y}_c) = (\underline{y} - X\underline{b})'\,(\underline{y} - X\underline{b}) \tag{19}$$

Expanding (19), we obtain

$$\underline{e}'\underline{e} = \underline{y}'\underline{y} - \underline{y}'\,X\,\underline{b} - \underline{b}'\,X'\,\underline{y} + \underline{b}'\,X'X\,\underline{b} \tag{20}$$

Substituting (17) for $\underline{b}$ in the last term of (20) gives

$$\underline{e}'\underline{e} = \underline{y}'\underline{y} - \underline{y}'X\underline{b} - \underline{b}'X'\underline{y} + \underline{b}'X'X(X'X)^{-1}X'\underline{y}$$

$$= \underline{y}'\underline{y} - \underline{y}'X\underline{b} - \underline{b}'X'\underline{y} + \underline{b}'X'\underline{y}$$

$$= \underline{y}'\underline{y} - \underline{y}'X\underline{b} \qquad (21)$$

Or, since $\underline{e}'\underline{e}$ is a scalar, all terms in (21) are scalars, and each is its own transpose, (21) can be alternatively written

$$\underline{e}'\underline{e} = \underline{y}'\underline{y} - \underline{b}'X'\underline{y} \qquad (22)$$

and, since $\underline{y} = X\underline{b}$, (22) becomes

$$\underline{e}'\underline{e} = \underline{y}'\underline{y} - \underline{b}'X'X\underline{b} \qquad (23)$$

Therefore (23) can be re-arranged to provide a partitioning of the sums of squares:

$$\underline{y}'\underline{y} = \underline{b}'X'X\underline{b} + \underline{e}'\underline{e} \qquad (24)$$

| Total SS | Regression SS | Residual SS |
|---|---|---|

The total and regression sums of squares can be straightforwardly computed, and the residual or error sum of squares can be computed by difference.

It is clear that if X'X is diagonal, the regression sum of squares is just a weighted sum of the squares of the regression coefficients. Further, if X'X is orthonormal, X'X is the identity matrix, and the model sum of squares is just the sum of the squared regression coefficients.

In the case of the eigenvector model, (1/N) X'X defines the correlation matrix of X, and when X is in standardized form, the weighting factors are just N times the eigenvalues of the correlation matrix.

The major consequences of the column orthogonality of X is that sums of squares for individual basis vectors are additive, so that estimates of the contribution of each basis vector to the model sum of squares can be computed as the weighted square of the coefficient of that vector.

## C.4 ADJUSTED SUMS OF SQUARES

There are instances in which it is necessary to examine subsets of the basis vectors that make up the X-matrix for the full regression model. Notable are those instances in which

there are effects caused by two distinct categories of predictor variables, such as engine effects and fuel effects influencing heavy duty diesel emissions. The partitioning of the sums of squares for such effects may not be amenable to the method used to isolate *individual* fuel effects from the aggregate of all fuel effects, because the two sources of variance may not be orthogonal to each other.

Moreover, since their contributions to the sum of squares are not additive, the partitioning must be regarded as conditional rather than absolute. It is achieved by computing the model sum of squares with both sources of variance included in the model, then removing one, then the other, and noting the changes in the model sum of squares.

For the full model, the solution of the least-squares normal equations is

$$\underline{b} = (X'X)^{-1} X' \underline{y}$$

The estimated or calculated responses are given by

$$\underline{y}_c = X \underline{b}$$

and the model sum of squares is given by

$$\text{Model SS} = \underline{b}' X' \underline{y}$$

Suppose, now, that we partition the X-matrix and the $\underline{b}$-vector as follows:

$$X = [X_1 \;\; X_2] \qquad \underline{b}' = [\underline{b}_1 \;\; \underline{b}_2]$$

to obtain $\underline{y}_c = X_1 \underline{b}_1 + X_2 \underline{b}_2$.

Now, consider a *restricted* model, in which the second subset of basis vectors, for whatever reason, was *not* included in the model. Then:

$$\underline{y}_c = X_1 \underline{b}_1$$

and the model sum of squares is computed as

$$\text{Model SS} = \underline{b}_1' X_1' \underline{y}$$

which is the sum of squares due to $\underline{b}_1$ unadjusted. Now consider:

(SS due to $\underline{b}$) - (SS due to $\underline{b}_1$ unadjusted)

and define the difference as

(SS due to $\underline{b}_2$ adjusted for $\underline{b}_1$)

Then,

SS due to $\underline{b}$ = $\underline{b}' X' \underline{y}$

SS due to $\underline{b}_1$, unadjusted = $\underline{b}_1' X' \underline{y}$

SS due to $\underline{b}_2$ adjusted for $\underline{b}_1$ = $\underline{b}' X' \underline{y} - \underline{b}_1' X' \underline{y}$

The sum of squares due to $\underline{b}_1$ adjusted for $\underline{b}_2$ can similarly be computed.

Thus in the context of diesel emissions, one can compute either:

(1) SS for fuels adjusted for engines, or

(2) SS for engines adjusted for fuels

Present interest is in (1), but there are occasions when both adjusted sums of squares might be of interest. The mechanism of adjustment is, of course, strongly dependent on the degree of correlation between the two sources of variance, in this case engine vectors and fuel vectors. For example, fuel-related emissions may be higher for one class of engines than for another, and this engine contribution becomes confounded or aliased with the actual contribution from fuels.

More light can be shed on this subject by computing the *alias* or *bias* matrix, which makes it possible to adjust regression coefficients for effects not originally included in the model.

Suppose we have data believed to fit the model

$$\underline{y}_1 = X_1 \underline{b}_1 + \underline{e} \qquad (25)$$

for which the least-squares best estimate of $\underline{b}_1$ is

$$\underline{b}_1 = (X_1' X_1)^{-1} X_1' \underline{y}$$

and the best estimate of $\underline{y}_1$ is $X_1 \underline{b}_1$.

Now suppose that upon further study it is indicated that additional predictors $X_2$ should be added to (25) to give an additional contribution $\underline{y}_2$. Then

$$\underline{y} = \underline{y}_1 + \underline{y}_2 \qquad (26)$$

where $\underline{y}$ is distinguished from $\underline{y}_1$ by virtue of the contribution of the additional predictors.

From the assumed model, (26) can be written

$$\underline{y} = X_1 \underline{b}_1 + X_2 \underline{b}_2 \qquad (27)$$

Since it is not known that (25) is "contaminated" or "aliased" by the predictors $X_2$, the solution for $\underline{b}$ is approached in the same manner as for (25). Thus,

$$\underline{b} = (X_1' X_1)^{-1} X_1' \underline{y}$$

$$= (X_1' X_1)^{-1} X_1' X^1 \underline{b}_1 + (X_1' X_1)^{-1} X_1 X_2 \underline{b}_2$$

$$= \underline{b}_1 + A \underline{b}_2$$

where $A = (X_1' X_1)^{-1} X_1' X_2$ is the alias matrix.

Suppose now that the vehicle predictors constitute the "forgotten" matrix $X_2$ and $X_1$ is the matrix of fuel predictors. Then the biased estimates of the fuel effects is

$$\underline{b} = \underline{b}_1 + A \underline{b}_2$$

where $\underline{b}_2$ is the least-squares estimate of the vehicle coefficient vector. Then,

$$\underline{b}_1 = \underline{b} - A \underline{b}_2$$

This equation allows for the direct computation of the fuel coefficients adjusted for vehicle effects in terms of the biased coefficients and the vehicle coefficients as operated on by the alias matrix. Thus it reveals the mechanism by which the fuel coefficients are modified by the vehicle coefficients.

# APPENDIX D: SIMPLIFICATION OF VECTOR-BASED REGRESSION MODELS

## D.1 INTRODUCTION

An important element in the eigenfuel methodology is the procedure for simplifying vector-based regression models by:

- Minimizing the number of eigenvectors in the model
- Minimizing the number of components in those vectors.

Minimizing or "pruning" the number of vectors is straightforward, inasmuch as the sum of squares contributions are additive. The criterion for dropping an eigenvector can be its failure to reach a specified significance level, the fact that it contributes minimally to the model sum of squares (meaning its elimination would have little effect on the predicted responses), or a combination of its significance and its substantiality.

The primary thrust of this appendix is to put forth a scheme for simplifying the internal structure of the eigenvectors by removing fuel variables across all eigenvectors. When one or more fuel variables are removed in this way, the dimensionality of the problem is reduced – for example, from the 12 fuel properties originally chosen to a smaller number that remain after pruning.

When an eigenvector is retained in the model, all fuel variables are retained with the weights, or numerical components, that make up that eigenvector. Fuel variables with small weights would seem to be likely candidates for removal. As will be shown, however, this conjecture does not necessarily follow. Though one might consider deleting components of individual eigenvectors, it is actually more logical to remove a component (fuel variable) only if its overall effect, as summed across all eigenvectors, is negligible. The reasoning behind this contention is as follows.

A predictive equation for $NO_x$, for example, consists of a set of regression coefficients that apply to respective eigenvectors in the model. These coefficients multiply every component of their associated eigenvectors. Thus the total contribution of a particular fuel variable is a function of the sum of a set of products that represents the overall influence of the variable on emissions. It is quite possible, therefore, for a fuel variable to appear very influential in a particular eigenvector and yet have only a minor effect on the final predicted emissions. For example, a large positive effect of a fuel variable in a particular eigenvector could be offset by negative effects coming from one or more other

eigenvectors. It is the combined effect of a fuel variable across all eigenvectors that determines whether that variable is influential enough to be retained in the model.

We propose a pruning procedure that combines both probabilistic and magnitude criteria as a preferred approach to the simplification of vector-based models. Under this procedure, eigenvectors would be tested first and rejected by a conventional t-test of the eigenvector regression coefficient. That having been done, the remaining eigenvectors would be transformed to compute the percent contributions of the fuel variables to the model sum of squares. Fuel variables that contribute little to the model sum of squares would then be eliminated on the basis that their omission will have negligible numerical effect on emissions predicted by the model. Iteration of the approach could provide additional refinement by re-computing the eigenvectors and eigenvalues of the reduced set of variables.

The approached developed above is preferred to stepwise regression. The eigenvector approach is an ordered system, whereas stepwise regression is only partially ordered. Moreover, tests of significance are found to be more discriminating when applied to eigenvectors than when applied to individual fuel properties. It is this characteristic that points to the testing of eigenvectors first and individual fuels last.

A MatLab program file called simplify.m has been written to combine the effects of fuel variables across eigenvectors and to compute the percent contribution of each fuel variable to the model sum of squares. The software has been applied to $NO_x$ regressed on the eigenvectors as determined earlier in this report. The simplification program may be found in Addendum D-I, and the application of this program in a demonstration of the transformations involved is presented in Addendum D-II. Addendum D-III examines the performance of stepwise regression, in comparison to the performance of the eigenfuel methodology, for selecting an optimal set of variables.

## D.2 THEORY

It is easily shown (see Appendix C) that the vector of coefficients $\underline{c}$, as computed for an eigenvector model of emissions, can be readily transformed into a vector of coefficients $\underline{cs}$ applicable to the corresponding model that is expressed in terms of the fuel-variable components that make up the eigenvectors.

Table D.1 shows the regression coefficient vectors $\underline{c}$ and $\underline{cs}$ for $\log(NO_x)$ emission models in which the 12 eigenvectors and the 12 fuel properties, respectively, are used as regressors. All 12 eigenvectors are included in the vector model (coefficient vector $\underline{c}$), and all 12 fuel properties are included in the fuel-variable model (coefficient vector $\underline{cs}$). The models are computed without removing vehicle effects, so that the regression equations are simplified in comparison to those presented in Section 2 of this report. The

**Table D.1. Log($NO_x$) Regressed on Eigenvectors and Fuel Variables**

| Eigenvector Coefficients, c | | | Fuel Variable Coefficients, cs | |
|---|---|---|---|---|
| Constant | 1.5372 | | Constant | 1.5372 |
| Eigenvector 1 | -0.0012 | | NatCetane | 0.0077 |
| Eigenvector 2 | 0.0335 | | CetImprv | -0.0115 |
| Eigenvector 3 | 0.0730 | | Density | 0.0458 |
| Eigenvector 4 | -0.0374 | | Viscosity | -0.0380 |
| Eigenvector 5 | -0.0002 | | Sulfur | -0.0020 |
| Eigenvector 6 | -0.0117 | | MonoArom | 0.0344 |
| Eigenvector 7 | 0.0021 | | PolyArom | -0.0013 |
| Eigenvector 8 | 0.0233 | | IBP | -0.0189 |
| Eigenvector 9 | 0.0124 | | T10 | 0.0493 |
| Eigenvector 10 | 0.0016 | | T50 | -0.0263 |
| Eigenvector 11 | -0.0329 | | T90 | -0.0263 |
| Eigenvector 12 | 0.0043 | | FBP | 0.0265 |

simplification is made here solely for purposes of presentation.

The constant term (intercept) is the same in both versions of the model, and the relationship between for the fuels-related coefficients is given by the transformation:

$$\underline{cs} = V * \underline{c}$$

where V is the 12 x 12 matrix of eigenvectors and c and cs are the eigenfuel and fuel-variable coefficients, respectively, constant term excluded.

Now let us turn our attention to the model sum of squares (SS). When computed from the regression model, one simply evaluates the regression equation for each of the observations in the data set to obtain the so-called estimated or predicted log emissions. These are then subtracted from the corresponding observed log emissions to obtain the residuals. The sum of squares of the residuals is what is labeled by that name in

conventional regression Analysis of Variance (ANOVA). The difference between the Total SS and the Residual SS is then said to be the Model SS.

For the orthogonal vector-based regression model, let us drop all eigenvector terms except one and compute the Model SS, and let us do this for each of the 12 eigenvectors in the model. Since the eigenvectors are orthogonal, one can add the 12 separately computed sums of squares to obtain the aggregate Model SS for the model incorporating all of the eigenvector terms. The regression coefficients in the aggregate model are exactly the same as those computed for corresponding individual terms.

An interesting and very useful observation can now be made, as illustrated in the following Table D.2. We regressed log($NO_x$) on all eigenvectors, then one at a time, to demonstrate that the same coefficients are obtained whether they are computed singly or collectively. It was seen that the Model SS are additive, as are also the R-squares. The right-most column in the table was computed from the conventional ANOVA.

Note the agreement of the calculation shown in the third column with the conventionally determined Model SS. Thus, the Model SS for any single term or any combination of

**Table D.2. Determination of Model SS for Eigenvector Terms**

| Eigenvector Term | Regression Coeff, c | Eigenvalue, e | Product c*c*e*279 | Model SS |
|---|---|---|---|---|
| 1 | 0.0335 | 0.0359 | 0.0113 | 0.0113 |
| 2 | -0.0012 | 0.0257 | 0.0000 | 0.0000 |
| 3 | 0.0730 | 0.0826 | 0.1229 | 0.1229 |
| 4 | -0.0374 | 0.1403 | 0.0547 | 0.0547 |
| 5 | -0.0002 | 0.2305 | 0.0000 | 0.0000 |
| 6 | -0.0117 | 0.3895 | 0.0148 | 0.0148 |
| 7 | 0.0021 | 0.5636 | 0.0007 | 0.0007 |
| 8 | 0.0233 | 0.6808 | 0.1034 | 0.1034 |
| 9 | 0.0124 | 1.1806 | 0.0506 | 0.0506 |
| 10 | 0.0016 | 1.5490 | 0.0010 | 0.0010 |
| 11 | -0.0329 | 2.5908 | 0.7805 | 0.7805 |
| 12 | 0.0043 | 4.5306 | 0.0236 | 0.0236 |

terms can be computed simply by squaring the coefficient, multiplying that square by the eigenvalue of the associated eigenvector, and then multiplying by the number of degrees of freedom (here 279).

Now comes the ***piece de resistance***. Consider a regression equation containing only one eigenvector as the basis (in addition to a constant term, of course). Let the coefficient for that vector be c, this time a scalar. Then,

$$\text{Model SS} = c * c * e * n$$

where e is the eigenvalue associated with the eigenvector in the regression equation, and n is the number of observations.

It is obvious that if this single-term regression model is re-expressed in terms of the original fuel variables, there will be 12 terms (in addition to the constant). It should also be clear that the Residual SS, and hence the Model SS, will be the same, whether computed from the original or transformed equations. Recall that

$$\underline{cs} = V\,\underline{c}$$

Multiply both sides of the equation by the transpose of V. Then,

$$V'\,\underline{cs} = V' * V * \underline{c}$$

and, since $V' * V = I$, the identity matrix, we have

$$\underline{c} = \underline{cs}\,V$$

Further,

$$\underline{c}^2 = (\underline{cs} * V) * (\underline{cs} * V) \qquad (1)$$

or, for each successive variable weight in the vector V, we have

$$\text{Model SS} = (c_i * v_i)^2$$

and, by adding these SS for all 12 vector components, one obtains the complete SS for that eigenvector.

It should be clear, then, that the relative "importance" of each component in a given vector can be obtained from (1) and that the contribution of each fuel variable can be expressed as a fraction or percent of the composite SS for that eigenvector.

The SS weights vary from one eigenvector to another, and when several eigenvectors are admitted into the model, (1) must be applied to each eigenvector and their results summed across those eigenvectors. Consequently, a variable that showed great "importance" in a single-vector model may be relatively unimportant when the model is expanded to include additional eigenvectors. Therefore, if the relative weights of individual fuel variables are to be the criterion for inclusion or exclusion of a variable in the set of eigenvectors comprising the model, that selection must take into account the concatenation of effects coming from individual eigenvectors.

## D.3 PERTINENT TRANSFORMATIONS: A TUTORIAL

To clarify some of the transformations involved in the statistical interpretation of vector-based regression, it will be informative to "walk through" a demonstration of procedure. First, we consider the transformation from the regression coefficients of eigenvectors to the corresponding coefficients of the fuel variables making up those eigenvectors. The transformation can be accomplished by multiplying the matrix of eigenvectors into the column vector of eigenvector regression coefficients.

How the transformation "plays out" from eigenvector regression coefficients to fuel-variable regression coefficients can be seen by writing out *in extenso* the matrix multiplication: V * $\underline{c}$. For demonstration purposes, it suffices to use only a few terms, and we choose to use the three eigenvectors with the largest eigenvalues.

| | Eig 1 | | Eig 2 | | Eig 3 |
|---|---|---|---|---|---|
| | 0.1633 | | 0.5556 | | 0.0612 |
| | -0.5490 | | -0.1435 | | 0.0344 |
| | 0.0492 | | -0.4488 | | 0.2849 |
| | -0.0169 | | 0.1196 | | 0.4323 |
| $\underline{c1}$ | 0.6362 | $\underline{c2}$ | -0.1799 | $\underline{c3}$ | 0.0017 |
| (0.0016)* | -0.2564 | +(-0.0329)* | -0.4640 | + (0.0043)* | 0.1456 |
| | 0.3851 | | -0.4177 | | 0.0758 |
| | 0.0521 | | 0.0725 | | 0.2619 |
| | 0.1278 | | 0.1134 | | 0.3988 |
| | 0.0641 | | 0.0828 | | 0.4309 |
| | -0.0746 | | 0.0743 | | 0.3921 |
| | -0.1472 | | 0.0482 | | 0.3647 |

The computation leads to the coefficients of the fuel variables, as follows:

| | |
|---|---|
| NatCetane | -0.0177 |
| CetImprv | 0.0040 |
| Density | 0.0161 |
| Viscosity | -0.0021 |
| Sulfur | 0.0069 |
| MonoArom | 0.0155 |
| PolyArom | 0.0147 |
| IBP | -0.0012 |
| T10 | -0.0018 |
| T50 | -0.0008 |
| T90 | -0.0009 |
| FBP | -0.0002 |

The procedure for computing the associated SS for the fuel properties is a bit more complicated, because it must incorporate, in addition to squaring the coefficients, the scaling effects of the associated eigenvalues.

For this purpose it is advantageous to multiply each vector by its *coefficient* and by the *square root of its eigenvalue* and then to square the resulting vector.

| | | | | |
|---|---|---|---|---|
| m1 = | ( 0.0016) | * 1.2446 | = | 0.0020 |
| m2 = | (-0.0324) | * 1.6096 | = | -0.0522 |
| m3 = | ( 0.0043) | * 2.1285 | = | 0.0092 |

These multipliers – m1, m2, and m3 – are multiplied into their respective eigenvectors:

| Eig 1 | Eig 2 | Eig 3 |
|---|---|---|
| 0.1633 | 0.5556 | 0.0612 |
| -0.5490 | -0.1435 | 0.0344 |
| 0.0492 | -0.4488 | 0.2849 |
| -0.0169 | 0.1196 | 0.4323 |
| 0.6362 | -0.1799 | 0.0017 |
| m1*-0.2564= q1 | m2*-0.4640= q2 | m3*0.1456= q3 |
| 0.3851 | -0.4177 | 0.0758 |
| 0.0521 | 0.0725 | 0.2619 |
| 0.1278 | 0.1134 | 0.3988 |
| 0.0641 | 0.0828 | 0.4309 |
| -0.0746 | 0.0743 | 0.3921 |
| -0.1472 | 0.0482 | 0.3647 |

| q1 | q2 | q3 |
|---|---|---|
| 0.0003 | -0.0294 | 0.0006 |
| -0.0011 | 0.0076 | 0.0003 |
| 0.0001 | 0.0237 | 0.0026 |
| -0.0000 | -0.0063 | 0.0040 |
| 0.0012 | 0.0095 | 0.0000 |
| -0.0005 | 0.0245 | 0.0013 |
| 0.0007 | 0.0221 | 0.0007 |
| 0.0001 | -0.0038 | 0.0024 |
| 0.0002 | -0.0060 | 0.0037 |
| 0.0001 | -0.0044 | 0.0040 |
| -0.0001 | -0.0039 | 0.0036 |
| -0.0003 | -0.0025 | 0.0034 |

Finally, these vectors must be squared by squaring each of their components and multiplied by the number of degrees of freedom (280 - 1 = 279) and then added. The end result of this procedure is as follows:

```
279*( q1.*q1 + q2.*q2 + q3.*q3 ) =
```

| | |
|---|---|
| NatCetane | 0.2411 |
| CetImprv | 0.0164 |
| Density | 0.1591 |
| Viscosity | 0.0156 |
| Sulfur | 0.0257 |
| MonoArom | 0.1686 |
| PolyArom | 0.1365 |
| IBP | 0.0057 |
| T10 | 0.0138 |
| T50 | 0.0097 |
| T90 | 0.0080 |
| FBP | 0.0050 |
| | |
| Total SS | 0.8051 |

The total sum of squares calculated from the above values checks against the Model SS as computed from the regression ANOVA.

## D.4 SUMMARY AND CONCLUSIONS

In summary, we believe that a solid and unique basis has been obtained for simplifying a regression model. It has a hierarchical structure, as follows:

1. Discard eigenvectors if they fail to meet a desired level of significance or if their contribution to the model SS is negligible or minimal – say less than few percent.

2. For the terms retained, carry through the simplification procedure shown here to estimate the sum of squares contributions for each fuel variable, considering for the concatenation of all eigenvector terms retained in the model.

3. Eliminate those components (fuel variables) that contribute little (say less than a few percent) to the final combined vector or whose omission will occasion only negligible change in model predictions.

What threshold to use at either Step 1 or Step 2 is something of a judgement call and may seem contentious. However, it should be based on the reality of the prediction equation and what its consequences are in the real world. Though a "percentage of SS" basis has been suggested above, one could just as well, and perhaps more meaningfully, determine how much "error" or, more importantly, how much "loss of information" can be tolerated when setting a threshold.

A major advantage of the procedure over alternatives such as stepwise regression is that the process is hierarchical and its outputs are numbers that can be interpreted in an *ordinal* sense. In contrast, stepwise regression can yield unstable results in that it is possible to end up with a variety of "best models" with seemingly equivalent performance .

Once a model has been selected by hierarchical means, it should be iterated to confirm the choice. Repeating the eigenstructure analysis with the "unimportant" bits removed would yield a possible refinement of estimation. For example, there should be no eigenvector carrying weights less than the threshold, and likewise for fuel variables within the eigenvectors considered collectively.

Finally, though no formal proof is offered here, it seems evident that the eigenvector approach is more discriminating than stepwise regression. In the eigenvector approach, the first eigenvector (the one associated with the largest eigenvalue) takes the largest "bite" out of the model sum of squares. This bite is also likely to be larger than the largest bite possible when emissions are regressed on the fuel variables, because the correlation with other variables will "dilute" the effects of any particular fuel variable. When arranged in descending order of their eigenvalues, the SS contributions of the eigenvalues should have the steepest gradient.

## Addendum D-I: A Program for Computing SS Contributions

```
function y=simplify(vecref,valref,coeff,s,n)
% Computes the SS contributions of the components of eigenfuels to
% the total contribution for any subset of the complete set of
% eigenfuels.
% vecref = complete set of eigenvectors, i.e., eigenvector matrix
% valref = complete set of eigenvalues, in column vector form
% coeff  = regression coefficients, intercept term excluded
% s = row vector denoting the eigenvectors to be retained in model
% n = the number of cases
[h,k] = size(coeff);
[a,b] = size(s);
vec = vecref(:,s);
val = valref(s);
coeffs = coeff(s);
z = zeros(h,1);
for i=1:b
      c = coeffs(i);
      cstar = c*vec(:,i);
      e = val(i);
      x = cstar*sqrt(e);
      z = z+x;
end
y1 = (n-1)*z.*z;
p = 100*y1/sum(y1);
y = [ [y1; sum(y1)] [p; sum(p)] ];
```

## Addendum D-II: Computing Model SS Contributions by Separate Fuel-Variable Components in a Vector-based Model

We begin by regressing $\log(NO_X)$ on the full set of 12 eigenvectors. The X-space inputs are the coefficients of the eigenvectors for each of the 280 observations. Regression of $\log(NO_X)$ on all eigenvectors:

| | Coeff | Std Err | t-ratio |
|---|---|---|---|
| Constant | 1.5372 | 0.0029 | 529.4058 |
| Eig 1 | 0.0043 | 0.0014 | 3.1639 |
| Eig 2 | -0.0329 | 0.0018 | 18.1829 |
| Eig 3 | 0.0016 | 0.0023 | 0.6663 |
| Eig 4 | 0.0124 | 0.0027 | 4.6290 |
| Eig 5 | 0.0233 | 0.0035 | 6.6174 |
| Eig 6 | 0.0021 | 0.0039 | 0.5303 |
| Eig 7 | -0.0117 | 0.0047 | 2.5006 |
| Eig 8 | -0.0002 | 0.0061 | 0.0304 |
| Eig 9 | -0.0374 | 0.0078 | 4.8132 |
| Eig 10 | 0.0730 | 0.0101 | 7.2147 |
| Eig 11 | -0.0012 | 0.0181 | 0.0689 |
| Eig 12 | 0.0335 | 0.0154 | 2.1833 |

**Regression Analysis of Variance**

| | SS | DF | MS |
|---|---|---|---|
| Intercept | 661.6004 | 1.0000 | 661.6004 |
| Model SS | 1.1633 | 12.0000 | 0.0969 |
| Residual SS | 0.6303 | 267.0000 | 0.0024 |
| Total SS | 663.3940 | 280.0000 | 2.3693 |

**R-square = 0.6486**

We now will apply the program simplify.m to determine the relative contributions of the fuel variables when all eigenvectors are retained in the model.

First, we define s to include all eigenvectors:

```
s = [1 2 3 4 5 6 7 8 9 10 11 12]
```

Then invoke simplify.m with vecref and velref coming from the eigenanalysis of all data, and with coeff being the coefficients obtained in the above regression:

```
y=simplify(vecref,valref,coeff,s,280)
```

This returns the result shown below. Note that the total SS is the same as obtained from conventional regression analysis.

| | SS | Percent SS |
|---|---|---|
| NatCetane | 0.1614 | 13.8723 |
| CetImprv | 0.0264 | 2.2661 |
| Density | 0.3396 | 29.1929 |
| Viscosity | 0.0252 | 2.1622 |
| Sulfur | 0.0033 | 0.2870 |
| MonoArom | 0.4058 | 34.8852 |
| PolyArom | 0.0543 | 4.6667 |
| IBP | 0.0073 | 0.6310 |
| T10 | 0.0629 | 5.4108 |
| T50 | 0.0256 | 2.2045 |
| T90 | 0.0374 | 3.2180 |
| FBP | 0.0140 | 1.2033 |
| Total | 1.1633 | 100.0000 |

Simplification of the regression model is accomplished by "pruning" either entire eigenvectors or selected components of those vectors. Pruning consists of removing terms considered to be in some sense "inappropriate' to include in the model. The basis for removal can be based on statistical, magnitude or judgemental considerations.

Logic seems to indicate that pruning begin with eigenvectors. The percent contributions of the 12 eigenvectors, arranged in decreasing order of magnitude, are as follows:

| | | | --- Sorted --- | | |
|---|---|---|---|---|---|
| Vector | SS | Percent SS | Vector | SS | Cum Percent |
| 1 | 0.0236 | 2.0284 | 2 | 0.7805 | 67.08 |
| 2 | 0.7805 | 67.0821 | 10 | 0.1229 | 77.65 |
| 3 | 0.0010 | 0.0859 | 5 | 0.1034 | 86.53 |
| 4 | 0.0506 | 4.3489 | 9 | 0.0547 | 91.23 |
| 5 | 0.1034 | 8.8870 | 4 | 0.0506 | 95.58 |
| 6 | 0.0007 | 0.0602 | 1 | 0.0236 | 97.61 |
| 7 | 0.0148 | 1.2720 | 7 | 0.0148 | 98.88 |
| 8 | 0.0000 | 0.0000 | 12 | 0.0113 | 99.85 |
| 9 | 0.0547 | 4.7013 | 3 | 0.0010 | 99.94 |
| 10 | 0.1229 | 10.5630 | 6 | 0.0007 | 100.00 |
| 11 | 0.0000 | 0.0000 | 8 | 0.0000 | 100.00 |
| 12 | 0.0113 | 0.9712 | 11 | 0.0000 | 100.00 |

The "best 7" eigenvectors, [1 2 4 5 7 9 10], cover nearly 99 percent (98.88) of the sum of squares. This is an arbitrary cut-off, but it ignores only 1 percent of the SS yet eliminates 5 eigenvectors.

Next, we regress log($NO_X$) on these 7 eigenvectors. Here are the coefficients, their standard errors and the associated t-values. Note that the regression coefficients are the same as the corresponding coefficients when all eigenvectors are used. Note, also, that the t-ratios are reduced only minimally by the elimination of the other five eigenvectors, thereby indicating that elimination of the five eigenvectors does little to change either the magnitude of the predictions or their statistical significance.

| | Coeff | Std Err | t-ratio |
|---|---|---|---|
| Intercept | 1.5372 | 0.0029 | 528.9223 |
| Eig 1 | 0.0043 | 0.0014 | 3.1611 |
| Eig 2 | -0.0329 | 0.0018 | 18.1663 |
| Eig 3 | 0.0124 | 0.0027 | 4.6248 |
| Eig 4 | 0.0233 | 0.0035 | 6.6113 |
| Eig 5 | -0.0117 | 0.0047 | 2.4983 |
| Eig 6 | -0.0374 | 0.0078 | 4.8088 |
| Eig 7 | 0.0730 | 0.0101 | 7.2081 |

Analysis of Variance

| | SS | DF | MS |
|---|---|---|---|
| Intercept | 661.6004 | 1.0000 | 661.6004 |
| Model SS | 1.1504 | 7.0000 | 0.1643 |
| Error SS | 0.6433 | 272.0000 | 0.0024 |
| Total SS | 663.3940 | 280.0000 | 2.3693 |

Note that 100 * 1.1504/1.1633 = 98.89 percent as was foreseen. We now define the selector variable s to include the "7 best" eigenvectors used in the regression above:

s = [1 2 4 5 7 9 10]

Then we invoke simplify.m: y=simplify(vecref,valref,coeff,s,280)

| | SS | Percent SS |
|---|---|---|
| NatCetane | 0.1609 | 13.9901 |
| CetImprv | 0.0228 | 1.9863 |
| Density | 0.2983 | 25.9304 |
| Viscosity | 0.0099 | 0.8647 |
| Sulfur | 0.0010 | 0.0873 |
| MonoArom | 0.4413 | 38.3644 |
| PolyArom | 0.0636 | 5.5299 |
| IBP | 0.0046 | 0.3982 |
| T10 | 0.0404 | 3.5139 |
| T50 | 0.0260 | 2.2638 |
| T90 | 0.0576 | 5.0045 |
| FBP | 0.0238 | 2.0665 |
| Total | 1.1504 | 100.0000 |

The table gives the relative percent of the sum of squares accounted for by each of the fuel variables. The fact that the sum (1.1504) checks against the value of SS obtained by direct computation assures that the break-down is correct.

We are now in a position to "prune" non-contributing variables. These include viscosity, sulfur and IBP on the basis of their minimal contribution to the model SS. Because both IBP and FBP are believed to be less reliably measured than T10, T50 and T90, FBP is also eliminated. (This consideration is included to illustrate the intervention of engineering judgement in eliminating terms.) This results in eight eigenvectors, each with eight components.

An iteration of the eigenanalysis is now in order. The eigenvectors and eigenvalues are recomputed for the new set of eight eigenvectors.

| | Eigenvectors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.4622 | -0.3652 | 0.4628 | 0.1238 | 0.2306 | -0.0237 | 0.5828 | -0.1734 |
| CetImprv | 0.0627 | -0.1062 | 0.0349 | -0.3969 | 0.2705 | -0.8452 | -0.1743 | 0.0901 |
| Density | 0.2044 | 0.3241 | 0.6751 | -0.1943 | -0.1833 | 0.1354 | -0.2250 | 0.5070 |
| MonoArom | 0.2304 | -0.4936 | -0.0848 | 0.5955 | -0.2276 | -0.1639 | -0.3589 | 0.3660 |
| PolyArom | 0.0423 | -0.2881 | -0.1170 | -0.2004 | 0.7218 | 0.4432 | -0.2655 | 0.2708 |
| T10 | -0.3824 | -0.5132 | -0.0388 | -0.4166 | -0.3637 | 0.0799 | 0.3617 | 0.3618 |
| T50 | 0.5111 | 0.3218 | -0.5376 | -0.0693 | -0.0189 | -0.0284 | 0.3767 | 0.4454 |
| T90 | -0.5275 | 0.2381 | 0.1321 | 0.4650 | 0.3669 | -0.1903 | 0.3275 | 0.3915 |
| | Eigenvalues | | | | | | | |
| | 0.1272 | 0.1912 | 0.0502 | 0.4961 | 0.7006 | 1.0805 | 2.3580 | 2.9961 |

Log($NO_X$) is now regressed on the complete set of eight eigenvectors retained from the previous operations.

| | Coeff | Std Err | t-ratio |
|---|---|---|---|
| Intercept | 1.5372 | 0.0031 | 496.5049 |
| Eig 1 | 0.0121 | 0.0081 | 1.4909 |
| Eig 2 | -0.0121 | 0.0034 | 3.5534 |
| Eig 3 | 0.0466 | 0.0099 | 4.7071 |
| Eig 4 | 0.0352 | 0.0051 | 6.8376 |
| Eig 5 | 0.0044 | 0.0041 | 1.0513 |
| Eig 6 | 0.0148 | 0.0056 | 2.6572 |
| Eig 7 | -0.0411 | 0.0090 | 4.5428 |
| Eig 8 | -0.0062 | 0.0059 | 1.0596 |

Analysis of Variance

| | SS | DF | MS |
|---|---|---|---|
| Intercept | 661.6004 | 1.0000 | 661.6004 |
| Model SS | 1.0663 | 8.0000 | 0.1333 |
| Error SS | 0.7273 | 271.0000 | 0.0027 |
| Total SS | 663.3940 | 280.0000 | 2.3693 |

R-square = 0.5945

For comparison, we compute the regression equation by regressing $\log(NO_x)$ directly on the eight retained fuel variables.

| | Coeff | Std Err | t-ratio |
|---|---|---|---|
| Intercept | 1.5372 | 0.0031 | 496.5049 |
| NatCetane | -0.0007 | 0.0087 | 0.0768 |
| CetImprv | -0.0289 | 0.0071 | 4.0754 |
| Density | 0.0539 | 0.0138 | 3.8915 |
| MonoArom | 0.0111 | 0.0044 | 2.5257 |
| PolyArom | -0.0208 | 0.0037 | 5.6034 |
| T10 | 0.0160 | 0.0030 | 5.3578 |
| T50 | -0.0272 | 0.0020 | 13.4837 |
| T90 | 0.0194 | 0.0018 | 10.8286 |

Analysis of Variance

| | SS | DF | MS |
|---|---|---|---|
| Intercept | 661.6004 | 1.0000 | 661.6004 |
| Model SS | 1.0663 | 8.0000 | 0.1333 |
| Error SS | 0.7273 | 271.0000 | 0.0027 |
| Total SS | 663.3940 | 280.0000 | 2.3693 |

R-square = 0.5945

Now, using simplify.m, we compute the percent contribution to the model SS accorded to each of the fuel variables: s = [1 2 3 4 5 6 7 8].

| | SS | Percent SS | Prior Percent |
|---|---|---|---|
| NatCetane | 0.1546 | 14.4966 | 13.9901 |
| CetImprv | 0.0267 | 2.5082 | 1.9863 |
| Density | 0.3293 | 30.8809 | 25.9304 |
| MonoArom | 0.4105 | 38.5004 | 38.3644 |
| PolyArom | 0.0685 | 6.4255 | 5.5299 |
| T10 | 0.0187 | 1.7561 | 3.5139 |
| T50 | 0.0412 | 3.8673 | 2.2638 |
| T90 | 0.0167 | 1.5650 | 5.0045 |
| Total | 1.0663 | 100.0000 | 100.0000 |

It is evident that by pruning from the "low end" of both eigenvectors and components of eigenvectors, there is a certain sense of stability and directionality in the results. The fuel variable contributions as computed by way of the original 12 eigenvectors do not differ radically from their contributions as computed from the revised set of eight eigenvectors.

It is believed that the scheme proposed above is more systematic and less erratic than results obtained by way of stepwise regression. It has the advantage of ordering the terms in the model from greatest to least contribution, and is subject to less variation than results obtained stepwise. By stepwise regression, one can obtain a large number of solutions giving essentially the same "performance" as judged by R-square. One can argue that the eigenvector approach gives a unique result, once the thresholds for acceptance or rejection of terms is established.

The results of this paper do not necessarily define a precise protocol for simplifying a regression model that contains correlated predictor variables, but it outlines an approach and methodology for such a protocol. Further study and exercising of the models are necessary. The outlined approach, however, is believed to be innovative and to have the advantages indicated above.

**Addendum D-III: Some Examples of Stepwise Regression**

Stepwise regression is often used to select the "best" set of terms to include in a regression equation. Parameters for the procedure include "alpha" – i.e., the significance level for entering or removing a term – and a quantity called "tolerance." Variables can also be forced into the equation regardless of what happens in the selection process.

In SYSTAT, default values are alpha = 0.15 and tolerance = 0.01. The SYSTAT manual says that the default alpha value is acceptable for predictors that are "relatively independent." It goes on to say that "... if your predictors are correlated, you should set the alphas to 0.05 or less."

Tolerance is defined as "... 1 minus the squared multiple correlation between a predictor and the remaining predictors in the model." Any variable with a tolerance less than the one specified will not be allowed to enter the equation. This constraint is said to protect

against "highly multi-collinear models that tend to have unstable regression coefficient estimates."

Stepwise regression can be exercised either automatically or interactively. In the automatic mode, one simply specifies the constraints and lets the program seek what it sees as the optimum, without any interference from the analyst. In the interactive mode, the analyst can examine results step by step and override automatic choices if he sees fit to do so.

With as many options as are possible in the stepwise procedure, it should be apparent that a variety of "best" regression models are possible and that there is no way to order the choices in an ascending or descending sequence. To illustrate the stepwise procedure and how it might apply to the diesel data, ten options are explored. Five values of alpha – 0.15, 0.10, 0.05, 0.01, and 0.001 – were used in both forward and backward mode. Only one tolerance value – the default value of 0.01 – was used. A summary of the results for these trials is given below.

For comparison, a procedure was developed to compute all possible regressions based on 1 or more of the 12 regressors. When applied to the data, it is easy to see that by the usual method of computation there may be no sharp optimum that represents the "best" selection of variables. Indeed, many, sometimes hundreds, of cases can be found that have essentially the same R-square yet quite different choices of variables to include in

Variables Selected and R-Squares Attained
In Ten Trial Stepwise Regressions
of log($NO_x$) on Twelve Diesel Fuel Variables

| | <---- Forward Selection ----> | | | | | <--- Backward Selection ---> | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Alpha | 0.150 | 0.100 | 0.050 | 0.010 | 0.001 | 0.150 | 0.100 | 0.050 | 0.010 | 0.001 |
| N* | 9 | 9 | 9 | 7 | 3 | 9 | 9 | 9 | 8 | 7 |
| R-square | 0.647 | 0.647 | 0.647 | 0.626 | 0.514 | 0.647 | 0.647 | 0.647 | 0.640 | 0.624 |
| NatCetane | | | | | X | | | | | |
| CetImprv | X | X | X | X | X | X | X | X | X | X |
| Density | X | X | X | X | | X | X | X | X | |
| Viscosity | X | X | X | X | | X | X | X | X | X |
| Sulfur | | | | | | | | | | |
| MonoArom | X | X | X | X | X | X | X | X | X | X |
| PolyArom | | | | | | | | | | |
| IBP | X | X | X | X | | X | X | X | X | X |
| T10 | X | X | X | X | | X | X | X | X | X |
| T50 | X | X | X | X | | X | X | X | | |
| T90 | X | X | X | | | X | X | X | X | X |
| FBP | X | X | X | | | X | X | X | X | X |

* N denotes the number of variables retained. Tolerance = 0.01 for all cases.

the model. The approach described here provides a unique way to approach the optimum by means of a maximum-discrimination protocol. It is readily seen that a variety of results can be obtained, and there would be many more if the tolerance were varied.

Note that, in comparison with the eigenvector approach, natural cetane occurs in only one case, and that is when only two other variables, mono-aromatics content and cetane improvement are included. (Evidently the three variables are sufficiently uncorrelated to meet the tolerance criterion.) Also, it appears that in other cases, the multiple correlation with natural cetane is high enough to prevent that variable from entering the model.

One needs to reflect on how the eigenvector approach differs from the stepwise approach. The number of variables retained in the eigenvector approach depends on the percentage SS or other criterion, and thus would seem to be subject to the same arbitrariness as in stepwise regression. However, branching options would seem to be fewer for eigenvectors than for stepwise selection.

Most important, though, is the matter of correlation of predictor variables. Tolerance is not an issue in the eigenvector approach, because correlation among variables was eliminated at the outset. Moreover, the logic of eigenvectors negates the validity of entering or removing variables one at a time, unless they are independent. Rather, combinations of predictor variables are entered or removed one at a time, and in such a way that the predictors are represented in their proper context. There can be no "reversals," such as entering a variable at one stage and removing it at a later stage, because contributions to sums of squares are additive.

Finally, though, it would appear that the stepwise approach could be used quite effectively by applying it to the eigenvectors rather than to the predictor variables. However, there would be no need to do so, because the elimination of one eigenvector would not have any influence on the removal of another.

In stepwise regression, the various choices are just different ways of aliasing correlated variables. With eigenvectors, there are *no aliases*.

# APPENDIX E: TESTS OF STATISTICAL SIGNIFICANCE

In a conventional regression analysis, tests of significance are customarily of two types:

- t-tests of individual regression coefficients
- F-test of the significance of the model as a whole

The t-tests involve the ratio of a selected coefficient to its standard error. The F-test involves the ratio of the model mean square to the error mean square. These two tests are viewed in perspective in the following discussion.

An important point of departure is to note that the F-test is applicable to the ratio of any two random variables distributed as Chi-square, such as mean squares in a regression equation. Therefore, the test is applicable to testing the significance of the role played by any term in a regression model, providing one can compute a legitimate mean square for that term. That is clearly possible in cases in which the predictors are orthogonal to each other, as is the case when the regression is computed in the space of eigenvectors (E-Space). In the space of fuel property variables (P-Space), however, significance presents something of a dilemma. This point is made evident by Tables E.1 and E.2 below, which address, respectively, the E-Space (orthogonal) and P-Space (non-orthogonal) scenarios.

First, we examine E-Space, as shown in Table E.1. Since the SS for each of the eigenvectors has only one degree of freedom (DF), the mean square (MS) for each eigenvector is the same as the its SS. Also, it is a theoretical fact that for F-tests involving only a single DF in the numerator, the observed value of F is exactly the same as the t-ratio squared. In the table, F is computed as the ratio of the eigenvector mean square to the error mean square, computed as 0.0024.

Table E.1

Comparison of F and t Tests for Regression on Eigenfuels - No Correction for Vehicles

| Eig | Reg Coef | Std Err | t | SS/MS | Percent SS | F | $t^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.0043 | 0.0014 | 3.16 | 0.0236 | 2.0 | 10.0 | 10.0 |
| 2 | -0.0329 | 0.0018 | 18.18 | 0.7805 | 67.1 | 330.6 | 330.6 |
| 3 | 0.0016 | 0.0023 | 0.67 | 0.0010 | 0.1 | 0.4 | 0.4 |
| 4 | 0.0124 | 0.0027 | 4.63 | 0.0506 | 4.3 | 21.4 | 21.4 |
| 5 | 0.0233 | 0.0035 | 6.62 | 0.1034 | 8.9 | 43.8 | 43.8 |
| 6 | 0.0021 | 0.0039 | 0.53 | 0.0007 | 0.1 | 0.3 | 0.3 |
| 7 | -0.0117 | 0.0047 | 2.50 | 0.0148 | 1.3 | 6.3 | 6.3 |
| 8 | -0.0002 | 0.0061 | 0.03 | 0.0000 | 0.0 | 0.0 | 0.0 |
| 9 | -0.0374 | 0.0078 | 4.81 | 0.0547 | 4.7 | 23.2 | 23.2 |
| 10 | 0.0730 | 0.0101 | 7.21 | 0.1229 | 10.6 | 52.1 | 52.1 |
| 11 | 0.0335 | 0.0154 | 2.18 | 0.0113 | 1.0 | 4.8 | 4.8 |
| 12 | -0.0012 | 0.0181 | 0.07 | 0.0000 | 0.0 | 0.0 | 0.0 |

Now consider the non-orthogonal case, in which the response variable $\log(NO_x)$ is regressed directly on the standardized fuel-property variables, as shown in Table E.2 below.

Table E.2

Comparison of F and t Tests for Regression on Fuel Properties - No Correction for Differences Among Vehicles

| Property | Reg Coef | Std Err | t | SS/MS | Percent SS | F |
|---|---|---|---|---|---|---|
| NatCetane | 0.0077 | 0.0091 | 0.85 | 0.1614 | 13.9 | 68.4 |
| CetImprv | -0.0115 | 0.0033 | 3.47 | 0.0264 | 2.3 | 11.2 |
| Density | 0.0458 | 0.0109 | 4.18 | 0.3396 | 29.2 | 143.9 |
| Viscosity | -0.0380 | 0.0124 | 3.06 | 0.0252 | 2.2 | 10.7 |
| Sulfur | -0.0020 | 0.0040 | 0.50 | 0.0033 | 0.3 | 1.4 |
| Mono Arom | 0.0344 | 0.0051 | 6.69 | 0.4058 | 34.9 | 171.9 |
| Poly Arom | -0.0013 | 0.0054 | 0.23 | 0.0543 | 4.7 | 23.0 |
| IBP | -0.0189 | 0.0040 | 4.76 | 0.0073 | 0.6 | 3.1 |
| T10 | 0.0493 | 0.0089 | 5.51 | 0.0629 | 5.4 | 26.7 |
| T50 | -0.0263 | 0.0112 | 2.35 | 0.0256 | 2.2 | 10.9 |
| T90 | -0.0263 | 0.0105 | 2.51 | 0.0374 | 3.2 | 15.9 |
| FBP | 0.0285 | 0.0078 | 3.37 | 0.0140 | 1.2 | 5.9 |

Note: SS attributed to the property variables as computed from components of the eigenvector sum of squares.

First, let it be noted that there are many ways in which a SS can be assigned to a given predictor variable. For example, if the response is regressed on Natural Cetane only, the SS for that variable should be the same as the model SS computed in the usual manner. (Note that this is true in the orthogonal case.) However, it would seem just as reasonable to compute the Natural Cetane SS as follows:

1. Compute the model SS with all predictors included

2. Compute the model SS with Natural Cetane excluded

3. Compute the difference in model SS between 1 and 2.

Certainly this computation is legitimate in the orthogonal case and gives exactly the same estimate as in the case when only a single eigenvector is included in the model. Even though the difference method is the one usually employed in conventional analysis of variance, there would seem to be no reason why a similar computation should not be applied to *any submodel* with and without the selected predictor variable. (This approach gives the same estimate for all submodels in the orthogonal case.) If applied to a non-orthogonal model, however, the various approaches can give different estimates of the SS for Natural Cetane. Indeed, it is this fact that necessitates, and at the same time complicates, the search for the "best" submodel via a stepwise procedure. The discrepancies become more severe the more the design matrix departs from orthogonality.

Yet, in Table E.2 we have assigned what we claim to be a "unique" and allegedly "correct" way to compute the SS associated with a non-orthogonal predictor variable. How can such a claim be supported?

For purposes of argument, consider the first predictor variable in the list, Natural Cetane. The t-test and F-test for the significance estimate are very different. According to the t-test, Natural Cetane is not statistically significant, but, according to the F-test, that variable is the third most important predictor and accounts for nearly 14 percent of the model sum of squares.

Now consider all submodels containing Natural Cetane. There are 2048 such cases, and an equal number of corresponding cases with Natural Cetane excluded. The submodels range from the all-inclusive to the all-exclusive cases:

| | Nat Cet | Cet Imp | Den | Vis | Sul | Mono Arom | Poly Arom | IBP | T10 | T50 | T90 | FBP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | All-inclusive Case | | | | | | |
| With Cet | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| W/O Cet | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | | | | | All-exclusive Case | | | | | | |
| With Cet | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| W/O Cet | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Normally, the SS attributed to Natural Cetane would be computed as the difference between the model SS for A (with) and B (without) in the all-inclusive case. It would seem just as reasonable to compute the model SS with only Natural Cetane included in the model. In the orthogonal case, both approaches would yield the same number. However, quite different results are obtained in the present non-orthogonal case.

It is important to note that the case with only Natural Cetane as a variable is one of the 4095 submodels possible (excluding the null case in which no variables are in the model). Therefore, one can consider the contribution to the SS when only Natural Cetane is present as the difference between the all-exclusive case with and without Natural Cetane.

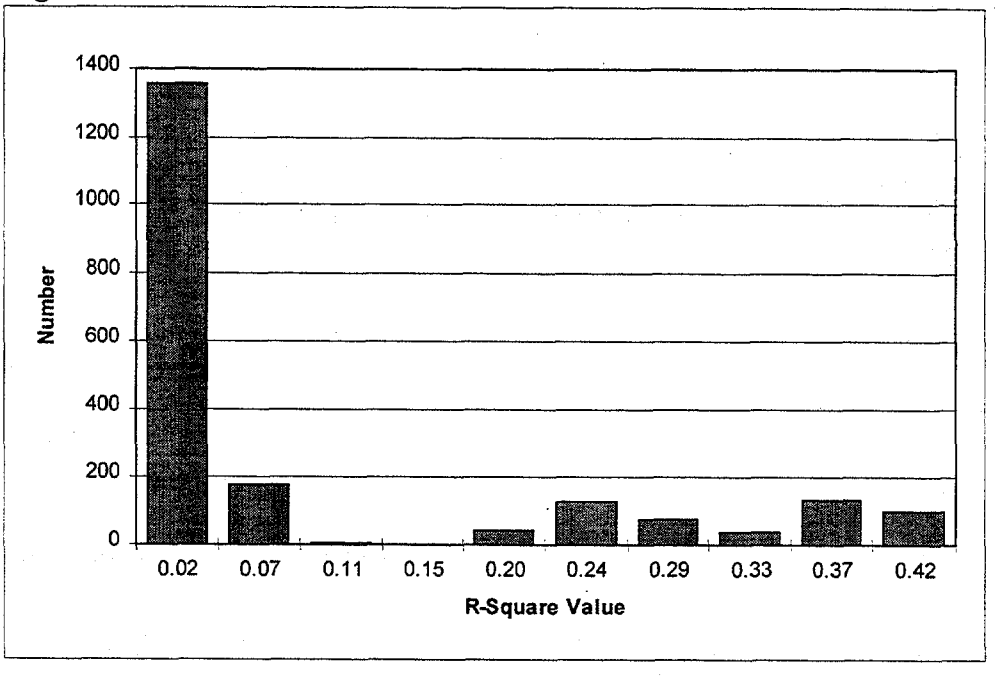
This argument generalizes to comparison of the "with" and "without" scenarios for *all submodels containing a particular variable*, any one of which could be encountered as a "way point" in a stepwise search. The difference between $R^2$ with and without Natural Cetane as a predictor range from zero to 0.4393. The maximum value is attained when the variables for inclusion and exclusion cases are as follows:

| | Nat Cet | Cet Imp | Den | Vis | Sul | Mono Arom | Poly Arom | IBP | T10 | T50 | T90 | FBP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| With Cet | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| W/O Cet | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

The $R^2$ for the case including Natural Cetane is 0.4489, while the $R^2$ for the case excluding it is 0.0096. Thus, the differential $R^2$ associated with Natural Cetane for this case in 0.4393. Note that the resulting $R^2$ is about two-thirds of the maximum $R^2$ achievable by any of the possible scenarios and is a far departure from the nonsignificant result indicated by t = 0.85 in Table E.2.

The distribution of all possible $R^2$ values for Natural Cetane is shown in Figure E.1 and is identical, except for scale differences, to the distribution of all possible sums of squares. The histograms were constructed by finding the half of the 4096 subsets that contain Natural Cetane and comparing that with the half that does not contain the variable. The difference between corresponding "with" and "without" scenarios was computed for each pair, and it is these quantities that make up the histograms.

**Figure E.1. Natural Cetane Contributions to $R^2$**




Summary statistics, based on Total SS = 1.7936, also were computed as follows:

| | SS | $R^2$ |
|---|---|---|
| Max | 0.7879 | 0.4393 |
| Min | 0.0000 | 0.0000 |
| Mean | 0.1652 | 0.0921 |

It may be more than coincidence that these mean values compare favorably with the values of 0.1614 and 0.0900 obtained by eigenvector analysis. However, we need to look further for a more rational support for the claim that, of all the ways to partition the model SS, the partitioning that derives from eigenvector analysis is the "Gold Standard," or the preferred partitioning.

First, we need to ask what the various partitionings mean? Certainly, each scenario implies a different interpretation from all others. In the case of the partitionings arising from all possible subsets of the predictor variables, the implications of the various scenarios are fairly evident, as is shown below. How these relate to our so-called "Gold Standard" partitioning, however, needs to be examined in more detail.

Suppose that *all* predictors are included in the model and that a single one of them – say, Natural Cetane – is dropped. The difference between the model sum of squares for the inclusion and exclusion cases connotes the contribution made by Natural Cetane *if it is our intent to keep all the predictors in the model.* If the one variable is dropped out, that deletion may have little effect on the quality of prediction, especially if the predictors are highly correlated, because each predictor is doing part of the work for all the others, via a process known as aliasing (see Appendix C, Theory).

Now suppose that we choose to include only half of the predictor variables in the model, one of which is Natural Cetane. When that variable is dropped out, whatever contribution it is seen to make is the contribution it would make *only if the six selected variables are the ones that would prevail in the final model.* Aliasing will be very different from that associated with the 12-variable case, if for no other reason than that there are fewer variables to "share the load." In the most degenerative case, only Natural Cetane is included in the model, and the only contribution to the model SS is that made by Natural Cetane.

In short, the contribution that any particular predictor variable makes to the model SS is determined by the model environment – that is, the number of variables included in the model and the extent to which those variables are correlated. Only in the orthogonal case are the contributions invariant with respect to the number of variables retained in the model.

How the various choices for inclusion of a predictor variable affect the computed model SS can be made more explicit by what may seem to be an artifice – namely, regressing one of the predictor variables on all the others. In subsequent discussion, we shall refer to the outcomes of such regressions as interdependency (ID) equations.

In ID regression, one of the predictor variables, rather than emissions, plays the role of response variable and is regressed on all the other predictor variables in the set. The outcome of the ID regression is fully analogous to the outcome when it is $NO_x$ or PM that is being modeled. There is a set of regression coefficients, a model SS, an error SS and an $R^2$ statistic, just as in any other regression analysis. Further, one can compute analogs

of the "calculated response" and the "residuals" representing the difference between the calculated and observed response.

For each of the 12 candidate predictors, therefore, there exist three associated "response" vectors:

1. The original predictor variable, as reported in the data base

2. The calculated response, as computed from the ID regression equation

3. The residual variable, computed as the difference between (1) and (2).

Vector (2) represents that part of the selected predictor variable that can be "explained by" the other predictors in the set. Vector (3) represents that part of the selected predictor variable "not explained by" other predictors.

We now wish to regress emissions – e.g., $\log(NO_x)$ – on one of the modified predictor variables defined above. It would seem reasonable to start with the residual, or "not explained by" vector, inasmuch as that vector should be uncontaminated by the other predictor variables. Moreover, if that is the only variable included in the model, the model SS provides an estimate of the SS that might be attributed to that predictor.

When Natural Cetane is regressed on all the other P-variables, the resulting $R^2$ is 0.8974. This number indicates that nearly 90 percent of the variance among levels of Natural Cetane is "explained" by the other predictor variables – e.g., cetane improvement, density, aromatics, etc. Similar results are found for each of the other predictor variables when they are regressed on all predictors other than the one playing the role of the "response" variable. The corresponding $R^2$ values are as follows:

Table E.3

Extent to Which a Selected Predictor Variable Can Be Explained by Other Variables in the Predictor Set

| | |
|---|---|
| Natural Cetane | 0.8974 |
| Cetane Improvement | 0.2286 |
| Density | 0.9293 |
| Viscosity | 0.9451 |
| Sulfur | 0.4821 |
| Mono Aromatics | 0.6802 |
| Poly Aromatics | 0.7062 |
| IBP | 0.4665 |
| T10 | 0.8944 |
| T50 | 0.9320 |
| T90 | 0.9230 |
| FBP | 0.8626 |

It would seem that when log emissions are regressed on, say, Natural Cetane, it should be regressed on only that part of Natural Cetane that is not explained by other predictor

variables. That part, of course, is represented by the residuals left when Natural Cetane is regressed on the other predictor variables. Let us concentrate our attention, then, on these residuals and think of them as "laundered" values, cleansed of all fragments of the other variables in the predictor set.

The following results are obtained when $\log(NO_x)$ is regressed on that residual variable only:

| Reg Coefficient Value | Std Err | t Ratio | Model SS | F Ratio | $t^2$ | $R^2$ |
|---|---|---|---|---|---|---|
| 0.0077 | 0.0150 | 0.52 | 0.0017 | 0.2655 | 0.2655 | 0.0010 |

It is interesting to compare these results with those obtained when $\log(NO_x)$ is regressed on *all* the *unmodified* predictor variables, as shown in Table E.4. It is seen that exactly the same regression coefficient is obtained for Natural Cetane but that the composite-based estimate has a lower coefficient standard error and, consequently, a higher t-ratio.

Table E.4

Comparison of Two Ways to Regress $\log(NO_x)$ on Fuel Variables

| Composite Regression on Original, Unmodified Predictors | | | One-variable-at-a-time Regression on ID Residuals | | |
|---|---|---|---|---|---|
| Reg Coefficient Value | Std Err | t Ratio | Reg Coefficient Value | Std Err | t Ratio |
| 0.0077 | 0.0091 | 0.85 | 0.0077 | 0.0150 | 0.52 |
| -0.0115 | 0.0033 | 3.47 | -0.0115 | 0.0054 | 2.12 |
| 0.0458 | 0.0109 | 4.18 | 0.0458 | 0.0179 | 2.56 |
| -0.0380 | 0.0124 | 3.06 | -0.3800 | 0.0204 | 1.86 |
| -0.0020 | 0.0040 | 0.50 | -0.0020 | 0.0067 | 0.30 |
| 0.0344 | 0.0051 | 6.69 | 0.0344 | 0.0082 | 4.17 |
| -0.0013 | 0.0054 | 0.23 | -0.0013 | 0.0089 | 0.14 |
| -0.0189 | 0.0040 | 4.76 | -0.0189 | 0.0065 | 2.92 |
| 0.0493 | 0.0089 | 5.51 | 0.0493 | 0.0145 | 3.40 |
| -0.0263 | 0.0112 | 2.35 | -0.0263 | 0.0184 | 1.43 |
| -0.0263 | 0.0105 | 2.51 | -0.0263 | 0.0173 | 1.52 |
| 0.0265 | 0.0078 | 3.37 | -0.0265 | 0.0129 | 2.06 |

The difference in indicated statistical significance comes as no surprise when it is recalled that many of the "laundered" predictor variables account for only about 10 percent of the information contained in the predictor set (see Table E.3). Despite the evident degradation in statistical precision, however, the one-variable-at-a-time regression on ID residuals offers an advantage not available via the composite regression on the unmodified predictor variables.

In a nonorthogonal regression framework, it is not possible to obtain SS estimates for individual predictor variables except by the inclusion/exclusion routine, and this approach provides only conditional partitions dependent on what variables are included in the model. Consequently, it is not possible to obtain a unique, unconditional partitioning of the model SS to serve as the basis for ranking individual fuel variables according to their relative importance in prediction.

The single-variable partitioning based on ID residuals seems to make such ranking possible and to offer exactly the interpretation desired. Consequently, we have plotted in Figure E.2 the relative SS as computed from the regression on the ID residuals. For comparison, we have plotted the so-called "Gold Standard" SS as Figure E.3. There is little similarity.

With closer examination comes further complication. Just as Natural Cetane is expressed in terms of Cetane Improvement and the other predictor variables, Natural Cetane excepted, so is Cetane Improvement expressed in terms of Natural Cetane and the other predictor variables, Cetane Improvement excepted. Clearly, there is a redundancy here that continues to hamper the testing of truly meaningful hypotheses.

What about the other part of the ID decomposition, the part that *is* explainable in terms of the other predictor variables in the set? This part, of course, stems from the "calculated" responses as computed from the ID equation. Accordingly, one can regress $\log(NO_x)$ singly on each of the ID "calculated" vectors just as was done for the ID residuals. The results are shown in Figure E.4. When compared with the "Gold Standard" SS, the similarity is unquestionable and brings us to our denouement.

We conclude that the similarity of the two plots in Figures E.3 and E.4 is no mere coincidence, but is the result of two related approaches to dealing with the correlation of predictor variables in a non-orthogonal environment. Figure E.4 plots the consequences of regression $\log(NO_x)$ on a set of "surrogate variables" designed, via ID regression, to take into account the interrelationship among predictor variables. Figure E.3 uses a similar ploy, but one that is more direct, more exact, and more elegant. It, too, yields a set of surrogate variables called eigenvectors that expresses relationships among the P-Space variables just as the "calculated" responses from the ID equations do.

The two sets of surrogate variables are just two alternative bases for spanning the space of the predictor variables, and both can be displayed in vector form. The components of the eigenvectors are the P-variable loadings, whereas the components of the ID-derived vectors are just the ID equation coefficients with zero filling the slot for the P-variable playing the role of the response variable. One might think of the ID set as a "first approximation" to the eigenvector set; the two are similar, but only the eigenvector set is exact.

**Figure E.2. SS Obtained from "Not Explained" ID Regressions**

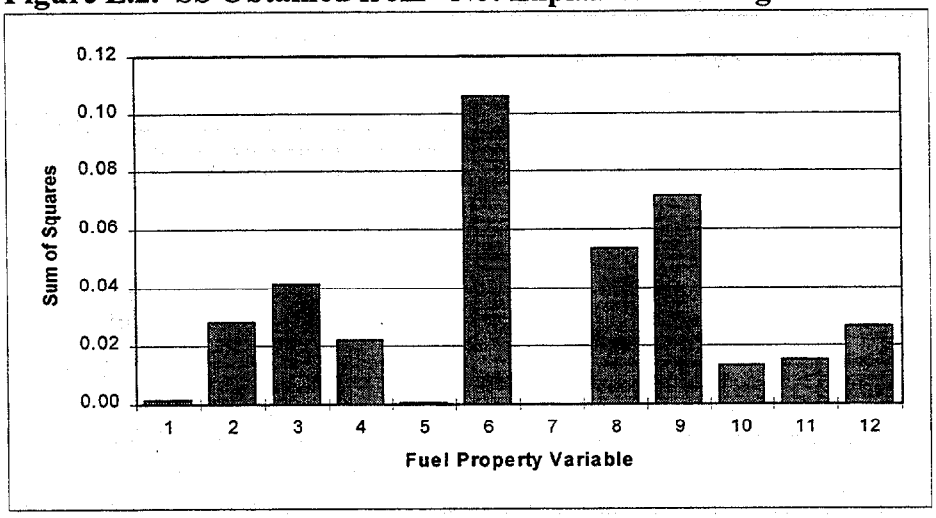



**Figure E.3. SS Obtained from Eigenvector Analysis**

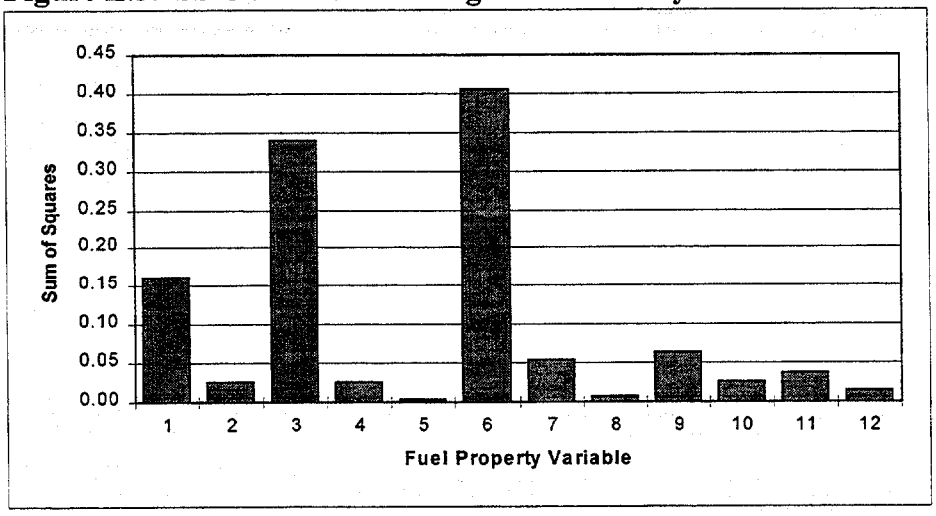
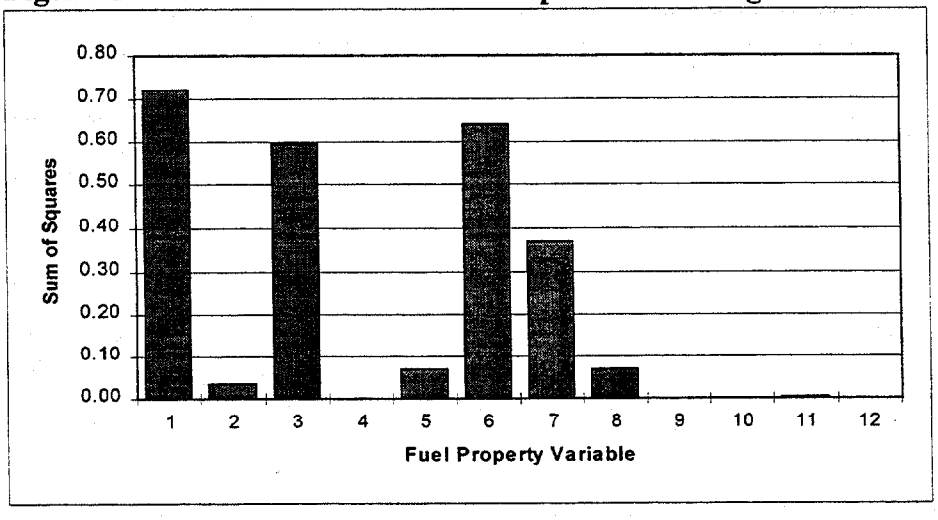



**Figure E.4. SS Obtained from the "Explained" ID Regressions**



Further support for the superiority of the Gold Standard over other proposed bases becomes evident when one compares termwise regression with regression on all terms simultaneously. Only in the E-Space of eigenvectors do the two approaches yield identical results.

On the other hand, we have seen elsewhere in this report that an emission model expressed in terms of the design-space eigenvectors has an exact equivalent in terms of the original, fuel-property variables. Similarly, the model can be expressed in terms of either the "calculated" or "residual" response from ID regression. Indeed, the regression ANOVA table shown below is applicable whether the response variable is regressed on the original predictor variables, the ID "calculated" variables, the ID residuals, or the Gold Standard eigenvectors.

| Source | SS | DF | MS |
|---|---|---|---|
| Intercept | 661.6004 | 1 | 661.6004 |
| Model | 1.1633 | 12 | 0.0969 |
| Error | 0.6303 | 267 | 0.0024 |
| Total | 663.3940 | 280 | |

Likewise, it can be shown that the predictions from all four versions of the model are point-by-point identical. All that is involved is a transformation of basis. All is well so long as the model is considered *in toto*. Only the eigenvector basis, however, provides for termwise fitting consistent with the overall model.

Two points are made clear by the above analysis. One is that there is a mechanism for isolating a SS for each predictor and testing that SS for statistical significance by means of an F-test. However, the corresponding SS derived from eigenvector analysis differs markedly from the SS derived from regressing emissions on ID equation residuals. Secondly, the analysis reinforces the perception that conventional regression analysis can lead to erroneous interpretation when applied in a nonorthogonal environment.

Lack of orthogonality and the associated non-additivity of sums of squares has far reaching consequences, some of which may not be overt to the data analyst. For example, it is well known that non-additivity is the stimulus for computing certain "sums of squares adjusted for ..." (see Appendix C, Theory). That is why it is customary to compute SS by the inclusion/exclusion procedure discussed above. However, analysts seem to have no problem in judging significance of a predictor by its t-ratio, and they place no restrictions similar to additivity on the t-ratio statistics.

In the application of the t-test, the conventional approach is simply to compute t as the ratio of the computed regression coefficient to its standard error. However, the standard error is computed by multiplying the residual mean square by the appropriate diagonal element of inverse moment matrix $(X'X)^{-1}$. Unless that matrix is diagonal, the off-diagonal terms, when multiplied by the residual mean square, provide estimates of the

covariance of pairs of predictor variables. The existence of these finite covariances clearly affects how the t-statistics are computed.

In short, the t-test is subject to the same non-additivity considerations as the F-test, but that fact is well disguised. On the contrary, the sums of squares derived from E-Space analysis are additive, because they are derived from an orthogonal decomposition of the sums of squares associated with uncorrelated eigenvectors. In the case of E-Space analysis, therefore, it is appropriate to test the significance of eigenvectors by either the t-test or the F-test. For the summed components of these eigenvectors, however, it is appropriate to pool the SS contributions made by a fuel variable to each of the eigenvectors and use that sum in a F-test. The numerator has one degree of freedom, while the denominator is the number of error degrees of freedom.

In conclusion, the eigenvector derivation of P-variable sums of squares is unique, and we believe that it is the best choice for arranging predictor variables in order of their importance to prediction. We recommend that statistical tests of significance be based on F-ratios involving those sums of squares. Further, we propose that the search for the optimum submodel be based on the outcome of these tests rather than on the more usual t-tests, which are subject to changing aliasing systems as one proceeds stepwise from one submodel to another.

# APPENDIX F: COMPARING THE RESULTS OF INDEPENDENT STUDIES

## F.1 INTRODUCTION

The scientific method requires that one evaluate research questions based on the preponderance of the evidence in repeated, independent studies. One study is inadequate, independent studies frequently disagree and, therefore, one must search for a consensus. When using conventional regression analysis, we can apply well known procedures to determine whether the differences between two studies are statistically significant – typically by testing whether the difference in regression coefficients for, say, the impact of sulfur on PM emissions, is sufficiently large that it is unlikely to arise solely by chance.

How does one make similar comparisons when using the eigenfuel methodology developed here? After all, two researchers using independent Data sets will estimate eigenfuel slates that differ from each other to at least some extent. Since these are the independent variables, can one even formulate a proper comparison between the regression coefficient estimated for eigenvector 1 in the first data set and a different eigenvector 1 in the second data set? As will be shown below, comparisons can properly and readily be made by transforming the results to a common basis, and then applying the customary tests of statistical significance.

The need to transform to a common basis may appear to be a limitation or complication of the eigenfuel methodology. In conventional regression analysis, the independent variables are fixed, and we can compare, for example, a sulfur effect in one study to a sulfur effect in another. Yet, we should remember that variables perceived as being the primary predictors in a regression model will typically be correlated and that the resulting regression coefficients will reflect that correlation. The extent of the correlations will differ in another data set and study, as will the variables included in the regression models.

Thus, a variable such as sulfur content will actually stand for a unique set of influences in each independent data set and study, and the comparison between two studies will be confounded by these differences. This problem is merely more apparent in the eigenfuel methodology and less often acknowledged in conventional regression analysis.

## F.2 METHODS FOR COMPARISON

To demonstrate the methods for comparing eigenfuel studies, we have divided the full data set into two subsets of similar size. If done by random selection, it would be no surprise to find that the subsets are similar, on average, in terms of the fuels and engines

tested and the sources from which they are drawn. We have, therefore, attempted to maximize the differences and strengthen the comparison through a systematic method of subsetting. Subset A (158 tests in total) was defined to include all data from four sources that report work by Southwest Research Institute (SWRI) for the Coordinating Research Council (CRC) engine test series, including:

- CRC VE-1 (SAE 892072)
- CRC VE-1 Phase II (SAE 902171)
- CRC VE-10 (SAE 941020 and SAE 950250)

Some fuels were used in more than one phase of the CRC testing and, where new fuels were formulated for the later testing, it is likely that similar blend stocks or procedures were used in their creation. Thus, we expect the fuels in subset A to be more uniform than we might find in a diverse collection of independent sources.

Subset B (122 tests) was formed from the 5 sources that comprise the remainder of the data set. One source used in this subset reports testing performed at SWRI and, therefore, may not be completely independent of subset A. Nevertheless, subset B testing includes 4 apparently independent testing programs and represents a more diverse, and less systematic, collection of engine/fuels testing.

We will treat subsets A and B as independent Data sets used in two studies A and B. Each study is assumed to standardize its data independently, to conduct a principal components analysis to determine the eigenvectors of its data set, and to use the 12 eigenvectors, along with controls for individual engines, in a regression analysis similar to that shown for the full data set in Section 2.3.3.

In the real world, the researchers would probably apply methods of simplification to reduce the number of eigenvectors and individual fuel property variables retained in the final emissions models. For purposes of this demonstration, however, we will retain all eigenvectors and fuel properties in the comparison. Our primary objective is to show the process for making the comparisons and not to interpret or resolve any differences that are found.

### F.2.1 The Independent Studies

The tables found at the close of this appendix summarize the basic results of studies A and B, including:

- Subset eigenvectors (F.A.1 and F.B.1)
- NOx regression results (F.A.2 and F.B.2)
- PM regression results (F.A.3 and F.B.3)

The studies report similar results in many respects:

- A relatively few eigenvectors account for most of the variance in the fuels – 7 account for 98 percent in subset A, while 8 account for the same amount in subset B.

- One eigenvector (number 2 in each subset) is most important in explaining NOx emissions, and three are enough to account for 90 percent of the total emissions effect that is related to the fuels. However, the three most important eigenvectors differ between the subsets – numbers 2, 1, and 3 in subset A, and numbers 2, 1, and 6 in subset B.

- In subset A, three eigenvectors (numbers 3, 2, and 4) are important in explaining the fuels-related effect on PM. In subset B, one eigenvector (number 1) is dominant, while vectors 2, 3, and 9 also contribute.

The two studies also generally identify a similar set of fuel properties as being important to emissions as summarized in Table F.1. The set includes cetane rating, density, sulfur content, mono- and poly-aromatics content and, in some instances, one or more points on the distillation curve.

**Table F.1: Important Fuel Properties to Emissions Explanation**

| | **Subset A** | | **Subset B** | |
|---|---|---|---|---|
| | **NOx** | **PM** | **NOx** | **PM** |
| Natural Cetane | Yes | Yes | Yes | No |
| Cetane Improvement | No | Yes | Yes | No |
| Viscosity | No | No | No | Yes |
| Density | Yes | Yes | Yes | Yes |
| Sulfur Content | No | Yes | Yes | Yes |
| Mono-Arom | Yes | Yes | Yes | Yes |
| Poly-Arom | Yes | Yes | No | Yes |
| Distillation Curve | T90 | T10 | No | FBP<br>T10<br>T50<br>T90 |

At the same time, there are easily apparent differences. The eigenvectors often have dissimilar internal structures and, where structures are similar, the component weights are different. Further, there are differences in which eigenvectors and fuel properties are found to be important to emissions.

### F.2.2 Comparing the Studies

We have several choices in seeking a common basis for comparison. First, researchers could transform their results into emissions coefficients and sum-of-squares partitioning by fuel property and conduct the comparison in "P-space". This comparison could always be made if researchers were to adopt the convention of publishing the P-space transformations of their results. It will, however, lose information linked to the eigenvectors, and not their individual components.

To retain this information, one could chose to transform the results of study A into the eigenvector space of study B (or vice versa), or one could transform both studies into the eigenvector space of a third, reference study. We will adopt the latter approach and use the eigenvectors of the full data set as a reference space for the purposes of this demonstration. However, the reference space could as well be another, completely independent study.

#### F.2.2.1 Comparing the Eigenvectors

As a first step, let us express the eigenvectors of each subset in terms of the reference eigenvectors from the full data set. First, we define the following matrices:

fuel_pc_ref = reference eigenvectors

fuel_pcA = eigenvectors from subset A

fuel_pcB = eigenvectors from subset B

where eigenvectors form the columns and the fuel properties form the rows. The transpose of any such matrix places the eigenvectors in the form of a row-oriented fuels data set. This can then be expressed in terms of an eigenvector basis by applying the matrix operation (fuel matrix) * (eigenvector matrix). Hence, subset A eigenfuels are expressed in terms of the reference eigenvectors by the operation

fuel_pcA' * fuel_pc_ref,

and subset B eigenfuels are similarly expressed in reference terms by:

fuel_pcB' * fuel_pc_ref.

Having done so, each subset eigenvector is represented as a weighted sum of the reference eigenvectors. The coefficients of this representation contain the information needed to map the results of each study into the reference space, where they can be compared on a common basis. The *squares* of the weighting coefficients are convenient measures of the relationship between subset and reference eigenvectors because they state the proportion of the variation in the subset vector that maps into each of the reference vectors. Thus, Table F.2 summarizes the squared coefficients for subset A vectors, while Table F.3 summarizes the same for subset B. The largest squared coefficients have been highlighted.

Perfect correspondence between the subset and reference eigenvectors would result in a diagonal, unit matrix, and we see, in general, that the diagonal elements for the two subsets do not depart greatly from it. It is also true that the subset A eigenfuels tend to map into 2 or 3 different reference vectors in most cases. Subset B eigenfuels map more frequently into a single, corresponding reference vector, although B4 and B5 map into a mixture of 3 reference vectors and B7, B11, and B12 into two. This is consistent with the notion that fuels in subset A share an experimental design selected for the SWRI/CEC testing, which was possibly dissimilar from the existing database of fuels overall, while fuels in subset B are more diverse and reflective of fuels in general.

### F.2.2.2 Transforming to the Reference Space

Let us now transform the results of study A into equivalent results in the reference eigenvector space. The transformation matrix is defined as follows:

```
PCA_transform=fuel_pc' *fuel_pc_ref;
PCA_transform=PCA_transform'
```

Then, the NOx regression coefficients and standard errors are transformed as follows:

```
coeffs=NOx_coeffs(7:18,1)
       -0.0089
        0.0211
       -0.0120
        0.0021
       -0.0060
       -0.0108
       -0.0015
       -0.0214
        0.0279
        0.0245
        0.0199
       -0.0131
```

**Table F.2. Subset A Eigenvectors in Terms of Reference Vectors (squared coefficients)**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref EV 1 | **0.805** | **0.157** | 0.014 | 0.003 | 0.001 | 0.012 | 0.000 | 0.005 | 0.003 | 0.000 | 0.000 | 0.000 |
| Ref EV 2 | **0.187** | **0.721** | 0.059 | 0.003 | 0.002 | 0.009 | 0.005 | 0.005 | 0.006 | 0.003 | 0.000 | 0.001 |
| Ref EV 3 | 0.001 | 0.007 | **0.347** | **0.616** | 0.001 | 0.001 | 0.000 | 0.000 | 0.020 | 0.000 | 0.004 | 0.003 |
| Ref EV 4 | 0.000 | 0.071 | **0.561** | **0.335** | 0.006 | 0.001 | 0.012 | 0.000 | 0.005 | 0.001 | 0.008 | 0.000 |
| Ref EV 5 | 0.002 | 0.021 | 0.006 | 0.000 | **0.748** | **0.156** | 0.021 | 0.034 | 0.012 | 0.000 | 0.000 | 0.000 |
| Ref EV 6 | 0.004 | 0.005 | 0.002 | 0.000 | **0.207** | **0.409** | **0.292** | 0.036 | 0.008 | 0.000 | 0.036 | 0.002 |
| Ref EV 7 | 0.000 | 0.013 | 0.009 | 0.015 | 0.031 | **0.214** | **0.550** | 0.067 | 0.039 | 0.001 | 0.061 | 0.000 |
| Ref EV 8 | 0.000 | 0.002 | 0.000 | 0.023 | 0.002 | 0.116 | 0.052 | 0.070 | **0.631** | 0.007 | 0.086 | 0.012 |
| Ref EV 9 | 0.000 | 0.000 | 0.000 | 0.003 | 0.001 | 0.061 | 0.002 | **0.754** | **0.158** | 0.006 | 0.000 | 0.013 |
| Ref EV 10 | 0.001 | 0.003 | 0.000 | 0.000 | 0.000 | 0.014 | 0.061 | 0.018 | 0.040 | 0.014 | **0.645** | **0.204** |
| Ref EV 11 | 0.000 | 0.001 | 0.001 | 0.001 | 0.000 | 0.007 | 0.001 | 0.008 | 0.038 | **0.680** | 0.077 | **0.186** |
| Ref EV 12 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.004 | 0.002 | 0.040 | **0.288** | 0.084 | **0.579** |

**Table F.3. Subset B Eigenvectors in Terms of Reference Vectors (squared coefficients)**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ref EV 1 | **0.921** | 0.041 | 0.016 | 0.008 | 0.003 | 0.010 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| Ref EV 2 | 0.048 | **0.752** | 0.050 | 0.088 | 0.000 | 0.038 | 0.007 | 0.002 | 0.009 | 0.001 | 0.000 | 0.004 |
| Ref EV 3 | 0.011 | 0.052 | **0.845** | 0.000 | 0.017 | 0.043 | 0.019 | 0.005 | 0.000 | 0.000 | 0.001 | 0.007 |
| Ref EV 4 | 0.000 | 0.052 | 0.008 | **0.286** | **0.554** | 0.093 | 0.000 | 0.002 | 0.001 | 0.002 | 0.001 | 0.000 |
| Ref EV 5 | 0.010 | 0.005 | 0.058 | 0.113 | 0.010 | **0.665** | 0.125 | 0.002 | 0.000 | 0.010 | 0.000 | 0.000 |
| Ref EV 6 | 0.001 | 0.020 | 0.007 | **0.323** | **0.177** | 0.019 | **0.408** | 0.015 | 0.021 | 0.003 | 0.006 | 0.001 |
| Ref EV 7 | 0.006 | 0.059 | 0.000 | **0.146** | **0.224** | 0.112 | **0.412** | 0.020 | 0.000 | 0.000 | 0.020 | 0.000 |
| Ref EV 8 | 0.001 | 0.013 | 0.007 | 0.020 | 0.006 | 0.005 | 0.000 | **0.921** | 0.005 | 0.019 | 0.002 | 0.001 |
| Ref EV 9 | 0.000 | 0.001 | 0.001 | 0.002 | 0.000 | 0.008 | 0.004 | 0.027 | 0.033 | **0.884** | 0.000 | 0.041 |
| Ref EV 10 | 0.000 | 0.004 | 0.001 | 0.015 | 0.000 | 0.001 | 0.009 | 0.004 | **0.845** | 0.036 | 0.083 | 0.001 |
| Ref EV 11 | 0.000 | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0.015 | 0.000 | 0.044 | 0.033 | **0.497** | **0.405** |

```
new_coeffs=PCA_transform*coeffs
       0.0000
       0.0235
       0.0059
       0.0058
      -0.0097
       0.0018
      -0.0036
      -0.0232
      -0.0095
      -0.0178
       0.0385
       0.0078

 stderrs=NOx_coeffs(7:18,2)

       0.0008
       0.0009
       0.0016
       0.0023
       0.0019
       0.0027
       0.0040
       0.0060
       0.0073
       0.0118
       0.0126
       0.0172

new_stderrs=sqrt( (PCA_transform.^2) * (stderrs.^2) )

       0.0011
       0.0015
       0.0026
       0.0023
       0.0025
       0.0040
       0.0050
       0.0075
       0.0064
       0.0130
       0.0128
       0.0150
```

The same procedures would be used to transform PM coefficients and standard errors.

When all regressors are independent, as in a regression model containing only eigenvector variables, the transformed coefficients and standard errors will be identical to those that would be estimated in a regression using the reference-space eigenvectors. (See Appendix D, section D.3). In the emission models considered here, the eigenvector variables are combined with dummy variables representing individual engine effects. This slate of regressors is not completely independent because of non-zero correlations that exist between dummy variables and the eigenvectors. The coefficients still transform

properly in this circumstance, but the transformed standard errors are only approximate. It is thought that the approximate standard errors are acceptable for purposes of comparing studies.

### F.2.2.3 Comparing the Study Results

The procedures shown above have been applied to the NOx and PM regressions from each subset to transform the results to the common basis of the reference eigenvectors. The subset regressions are compared directly in the reference space, and are then transformed to the P-Space of the individual fuel properties for further, but brief, comparison. Highlights are discussed below.
Several comparisons may be appropriate to make when studies are placed on a common basis in the reference space. We may be interested to know which of the effects are statistically significant, whether the effects are of similar size, and how much they contribute to the fuels-related SS. The first two questions taken together relate to whether the studies concur on the identity and magnitude of the causal factors and are the ones on which we will focus.

The third question, related to the SS partitioning, adds information on the relative importance of the variables in each study, but is a less direct means of comparing the studies. The SS contributions of any explanatory variable will depend on the estimated magnitude of the effect and the variation of the variable in the data set. Two studies may agree on the magnitude of an effect, but one study may show a much smaller contribution to the SS because that variable was varied much less in its database. We will consider, but not emphasize, this comparison.

Comparing the magnitude of effects estimated in the two studies can be treated by testing whether the observed differences in magnitude are statistically significant. This is the Behrens-Fisher problem of testing the hypothesis that two means are equal in the circumstance where the population variances (equal to the square of the standard errors) are unknown and may be unequal. Several treatments have been proposed for this problem, although its solution remains controversial. For purposes of this paper, the test will be made using the Smith-Satterthwaite procedure[1], which is labeled the "SmSw" test in the tables. This procedure is said to "perform well when variances are unequal, but it yields results that are virtually equivalent to those obtained with the pooled t-test when variances are equal." Thus, it can be applied without first testing whether the variances can be pooled.

---

[1] *Introduction to Probability and Statistics: Principals and Applications for Engineering and the Computer Sciences*, Milton and Arnold, Irwin McGraw Hill, 1995. pp. 351-353.

Table F.4 presents the $NO_x$ regression coefficients transformed to the space of the reference eigenvectors. (This tabulation, like its companions, omits the dummy variable effects estimated for the individual engines.) The t-ratio and the percentage contribution to the SS are also given. The center column identifies whether the difference in coefficient values between the Data sets is statistically significant at the 0.05 level. We see the following major points of comparison in the table:

- Both subsets agree that reference vector number 2, which was previously identified as representing high-aromatic cracked stocked in diesel fuel blends, is the dominant effect on NOx emissions. It is highly significant and accounts for approximately 80 percent of the fuels-related SS in both subsets. The subsets concur that reducing the proportion of reference vector 2 in a diesel fuel blend will reduce $NO_x$ emissions.

- The subsets disagree, however, on the magnitude of the effect for reference vector 2. Subset B estimates a magnitude that is twice as large as in Subset A, and the difference is found to be significant at better than the 0.05 level (SmSw test ratio of 7.0). This difference could be related to the model year of the vehicles involved, since some investigators believe the effect of fuels on emissions is lessened in vehicles certified to tighter standards. We have not attempted to identify the cause for this difference.

**Table F.4. $NO_x$ Regression Coefficients (excluding engine effects)**

| | **Subset A Regression** | | | SmSw test | **Subset B Regression** | | |
|---|---|---|---|---|---|---|---|
| Reference Eigenvector | Coefficient | t-ratio | SS Part percent | | Coefficient | t-ratio | SS Part percent |
| 1 | 0.0000 | 0.04 | 0.00 | Dif | 0.0074 | 4.97 | 3.56 |
| 2 | 0.0235 | 15.97 | 78.98 | Dif | 0.0472 | 15.54 | 82.38 |
| 3 | 0.0059 | 2.25 | 2.94 | – | 0.0031 | 0.89 | 0.21 |
| 4 | 0.0058 | 2.53 | 2.16 | – | 0.0133 | 3.99 | 2.98 |
| 5 | -0.0097 | -3.93 | 3.55 | Dif | -0.0282 | -7.73 | 7.75 |
| 6 | 0.0018 | 0.44 | 0.10 | – | 0.0073 | 1.65 | 0.43 |
| 7 | -0.0036 | -0.71 | 0.27 | – | 0.0073 | 1.58 | 0.30 |
| 8 | -0.0232 | -3.10 | 6.85 | Dif | 0.0153 | 2.25 | 0.77 |
| 9 | -0.0095 | -1.49 | 0.70 | – | 0.0144 | 0.96 | 0.42 |
| 10 | -0.0178 | -1.37 | 1.43 | – | -0.0007 | -0.06 | 0.00 |
| 11 | 0.0385 | 3.01 | 2.93 | Dif | -0.0452 | -2.38 | 1.05 |
| 12 | 0.0078 | 0.52 | 0.09 | – | 0.0206 | 1.08 | 0.16 |

- The subsets agree that vectors 4, 5, 8, and 11 make meaningful contributions to explaining the 20 percent of the SS remaining after reference vector 2 is accounted for. They differ, however, in the order of importance in partitioning the SS on the directional impact on emissions for vectors 8 and 11. Subset A would add vector 3 to the list of contributors, while subset B would add vector 1. Where the comparisons are meaningful, it is generally true that subset B estimates a larger impact on NOx emissions.

Thus, the two subsets concur on the overriding importance of reducing eigenvector 2 as a means of reducing NOx emissions, although they disagree on the magnitude of the effect, and they also disagree to some extent on the smaller influences.

We see much less apparent agreement when the results for NOx emissions are stated in terms of the individual fuel properties, as shown in Table F.5. Here, we will examine differences in the size of coefficients and the SS partitioning, but we will not attempt to apply the Smith-Satterthwaite test to the differences in the coefficients.

**Table F.5. NOx Regression Coefficients for Fuel Properties**

| | **Subset A Regression** | | | SmSw test | **Subset B Regression** | | |
|---|---|---|---|---|---|---|---|
| Fuel Property | Coefficient | t-ratio | SS Part percent | | Coefficient | t-ratio | SS Part percent |
| NatCetane | 0.0079 | 1.07 | 11.27 | n/a | -0.0288 | -2.66 | 31.53 |
| CetImprv | -0.0069 | -3.02 | 1.34 | n/a | -0.0186 | -5.04 | 2.07 |
| Density | 0.0048 | 0.57 | 14.35 | n/a | 0.0583 | 4.23 | 36.08 |
| Viscosity | 0.0179 | 1.36 | 0.12 | n/a | -0.0221 | -1.61 | 0.30 |
| Sulfur | -0.0057 | -1.54 | 0.63 | n/a | 0.0034 | 1.09 | 1.02 |
| MonoArom | 0.0309 | 8.36 | 33.75 | n/a | 0.0175 | 2.55 | 23.65 |
| PolyArom | 0.0292 | 4.76 | 29.18 | n/a | -0.0078 | -1.54 | 5.14 |
| IBP | -0.0057 | -1.41 | 0.56 | n/a | -0.0038 | -0.80 | 0.04 |
| T10 | -0.0049 | -0.59 | 0.00 | n/a | 0.0025 | 0.18 | 0.01 |
| T50 | -0.0022 | -0.21 | 0.13 | n/a | -0.0091 | -0.69 | 0.00 |
| T90 | -0.0294 | -2.92 | 6.38 | n/a | 0.0239 | 1.90 | 0.06 |
| FBP | -0.0003 | -0.04 | 2.28 | n/a | -0.0128 | -1.27 | 0.10 |

Subset A shows mono- and poly-aromatics content and T90 to have the largest effects, based on the size of the regression coefficient, with mono-aromatics content most important when viewed in terms of the SS partitioning. Subset B agrees on the magnitude of the effect for mono-aromatics content, but disagrees on the algebraic sign and magnitude of the other two effects.

Conversely, subset B shows density to have the largest effect, followed by natural cetane and T90, with density and natural cetane being most important in terms of SS. Subset A disagrees on the magnitude of the density coefficient, given the standard errors, and both the algebraic signs and magnitudes of the other two effects.

Table F.6 presents the PM regression coefficients transformed to the reference eigenvector space. In subset A, reference eigenvectors 2 and 3 are highly significant and account for 88 percent of the fuels-relate SS. Reference vectors 1, 9, and 10 also contribute, although vector 10 fails the significance test. All other reference vectors both fail the test of significance and make negligible contributions to the SS.

**Table F.6. PM Regression Coefficients (excluding engine effects)**

| | **Subset A Regression** | | | SmSw test | **Subset B Regression** | | |
|---|---|---|---|---|---|---|---|
| Reference Eigenvector | Coefficient | t-ratio | SS Part percent | | Coefficient | t-ratio | SS Part percent |
| 1 | 0.0142 | 2.86 | 4.20 | Dif | 0.0294 | 10.31 | 31.80 |
| 2 | 0.0616 | 9.36 | 45.15 | – | 0.0467 | 8.05 | 46.06 |
| 3 | 0.0778 | 6.67 | 43.06 | Dif | 0.0227 | 3.48 | 6.50 |
| 4 | 0.0014 | 0.14 | 0.01 | – | -0.0200 | -3.14 | 3.84 |
| 5 | 0.0131 | 1.19 | 0.54 | – | -0.0053 | -0.76 | 0.15 |
| 6 | -0.0015 | -0.08 | 0.01 | – | -0.0177 | -2.09 | 1.44 |
| 7 | -0.0046 | -0.21 | 0.04 | – | 0.0172 | 1.94 | 0.94 |
| 8 | -0.0267 | -0.80 | 0.76 | – | 0.0083 | 0.64 | 0.13 |
| 9 | -0.0640 | -2.25 | 2.64 | – | -0.0285 | -1.00 | 0.93 |
| 10 | 0.0928 | 1.60 | 3.27 | Dif | -0.1038 | -4.37 | 7.24 |
| 11 | 0.0421 | 0.74 | 0.29 | – | 0.0533 | 1.47 | 0.83 |
| 12 | 0.0195 | 0.29 | 0.05 | – | 0.0251 | 0.69 | 0.13 |

In subset B, reference vector 2 is, again, most important, but is accompanied by vector 1 (not 3) in making the largest SS contribution. Reference vectors 10, 3, 4, and 6 also contribute, while the remaining vectors fail on significance and SS contributions. The major points of comparison for PM are:

- Reference vector number 2 is the most important contributor to emissions. In addition, the subsets agree on the magnitude of the effect on PM emissions, with the difference between the coefficients found not to be significant at the 0.05 level.

- Vectors 1 and 3 are also important, as was found in the analysis of the full data set, but the subsets disagree on the relative importance. The magnitude of the effect for reference vector 1 differs (weakly) between the subsets (t-ratio of 2.65), while the difference for vector 3 is both larger and more strongly significant (t-ratio of 4.12).

- Both subsets suggest that reference vector 10 may have an impact on PM emissions, but disagree on relatively small effects among vectors 4, 5, and 9.

Transformed to P-space (see Table F.7), subset A indicates that sulfur content has the largest effect on PM emissions, measured by the SS explanation, followed by poly-aromatics content, natural cetane, density, cetane improvement, and mono-aromatics content. Subset B indicates that density and poly-aromatics content have the largest effects, again measured by SS explanation, followed by mono-aromatics content, sulfur content and FBP. Hence, there is substantial discord among the two subsets when viewed in terms of the effects attributed to fuel properties. The subsets also disagree on the algebraic sign of the emissions effect for 5 of the 12 fuel properties, although they agree on the magnitude of the individual effects for all but one fuel property (FBP) within the bound determined by the standard errors,.

The finding of conflicting results when emissions effects are compared in P-space has been one of the recurring themes in the study of fuel effects on emissions. The two hypothetical studies created by subsetting the database tend to show more agreement on the fuel factors affecting emissions when they are described as eigenfuels, rather than individual fuel properties. This is thought to reflect, as seen elsewhere in this report, the reality that fuel properties tend to act in concert through the components present in the diesel fuel blends. Fuels formed of similar components in two samples may yield similar test results when viewed in eigenfuel terms. Yet, to the extent that the components are combined in different ways and proportions in the samples, the individual properties will be varied by different amounts, and the correlations among properties may lead to apparently different impacts on emissions.

| Table F.7. PM Regression Coefficients for Fuel Properties | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Subset A Regression | | | SmSw test | Subset B Regression | | |
| Fuel Property | Coefficient | t-ratio | SS Part percent | | Coefficient | t-ratio | SS Part percent |
| NatCetane | -0.0478 | -1.45 | 9.11 | n/a | 0.0400 | 1.93 | 1.54 |
| CetImprv | -0.0276 | -2.70 | 4.19 | n/a | -0.0039 | -0.56 | 0.02 |
| Density | -0.0339 | -0.90 | 5.51 | n/a | 0.0555 | 2.11 | 24.69 |
| Viscosity | 0.0655 | 1.12 | 1.21 | n/a | 0.0168 | 0.64 | 1.09 |
| Sulfur | 0.0596 | 3.62 | 29.55 | n/a | 0.0237 | 3.94 | 11.08 |
| MonoArom | 0.0252 | 1.52 | 4.01 | n/a | 0.0464 | 3.55 | 17.77 |
| PolyArom | 0.0938 | 3.43 | 41.20 | n/a | 0.0432 | 4.45 | 22.44 |
| IBP | 0.0057 | 0.32 | 0.29 | n/a | -0.0121 | -1.34 | 0.12 |
| T10 | 0.0173 | 0.47 | 2.46 | n/a | 0.0081 | 0.31 | 2.15 |
| T50 | -0.0304 | -0.65 | 0.00 | n/a | -0.0261 | -1.04 | 0.68 |
| T90 | 0.0276 | 0.61 | 0.19 | n/a | -0.0582 | -2.42 | 0.61 |
| FBP | -0.0481 | -1.45 | 2.28 | n/a | 0.0767 | 3.98 | 17.79 |

**Table F.A.1: Subset A Eigenvectors**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.283 | -0.385 | 0.197 | -0.238 | 0.012 | -0.273 | 0.048 | -0.053 | 0.618 | -0.298 | 0.353 | -0.034 |
| CetImprv | -0.079 | 0.142 | 0.386 | 0.333 | 0.753 | -0.337 | 0.105 | 0.123 | -0.064 | -0.055 | 0.012 | 0.006 |
| Density | 0.049 | 0.500 | -0.321 | 0.087 | 0.044 | 0.254 | 0.131 | 0.334 | 0.059 | -0.444 | 0.489 | -0.039 |
| Viscosity | 0.467 | 0.013 | -0.118 | 0.061 | 0.204 | 0.129 | -0.159 | -0.118 | -0.050 | 0.363 | 0.169 | -0.709 |
| Sulfur | -0.074 | 0.031 | -0.237 | -0.700 | 0.293 | -0.271 | -0.399 | 0.260 | -0.223 | 0.049 | 0.065 | 0.069 |
| MonoArom | -0.064 | 0.464 | 0.078 | 0.256 | -0.321 | -0.444 | -0.566 | -0.135 | 0.224 | 0.097 | 0.102 | 0.010 |
| PolyArom | -0.104 | 0.462 | -0.198 | -0.326 | 0.093 | -0.183 | 0.492 | -0.292 | 0.362 | 0.156 | -0.292 | -0.139 |
| IBP | 0.207 | -0.214 | -0.477 | 0.293 | -0.167 | -0.543 | 0.243 | 0.423 | -0.060 | 0.057 | -0.169 | -0.041 |
| T10 | 0.409 | 0.005 | -0.347 | 0.099 | 0.204 | -0.090 | -0.082 | -0.615 | -0.230 | -0.288 | -0.026 | 0.362 |
| T50 | 0.452 | 0.143 | 0.011 | 0.031 | 0.153 | 0.285 | -0.118 | 0.309 | 0.368 | 0.379 | -0.191 | 0.495 |
| T90 | 0.387 | 0.206 | 0.321 | -0.152 | -0.149 | -0.003 | -0.092 | 0.180 | -0.123 | -0.496 | -0.549 | -0.237 |
| FBP | 0.329 | 0.207 | 0.378 | -0.197 | -0.278 | -0.206 | 0.367 | -0.004 | -0.412 | 0.258 | 0.374 | 0.192 |
| Eigenvalues | 4.187 | 3.009 | 1.676 | 1.519 | 0.785 | 0.410 | 0.176 | 0.096 | 0.078 | 0.039 | 0.017 | 0.009 |
| Pct Variance | 34.893 | 25.075 | 13.964 | 12.656 | 6.542 | 3.416 | 1.466 | 0.797 | 0.646 | 0.326 | 0.144 | 0.074 |
| Cumulative Pct | 34.893 | 59.968 | 73.932 | 86.589 | 93.130 | 96.546 | 98.012 | 98.809 | 99.455 | 99.782 | 99.926 | 100.000 |

**Table F.A.2: Subset A NOx Regression**

Ln(NOx) Emissions

R2 = 0.8728

| Coeff | Std Err | t-ratio | |
|---|---|---|---|
| | Intercept | | |
| 1.5397 | 0.0047 | 326.0628 | |
| | Engine Effects | | |
| 0.0213 | 0.0055 | 3.8729 | |
| -0.0311 | 0.0054 | 5.7813 | |
| 0.0189 | 0.0077 | 2.4617 | |
| 0.0483 | 0.0097 | 4.9769 | |
| 0.0215 | 0.0099 | 2.1747 | Fuels SS |
| | Eigenfuel Effects | | (percent) |
| -0.0089 | 0.0008 | 10.5253 | 15.5 |
| 0.0211 | 0.0009 | 22.7107 | 62.9 |
| -0.0120 | 0.0016 | 7.5935 | 11.3 |
| 0.0021 | 0.0023 | 0.8835 | 0.3 |
| -0.0060 | 0.0019 | 3.1693 | 1.3 |
| -0.0108 | 0.0027 | 4.0438 | 2.2 |
| -0.0015 | 0.0040 | 0.3890 | 0.0 |
| -0.0214 | 0.0060 | 3.5905 | 2.1 |
| 0.0279 | 0.0073 | 3.8041 | 2.8 |
| 0.0245 | 0.0118 | 2.0725 | 1.1 |
| 0.0199 | 0.0126 | 1.5768 | 0.3 |
| -0.0131 | 0.0172 | 0.7606 | 0.7 |

**Table F.A.3: Subset A PM Regression**

Ln(PM) Emissions

R2 = 0.9783

| Coeff | Std Err | t-ratio | |
|---|---|---|---|
| | Intercept | | |
| -0.7376 | 0.0211 | 34.9909 | |
| | Engine Effects | | |
| -0.6060 | 0.0245 | 24.7369 | |
| -0.5933 | 0.0240 | 24.7358 | |
| -0.9743 | 0.0342 | 28.5010 | |
| -1.4283 | 0.0433 | 32.9788 | |
| -1.6054 | 0.0442 | 36.3219 | Fuels SS |
| | Eigenfuel Effects | | (Percent) |
| -0.0088 | 0.0038 | 2.3487 | 1.7 |
| 0.0456 | 0.0041 | 10.9961 | 32.1 |
| -0.0660 | 0.0071 | 9.3598 | 37.5 |
| -0.0549 | 0.0104 | 5.2863 | 23.5 |
| 0.0158 | 0.0085 | 1.8674 | 1.0 |
| -0.0258 | 0.0119 | 2.1751 | 1.4 |
| -0.0286 | 0.0177 | 1.6123 | 0.7 |
| -0.0477 | 0.0266 | 1.7957 | 1.1 |
| -0.0059 | 0.0327 | 0.1811 | 0.1 |
| 0.0324 | 0.0527 | 0.6138 | 0.2 |
| -0.0724 | 0.0563 | 1.2854 | 0.5 |
| -0.0771 | 0.0767 | 1.0057 | 0.3 |

**Table F.B.1: Subset B Eigenvectors**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.056 | -0.645 | 0.017 | -0.034 | 0.172 | -0.115 | 0.079 | -0.457 | 0.276 | -0.100 | 0.427 | -0.232 |
| CetImprv | -0.079 | 0.235 | 0.685 | 0.402 | 0.172 | -0.403 | -0.261 | -0.183 | 0.066 | -0.094 | -0.023 | -0.009 |
| Density | -0.347 | 0.324 | -0.036 | -0.148 | -0.116 | 0.142 | -0.249 | 0.107 | 0.434 | 0.181 | 0.194 | -0.621 |
| Viscosity | -0.385 | -0.054 | 0.131 | -0.255 | -0.037 | -0.009 | -0.269 | 0.122 | -0.564 | -0.023 | 0.579 | 0.154 |
| Sulfur | -0.197 | 0.168 | -0.499 | -0.011 | 0.607 | -0.549 | 0.051 | 0.105 | -0.015 | 0.011 | 0.014 | -0.014 |
| MonoArom | -0.212 | 0.496 | 0.014 | -0.120 | 0.102 | 0.260 | 0.461 | -0.572 | 0.017 | -0.132 | 0.158 | 0.185 |
| PolyArom | -0.169 | 0.002 | -0.450 | 0.730 | -0.166 | 0.211 | -0.295 | -0.175 | -0.039 | -0.131 | 0.136 | 0.092 |
| IBP | -0.292 | -0.061 | -0.048 | 0.112 | -0.625 | -0.526 | 0.418 | -0.030 | -0.004 | 0.231 | 0.017 | 0.001 |
| T10 | -0.379 | -0.132 | -0.051 | -0.267 | -0.169 | -0.091 | -0.163 | -0.043 | 0.114 | -0.739 | -0.375 | 0.026 |
| T50 | -0.370 | -0.213 | -0.004 | -0.164 | 0.098 | 0.080 | -0.316 | -0.332 | 0.113 | 0.557 | -0.362 | 0.333 |
| T90 | -0.348 | -0.225 | 0.141 | 0.231 | 0.237 | 0.238 | 0.317 | 0.024 | -0.441 | 0.047 | -0.318 | -0.493 |
| FBP | -0.349 | -0.172 | 0.189 | 0.196 | 0.186 | 0.208 | 0.296 | 0.499 | 0.436 | -0.040 | 0.166 | 0.374 |
| Eigenvalues | 5.621 | 2.111 | 1.173 | 0.966 | 0.734 | 0.519 | 0.365 | 0.277 | 0.109 | 0.081 | 0.026 | 0.019 |
| Pct Variance | 46.845 | 17.592 | 9.771 | 8.048 | 6.113 | 4.325 | 3.041 | 2.309 | 0.907 | 0.674 | 0.219 | 0.157 |
| Cumulative Pct | 46.845 | 64.437 | 74.208 | 82.256 | 88.369 | 92.694 | 95.735 | 98.044 | 98.951 | 99.624 | 99.843 | 100.000 |

**Table F.B.2: Subset B NOx Regression**

Ln(NOx) Emissions

R2 = 0.9544

| Coeff | Std Err | t-ratio | |
|---|---|---|---|
| -------- Intercept ------- | | | |
| 1.5471 | 0.0100 | 154.3597 | |
| ----- Engine Effects ----- | | | |
| 0.0006 | 0.0127 | 0.0508 | |
| -0.1617 | 0.0187 | 8.6526 | |
| -0.1594 | 0.0179 | 8.9175 | |
| 0.0099 | 0.0160 | 0.6193 | Fuels SS |
| ---- Eigenfuel Effects --- | | | (Percent) |
| -0.0153 | 0.0013 | 12.0045 | 21.9 |
| 0.0422 | 0.0024 | 17.4835 | 62.8 |
| -0.0151 | 0.0030 | 5.1215 | 4.5 |
| -0.0139 | 0.0033 | 4.2184 | 3.1 |
| -0.0046 | 0.0032 | 1.4407 | 0.3 |
| 0.0244 | 0.0031 | 7.7331 | 5.1 |
| 0.0092 | 0.0042 | 2.1999 | 0.5 |
| 0.0091 | 0.0066 | 1.3838 | 0.4 |
| 0.0123 | 0.0115 | 1.0653 | 0.3 |
| 0.0083 | 0.0152 | 0.5464 | 0.1 |
| -0.0190 | 0.0201 | 0.9479 | 0.2 |
| -0.0498 | 0.0190 | 2.6192 | 0.8 |

**Table F.B.3: Subset B PM Regression**

Ln(PM) Emissions

R2 = 0.9918

| Coeff | Std Err | t-ratio | |
|---|---|---|---|
| -------- Intercept ------- | | | |
| -1.5968 | 0.0192 | 83.3754 | |
| ----- Engine Effects ----- | | | |
| -0.9304 | 0.0243 | 38.3114 | |
| -0.1644 | 0.0357 | 4.6044 | |
| -0.8385 | 0.0341 | 24.5574 | |
| -0.0282 | 0.0305 | 0.9229 | Fuels SS |
| ---- Eigenfuel Effects - | | | (Percent) |
| -0.0413 | 0.0024 | 16.9997 | 73.8 |
| 0.0226 | 0.0046 | 4.9006 | 8.3 |
| -0.0259 | 0.0056 | 4.5970 | 6.1 |
| 0.0126 | 0.0063 | 1.9955 | 1.2 |
| 0.0152 | 0.0061 | 2.5009 | 1.3 |
| 0.0186 | 0.0060 | 3.0899 | 1.4 |
| 0.0018 | 0.0080 | 0.2234 | 0.0 |
| 0.0045 | 0.0126 | 0.3585 | 0.0 |
| 0.0813 | 0.0220 | 3.6917 | 5.5 |
| -0.0347 | 0.0290 | 1.1949 | 0.7 |
| 0.0887 | 0.0384 | 2.3110 | 1.6 |
| 0.0200 | 0.0364 | 0.5489 | 0.1 |

# APPENDIX G: BOOTSTRAP SAMPLING

This appendix summarizes the use of bootstrap sampling to understand the behavior of diesel fuel eigenvectors. It covers these topics:

- The generation of bootstrap samples to simulate the variability of fuel eigenvectors
- The determination of variability for the eigenvalues
- Whether eigenvectors of substantially similar composition reoccur across the samples
- The variation of the coefficients that make up the eigenvectors.

## G.1 Generation of Bootstrap Samples

A program for the bootstrap analysis is attached as an addendum (Exhibit G.1) to the appendix. The eigenfuel analysis is done by sampling from the original (unstandardized) testfuels data set, and the results are saved in arrays with the "_ref" suffix. Nboot=1000 samples are taken directly from testfuels, each sample is standardized to mean=0 and variance=1, and the eigenfuels analysis is done on the correlation matrix.

The eigenvalue and eigenvector results from the samples are stored in the 3-dimensional matrices Eigval_samp, Eigvec_samp, Explain_samp. The rightmost subscript indexes a third dimension representing the samples, so that, for example, Eigvec_samp( : , : , n) is the plane representing the $n^{th}$ sample. The eigenvectors form the columns of this plane, while their coefficients go down the rows.

## G.2 SUMMARY STATISTICS ON THE BOOTSTRAP SAMPLES

The first step in analysis is to generate basic descriptive statistics on the variation of the eigenvalues. The lower and upper limits of the 95 percent confidence interval are determined by reading sampled eigenvalues at the $25^{th}$ and $975^{th}$ entries (in sort order), respectively:

```
Eigenvalue  1 Descriptive Statistics
Minimum      =          3.9474
Maximum      =           5.169
Not a Number=               0

Mean         =          4.5755
Median       =          4.5779
Std Dev      =          0.1904
Coeff of Var=        0.041614
```

```
95 percent CL Lower=   4.2257
95 percent CL Upper=   4.9494
Eigenvalue  2 Descriptive Statistics
Minimum     =           2.196
Maximum     =          3.0288
Not a Number=               0

Mean        =          2.5927
Median      =          2.5948
Std Dev     =         0.12173
Coeff of Var=         0.04695

95 percent CL Lower=   2.3430
95 percent CL Upper=   2.8224

Eigenvalue  3 Descriptive Statistics
Minimum     =          1.3406
Maximum     =          1.7258
Not a Number=               0

Mean        =          1.5434
Median      =          1.5424
Std Dev     =         0.05633
Coeff of Var=        0.036496

95 percent CL Lower=   1.4318
95 percent CL Upper=   1.6567

Eigenvalue  4 Descriptive Statistics
Minimum     =         0.86965
Maximum     =          1.5141
Not a Number=               0

Mean        =          1.1766
Median      =          1.1786
Std Dev     =         0.11881
Coeff of Var=         0.10098

95 percent CL Lower=  0.94929
95 percent CL Upper=  1.40170

Eigenvalue  5 Descriptive Statistics
Minimum     =         0.58684
Maximum     =         0.81755
Not a Number=               0

Mean        =         0.68577
Median      =         0.68541
Std Dev     =        0.036026
Coeff of Var=        0.052533

95 percent CL Lower=  0.61639
95 percent CL Upper=  0.75805
```

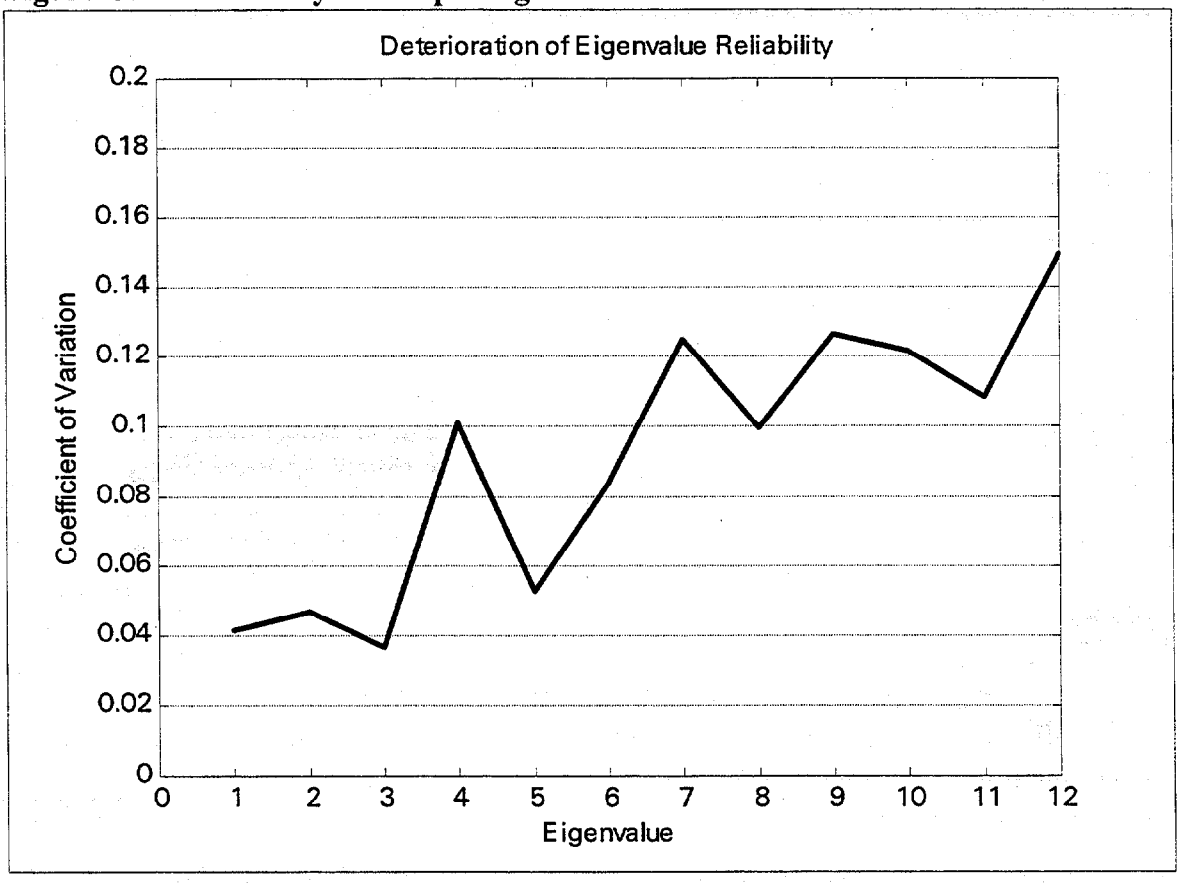
```
Eigenvalue  6 Descriptive Statistics
Minimum      =          0.40955
Maximum      =          0.70387
Not a Number=                0

Mean         =          0.55319
Median       =          0.55284
Std Dev      =         0.046298
Coeff of Var=         0.083693

95 percent CL Lower=   0.46038
95 percent CL Upper=   0.63950

Eigenvalue  7 Descriptive Statistics
Minimum      =          0.23415
Maximum      =          0.49958
Not a Number=                0

Mean         =          0.37349
Median       =          0.37401
Std Dev      =         0.046582
Coeff of Var=          0.12472

95 percent CL Lower=   0.28450
95 percent CL Upper=   0.46108

Eigenvalue  8 Descriptive Statistics
Minimum      =           0.1588
Maximum      =          0.29914
Not a Number=                0

Mean         =          0.22374
Median       =           0.2226
Std Dev      =         0.022264
Coeff of Var=         0.099504

95 percent CL Lower=   0.18312
95 percent CL Upper=   0.26969

Eigenvalue  9 Descriptive Statistics
Minimum      =         0.087113
Maximum      =          0.18513
Not a Number=                0

Mean         =          0.13669
Median       =          0.13694
Std Dev      =         0.017244
Coeff of Var=          0.12615

95 percent CL Lower=   0.10262
95 percent CL Upper=   0.16917

Eigenvalue 10 Descriptive Statistics
Minimum      =         0.050989
Maximum      =          0.12456
Not a Number=                0
```

```
Mean          =          0.07988
Median        =         0.079549
Std Dev       =        0.0097042
Coeff of Var=            0.12148

95 percent CL Lower= 0.062413
95 percent CL Upper= 0.100410

Eigenvalue 11 Descriptive Statistics
Minimum       =         0.023847
Maximum       =         0.051324
Not a Number=                 0

Mean          =         0.034859
Median        =         0.034749
Std Dev       =        0.0037768
Coeff of Var=            0.10835

95 percent CL Lower= 0.028110
95 percent CL Upper= 0.042678

Eigenvalue 12 Descriptive Statistics
Minimum       =          0.01452
Maximum       =         0.041439
Not a Number=                 0

Mean          =         0.024167
Median        =          0.02416
Std Dev       =        0.0036131
Coeff of Var=             0.1495

95 percent CL Lower= 0.017025
95 percent CL Upper= 0.031887
```

Figure G.1 shows how the reliability of the eigenvalues – defined in terms of the coefficient of variation (COV) – generally deteriorates as one goes to the smaller eigenvalues. The most reliable are eigenvalues 1, 2 3, and 5 with COVs in the range of 0.04 to 0.05. Eigenvalue 6 is intermediate (COV = 0.084), while the remaining eigenvalues 4 and 7-12 are at COV = 0.10 or larger.

These results are intuitively satisfying, because we tend to believe that the eigenvectors associated with the largest eigenvalues are most likely to represent general properties of diesel fuels and thereby be present in nearly all samples of those fuels. However, the chart does not itself indicate how to segregate eigenvalues (and associated eigenvectors) that are "reliable" from those that are less so.

**Figure G.1. Variability of Sample Eigenvalues**




## G.3 ROBUSTNESS OF THE EIGENVECTORS

A greater interest lies in whether the eigenvectors tend to repeat themselves from one sample to the next. If so, we gain confidence that they represent recurring features of diesel fuels. A related question, assuming they do repeat, is whether they tend to be found in the same eigenvalue-determined order in each sample, or are subject to switching order between samples. If the ordering is relatively stable, then we gain confidence that the relative importance of the eigenvectors is also a recurring feature. We may also directly compare the variability from sample to sample without being concerned that eigenvector N is "apples" in one sample, but "oranges" in another.

A second program (Exhibit G.2) expresses the eigenvectors from each sample (Eigvec_samp) in terms of the reference eigenvectors (Eigvec_ref). The coefficients are stored in RefCoeff_samp, where the array RefCoeff_samp( v , n , k ) is the nth coefficient of vector v generated in sample k. Two output arrays are set up in this program. For each eigenvector, the reference vector coefficients are squared, the largest squared weight is

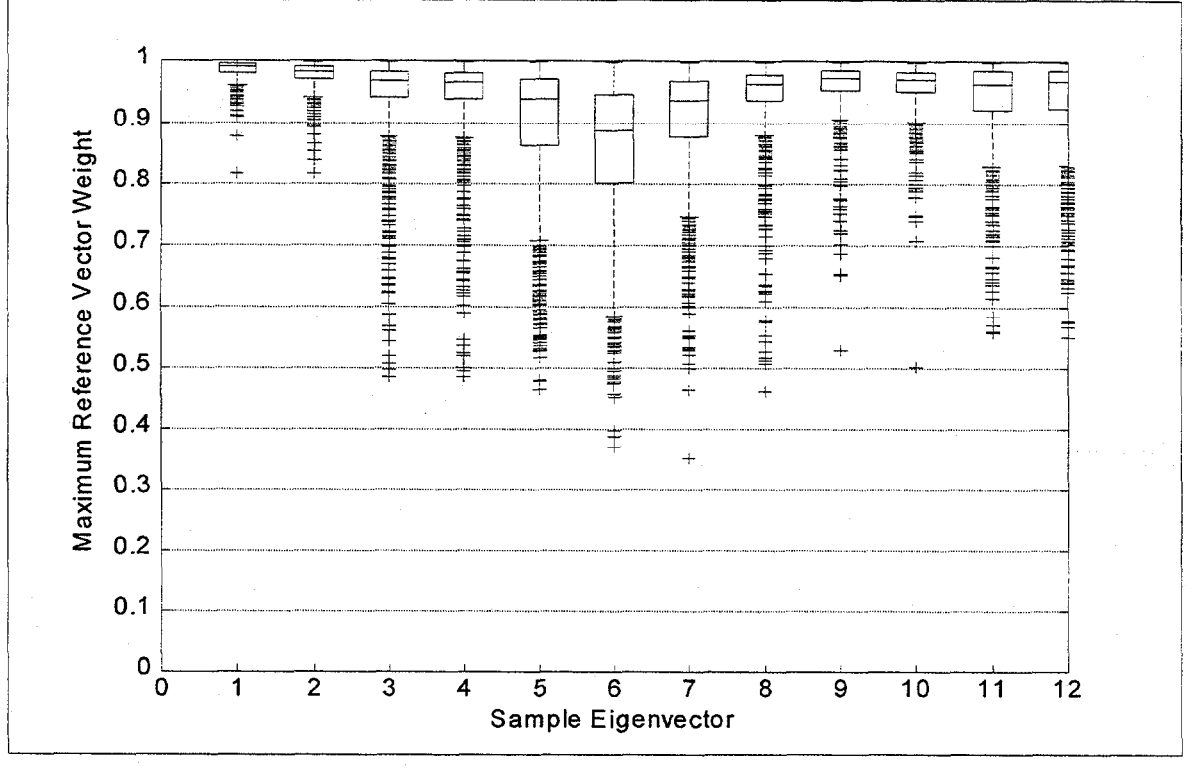
stored in array maxweights, and the reference vector associated with that weight is stored in array pointers:

- pointers( k , v ) gives the reference vector number (1:12) for eigenvector v in sample k
- maxweights( k, v ) gives the squared weight associated with the reference vector.

The term "max vector" is used to refer to the reference vector that gains the largest squared weight. The first question is how purely the sampled vectors load on the "max vector". This is answered in the box plot of the maximum squared-weights for the sample eigenvectors shown in Figure G.2. The mean values are above 0.90 for all sample eigenvectors except for vector 6, for which the mean falls slightly below 0.90. It is a bit of a surprise to find that the mean value declines, and the distribution of the maximum squared-weight broadens, in the middle of the eigenvector range. It will be seen that this middle territory is associated with the most "order switching". Overall, these results indicate that the eigenvectors tend to repeat themselves in each sample with a high degree of consistency.

Table G.1 shows how little order switching exists across the samples. Most of what exists occurs in the middle range of the sample eigenvectors, where vectors 5 and 6 switch their order 3.3 percent of the time. These two vectors also have relatively smaller

**Figure G.2. Maximum Squared Loadings for Sampled Eigenvectors**



mean values for the maximum squared-weights, although note that the mean value is also relatively smaller (and the distribution also broadens) for sample eigenvector 7, which has very little switching.

With one exception the switching happens in pairs – vectors 3 and 4 switch, as do vectors 5 and 6, 7 and 8, 9 and 10, and 11 and 12. The relative prevalence of the fuel features represented by the eigenvectors is apparently quite stable from one sample to another, with the exception of a few samples in which a pair of vectors swap order.

An interesting insight is shown in the 33rd sample, which is the first to show switching between vectors 5 and 6. The box to the right shows the coefficients for sample vectors 5 and 6. Note that the switching occurs on a small margin – in sample vector 5, reference vector 6 beats out reference vector 5 by a small amount, and the reverse is true for sample vector 6. Note also that the highlighted signs are opposite (one plus, one minus) for sample vector 5, but are concurrent (both plus) for sample vector 6. It seems that this sample cannot resolve the features represented by reference vectors 5 and 6, but gives us a sum (5,6) feature and a difference (5,6) feature instead. This looks much like a rotation of the axes at an angle close to 45 degrees.

**Figure G.3. What Order Switching Looks Like**

| | Sample Eigenvector 5 | 6 |
|---|---|---|
| 1 | 0.0155 | 0.0033 |
| 2 | 0.0841 | 0.0273 |
| 3 | -0.0461 | -0.0609 |
| 4 | 0.0487 | 0.0518 |
| 5 | 0.6162 | 0.7520 |
| 6 | -0.7728 | 0.6254 |
| 7 | 0.0862 | 0.1808 |
| 8 | -0.0223 | 0.0239 |
| 9 | 0.0531 | -0.0398 |
| 10 | -0.0186 | 0.0340 |
| 11 | 0.0055 | -0.0123 |
| 12 | -0.0108 | -0.0107 |

It should also be noted that there are a fairly large number of "blurred" cases, primarily in the middle range of eigenvectors, where the vectors do not load purely on their corresponding reference vector, but the same-vector loading is not reduced so much that we have a case of order switching. For example, vector 6 in such a case might have a 0.80 loading on reference vector 6 and a 0.60 loading on reference vector 5. The box plot chart in Figure G.2 shows this to happen most often in vectors 5, 6, and 7. This suggests that simple order switching – where the vectors are essentially unchanged and just swap order – is very uncommon.

We interpret the foregoing results to mean that the eigenvectors are actually quite repeatable from one sample to another and, with only a few exceptions, appear in essentially the same order. This is as true for the eigenvectors with the smallest eigenvalues as for those with the largest. The few exceptions aside, this tends to indicate that the eigenvectors represent recurring features of this group of diesel fuels.

## G.4. CONFIDENCE INTERVALS FOR THE COEFFICIENT VALUES

A goal for the bootstrapping method was to develop the empirical range of values for the coefficients that make up the eigenvectors and use this to define the confidence intervals.

**Table G.1 Order Switching in 1000 Repeated Samples**

| Sample Eigen-vector | Reference Eigenvector | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 1000 | | | | | | | | | | | |
| 2 | | 1000 | | | | | | | | | | |
| 3 | | | 986 | 14 | | | | | | | | |
| 4 | | | 14 | 986 | | | | | | | | |
| 5 | | | | | 967 | 33 | | | | | | |
| 6 | | | | | 33 | 966 | 1 | | | | | |
| 7 | | | | | | 1 | 993 | 6 | | | | |
| 8 | | | | | | | 5 | 995 | | | | |
| 9 | | | | | | | | | 998 | 2 | | |
| 10 | | | | | | | | | 2 | 998 | | |
| 11 | | | | | | | | | | | 995 | 5 |
| 12 | | | | | | | | | | | 5 | 995 |

The "order switching" issue was raised because, if switching happened often enough, it would tend to inflate the confidence intervals. Based on the results presented above, it seems that order switching occurs so infrequently that we can ignore it in what follows.

What cannot be ignored, however, is that the eigenvectors appear with an arbitrary algebraic sign. To remove this trivial source of difference, the eigenvectors were standardized so that the same-vector coefficient is positive in the reference vector expansion. That is, if vector 6, say, has a negative loading on reference vector 6, all component coefficients in vector 6 are multiplied by -1. Having done this, the 1000 bootstrap samples were analyzed to generate the following statistics for each component coefficient: mean, median, standard deviation, the coefficient of variation, and the lower

2.5 percent and upper 97.5 percent cutpoints forming the 95 percent confidence intervals. These results are all tabulated as addendums to this appendix.

The mean and median coefficient values are very similar to the coefficients of the reference vectors. The mean and median values are also very similar to each other, implying that component coefficients are distributed in a fairly symmetric way. Empirical confidence intervals can also be developed using these statistics, as follows: a component coefficient is set to zero if its 95 percent confidence interval contains the value zero. Thus, a non-zero coefficient is retained when it can be said with at least 95 percent confidence that it differs from zero.

The result of this is shown on the following page. The highlighted components are those previous highlighted in Table 2.3 using the judgmental rule of "the three largest plus near ties." This process can be used in simplifying the internal structure, but it still leaves a number of relatively small components. It would probably be desirable in many circumstances to further reduce the number of component coefficients. This goal will be pursued by developing criteria that exclude components (the original fuel property variables) from *all vectors*, when the variables are shown to have negligible impact on emissions.

**Table G.2. Mean Eigenvectors Simplified based on 95 Percent Confidence Interval**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.000 | **-0.550** | 0.000 | -0.221 | 0.000 | 0.000 | 0.000 | **-0.446** | 0.000 | **-0.447** | 0.000 | **0.375** |
| CetImprv | 0.000 | 0.142 | **-0.541** | 0.000 | **0.724** | 0.000 | 0.000 | 0.000 | 0.000 | -0.054 | -0.048 | 0.000 |
| Density | **0.281** | **0.443** | 0.000 | 0.160 | 0.000 | 0.151 | -0.114 | 0.226 | **0.340** | **-0.414** | -0.262 | **0.457** |
| Viscosity | **0.430** | -0.117 | 0.000 | 0.133 | 0.000 | **0.280** | 0.000 | 0.151 | -0.165 | 0.285 | **0.612** | **0.365** |
| Sulfur | 0.000 | 0.175 | **0.605** | 0.000 | **0.366** | 0.000 | **0.521** | 0.186 | 0.000 | 0.000 | 0.000 | -0.097 |
| MonoArom | 0.143 | **0.461** | -0.246 | 0.000 | **-0.288** | 0.000 | **0.534** | **-0.472** | -0.131 | 0.000 | 0.158 | 0.000 |
| PolyArom | 0.000 | **0.411** | **0.366** | -0.277 | 0.000 | **-0.272** | **-0.512** | -0.350 | -0.173 | 0.000 | 0.215 | 0.000 |
| IBP | **0.260** | 0.000 | 0.000 | **0.515** | 0.000 | **-0.626** | 0.000 | 0.000 | 0.230 | 0.126 | 0.000 | 0.000 |
| T10 | **0.396** | -0.111 | 0.134 | 0.297 | 0.000 | 0.173 | 0.000 | 0.000 | **-0.622** | -0.184 | **-0.387** | -0.201 |
| T50 | **0.428** | 0.000 | 0.000 | 0.000 | 0.000 | **0.291** | 0.000 | -0.233 | **0.543** | 0.000 | 0.000 | **-0.528** |
| T90 | **0.390** | 0.000 | 0.000 | **-0.410** | -0.148 | 0.000 | 0.000 | 0.000 | 0.000 | **0.536** | **-0.470** | 0.244 |
| FBP | **0.362** | 0.000 | -0.159 | **-0.388** | 0.000 | **-0.325** | 0.000 | **0.455** | -0.156 | -0.396 | 0.199 | -0.250 |

**MEAN VALUES**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.062 | -0.550 | 0.147 | -0.221 | 0.061 | -0.060 | 0.141 | -0.446 | 0.099 | -0.447 | 0.042 | 0.375 |
| CetImprv | 0.033 | 0.142 | -0.541 | -0.195 | 0.724 | 0.055 | -0.019 | -0.103 | 0.003 | -0.054 | -0.048 | 0.004 |
| Density | 0.281 | 0.443 | 0.055 | 0.160 | -0.044 | 0.151 | -0.114 | 0.226 | 0.340 | -0.414 | -0.262 | 0.457 |
| Viscosity | 0.430 | -0.117 | -0.014 | 0.133 | 0.078 | 0.280 | -0.003 | 0.151 | -0.165 | 0.285 | 0.612 | 0.365 |
| Sulfur | 0.002 | 0.175 | 0.605 | -0.218 | 0.366 | 0.117 | 0.521 | 0.186 | 0.013 | 0.054 | -0.022 | -0.097 |
| MonoArom | 0.143 | 0.461 | -0.246 | 0.055 | -0.288 | -0.027 | 0.534 | -0.472 | -0.131 | -0.075 | 0.158 | -0.038 |
| PolyArom | 0.075 | 0.411 | 0.366 | -0.277 | 0.088 | -0.272 | -0.512 | -0.350 | -0.173 | 0.032 | 0.215 | 0.049 |
| IBP | 0.260 | -0.071 | 0.065 | 0.515 | 0.242 | -0.626 | 0.118 | -0.032 | 0.230 | 0.126 | 0.016 | -0.022 |
| T10 | 0.396 | -0.111 | 0.134 | 0.297 | 0.102 | 0.173 | -0.088 | -0.120 | -0.622 | -0.184 | -0.387 | -0.201 |
| T50 | 0.428 | -0.081 | 0.058 | -0.057 | -0.029 | 0.291 | -0.141 | -0.233 | 0.543 | 0.017 | 0.073 | -0.528 |
| T90 | 0.390 | -0.070 | -0.089 | -0.410 | -0.148 | -0.124 | 0.075 | -0.031 | 0.024 | 0.536 | -0.470 | 0.244 |
| FBP | 0.362 | -0.045 | -0.159 | -0.388 | -0.132 | -0.325 | 0.057 | 0.455 | -0.156 | -0.396 | 0.199 | -0.250 |

**MEDIAN VALUES**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.064 | -0.551 | 0.152 | -0.220 | 0.060 | -0.064 | 0.138 | -0.450 | 0.102 | -0.450 | 0.039 | 0.377 |
| CetImprv | 0.036 | 0.142 | -0.546 | -0.213 | 0.749 | 0.054 | -0.017 | -0.104 | 0.003 | -0.055 | -0.048 | 0.003 |
| Density | 0.287 | 0.444 | 0.053 | 0.161 | -0.041 | 0.155 | -0.113 | 0.229 | 0.337 | -0.414 | -0.270 | 0.459 |
| Viscosity | 0.430 | -0.117 | -0.017 | 0.137 | 0.087 | 0.288 | -0.001 | 0.149 | -0.162 | 0.285 | 0.621 | 0.385 |
| Sulfur | 0.003 | 0.182 | 0.621 | -0.213 | 0.378 | 0.109 | 0.532 | 0.173 | 0.013 | 0.054 | -0.021 | -0.098 |
| MonoArom | 0.146 | 0.463 | -0.248 | 0.051 | -0.291 | -0.026 | 0.531 | -0.497 | -0.134 | -0.075 | 0.159 | -0.037 |
| PolyArom | 0.078 | 0.414 | 0.375 | -0.273 | 0.084 | -0.279 | -0.520 | -0.342 | -0.174 | 0.032 | 0.217 | 0.053 |
| IBP | 0.261 | -0.072 | 0.059 | 0.524 | 0.251 | -0.660 | 0.120 | -0.034 | 0.230 | 0.126 | 0.016 | -0.022 |
| T10 | 0.397 | -0.113 | 0.125 | 0.304 | 0.108 | 0.179 | -0.093 | -0.117 | -0.629 | -0.185 | -0.388 | -0.205 |
| T50 | 0.428 | -0.083 | 0.059 | -0.057 | -0.029 | 0.298 | -0.146 | -0.239 | 0.544 | 0.020 | 0.079 | -0.535 |
| T90 | 0.390 | -0.071 | -0.081 | -0.417 | -0.152 | -0.129 | 0.072 | -0.031 | 0.015 | 0.542 | -0.483 | 0.235 |
| FBP | 0.362 | -0.048 | -0.154 | -0.395 | -0.139 | -0.329 | 0.071 | 0.460 | -0.158 | -0.398 | 0.208 | -0.247 |

**STANDARD DEVIATION**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.054 | 0.023 | 0.062 | 0.051 | 0.049 | 0.065 | 0.078 | 0.052 | 0.077 | 0.045 | 0.099 | 0.042 |
| CetImprv | 0.038 | 0.052 | 0.052 | 0.144 | 0.095 | 0.241 | 0.064 | 0.053 | 0.033 | 0.023 | 0.016 | 0.015 |
| Density | 0.040 | 0.040 | 0.044 | 0.039 | 0.053 | 0.050 | 0.052 | 0.056 | 0.073 | 0.068 | 0.110 | 0.061 |
| Viscosity | 0.011 | 0.039 | 0.048 | 0.032 | 0.094 | 0.050 | 0.079 | 0.038 | 0.055 | 0.057 | 0.087 | 0.149 |
| Sulfur | 0.055 | 0.077 | 0.076 | 0.126 | 0.094 | 0.186 | 0.066 | 0.091 | 0.034 | 0.025 | 0.026 | 0.017 |
| MonoArom | 0.042 | 0.031 | 0.057 | 0.068 | 0.072 | 0.144 | 0.080 | 0.111 | 0.059 | 0.047 | 0.023 | 0.034 |
| PolyArom | 0.052 | 0.041 | 0.085 | 0.077 | 0.106 | 0.127 | 0.092 | 0.095 | 0.048 | 0.041 | 0.025 | 0.049 |
| IBP | 0.018 | 0.047 | 0.130 | 0.081 | 0.220 | 0.129 | 0.148 | 0.061 | 0.060 | 0.044 | 0.018 | 0.014 |
| T10 | 0.011 | 0.040 | 0.076 | 0.049 | 0.069 | 0.068 | 0.075 | 0.074 | 0.039 | 0.088 | 0.050 | 0.095 |
| T50 | 0.011 | 0.041 | 0.034 | 0.033 | 0.102 | 0.061 | 0.092 | 0.066 | 0.038 | 0.106 | 0.137 | 0.045 |
| T90 | 0.009 | 0.039 | 0.080 | 0.041 | 0.061 | 0.066 | 0.051 | 0.052 | 0.079 | 0.053 | 0.084 | 0.126 |
| FBP | 0.010 | 0.043 | 0.077 | 0.054 | 0.120 | 0.085 | 0.145 | 0.054 | 0.075 | 0.046 | 0.076 | 0.065 |

**COEFFICIENT OF VARIATION**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.868 | -0.042 | 0.423 | -0.231 | 0.800 | -1.070 | 0.554 | -0.117 | 0.777 | -0.100 | 2.341 | 0.113 |
| CetImprv | 1.138 | 0.364 | -0.095 | -0.736 | 0.131 | 4.411 | -3.325 | -0.517 | 9.999 | -0.417 | -0.331 | 4.066 |
| Density | 0.143 | 0.090 | 0.790 | 0.244 | -1.221 | 0.333 | -0.460 | 0.248 | 0.214 | -0.164 | -0.421 | 0.133 |
| Viscosity | 0.025 | -0.332 | -3.539 | 0.242 | 1.194 | 0.178 | -24.800 | 0.251 | -0.335 | 0.200 | 0.142 | 0.407 |
| Sulfur | 24.723 | 0.437 | 0.125 | -0.579 | 0.255 | 1.596 | 0.127 | 0.489 | 2.671 | 0.470 | -1.187 | -0.170 |
| MonoArom | 0.291 | 0.067 | -0.232 | 1.250 | -0.252 | -5.324 | 0.150 | -0.236 | -0.448 | -0.622 | 0.145 | -0.911 |
| PolyArom | 0.698 | 0.100 | 0.233 | -0.277 | 1.207 | -0.467 | -0.179 | -0.272 | -0.278 | 1.272 | 0.118 | 1.014 |
| IBP | 0.071 | -0.666 | 2.007 | 0.158 | 0.909 | -0.206 | 1.252 | -1.873 | 0.263 | 0.347 | 1.115 | -0.615 |
| T10 | 0.027 | -0.361 | 0.569 | 0.164 | 0.671 | 0.393 | -0.851 | -0.619 | -0.063 | -0.480 | -0.128 | -0.471 |
| T50 | 0.025 | -0.507 | 0.581 | -0.583 | -3.525 | 0.210 | -0.649 | -0.281 | 0.070 | 6.383 | 1.873 | -0.085 |
| T90 | 0.022 | -0.557 | -0.904 | -0.101 | -0.414 | -0.533 | 0.675 | -1.648 | 3.309 | 0.100 | -0.179 | 0.516 |
| FBP | 0.026 | -0.950 | -0.483 | -0.140 | -0.912 | -0.263 | 2.519 | 0.118 | -0.479 | -0.117 | 0.379 | -0.261 |

**LOWER 0.025 PERCENTILE**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | -0.039 | -0.594 | -0.006 | -0.319 | -0.039 | -0.179 | -0.014 | -0.549 | -0.064 | -0.525 | -0.147 | 0.286 |
| CetImprv | -0.042 | 0.038 | -0.623 | -0.427 | 0.465 | -0.447 | -0.146 | -0.202 | -0.062 | -0.095 | -0.081 | -0.023 |
| Density | 0.183 | 0.360 | -0.023 | 0.084 | -0.152 | 0.042 | -0.222 | 0.110 | 0.192 | -0.542 | -0.456 | 0.335 |
| Viscosity | 0.409 | -0.187 | -0.099 | 0.056 | -0.115 | 0.155 | -0.160 | 0.079 | -0.276 | 0.165 | 0.432 | 0.010 |
| Sulfur | -0.114 | 0.003 | 0.393 | -0.507 | 0.153 | -0.215 | 0.344 | 0.045 | -0.056 | -0.002 | -0.073 | -0.129 |
| MonoArom | 0.058 | 0.398 | -0.347 | -0.073 | -0.416 | -0.298 | 0.384 | -0.624 | -0.234 | -0.163 | 0.110 | -0.104 |
| PolyArom | -0.033 | 0.316 | 0.165 | -0.436 | -0.118 | -0.498 | -0.661 | -0.583 | -0.264 | -0.043 | 0.157 | -0.068 |
| IBP | 0.219 | -0.162 | -0.169 | 0.307 | -0.243 | -0.767 | -0.165 | -0.152 | 0.106 | 0.039 | -0.018 | -0.050 |
| T10 | 0.373 | -0.188 | 0.011 | 0.180 | -0.040 | 0.025 | -0.221 | -0.267 | -0.678 | -0.354 | -0.484 | -0.390 |
| T50 | 0.407 | -0.154 | -0.010 | -0.122 | -0.236 | 0.151 | -0.310 | -0.349 | 0.467 | -0.195 | -0.204 | -0.593 |
| T90 | 0.372 | -0.142 | -0.279 | -0.464 | -0.264 | -0.243 | -0.027 | -0.130 | -0.106 | 0.419 | -0.600 | 0.013 |
| FBP | 0.342 | -0.127 | -0.340 | -0.474 | -0.346 | -0.465 | -0.271 | 0.336 | -0.307 | -0.482 | 0.028 | -0.382 |

**UPPER 0.975 PERCENTILE**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NatCetane | 0.170 | -0.504 | 0.246 | -0.121 | 0.156 | 0.076 | 0.295 | -0.326 | 0.242 | -0.348 | 0.242 | 0.448 |
| CetImprv | 0.103 | 0.239 | -0.444 | 0.136 | 0.810 | 0.520 | 0.110 | 0.007 | 0.067 | -0.012 | -0.018 | 0.033 |
| Density | 0.343 | 0.520 | 0.150 | 0.233 | 0.056 | 0.237 | -0.018 | 0.332 | 0.479 | -0.287 | -0.034 | 0.570 |
| Viscosity | 0.451 | -0.032 | 0.099 | 0.189 | 0.255 | 0.349 | 0.138 | 0.225 | -0.061 | 0.394 | 0.746 | 0.600 |
| Sulfur | 0.105 | 0.314 | 0.689 | 0.012 | 0.497 | 0.500 | 0.619 | 0.411 | 0.078 | 0.101 | 0.031 | -0.063 |
| MonoArom | 0.217 | 0.516 | -0.124 | 0.201 | -0.143 | 0.253 | 0.692 | -0.169 | -0.010 | 0.021 | 0.202 | 0.028 |
| PolyArom | 0.168 | 0.479 | 0.496 | -0.147 | 0.312 | -0.009 | -0.302 | -0.195 | -0.078 | 0.120 | 0.257 | 0.129 |
| IBP | 0.292 | 0.021 | 0.351 | 0.638 | 0.636 | -0.298 | 0.410 | 0.089 | 0.347 | 0.211 | 0.052 | 0.004 |
| T10 | 0.416 | -0.029 | 0.310 | 0.363 | 0.221 | 0.287 | 0.076 | 0.012 | -0.530 | -0.002 | -0.295 | -0.009 |
| T50 | 0.448 | 0.009 | 0.120 | 0.006 | 0.175 | 0.388 | 0.054 | -0.100 | 0.615 | 0.227 | 0.347 | -0.416 |
| T90 | 0.406 | 0.007 | 0.036 | -0.318 | -0.011 | 0.033 | 0.174 | 0.070 | 0.200 | 0.621 | -0.270 | 0.488 |
| FBP | 0.379 | 0.043 | -0.035 | -0.264 | 0.124 | -0.124 | 0.309 | 0.536 | -0.005 | -0.298 | 0.315 | -0.132 |

Exhibit G.1. Program to Generate Bootstrap Samples

```
% This m-file conducts a bootstrap simulation of the fuel eigenvectors
% load emissions and fuel data, setting up workspace
filename='c:\matlab\studies\bootstrap\HCNPFuel.csv';
[HC,CO,NOx,PM,testfuels,fuelnames,xdata,NRepl,weights] =
   loader(filename,'comma');
% standardize variables to mean 0 and variance 1
[xvariables, testfuels_mean, testfuels_std] =
standardform(testfuels,NRepl);
% conduct the principal components analysis
[fuel_pc,latent,variance,explained,PCA_coeffs]=
PCAfunction(xvariables,1,fuelnames);

X_ref = xvariables;
Eigvec_ref = fuel_pc;
Eigval_ref = latent;
Explain_ref = explained;
PCACoeff_ref = PCA_coeffs;
NBoot = 1000
[obs, vars] = size(xvariables);
NSize = obs

Seed = 0
rand('seed',Seed)
for bootsample = 1:NBoot
   for i = 1:NSize
      z(i) = 1 + fix(NSize*rand(1,1));
   end
   % select sample, re-standardize to (0,1), and do PCA analysis
   xvariables = testfuels(z,:);
   [xvariables, xvariables_mean,
xvariables_std]=standardform(xvariables,NRepl);

[fuel_pc,latent,variance,explained,PCA_coeffs]=PCAfunction(xvariables,0,
fuelnames);

   if (bootsample==1)
      Eigvec_samp = fuel_pc;
      Eigval_samp = latent';
      Explain_samp = explained;
   else
      Eigvec_samp = cat(3, Eigvec_samp, fuel_pc);
      Eigval_samp = [Eigval_samp ; latent'];
      Explain_samp = cat(3, Explain_samp, explained);
   end
end

% Generate basic summary statistics
bootsummary
```

Exhibit G.2. MatLab Program to Examine Repeatability and Order Switching

```
% This script expresses the simulated eigenvectors in terms of the
% original (reference) eigenvectors.

clear
load pcabootstrap Eigvec_samp Eigvec_ref
whos

[rows,cols,reps]=size(Eigvec_samp);

lastsample=reps;
for bootsample = 1:lastsample

   Coeffs = Eigvec_samp(:,:,bootsample)'*Eigvec_ref;

   if (bootsample==1)
      RefCoeff_samp = Coeffs;
   else
      RefCoeff_samp = cat(3,RefCoeff_samp,Coeffs);
   end

   coefficients = RefCoeff_samp(:,:,bootsample)';
   coeffsquared = coefficients.^2;
   a=max(coeffsquared);
   b=kron(ones(rows,1),a);
   c=repmat([1:12],cols,1)';
   d=c(coeffsquared==b);

   e=diag(coefficients);
   s=e./abs(e);

   if (bootsample==1)
      pointers = d';
      maxweights = a;
      signs = s';
   else
      pointers = [pointers; d'];
      maxweights = [maxweights; a];
      signs = [signs; s'];
   end

end
```

# APPENDIX H: INCORPORATING NON-LINEAR TERMS IN A VECTOR MODEL

Nonlinear basis elements, such as higher powers of a variable or a product of two variables (interaction) can be incorporated in an eigenvector regression in much the same way as linear terms are introduced. However, there are some nuances to be observed and understood and some interpretational aspects that need to be examined. This Appendix aims to cover these points and to illustrate them computationally, albeit with hypothetical data.

As a preliminary, it will be insightful to examine some of the relations that obtain between predictor and response variables and how these relations affect the partitioning of the response sum of squares.

## H.1 PARTITIONING VARIANCE IN X AND Y SPACE

One of the misunderstandings that has hampered the application of Principal Component Regression (PCR) has been the misconception that the variance partitioning of the design matrix can be used to select the eigenvectors to use in the regression model. The difficulties that can be encountered as a consequence of this assumption have been pointed out in the statistical literature, but there seems to be no explicit pronouncement on how the *response variable* alters the relative importance of the *design matrix components*. That relationship will be clearly explicated in this section, even to the point of how to formulate a response vector to yield a specified partitioning of the model sum of squares for any given design matrix.

For our illustration, we take as our "given design matrix" the following data from Cooley and Lohnes[1]. The term "treatment" denotes combinations of values of the three variables $x_1$, $x_2$, and $x_3$.

| Treatment | $x_1$ | $x_2$ | $x_3$ | Treatment | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 4 | 3 | 6 | 7 | 2 | 9 |
| 2 | 4 | 1 | 8 | 7 | 5 | 3 | 3 |
| 3 | 6 | 3 | 5 | 8 | 9 | 5 | 8 |
| 4 | 8 | 6 | 1 | 9 | 7 | 4 | 5 |
| 5 | 8 | 5 | 7 | 10 | 8 | 2 | 2 |

[1] Cooley, W.W., P.R. Lohnes, *Multivariate Data Analysis*, John Wiley & Sons, Inc., New York, 1971, p. 110.

First, we transform these predictor vectors to standardized form, with zero mean and unit variance. Then we compute the eigenvectors and eigenvalues of the correlation matrix for these vectors. The percent of the variance among treatments attributable to each of these vectors can be computed directly from the eigenvalues and is found to be:

| Percent Design Matrix SS | |
|---|---|
| Eigenvector 1 | 58.96 |
| Eigenvector 2 | 30.90 |
| Eigenvector 3 | 10.14 |

However, this partitioning has no obvious relation to the response vector; indeed, a response vector has not yet been introduced. Inasmuch as *any* set of 10 numbers could be used as the response vector, it is evident that there is nothing necessarily *special or insightful* about the design-matrix partitioning so far as the response is concerned.

We present, now, three candidate response vectors, together with the partitioning of their model sums of squares.

| Treatment | $Y_1$ | $Y_2$ | $Y_3$ |
|---|---|---|---|
| 1 | 1.1121 | 1.6759 | 0.9815 |
| 2 | -3.0537 | -3.7783 | -2.6654 |
| 3 | -0.4500 | -0.4726 | -0.4213 |
| 4 | 2.5911 | 2.7682 | 1.8204 |
| 5 | -0.4495 | -2.1237 | -0.9846 |
| 6 | -1.8453 | -2.1597 | -1.2595 |
| 7 | -0.0128 | 0.2881 | -0.1421 |
| 8 | -0.3660 | -1.8930 | -0.7023 |
| 9 | 0.0709 | -0.3068 | -0.1414 |
| 10 | 2.4032 | 6.0020 | 3.5147 |

| Percent of Response SS | | | |
|---|---|---|---|
| Eigenvector 1 | 58.9591 | 21.5326 | 33.3333 |
| Eigenvector 2 | 30.9025 | 45.1438 | 33.3333 |
| Eigenvector 3 | 10.1383 | 33.3236 | 33.3333 |

It is evident that the first response vector has a model SS partitioning that exactly duplicates the partitioning of the design matrix. In the second, however, the eigenvector that contributes most to the partitioning of the X-matrix is the one that contributes least to the partitioning of the response SS. Finally, the third Y-vector yields a model SS for which all three eigenvectors make equal contributions to the response SS.

The regression coefficients for each of these three cases are displayed below. No error contribution was added to the response vectors as formulated, so the coefficients represent an *exact* fit to the response data.

| | Regression Coefficients for: | | |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | $Y_3$ |
| Eigenvector 1 | 1.0000 | 1.0000 | 0.7519 |
| Eigenvector 2 | 1.0000 | 2.0000 | 1.0386 |
| Eigenvector 3 | 1.0000 | 3.0000 | 1.8133 |

No smoke and mirrors is involved in defining response vectors to achieve a pre-determined partitioning of the sum of squares. The response SS attributable to each eigenvector can be computed as

$$\underline{c}^2 * \underline{e} * (n-1)$$

where $\underline{c}$ is the vector of regression coefficients, $\underline{e}$ is the vector of eigenvalues for the correlation matrix and n is the number of observations. Multiplication here is to be understood as the scalar multiplication of corresponding elements of the two vectors, *not* matrix multiplication.

It is clear, then, that if $\underline{c}$ = [1 1 1]', the sum of squares partitioning will be proportional to the eigenvalues and, accordingly, to the percent breakdown of the X-matrix variance. In the second case, the coefficients [1 2 3]' weight the eigenvalues respectively by 1, 4 and 9. That leads to the second partitioning as tabulated above. In the third case, one simply asks what numbers are needed to multiply into the eigenvalue matrix so as to produce equal partitioning of the response SS. One then simply extracts the square roots of those numbers, and the resulting values are the required regression coefficients.

This preliminary presentation has two purposes. First, it destroys once and for all the myth that PCR can lead to erroneous selection of eigenvectors. It is not PCR that should be implicated, but the past practice of pre-selecting eigenvectors based on the design matrix SS partitioning. When all vectors are initially retained and related to the response variable, then simplification can subsequently be done without risk of unforseen error. Second, it provides a procedure for constructing artificial data to illustrate specific effects, including those that are both linear and nonlinear and will be used in the discussion that follows.

## H.2 NONLINEAR BASIS ELEMENTS

Two problems face a data analyst when he or she inserts a nonlinear basis vector in a regression model:

25. How to incorporate the nonlinear element into the eigenvector environment.

26. How to normalize the nonlinear term

The following sections consider these problems as they pertain to quadratic and interaction terms in the regression equation.

## H.2.1 Adding a Square Term to the Basis

To annex a quadratic vector to the design matrix, it is necessary only to construct a vector, the elements of which are the squares of the original variable. Below we annex such a vector to the original X-matrix. Note that the elements of column 4 are simply the squares of the elements of column 3.

| Treatment | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| 1 | 7 | 4 | 3 | 9 |
| 2 | 4 | 1 | 8 | 64 |
| 3 | 6 | 3 | 5 | 25 |
| 4 | 8 | 6 | 1 | 1 |
| 5 | 8 | 5 | 7 | 49 |
| 6 | 7 | 2 | 9 | 81 |
| 7 | 5 | 3 | 3 | 9 |
| 8 | 9 | 5 | 8 | 64 |
| 9 | 7 | 4 | 5 | 25 |
| 10 | 8 | 2 | 2 | 4 |

If OLS regression were to be performed in the usual way, no further processing of the above matrix would be necessary. In keeping with the notion of "a level playing field" for all variables, one would first "normalize" the design (X) matrix to zero mean and unit standard deviation.

In the process of annexing quadratic or other nonlinear terms to the regression equation, one can perform this normalization in two distinct ways. These two approaches are discussed below under the headings "Postnorm" and "Prenorm."

### H.2.1.1 Postnorm

As used here, the term *postnorm* refers to normalizing the X-matrix *after* computing the nonlinear vector from one or more of the "raw" data variables. To illustrate, the above matrix is treated just as if it consisted of four rather than three variables, the fourth being the column of squared elements of $x_3$. This four-variable set of treatments is listed below in standard (normalized) form (zero mean, unit variance).

| Treatment | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| 1 | 0.0656 | 0.3162 | -0.7482 | -0.8279 |
| 2 | -1.9030 | -1.5811 | 1.0332 | 1.0615 |
| 3 | -0.5906 | -0.3162 | -0.0356 | -0.2782 |
| 4 | 0.7218 | 1.5811 | -1.4608 | -1.1027 |
| 5 | 0.7218 | 0.9487 | 0.6769 | 0.5462 |
| 6 | 0.0656 | -0.9487 | 1.3895 | 1.6454 |
| 7 | -1.2468 | -0.3162 | -0.7482 | -0.8279 |
| 8 | 1.3781 | 0.9487 | 1.0332 | 1.0615 |
| 9 | 0.0656 | 0.3162 | -0.0356 | -0.2782 |
| 10 | 0.7218 | -0.9487 | -1.1045 | -0.9996 |

Upon computing the correlation matrix we note that the fourth column vector is highly correlated with the third.

**Correlation Matrix**

| | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| $x_1$ | 1.0000 | 0.6687 | -0.1013 | -0.0498 |
| $x_2$ | 0.6687 | 1.0000 | -0.2879 | -0.2788 |
| $x_3$ | -0.1013 | -0.2879 | 1.0000 | 0.9803 |
| $x_3^2$ | -0.0498 | -0.2788 | 0.9803 | 1.0000 |

We then compute the eigenvectors and eigenvalues of this matrix.

**Eigenvectors**

| | | | | |
|---|---|---|---|---|
| $x_1$ | 0.3137 | -0.6664 | 0.6735 | -0.0619 |
| $x_2$ | 0.4434 | -0.5267 | -0.7242 | 0.0373 |
| $x_3$ | -0.5982 | -0.3592 | -0.1412 | -0.7022 |
| $x_3^2$ | -0.5891 | -0.3866 | -0.0430 | 0.7083 |

**Eigenvalues**

| | | | |
|---|---|---|---|
| 2.2319 | 1.4450 | 0.3054 | 0.0176 |

Note that the design space still has essentially only three dimensions; apparently the annexation of the square term has made little addition to the variation among treatments. This phenomenon is the result of the fact that, in the range of the values of the variable, the two curves are closely related.

Next, we convert the above matrix to the "vectorized" version shown below. The method used to perform the conversion is that given by Equation 6 in Appendix D.

| Treatment | Vector 1 | Vector 2 | Vector 3 | Vector 4 |
|---|---|---|---|---|
| 1 | 1.0961 | 0.3785 | -0.0435 | -0.0532 |
| 2 | -2.5416 | 1.3195 | -0.3283 | 0.0850 |
| 3 | -0.1403 | 0.6805 | -0.1518 | -0.1473 |
| 4 | 2.4511 | -0.3629 | -0.4052 | 0.2592 |
| 5 | -0.0796 | -1.4350 | -0.3200 | -0.0978 |
| 6 | -2.2007 | -0.6792 | 0.4642 | 0.1501 |
| 7 | 0.4040 | 1.5862 | -0.4694 | 0.0044 |
| 8 | -0.3905 | -2.1995 | 0.0495 | -0.0237 |
| 9 | 0.3460 | -0.0899 | -0.1678 | -0.1643 |
| 10 | 1.0554 | 0.8018 | 1.3723 | -0.0125 |

We are now ready to do regression if only we had a response vector corresponding to the above X-matrix.

As indicated above, we are at liberty, for demonstration purposes, to construct an arbitrary response vector. Accordingly, we hope to choose responses in such a way that the quadratic

term contributes materially to the response sum of squares. In the real world, however, quadratic curvature may play only a minor or negligible role in prediction.

A multitude of response vectors can be generated simply as linear combinations of the columns of the above matrix. Five such response vectors are listed below, together with the weighting factors applied to the normalized X-matrix vectors.

**Formulation of Five Hypothetical Response Vectors**

| Case: | A | B | C | D | E |
|---|---|---|---|---|---|
| | Weights Applied to Basis Vectors | | | | |
| | 1.0000 | 0.2000 | 0.5000 | 0.5000 | 1.0000 |
| | 1.0000 | 0.5000 | 0.5000 | 0.5000 | 1.3156 |
| | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.4406 |
| | 1.0000 | 5.0000 | 5.0000 | 4.0000 | 1.4958 |
| | Hypothetical Response Vectors | | | | |
| Treatment | | | | | |
| 1 | 1.3779 | 0.0990 | 0.4279 | 0.4810 | 1.4518 |
| 2 | -1.4653 | 0.2481 | -0.5144 | -0.5993 | -1.1514 |
| 3 | 0.2412 | -0.5761 | -0.6181 | -0.4708 | 0.3161 |
| 4 | 1.9422 | 1.1994 | 1.9348 | 1.6756 | 1.7776 |
| 5 | -1.9325 | -1.5425 | -1.5664 | -1.4686 | -2.5749 |
| 6 | -2.2656 | 0.4352 | -0.2250 | -0.3752 | -2.2010 |
| 7 | 1.5252 | 0.4267 | 0.5478 | 0.5434 | 1.8212 |
| 8 | -2.5642 | -1.2468 | -1.3639 | -1.3402 | -3.2483 |
| 9 | -0.0760 | -0.9651 | -0.8613 | -0.6970 | -0.2598 |
| 10 | 3.2171 | 1.9220 | 2.2387 | 2.2511 | 4.0686 |

Since no random or *error* component has been added to these fictional responses, regression analysis simply returns, as the regression coefficients, the weighting factors used in generating the responses. Because of the orthogonality of the transformed design matrix, the contribution that each basis *eigenvector* makes to the Model SS is readily computed. Also, by virtue of simplification procedures illustrated elsewhere in this report, contributions to the SS attributable to each original *property* variable can also be computed. These contributions, denoted *E-Space* and *P-Space* respectively, are tabulated below.

SS Partitioning in E-Space

| | Case A | Case B | Case C | Case D | Case E |
|---|---|---|---|---|---|
| Vector 1 | 55.7983 | 7.4587 | 33.4991 | 37.0284 | 1.2841 |
| Vector 2 | 36.1248 | 30.1804 | 21.6878 | 23.9728 | 46.2609 |
| Vector 3 | 7.6359 | 25.5176 | 18.3372 | 20.2691 | 11.7249 |
| Vector 4 | 0.4410 | 36.8434 | 26.4759 | 18.7298 | 0.7300 |

SS Partitioning in P-Space

| | Case A | Case B | Case C | Case D | Case E |
|---|---|---|---|---|---|
| $x_1$ | 0.0250 | 0.0495 | 1.6334 | 1.9900 | 0.0694 |
| $x_2$ | 3.3491 | 26.1599 | 7.8168 | 8.8793 | 10.1218 |
| $x_3$ | 56.0131 | 73.6689 | 87.4869 | 82.3380 | 54.3237 |
| $x_3^2$ | 40.6128 | 0.1217 | 3.0628 | 6.7927 | 35.4851 |

Interesting, but no surprise, is the fact that the contributions for Case A agree exactly with the percent contributions to the treatment variance as computed from the eigenvalues of the correlation matrix. In this instance, the last eigenvector contributes little to the predictive capability of the model. Cases B, C, D and E, however, illustrate how different response vectors can alter the contributions made by the basis vectors, *in spite* of their relative importance in explaining the variability among treatments.

Several observations can be made with regard to the P-Space contributions and particularly the contribution of the nonlinear term. One sees that $x_1$ contributes little to the response SS in all five cases but that the contribution of the quadratic term ranges from little more than 0.1 percent to more than 40 percent. Note that, in Case B, the quadratic component of $x_3^2$ makes a negligible contribution, even though the linear component of $x_3$ accounts for nearly 75 percent of the response SS. In cases A and E, however, the quadratic contributions are nearly comparable to the linear contributions.

Interpretation of the roles played by the four vectors and by their components need to be guided by physical knowledge of the system under study. In this demonstration exercise, no such knowledge is available, the purpose of the demonstration being primarily to illustrate formulation and computation procedures.

## H.2.1.2 Prenorm

Let us now consider an alternative approach, in which the original X-matrix is normalized *before* computing the square term. To generate the quadratic term, one squares each transformed value in the third column to produce a fourth column, as shown below.

Normalized X-Matrix

| Treatment | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| 1 | 0.0656 | 0.3162 | -0.7482 | 0.5598 |
| 2 | -1.9030 | -1.5811 | 1.0332 | 1.0676 |
| 3 | -0.5906 | -0.3162 | -0.0356 | 0.0013 |
| 4 | 0.7218 | 1.5811 | -1.4608 | 2.1339 |
| 5 | 0.7218 | 0.9487 | 0.6769 | 0.4583 |
| 6 | 0.0656 | -0.9487 | 1.3895 | 1.9307 |
| 7 | -1.2468 | -0.3162 | -0.7482 | 0.5598 |
| 8 | 1.3781 | 0.9487 | 1.0332 | 1.0676 |
| 9 | 0.0656 | 0.3162 | -0.0356 | 0.0013 |
| 10 | 0.7218 | -0.9487 | -1.1045 | 1.2199 |
| Mean | 0 | 0 | 0 | 0.9000 |
| Std Dev | 1 | 1 | 1 | 0.7292 |

Note, however, that the fourth column is now *not* normalized. Consequently, renormalization is required because the eigenvector procedures, operating on the *correlation matrix*, is based on zero mean and unit variance. If these eigenvectors are applied to the matrix in which the fourth column is *not* in standard measure, the resulting X-matrix is *not* orthogonal.

The *renormalized* X-matrix is shown below, together with its mean and standard deviation vectors.

**Re-normalized X-Matrix**

| Treatment | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| 1 | 0.0656 | 0.3162 | -0.7482 | -0.4665 |
| 2 | -1.9030 | -1.5811 | 1.0332 | 0.2298 |
| 3 | -0.5906 | -0.3162 | -0.0356 | -1.2325 |
| 4 | 0.7218 | 1.5811 | -1.4608 | 1.6921 |
| 5 | 0.7218 | 0.9487 | 0.6769 | -0.6058 |
| 6 | 0.0656 | -0.9487 | 1.3895 | 1.4136 |
| 7 | -1.2468 | -0.3162 | -0.7482 | -0.4665 |
| 8 | 1.3781 | 0.9487 | 1.0332 | 0.2298 |
| 9 | 0.0656 | 0.3162 | -0.0356 | -1.2325 |
| 10 | 0.7218 | -0.9487 | -1.1045 | 0.4387 |
| Mean | 0 | 0 | 0 | 0 |
| Std Dev | 1 | 1 | 1 | 1 |

The correlation matrix is shown below.

**Correlation Matrix**

| Vector | $x_1$ | $x_2$ | $x_3$ | $x_3^2$ |
|---|---|---|---|---|
| $x_1$ | 1.0000 | 0.6687 | -0.1013 | 0.2523 |
| $x_2$ | 0.6687 | 1.0000 | -0.2879 | 0.0220 |
| $x_3$ | -0.1013 | -0.2879 | 1.0000 | -0.0157 |
| $x_3^2$ | 0.2523 | 0.0220 | -0.0157 | 1.0000 |

Note that correlation between $x_3$ and its square has disappeared, but there is another question to be considered. The correlation matrix shown above is no different than would have been obtained even if re-normalization had *not* been done, because the correlation computation automatically normalizes the input matrix to zero means and unit standard deviation. Consequently, the same eigenvectors and eigenvectors are obtained, as shown below.

**Eigenvectors**

| | | | | |
|---|---|---|---|---|
| $x_1$ | 0.6488 | 0.1964 | 0.2921 | -0.6747 |
| $x_2$ | 0.6537 | -0.2005 | 0.2587 | 0.6823 |
| $x_3$ | -0.3175 | 0.5486 | 0.7522 | 0.1801 |
| $x_3^2$ | 0.2257 | 0.7876 | -0.5310 | 0.2165 |

**Eigenvalues**

| | | | |
|---|---|---|---|
| 1.8111 | 1.0464 | 0.8727 | 0.2698 |

It is seen that now the fourth eigenvector contributes materially to the variation among treatments and that the dimensionality of the X-space is less degenerate than it was before the standardizing transformation. In this case, the transformed values in the third column are symmetrically distributed about zero, so that there is essentially no relation to the squared values.

However, if an attempt is made to "eigenize" the X-variables *without* renormalizing, the resulting X-matrix is found to be *not orthogonal*. But, upon renormalizing, the following X-matrix, representing the renormalized X-variables resolved into multipliers for the eigenvectors, is obtained.

**Re-normalized Matrix**

| Treatment | Vector 1 | Vector 2 | Vector 3 | Vector 4 |
|---|---|---|---|---|
| 1 | 0.3815 | -0.8284 | -0.2141 | -0.0642 |
| 2 | -2.5444 | 0.6910 | -0.3099 | 0.4409 |
| 3 | -0.8568 | -1.0428 | 0.3733 | -0.0905 |
| 4 | 2.3476 | 0.3560 | -1.3773 | 0.6950 |
| 5 | 0.7368 | -0.1541 | 1.2872 | 0.1511 |
| 6 | -0.6996 | 2.0787 | 0.0683 | -0.1353 |
| 7 | -0.8834 | -0.9594 | -0.7611 | 0.3897 |
| 8 | 1.2381 | 0.8283 | 1.3032 | -0.0466 |
| 9 | -0.0176 | -1.0407 | 0.7286 | -0.1017 |
| 10 | 0.2979 | 0.0715 | -1.0983 | -1.2383 |

As demonstrated below, that matrix *is* orthogonal, and the diagonal elements of the revised x'* x matrix are consistent with the eigenvalues computed from the correlation matrix:

$$9 * [1.8111 \quad 1.0464 \quad 0.9727 \quad 0.2698] = [16.2997 \quad 9.4174 \quad 7.8543 \quad 2.4285]$$

There *appears* to be a definite advantage in normalizing variables before squaring them. For example, we see that the correlation between the third and fourth columns is essentially eliminated.

However, appearances can deceive; upon further examination it is seen that the procedure brings with it some complications associated with the linear transformations employed and the fact that a *double* normalizing transformation is required. Clearly, this requirement adds further complication to decoding the data into the original, physical variables and to interpreting the regression results in terms of those variables.

The subtleties noted above are not to be ignored. More on this matter will be presented in the section H.4 Discussion.

### H.2.2 Adding a Product (Interaction) Term

Now consider an interaction term, introduced as the product of corresponding elements of column 2 and column 3.

| Treatment | $x_1$ | $x_2$ | $x_3$ | $x_2x_3$ |
|---|---|---|---|---|
| 1 | 7 | 4 | 3 | 12 |
| 2 | 4 | 1 | 8 | 8 |
| 3 | 6 | 3 | 5 | 15 |
| 4 | 8 | 6 | 1 | 6 |
| 5 | 8 | 5 | 7 | 35 |
| 6 | 7 | 2 | 9 | 18 |
| 7 | 5 | 3 | 3 | 9 |
| 8 | 9 | 5 | 8 | 40 |
| 9 | 7 | 4 | 5 | 20 |
| 10 | 8 | 2 | 2 | 4 |

First we compute the correlation matrix and note that columns two and three are positively correlated with column four, as might be expected.

Correlation Matrix

| | | | |
|---|---|---|---|
| 1.0000 | 0.6687 | -0.1013 | 0.5088 |
| 0.6687 | 1.0000 | -0.2879 | 0.4545 |
| -0.1013 | -0.2879 | 1.0000 | 0.6239 |
| 0.5088 | 0.4545 | 0.6239 | 1.0000 |

Upon computing the eigenvectors and eigenvalues of this matrix, we see that the fourth eigenvector contributes little to the variation among treatments.

Eigenvectors

| | | | |
|---|---|---|---|
| 0.5894 | -0.2245 | 0.7643 | -0.1340 |
| 0.5591 | -0.3712 | -0.6133 | -0.4165 |
| 0.1221 | 0.7938 | 0.0347 | -0.5948 |
| 0.5701 | 0.4262 | -0.1962 | 0.6744 |

Eigenvalues

| | | | |
|---|---|---|---|
| 2.1054 | 1.4982 | 0.3283 | 0.0681 |

Now suppose, however, that we first transform the three original variables to standard form and then compute, as the fourth column of the X-matrix, the products of corresponding elements in columns two and three.

| | $x_1$ | $x_2$ | $x_3$ | $x_2x_3$ |
|---|---|---|---|---|
| 1 | 0.0656 | 0.3162 | -0.7482 | -0.2366 |
| 2 | -1.9030 | -1.5811 | 1.0332 | -1.6337 |
| 3 | -0.5906 | -0.3162 | -0.0356 | 0.0113 |
| 4 | 0.7218 | 1.5811 | -1.4608 | -2.3097 |
| 5 | 0.7218 | 0.9487 | 0.6769 | 0.6422 |
| 6 | 0.0656 | -0.9487 | 1.3895 | -1.3182 |
| 7 | -1.2468 | -0.3162 | -0.7482 | 0.2366 |
| 8 | 1.3781 | 0.9487 | 1.0332 | 0.9802 |
| 9 | 0.0656 | 0.3162 | -0.0356 | -0.0113 |
| 10 | 0.7218 | -0.9487 | -1.1045 | 1.0478 |

The correlation matrix, shown below, shows that the correlations of $\underline{x}_2$ and $\underline{x}_3$ with their product is essentially eliminated.

Correlation Matrix

| | | | |
|---|---|---|---|
| 1.0000 | 0.6687 | -0.1013 | 0.3523 |
| 0.6687 | 1.0000 | -0.2879 | 0.0557 |
| -0.1013 | -0.2879 | 1.0000 | 0.0141 |
| 0.3523 | 0.0557 | 0.0141 | 1.0000 |

Similarly, upon computing the eigenvectors and eigenvalues of this matrix, one sees that the added column now contributes appreciably to the variation among treatments.

Eigenvectors

| | | | |
|---|---|---|---|
| 0.6544 | 0.1909 | 0.2441 | -0.6898 |
| 0.6295 | -0.2399 | 0.3456 | 0.6532 |
| -0.2853 | 0.6531 | 0.6848 | 0.1524 |
| 0.3068 | 0.6925 | -0.5933 | 0.2726 |

Eigenvalues

| | | | |
|---|---|---|---|
| 1.8526 | 1.0911 | 0.8063 | 0.2499 |

As in the case of the quadratic term, these changes are the result of the symmetry of the transformed values of columns 2 and 3.

It appears, therefore, that it would be advantageous to transform variables to standard measure before introducing a square term. As in the case of a quadratic basis element, however, the *double* normalizing transformation complicates interpretation.

Such double transformations are not required if normalization is delayed until *after* the product vector has been computed. Though other functional forms may require adaptations peculiar to the form of the function, for the most part, the general procedure described for quadratic and product terms can be expected to apply.

## H.3 INTERPRETATION

Interpreting the role played by a nonlinear element that is incorporated in *each* of the basis vectors may seem to pose a serious problem to the data analyst. It is the intent of this section of this document to dispel some of the trepidation associated with this problem.

Because of the correlation of the nonlinear variable with one or more of the *linear* variables, the nonlinear term may be significant and substantial in *some* eigenvectors but not in others, just as in a purely linear model. The effect is most likely to be induced by correlation between the nonlinear function and the argument or arguments of that function. However, it is entirely possible for the nonlinear vector to be associated with variables *other* than its

argument, in which case there would be aliasing that remains unsuspected without eigenvector decomposition.

A measure of the degree to which the nonlinear variable is associated with the eigenvectors of which it is a component can be had by examining the correlation between the nonlinear term and each of the eigenvectors. This measure should help identify the eigenvector or eigenvectors by which the nonlinear component shows its effect. That, combined with physical reasoning, should provide insight into the mechanism(s) giving rise to the nonlinear effect.

## H.4 DISCUSSION

The apparent dichotomy of pre- and post-normalization can be explained in terms of how nonlinear functions are affected by linear transformation. Two facets of the analysis require further explanation: (1) linear transformations in a nonlinear world, and (2) the non-specificity of sum-of-squares partitionings.

Let x be a variable to be standardized, let f(x) be some function of x, and let T(x) be a linear transformation of x. For standardization (zero mean, unit variance) of the variable x, the transformation T takes the following form:

$$T(x) = (x - m)/s = ax + b$$

where a = 1/s, b = -m/s, m being the sample mean of x and s being the sample standard deviation.

A basic question to be considered is this: under what conditions does $T( f(x) ) = f( T(x) )$? This question is the root of such observations as: (1) why the mean square of a set of integers is not the same as the square of their mean; or (2) why one cannot average a set of vulgar fractions by separately averaging their numerators and denominators.

Suppose that f(x) is a linear function of x, of the form:

$$f(x) = cx + d$$

Then,

$$f( T(x) ) = c (ax + b) + d = (ac) x + (bc + d)$$

which is still of the form Ax + B, where A and B are constants.

Further,

$$T( f(x) ) = a (cx + d) + b = (ac) x + (ad + b)$$

which, again, is of the linear form Cx + D.

Suppose, now, that f(x) is a nonlinear function, specifically $f(x) = x^2$. Then,

$$f( T(x) ) = (ax + b)^2 = a^2 x^2 + 2 abx + b^2$$

whereas,

$$T( f(x) ) = ax^2 + b$$

It is this difference between the two sequences that is the root of the two approaches to incorporating a nonlinear basis element in a regression model. In fact, the "standardize first, then square" approach suffers a "double whammy." After the variable has been standardized, it has to be *restandardized* in order to comply with the mean zero, unit standard deviation measure expected in the correlation matrix. Though both of these standardizing transformations can be retraced to the original variables, their results are not particularly transparent.

If standardization is performed *after* squaring, however, one obtains a straightforward result and one that is consistent with the treatment of the linear terms in the model:

$$f(x) = x^2$$

and

$$T( f(x^2) ) = ax^2 + b$$

Interpretation of eigenvectors containing one or more nonlinear elements presents a challenge to the data analyst. On the other hand, it forces him or her to recognize that the effect of the nonlinear element, either by chance or by design, may be associated with other variables in a way not anticipated and not evident without eigenvector resolution. Further, it may encourage the analyst to perform a thoughtful examination of variables during the model-simplification process.

The considerations presented above apply equally to interaction effects, expressed as a weighted *product* of two variables. Moreover, quadratic and product terms do not exhaust the variety of nonlinear elements that might be included in a regression model. Extremely complex terms are known to have been included in a model[1], though their *raison d'etre* may not be apparent. Our view is that a regression model should be as simple as possible and that

---

[1] Hewlett, R.F. *Computer Methods in Evaluation, Development, and Operations in an Ore Deposit.* American Institute of Mining, Metallurgical, and Petroleum Engineers, Inc. Dallas, TX, February 1963. Preprint No. 63151.

nonlinear effects may be better modeled by more conventional approaches such as transformation of variables.

It could well be that the most valuable contribution PCR can make is as a variable-screening procedure in a manner not unlike existing "screening designs," in which the primary purpose of the model is to identify predictor variables but not necessarily how they interact in their effects on response.

Clearly, more experience is needed in the application of nonlinear elements in regression analysis. In addition, one needs to consider carefully the prospect and consequences of nonlinear transformations, such as logarithmic, to either or both the predictor and response variables. This approach to nonlinearity is covered in Appendix I.

# APPENDIX I: TRANSFORMATION AS AN APPROACH TO NON-LINEAR REGRESSIONS

One of the assumptions of least squares regression is that errors are normally distributed and homoscedastic. Since regression deals with estimating the average response to the predictor variables, the errors are likely to be approximately normally distributed as a consequence of the very powerful Central Limit Theorem. Homoscedasticity, though, is another matter. If errors get larger or smaller as we go from lower to higher emissions, we are likely to get biased estimations if we assume that the error variance is constant over the whole range of emissions.

Usually, we make transformations for one or more of several reasons:

1. To linearize or otherwise simplify the form of the regression equation.

2. To "stabilize" the error variance – that is, to change a heteroscedastic error distribution to one that is at least approximately homoscedastic.

3. To produce a result that is more consistent with real-world experience.

Linearization is a convenient technique for converting a non-linear equation to one of linear form, so much so that one is tempted to make that conversion without regard to its consequences with regard to (2) and (3). Similarly, error stabilization may be indulged in quite legitimately but could distort the regression curve, surface or hypersurface to such an extent that (3) is violated.

Finally, whether a result is "consistent with real-world experience" is often a judgment call. However, there may be good reason to believe that the effect of an incremental change in a predictor on the response is not constant but is proportional to the value of the response at which the increment is applied. Again, a transformation made to implement that nugget of experience may not necessarily be consistent with (1) and (2).

In short, in making a transformation, we often deal with conflicting requirements and are lucky if we *do not* get an effect that we *do not want* along with one that we do. For example, if our error distribution really is homoscedastic and we make a log transformation for purposes of – say – linearization, we will produce a heteroscedastic distribution that works to our disadvantage, so far as goodness of fit is concerned, and that may yield biased predictions.

These concerns are in addition to concerns with regard to the scaling or re-parameterization of data by such means as removing the mean or normalizing to units of standard deviation

about the mean. Table I.1 below summarizes properties of transformation often applied to data, particularly in the area of emission and fuel-economy data.

| Table I.1. Typical Transformations and Associated Metameters | | | |
|---|---|---|---|
| **Transformation** | **Equation Form** | **Differential Effect** | **Error Weighting** |
| 1. Identity | $y = a + b\,x$ | $dy = b\,dx$ | $s^2 = k$ |
| 2. Mean removal | $y = a + b\,(x - m)$ | $dy = b\,dx$ | $s^2 = k$ |
| 3. Rescaling 1 | $y = a + b\,(x - m) / s$ | $dy = (b/s)\,dx$ | $s^2 = k$ |
| 4. Rescaling 2 | $y = a + b\,(x - m) / r$ | $dy = (b/k)\,dx$ | $s^2 = k$ |
| 5. Log y | $y = a\,e^{bx}$ | $dy/y = b\,dx$ | $s^2 = k / m^2$ |
| 6. Log x, log y | $y = a\,x^b$ | $dy/y = b\,dx/x$ | $s^2 = k / m^2$ |
| a, b, r, k = Arbitrary constants<br>s = Sample standard deviation<br>m = Sample mean | | | |

Transformations (1) and (2) are typically used in Ordinary Least Squares (OLS) regression, either with or without logarithmic transformation of the dependent variable. Removing the mean makes the x-variables symmetric about zero and may help to minimize computational difficulties. Transformation (2) was generally used in developing the Complex Model for Reformulated Gasoline.

Transformation (3) reduces data to so-called "standard scores" as used in psychological testing and has the advantage, in all applications, of placing all predictor variables on a *level playing field* in that all data in the design matrix span essentially the same range. Otherwise, there might be a large diversity of scale among the predictor variables, and this diversity could lead to misinterpretation. On the other hand, unless the original range of each predictor is representative of the viable range of that variable in the real world, a predictor that was varied only over a fraction of its real range could seem to have an exaggerated effect. That exaggeration comes about simply because its short range was reduced to the same scale as that of other predictors that were varied over a much larger range relative to what is realistically feasible.

Transformation (4) offers a way out of this difficulty by using, rather than the standard deviation of the sample, some number *r* more representative of the range of variation at the disposal of the experimenter. This is the type of scaling used in applying random balance methodology to assess the relative importance of predictor variables as measured by partitioning the sum of squares induced by random sampling over the design space.

Transformation (5) is probably the one most frequently used in regressions of emissions and fuel economy on fuel or vehicle predictors. It is widely believed that a unit increment in a predictor – say, one point in cetane number – has a larger effect when emissions are high than when small. As shown in Table I.1, logarithmic transformation of response accomplishes that end. It also is an easy dodge of negative emission predictions. Perhaps most important of all, though, it linearizes the equation so that all the theory of the General Linear Model is applicable.

It will be insightful to examine the mathematics involved in transformations (5) and (6). Consider an equation of the form

$$y = e^{a+bx}$$

which transforms to

$$\log y = a + b\,x$$

Then,

$$dy/dx = b\,e^{a+bx} \quad \text{or} \quad dy/y = b\,dx.$$

Thus, it is implied that an incremental change in the predictor variable produces a constant percent change in the response.

Now consider an equation of the form

$$y = a\,x^{b}$$

Logarithmic transformation gives

$$\log y = \log a + b \log x$$

Then,

$$dy/dx = a\,b\,x^{b-1} \quad \text{or} \quad dy/y = b\,dx/x.$$

In other words, a percent change in the predictor variable produces a percent change in response.

Next, let us consider how transformations (5) and (6) are related to error variance homogeneity-heterogeneity. A good approximation to the effect of a transformation on the variance of a random variable Y is provided by the equation:

$$\text{Var}[\,g(y)\,] = \text{Var}[\,Y\,] * \{\,g'(\,E[Y])\,\}^2$$

where g is the transforming function and g' is its derivative. For example, if we define g(y) as

$$g(y) = \log y$$

then

$$g'(y) = -1/y$$

and

$$\text{Var}[\log(Y)] = k\, m^2 / m^2 = k$$

Thus, the logarithmic transformation, in addition to linearizing the regression model, also "stabilizes" the variance and satisfies the least-squares assumption of homoscedasticity.

Suppose, however, that the response variance really is homoscedastic. In that case, the logarithmic transformation actually produces heteroscedasticity – in fact, the response error variance is inversely proportional to the square of the mean. Because the least-squares algorithm attempts to minimize the sum of all squared deviations from the regression line, it "tries harder" when the mean response is low than when the mean response is high. The result is that the lower end of the curve is weighted more heavily than the higher end. A possible result is that estimation may be biased relative to a regression equation for which errors are equally weighted over the range of the response variable.

A demonstration of this possibility is afforded by the following textbook example, in which both the predictor and response variables are log-transformed. The data set, obviously contrived for demonstration, is shown as Table I.2.

**Table I.2. Demonstration Data Base**

| Predictor Variable x | Response Variable y |
|---|---|
| 1 | 2.5 |
| 2 | 8.0 |
| 3 | 19.0 |
| 4 | 50.0 |

The results of regressing log(y) on log(x) are summarized below in Table I.3. In antilog space the regression equation takes the form

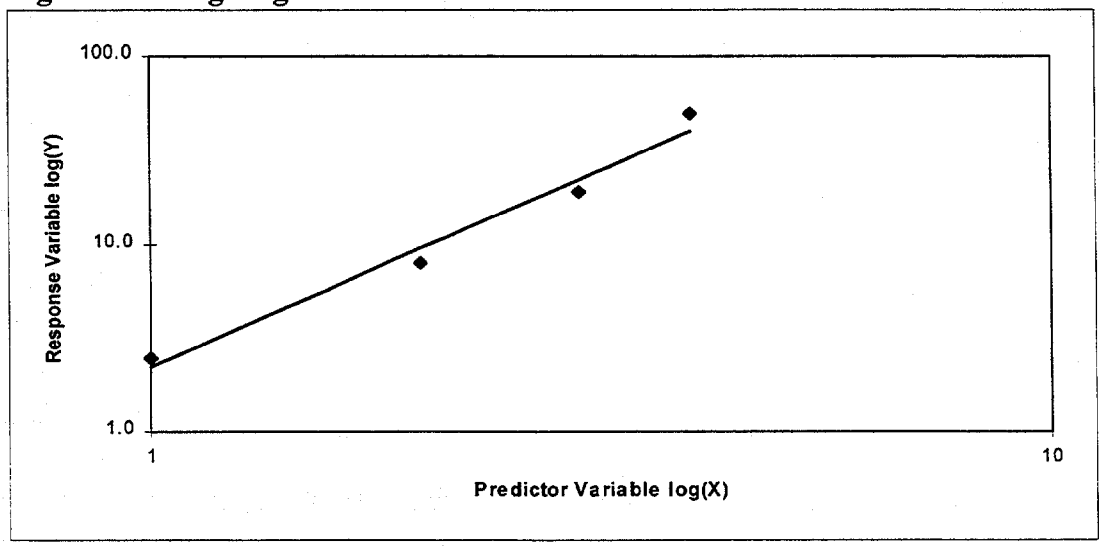
$y = 2.224\ x^{2.096}$

| Table I.3. Regression of log(y) on log(x) | | | |
|---|---|---|---|
| **Regression of log(y) on log(x)** | | | |
| | **Regression Coefficient** | **Coefficient Std. Error** | **t Ratio** |
| **Intercept** | 0.7984 | 0.2151 | 3.7115 |
| **log(x)** | 2.0952 | 0.2264 | 9.2527 |
| **Analysis of Variance: $R^2 = 0.9772$** | | | |
| | **SS** | **Df** | **MS** |
| **Model** | 4.7597 | 1 | 4.7597 |
| **Error** | 0.1112 | 2 | 0.0556 |
| **Total** | 4.8709 | | |

Figure I.1 is a scatter plot of log(y) versus log(x) with the line representing the fitted equation superimposed. The algebraic sum of the residuals in log space is zero, and it is readily seen that the log(y) observations are essentially symmetrically disposed about the least squares line.

One can compute the residuals in antilog space by taking the differences between the exponentiated logarithms of the observations and the corresponding exponentiated logarithms as computed from the regression equation:

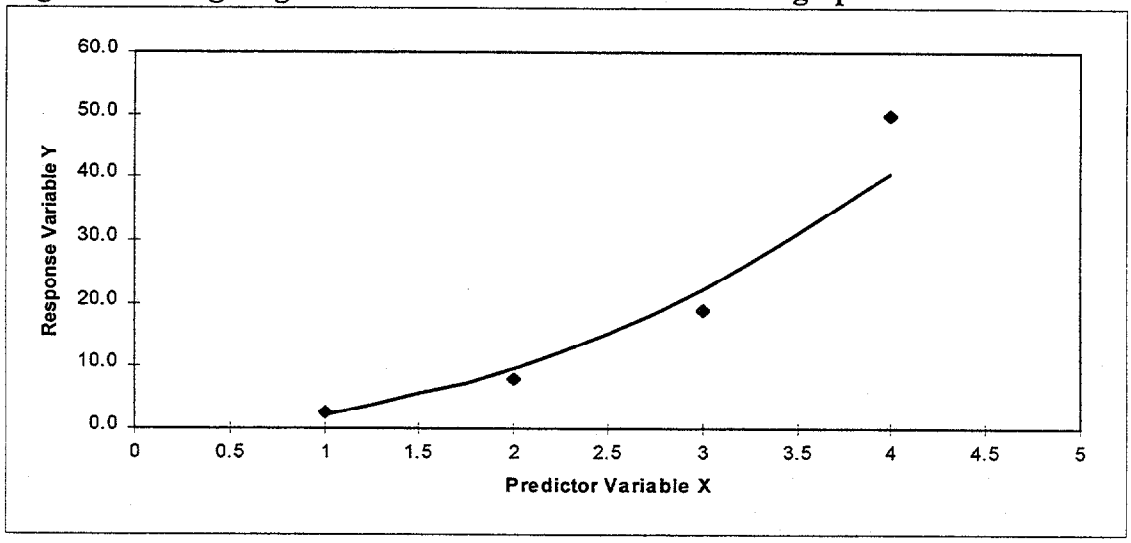
**Figure I.1. Log-Log Fit to Demonstration Data**



```
Sum of residuals          =      4.88
Mean residual             =      1.26

Residual SS
   Corrected for mean     =    100.30
   Not corrected          =    101.54
```

Considering that the regression equation displayed $R^2 = 0.98$, the residual sum of squares seems unreasonably high. Moreover, the sum of the residuals is not zero, as it should be for a well-fitted equation. The nature of the apparent misfit is shown clearly in Figure I.2. The heteroscedastic error distribution induced by the log transformation has placed greater weight on the lower end of the curve, resulting in a large residual error at its upper end.

**Figure I.2. Log-Log Fit to Demonstration Data in Antilog Space**




Other means exist for minimizing the sum of squares in exponentiated space, but they do not permit the nice linearization provided by logarithmic transformation. For example, starting with the parameters estimated from the log-log regression, one can explore other candidate

values of intercept and exponent in the vicinity of these initial estimates. To do so, one simply computes, for each candidate set, the sum of the residuals and the sum of the squared residuals and seeks that set of parameters that best satisfies the criterion of minimum sum of squares, with or without the additional requirement that the algebraic sum of the residuals be zero. It is evident, as well, that other "loss functions" can be used, such as minimizing the least absolute deviation, as might be preferred in some applications. One can draw on existing software for performing such nonlinear regression or can develop simple MatLab algorithms for that purpose.

Two solutions obtained by the above approach are compared below with the initial estimate as obtained by least-squares regression of log(y) on log(x):

| Intercept | Exponent | Sum of Residuals | Residual SS |
|---|---|---|---|
| 2.224 | 2.096 | 4.8763 | 100.3 (*) |
| 0.7385 | 3.0343 | 2.4395 | 10.0012 |
| 0.8950 | 2.9042 | 0.00 | 11.86 |

* From least-squares regression of log(y) on log(x).

The two "improved" solutions assume that the errors are homoscedastic, whereas the initial log-log approach, by its very nature, treats errors as if they change inversely as the magnitude of the response.

The first of the alternative solutions, the result of seeking to minimize the residual SS in antilog space, produces a residual SS value of 10.0012 versus 100.3 for the anti-log equivalent of the log-log fit, but it exhibits a continued bias as indicated by the non-zero sum of residuals. The second of the alternative solutions attempts to find the smallest residual SS among the solutions that have a zero sum of residuals. Its residual SS, at 11.86, is only slightly larger than that of the first alternative, and its sum of residuals is zero to two decimal places.

The second of these two alternatives, which could be viewed as the best, unbiased fit to the demonstration data in antilog space, is shown graphically in antilog and log-log forms in Figure I.3 and Figure I.4, respectively. The curve closely fits the demonstration data in antilog space and distributes its residuals evenly between plus and minus. But, optimizing the fit for antilog space necessarily implies an apparent misfit in log-log space, such that the points lowest on the curve are given lesser weight compared to those higher on the curve.

Which solution is to be believed? In spite of what may seem to be a poor fit in Figure I.1, the log(y) versus log(x) solution would be more appropriate than the others if the error distribution is actually such that the variance is proportional to the square of the mean.

**Figure I.3. Antilog of $y = 0.8950\ x^{2.9042}$**

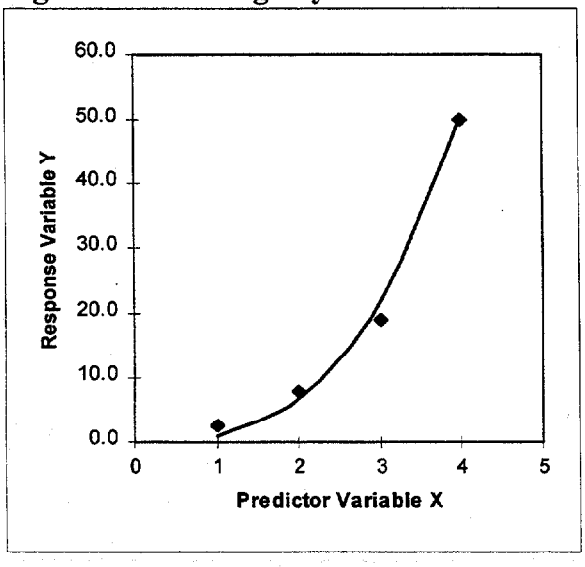
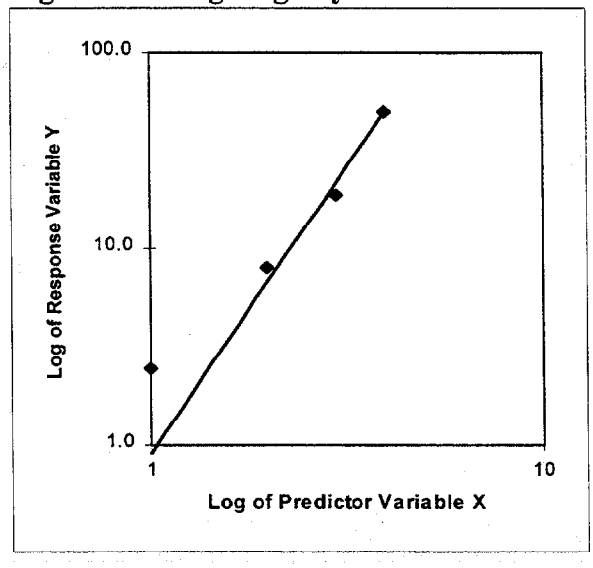



**Figure I.4. Log-Log of $y = 0.8950\ x^{2.9042}$**



However, if the errors are actually homoscedastic, the log transform version, in spite of its convenience, would not be appropriate, and another solution, directly optimized for antilog space, would be preferred.

The advantages and limitations of transformations are as relevant to vector-based regression as they are to conventional regression analysis. Every effort should be made to assure that the assumptions made with regard to error distribution and model correctness are appropriate before proceeding with regression analysis in either its conventional form or the modified PCR approach advanced in this report.

This report does not specifically address the effect of transformations on the modified version of PCR. Suffice it to say, however, that the same advantages, limitations, and sometime dilemmas apply in PCR as in conventional regression analysis. Inasmuch as transformation decisions are made at the P-Space level, any errors of judgment made at that level will be propagated to E-Space and will vitiate any advantages that might be gained by the eigenvector approach.

## INTERNAL DISTRIBUTION

1. G.E. Courville
2. T.R. Curlee
3. S. Das
4. S.C. Davis
5. Ronald L. Graves
6. D.L. Greene
7-31. G.R. Hadder
32. P.S. Hu
33. R.N. McGill
34. C.I. Moser
35. R.B. Shelton
36. Central Research Library
37. Laboratory Records

## EXTERNAL DISTRIBUTION

38. L.A. Abron, President, PEER Consultants, P.C., 1460 Gulf Blvd., 11th Floor, Clearwater, Florida 34630
39. Susan L. Cutter, Professor and Chair, Director, Hazards Research Lab, Department of Geography, University of South Carolina, Columbia, South Carolina 29208
40. S.G. Hildebrand, Director, Environmental Sciences Division, Oak Ridge National Laboratory, Post Office Box 2008, Oak Ridge, Tennessee 37831-6037
41. P.R. Rittelmann, FAIA, Executive Vice President, Burt Hill Kosar Rittelmann Associates, 400 Morgan Center, Butler, Pennsylvania 16001-5977
42. S.F. Tierney, The Economic Resource Group, Inc., One Mifflin Place, Cambridge, Massachusetts 02138
43. C.M. Walton, Ernest H. Cockrell Centennial Chair in Engineering and Chairman, Department of Civil Engineering, University of Texas at Austin, E Cockrell, Jr. Hall I, Suite 4210, Austin, Texas 78712-1075
44-49. Wendy Clark, National Renewable Energy Laboratory, 1617 Cole Blvd., Golden, CO 80401-3393.
50-59. R.W. Crawford, RW Crawford Energy Systems, 2853 S. Quail Trail, Tucson, AZ 85730-5627
60. P.R. Devlin, EE-32, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
61. K.G. Duleep, Energy and Environmental Analysis, Inc., 1655 North Fort Meyer Drive, Arlington, VA 22209
62-63. Robert G. Dulla, Sierra Research, 1521 I Street, Sacramento, CA 85914-3009.
64. J.A. Garbak, EE-32, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585

65. S.J. Goguen, EE-33, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
66. Michael S. Grabowski, Dept. of Chemical Engineering, Colorado School of Mines, Golden, CO 80401.
67. A.M. Hartstein, FE-3, U.S. Department of Energy, Germantown, 19901 Germantown Road, Germantown, MD 20874-1290
68. Robert Mason, Southwest Research Institute, 6220 Culebra Road, San Antonio, TX 78228-0510.
69-78. H.T. McAdams, AccaMath Services, Carrollton, IL 62016
79-83. B.D. McNutt, PO-62, Room 7H-021, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
84. W. Stuart Neill, National Research Council Canada, Montreal Road, Building M-9, Ottawa, Ontario, K1A 0R6.
85. P.D. Patterson, Jr., EE-30, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
86. Richard A. Rykowski, Assessments and Standards Division, U.S. Environmental Protection Agency, 2000 Traverwood Drive, Ann Arbor, MI 48105
87. Gurpreet Singh, EE-33, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
88. M. Singh, Argonne National Laboratory, 955 L'Enfant Plaza, SW, Suite 6000, Washington, DC 20024-2168