Underlying Construct Analysis (UCA):

Where Individual Systems End and Human Systems Begin

Alvin L. Killough, North Carolina Central University
Christopher L. Edwards, Duke University Medical Center
Donald W. Drewes, North Carolina State University

ABSTRACT

The association of antecedents to performance outcomes, including individual behaviors, interfaces and the environments in which engagements are performed and sustained as a behavioral system is certainly implied, but has not been a primary focus of traditional data mining techniques. The current paper introduces the concept of Underlying Construct Analysis (UCA) using techniques associated with confirmatory factor analysis, and demonstrates their utility for theory development. We provide step-by-step instructions for data mining beyond observable variables, to include larger units of behavior frameworks and demonstrate using a case example, unique variance associated with two distinct populations in a large data set (*Survey of Parents and Children*, 1990), populations that were not defined prior to the UCA. We lastly emphasize the potential global application of our proposed procedures for individuals with only moderate levels of understanding of statistics, and across multiple domains of exploration.

INTRODUCTION

Data mining is an analytical process, but moreover a reflection of traditional thinking designed to explore large aggregates of data in search of consistent patterns of relationships among variables. For many years, the concept of data mining has included the notion of exploring observable factors in three basic stages: (1) data exploration; (2) model building or pattern definition; and (3) model validation. The authors of the current paper argue that the utility of data mining techniques for many researchers in medicine, engineering, business, and the social sciences can be greatly enhanced by the use of multiple additive regression analyses that elucidate underlying constructs in the prediction of observed factors. Required is the re-framing of knowledge discovery, first at the conceptual level to include objective and subjective indicators of outcomes as geo-physical, -psychological, -physiological, or -sociological phenomena. Second, at the procedure level by taking advantage of rapid advancements in data mining techniques to better leverage pattern recognition. In the true spirit of "mining" defined as extracting underlying wealth, the current paper will present a rationale supporting SAS code and a case example of the utility of Underlying Construct Analysis (UCA). This new procedure, when combined with advanced data visualization techniques, can assist researchers to attain maximum meaning from their data.

We propose using UCA to explore the underlying systematic coherence that ties people, events, and situations together in a meaningful manner that drives measurable relationships. The justification of UCA using a mining analogy can be conceptualized as “why focus on nuggets of data when you can excavate core structure?” Because UCA focuses on mining beyond the observable, we believe that UCA provides data miners a more reliable representation of the relationship between variables in a large data set, and a great foundation from which to develop, not just respond to, theory. Lastly, we believe that UCA has increased utility over traditional data mining approaches because of its relative ease of application to multiple domains of science, medicine and industry, and the need for only a moderate level of knowledge of statistics to quickly and effectively gain benefit in the data mining process.

The current paper is organized to provide step-by-step instructions for the use of UCA procedures, and, using an example from the social sciences to guide the reader’s thinking and learning processes, each data mining procedural step is conceptualized in detail, along with specific examples of SAS code. The paper is designed to provide readers an enhanced understanding of the use and interpretation of UCA as a data mining tool. Code segments, and in some instances exemplary programs, are provided for user convenience. No specialized knowledge is required. For a supplementary reference, the reader is referred to Hatcher (1994).

DATA MINING USING UCA

In the following sub-sections, we describe a series of procedural steps required to mine a raw data set for the underlying factorial associative network implied by the covariances. For extended generalizability, we propose techniques that can be used to induce and compare the associative networks of multiple groups. Because UCA draws heavily on the theory and methodology of Confirmatory Factor Analysis (CFA), we couch much of the procedural description in CFA terminology.

SELECT DATA SET

In recognition of the CFA heritage, the observed data variables are referred to as manifest variables. Manifest variables may be drawn from either primary or secondary sources or a mixture of both. Regardless of source, the analyst should ensure that a complete data set is available for each of the groups to be modeled. The use of proxy variables across groups under the assumption of measurement equivalency is to be avoided, as CFA is extremely sensitive to violations of this assumption.

The presence of nonnormal manifest variables influences the tests of statistical significance used in CFA modeling. We therefore recommend that the data set for each group be simultaneously tested for skewness and kurtosis using multivariate procedures as exemplified by Mardia (1974), Mardia and Foster (1983) and Mardia (1985). In cases of significant departure from normality, the analyst has the following choices: transforming the data in hopes of achieving better approximations of multivariate normalcy, removal of outliers, or employing alternative parameter estimation techniques. Commonly employed transformations include logarithmic, square root, and other power functions that alter distributional shape. The stem-leaf tables produced by PROC UNIVARIATE are recommended for identification of outlier cases. Covariance matrices with and without outliers can be computed and compared to ascertain outlier effect. A more drastic solution is to use an available alternative estimator that does not depend upon multivariate normalcy (Bentler, 1989).

Missing data is an ongoing problem that is typically dealt with by a) pairwise deletions, b) listwise deletions, or c) estimation of missing values. Pairwise deletion is to be avoided in CFA modeling, as negative definite covariance matrices are possible and, when present, will abort the estimation procedure. Therefore, we recommend that the nomiss option be used in the PROC CORR statement. However, in some cases, listwise deletion can result in severe attenuation of sample size. As a result, the use of a reduced sample leads to less efficient CFA parameter estimates than would be obtained with the full sample (Bollen, 1989). Use of sample means or linear estimates has pitfalls, including increased possibilities of violation of multivariate normalcy assumptions and heterogeneity of error variances.

For the research herein reported, the data set was drawn from the *Survey of Parents and Children* (1990). This secondary source was selected because it provided a national sample of antecedent variables with potential relevance for educational achievement of youth from white and black families. Development of separate CFA models for black and for white families allows for a structural comparison of cross-cultural effects. Data screening and cleaning resulted in the selection of a data set consisting of 21 variables for each group.

DRAW RANDOM SAMPLE

At the present stage of development, CFA parameter estimation requires simple random sampling from each of the k population groups under modeling consideration. Current advances such as PROC SURVEYREG and PROC SURVEYSELECT suggest that procedures may soon be available to allow survey researchers to conduct CFA using complex sampling designs.

For our research project, we defined the sample to be the set of all observations having no missing data on the selected data set. As a result, we obtained a sample of 193 white families and 117 black families. To ensure that the reduced sample did not differ from the full sample, we compared the reduced and full samples on selected demographic characteristics. No significant differences were obtained. From this, we concluded that missing values were distributed randomly and thus that our selected sub- samples could be said to be representative of the full sample for each population group. Consistent with CFA requirements, we further assumed that each sample could be construed as a simple random sample drawn from a population as defined by the respective sampling frames used in the overall survey design.

CONDUCT EXPLORATORY FACTOR ANALYSIS

An initial exploratory factor analysis is helpful in making the preliminary assignment of manifest data set variables to an underlying factor, referred to as a latent variable in CFA terminology. A suggested command line is as follows:

```
   proc factor data=groupdataset priors=smc
   min=1 scree rotate=varimax;
```

These options instruct the procedure to select the appropriate date set for the group to be analyzed, use the prior multiple correlation of each manifest variable with all others as the communality estimate for that variable, retain only those factors whose eigenvalues equal or exceed 1, produce a scree plot of the retained eigenvalues, and rotate the factor loadings using an orthogonal transformation. Orthogonal rotation is preferred because its use tends to reduce the number of manifest variables that load on more than one factor.

Several heuristics are offered to guide the analysis. First, with respect to sample size, we concur with Hatcher's (1994) advice that the sample size should exceed either 100 or 5 times the number of manifest variables in the data set, whichever is greater. The requirement that only factors whose eigenvalues exceed 1 be retained for further consideration is admittedly arbitrary but commonly employed in the factor analytic literature. Use of the assignment rule that manifest variables are assigned to an underlying factor only if the absolute value of its rotated factor loading exceeds .40 enjoys wide usage. Manifest variables whose rotated loadings are in excess of .40 on several variables warrant special attention. These cases are termed cross-loaders and should initially be assigned to the factor with the highest rotated loading.

When applied to our data, exploratory factor analysis yielded three factors for the white family sample and four factors for the black family sample that met the minimum eigenvalue criterion. For the white family sample, 6 manifest variables were assigned to factor 1 (f1), 6 manifest variables to factor f2, and 5 manifest variables to factor f3, with no cross-loaders. In contrast, for the black families, 4 factors met the minimum eigenvalue criterion, with 4 manifest variables assigned to the first factor, 5 to the second factor, 4 to the third factor, and 5 to the fourth factor, with one manifest variable loading on two factors.

TEST THE SAMPLE COVARIANCE MATRICES FOR EQUIVALENCY

The fundamental hypothesis of CFA modeling is that the population covariance can be reproduced as a function of a set of parameters as represented by the matrix equation

$$\Sigma_l = \Sigma_i(\theta_l)$$

where Σ_l is the covariance matrix and θ_l is the vector of structural parameters for the i^{th} population. Thus, if $\Sigma_l = \ldots = \Sigma_{\mathbf{k}}$**,** the underlying structure as represented by the vector of model parameters must also be equivalent. What is needed, then, is a statistical test for equality of covariance matrices using k random samples, one from each population.

A SAS® IML program of a likelihood ratio test of the hypothesis of equality of covariances as described by Anderson (1984) is presented in the Appendix to this paper. The numerator of the likelihood ratio is proportionate to a power of the weighted geometric means of the sample generalized variances and the denominator is proportionate to a power of the determinant of the arithmetic means of the sample covariance matrices. The likelihood ratio criterion is asymptotically distributed as chi-square.

Use of this test criterion requires that the data sets for the k samples contain the same manifest variables. Sample sizes need not be equal. The common data set should include all manifest variables retained as a result of the exploratory factor analysis conducted separately for each sample. Covariance matrices can be computed separately for each sample using the command line

```
   proc corr data=commondatasetsamplei cov
   nocorr outp=sampleicovariance;
```

The cov option instructs the procedure to compute and print a covariance matrix. The nocorr option suppresses printing of the correlation matrix. The outp option creates a new SAS® data set containing the sample covariance matrix. The sample covariance matrices are used as input to the covariance equality test. Acceptance of the null hypothesis H_0: $\Sigma_l = \ldots = \Sigma_{\mathbf{k}}$ implies an equivalency of underlying parametric structure; hence, pooling of sample covariance matrices is permissible. Pooling is achieved by summing the corrected sums of cross-products matrices across the k samples and dividing each matrix entry by N – k. If H_0 is rejected, the covariance matrices must be analyzed separately for each sample in subsequent procedural steps.

As a result of exploratory factor analyses of our cross-cultural samples, 19 variables from the original set of 21 were retained for further analysis. Two manifest variables were unique to the black families sample and 1 to the white family sample. One common variable was found to cross-load on the black family sample but not on the white family sample. The presence of unique variable assignments suggests nonequivalent sample covariances. A common data set of 19 variables was created and covariance matrices computed separately for each sample. Testing the two covariance matrices for equality using the program as presented in the Appendix yielded $\chi^2 = 287.8$, which with 190 df resulted in p < .0001. The hypothesis of no differences in population covariances was consequently rejected in favor of the alternative hypothesis of differential parametric structure.

CREATE MEASUREMENT MODELS

The purpose of this step is to select subsets of manifest variables that are pure measures of a single latent variable. Sets of manifest variables that measure a single underlying construct (latent variable) are referred to as congeneric variables (Drewes, in press). The goal is to identify congeneric variables for every factor identified in the exploratory analysis conducted for each sample. The obvious candidates for initial consideration are the manifest variables assigned to a specific factor. These variables are submitted as a preliminary CFA model for confirmation. If the CFA model results in an acceptable fit of the covariances of the included variables, the measurement model is accepted. If not, manifest variables are sequentially dropped, and the reduced model refit until an acceptable model is found or the number of remaining variables is less than 4. With less than 4 manifest variables, model fit cannot be tested.

Measurement models are tested using PROC CALIS. The following prototype program is suggested:

```
   proc calis data=reducedcov edf=samplesize-1
   corr method=ml res se;
      lineqs
        v1 = gamma1 f1 + e1,
        v2 = gamma2 f1 + e2,
        .
        .
        vp = gammap f1 + ep;
      std
        e1 – ep = error: (p * .1),
        f1 = 1.0;
   run;
```

In the above code, v(i) refers to a manifest variable name, with p being the number of manifest variables assigned to the factor designated for notational simplicity as f1. Before the program is submitted, the appropriate integer value should be substituted for p. The measurement model is defined in the lineqs (line equations) section. Each manifest variable is assumed to be decomposable into an underlying true component designated as f1 and a measurement error component designated as e(i). The latent variable (f1) is weighted by a constant, gamma(i), which in a standardized model is interpreted as the correlation of the manifest variable with the underlying true component f1. The free parameters to be estimated are the p gamma values and the p error variances. As the variance of f1 is arbitrarily fixed at 1.0, these parameters are sufficient to estimate the observed covariance matrix. For further discussion and clarification, see Hatcher (1994).

The CFA measurement model when run returns 30 indices of model fit. For simplicity of interpretation, we recommend primary consideration be given to the following: a) the fit function, b) Pr > Chi-Square, c) RMSEA Estimate, and 4) Bentler & Bonett's (1980) NFI. In order to qualify as acceptable model fit, the fit criterion should be near 0, Pr > Chi-Square should be > .05, RMSEA should be < .08, and Bentler & Bonett's NFI should be > .90. If these multiple criteria are met, the measurement model can be said to have been confirmed and may be retained for subsequent use. Before a confirmed model is designated as the final choice, it is a good idea to test several alternate candidates to determine if an improved fit can be obtained. For direct model comparisons, the fit criterion and the RMSEA should be given the greatest decision weighting.

If a given assignment of manifest variables is rejected as a candidate measurement model, a variable from the set should be dropped and the model rerun. Which variable to drop requires a bit of detective work. A good place to start is the PROC CALIS output section titled "Rank Order of the 10 Largest Asymptotically Standardized Residuals." Residuals greater than 2.0 should be singled out and the row and column examined to determine if one manifest variable is contributing to multiple residuals greater than 2.0. If so, that variable is a prime candidate for exclusion. A companion tactic is to examine the standard errors (Std Err) found in "Manifest Variable Equations with Estimates" and in the "Variances of Exogenous Variables" for size consistency. A variable with an inordinately large standard error is suspect. In CFA modeling, there is always the possibility that dropping manifest variables will not guaranteed a good fitting measurement model. If all variables have been exhausted and no acceptable measurement model found, the factor should be removed from further consideration.

Once a measurement model has been found for each factor across all k samples, the factor should be named. Factor naming is critically important in that the name serves to identify and communicate the central construct being measured. Factor naming is an inferential process that begins with a close examination of the manifest variables of the measurement model to determine what it is that they share in common. Factors should be considered as theoretical identities drawing their meaning from the relevant physical, social, life, engineering, or management sciences. The gamma weights associated with each manifest variable are invaluable aids in factor naming. The closer the gamma weight approaches 1, the more the variable can be said to be "loaded" on the underlying factor. Consequently, those manifest variables with the highest loadings provide the strongest clues as to factor identity. Negative signs indicate an inverse relation between the manifest variable and the latent factor—the lower the manifest variable score, the higher the associated factor score and vice versa.

The measurement models for our white and black families are shown in Tables 1 and 2, respectively.

Table 1: Measurement models for white family sample

Factor	Gamma wt.	Fit criterion	RMSEA
f1		.0109	.0162
V4	.5877*		
V9	-.6848*		
V16	-.4553*		
V21	.-.7357*		
f2		.0025	.0000
V3	.4174*	.	
V8	.7006*		
V13	.9184*		
V20	.5808*		
f3		.0229	.0790
V7	.3408*		
V14	.7951*		
V15	-.4215*		
V18	.6178*		

* Significant at .01 level or beyond.

Table 2: Measurement models for black family sample

Factor	Gamma wt.	Fit criterion	RMSEA
f1		.0068	.0000
V4	.6010*		
V9	-.7620*		
V16	-.6221*		
V21	-.8439*		
f2		.0199	.0368
V3	.6388*		
V5	.5255*		
V13	.5657*		
V20	.3577*		
f3		.0035	.0000
V5	.4020*		
V7	-.6230*		
V18	-.3460*		
V19	.5866*		

* Significant at .01 level or beyond.

The measurement models for factor f1 contained identical manifest variables for both samples, although the black family sample exhibited consistently higher absolute gamma weights. Measurements models for f2 overlapped on three manifest variables (V3, V13, and V20), with the white family measurement model loading higher on V13 and V20 and the black family model loading higher on V3. The measurement models for f3 shared two manifest variables (V7and V18), with a sign reversal across samples. Manifest variable V5 loaded on both f2 and f3 in the black family sample.

Factor names, measurement variables, and their associated factor loadings for the white family sample are presented below:

f1: Neighborhood Context

Global evaluation	.5877
Neighborhood problems	-.6848
Neighborhood norms	-.4553
Economic opportunities/stability	-.7357

f2: Family educational expectations

Student educational aspirations	.4174
Parental educational aspirations	.7006
Parental educational expectations	.9184
Parents educational attainment	.5808

f3: External influences

Family norms	.3408
Peer norms	.7951
School practices	-.4215
School safety	.6178

Comparable information for the black family sample is as follows:

f1: Neighborhood Context

Global evaluation	.6010
Neighborhood problems	-.7620
Neighborhood norms	-.6221
Economic opportunitities/stability	-.8439

f2: Family educational expectations

Student educational aspirations	.6388
Peer expectancies	.5255
Parental educational expectations	.5657
Parents educational attainment	.3577

f3: External influences

Peer expectancies	.4020
Family norms	-.6230
School safety	-.3460
Maternal support	.5866

Comparison of factor composition across samples reveals subtle differences. Although the factors have been assigned identical names, their measurement models differ with respect to manifest variable composition, factor loadings, or both. For f1, the variable composition of the measurement models is identical for both samples; however, the black family sample exhibits consistently higher absolute factor loadings. For factor f2, parental expectation is the defining variable for the white family sample, but is of lesser importance in the black family sample. Peer expectancies replace parental educational aspirations in the black sample. Student educational aspiration is assigned a higher loading in the black family measurement model than in the white model. Parental education correlates higher with factor f2 (has a higher loading) in the white sample than in the black sample. Factor f3 shows the greatest sample differences, sharing only two common manifest variables---family norms and school safety. Interestingly enough, both variables change signs from the white to the black measurement models. The unique variable contributors are peer norms and school practices for the white family sample and peer expectancies and maternal support for the black family sample.

CREATE STRUCTURAL EQUATIONS MODELS

The remaining task is to meld the factors into an associative network. Interfactor covariances/correlations are used as the defining network relation and are estimated using multifactor CFA techniques. The number of samples and the factor set for each sample is dependent upon the empirical results of the previously detailed procedures. For each sample, a multfactor, noncorrelated error CFA model can be conducted using the suggested program code:

```
proc calis data=reducedcov edf=samplesize-1
corr method=ml res se;
   lineqs
     v1 = gamma1 f1 + e1,
     v2 = gamma2 f1 + e2,
     .
     vi = gammai f1 + ei,
     v(i+1) = gamma(i+1) f2 + e(i+1),
     v(i+2) = gamma(i+2) f2 + e(i+2),
     .
     v(i+j+1) = gamma(i+j+1) f3 + e(i+j+1),
     .
     v(i+j+…+m+1) = gamma((i+j+…+m+1)) fp +
e(i+j+…+m+1),
     .
     v(i+j+…+m+n) = gamma(i+j+…+m+n) fp +
e(i+j+…+m+n);
   std
     e1 – e(i+j+…+m+n) = error: ((i+j+…+m+n)
* .1),
     f1 = 1.0, f2 =1.0, …, fp=1.0;
   cov
f1 – fp = cov:((p(p-1)/2 * .1);
run;
```

For the above code, i refers to the number of variables in the f1 measurement model, j to the number in the f2 measurement model, and n to the number in the fp measurement model. Prior to running, symbols within the parentheses should be replaced with the appropriate integer values. The lineqs section, then, contains a sequential listing of the factorial measurement models with appropriate variable notation. The cov statement instructs the procedure to estimate the p(p-1)/2 factor correlations, where p is the total number of factors included in the model. Program code should be run separately for each of the samples previously deemed to have been drawn from populations with unequal covariance matrices.

The extent of model fit should be determined by employing the same criteria as that specified for measurement models. Because the multifactor CFA model is testing for cross factor as well as within factor order constraints, the test is more stringent than are single factor measurement models, leading to a greater expectation that moderate to poor fitting models will likely be encountered. Unlike measurement models, dropping variables is no longer an option for improving model fit. Nor, for that matter, is eliminating factors, as the underlying factor structure has been already established. The only viable option for improving structural model fit is to scan for correlated error and/or cross-factor loadings. Correlation between error variables ei and ej, where variables i and j are from different factor sets, implies the presence of a common contribution to both error variables. Correlated errors within a factor variable set are disallowed by virtue of the definition of a measurement model. Correlated errors would imply the contributing effects of other factors beyond that of the single factor being measured. Cross-loaders indicate that manifest variables are subject to the causal influences of multiple factors.

A procedure is required that is capable of differentiating the unique signatures of poor fitting structural models containing correlated error versus those containing cross loading variables. For the sake of explication, let p=4 and further suppose that each of the four factors has a measurement model containing 4 manifest variables. The complete data set then contains 4 x 4 = 16 manifest variables. The 16 x 16 correlation matrix can be partitioned as

$$\begin{array}{c|c|c|c|c}
 & F1 & F2 & F3 & F4 \\
\hline
F1 & R_{F1F1} & & & \\
\hline
F2 & R_{F2F1} & R_{F2F2} & & \\
\hline
F3 & R_{F3F1} & R_{F3F2} & R_{F3F3} & \\
\hline
F4 & R_{F4F1} & R_{F4F2} & R_{F4F3} & R_{F4F4}
\end{array}$$

where R_{FiFj} is the submatrice of the variables in the Fi measurement model crossed with the variables in the Fj measurement model. When i≠j, these blocks are termed hetro-factor blocks to emphasize that the variables cross factor pairs. When i≠j, R_{FiFj} is of unit rank for a perfectly fitting structural model. Anderson (1984) describes a test statistic

$$-\left[\mathrm{N} - \frac{1}{2}(p + 3)\right] \cdot \sum_{i=2}^{p_1} \ln(1 - r_i^2)$$

where p is the sum of the manifest variables in the measurement models of Fi and Fj, p_1 is the lesser number of variables in either measurement model, N is the sample size, and r_i^2 is the squared I^{th} canonical correlation of the variables in Fi measurement model with those in the Fj measurement model. Under the null hypothesis of unit rank, the test statistic is distributed as χ^2 with $(p_1 - 1)(p_2 - 1)$ df. We adopt the notation Fi*Fj to signify a canonical correlation computed using PROC CANCORR with the manifest variables in the Fi measurement model indicated by the *var* statement and those in the Fj measurement model by the *with* statement.

For a structural model with p factors, there are p(p-1)/2j canonical correlations of the form Fi*Fj to be computed. Test statistics should be computed and tested for significance for each Fi*Fj. A significant Fi*Fj pairing indicates the presence of correlated error between variables in the two measurement models for that hetro-factor block. Testing for cross-loaders requires a bit more notation. For each column in the partitioned matrix described above, we define the notation Fi*(Fj x Fk x Fl), where the hetro-factor block is the i^{th} column crossed with all rows excluding i. Again the notation signifies a canonical correlation between the variables in the i^{th} column with the variables in the remaining rows. In the above example, F1*(F2 x F3 x F4) signifies a canonical correlation with the variables in the F1 measurement model designated by the *var* statement and all others by the *with* statement. For each column of the partitioned matrice, compute the statistic

$$\gamma(\text{Fi * (Fj x Fk x Fl)}) = -\sum_{i=2}^{p_1} \ln(1 - r_i^2)$$

as well as the p-1 statistics

$$\gamma(\text{Fi * Fj}) = -\sum_{i=2}^{p_1} \ln(1 - r_i^2)$$

for j = 1, 2, …, p - 1. Under the null hypothesis, the test statistic

$$\gamma(\text{Fi * (Fj x Fk x Fl)} - \sum_{j}^{l} \gamma(\text{Fi * Fj})$$

is distributed as χ^2 with $(p_i - 1)(p - p_i - 1) - (p_i - 1)\Sigma(p_j - 1)$ df, where p_i is the number of variables in the Fi measurement model, p_j is the number of manifest variables in the Fj measurement model, and p is the total number of manifest variables in the structural model. If the statistic is significant, the interpretation is that one or more of the manifest variables comprising the measurement model of Fi cross load on at least one of the factors Fj, Fk, or Fl. Exactly which variables and factors are involved can be determined by using the modification option of PROC CALIS. Examination of the output section "Rank order of the 10 largest Lagrange multipliers in _GAMMA_" can be expected to identify which manifest variables in the Fi measurement model are cross loading on one of the remaining factors.

For our research data, initial CFA structural models were separately run for each sample. For the black family sample, manifest variable V5 was dropped from the f2 measurement so as to avoid having cross loaders in the initial model. The structural CFA models when run yielded the following fit statistics for the white family sample: fit criterion = .4852, Pr > Chi-Square < .0003, RMSEA = .0656, and NFI = .8294; for the black family sample: fit criterion = .4903, Pr > Chi-Square = .0506, RMSEA = .0578, and NFI = .7918. Based on assessment of fit statistics for each group, we decided to analyze the data further to see if improved models could be specified. The results are presented in Table 3.

Table 3: Fit improvement statistics

Test statistic	Numeric value	Chi-square (df)
White family sample		
γ(F1*F2)	.06715	12.591 (9)
γ(F1*F3)	.05649	10.592 (9)
γ(F2*F3)	.09088	17.040 (9)*
γ(F1*(F2*F3)) - γ(F1*F2) - γ(F1*F3)	.04491	8.330* (3)
γ(F2*(F1*F3))) - γ(F1*F2) - γ(F2*F3)	.02360	4.378 (3)
γ(F3*(F1*F2)) - γ(F1*F3) - γ(F2*F3)	.03482	6.459 (3)
Black family sample		
γ(F1*F2)	.02575	2.884 (6)
γ(F1*F3)	.04668	5.204 (9)
γ(F2*F3)	.07147	8.004 (6)
γ(F1*(F2*F3)) - γ(F1*F2) - γ(F1*F3)	.03964	4.360 (3)
γ(F2*(F1*F3))) - γ(F1*F2) - γ(F2*F3)	.02546	2.800 (2)
γ(F3*(F1*F2)) - γ(F1*F3) - γ(F2*F3)	.10976	12.073 **(3)

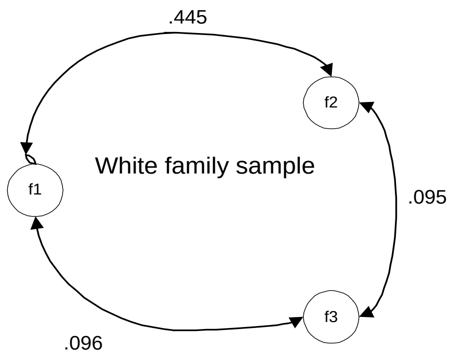
* Significant at .05 level.

** Significant at .01 level.

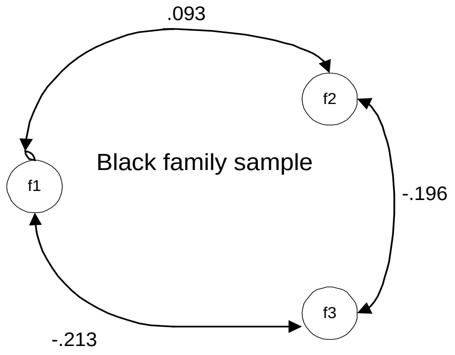
The tabular results suggest that the initial fit of the white family model may be improved by allowing for correlated errors between measurement model variables for f2 and f3 and cross loaders between f1 manifest variables and factors f2 and/or f3. For the black family sample, one or more f3 measurement variables may be expected to cross load on factors f1 and/or f2.

Examination of the modification output for the white family sample did not show any significant correlated errors between f2 and f3 measurement model variables as might be expected from the fit improvement statistics. However, V2 (neighborhood problems) was revealed to be significantly cross loaded on f2 and f3. For the black family model, V7 (peer expectations) was shown to be cross loaded on f2, as would be expected from the fit improvement results. Consequently, the white model was revised by allowing variable V2 to cross load on factors f2 and f3, and when rerun yielded fit criterion = .4121, Pr > Chi Square = .0041, RMSEA = .0566, and NFI = .8551. The black family model was revised by allowing V7 to cross load on f2. The revised model when run produced fit criterion = .3938, Pr > Chi Square = .2480, RMSEA = .0350, and NF1 = .8328. In that these models were derived using an explicit inductive procedure grounded in statistical and structural modeling theory, we consider them to be superior to those that may be derived by a pure ad hoc approach.

The associative factor network for white family groups is shown below:



and for the black family group



DISCUSSION

Associated Structural Networks in the Context of Academic Performance.

The idea that African American and Caucasian parents and their children can be segmented differentially as geo-psychological and –sociological, behaviorally unique entities using underlying construct analysis suggests the utility of alternate approaches to defining and capturing methodologically user and customer sub-populations as systems. In the context of academic achievement, what is evident are two different societal levels of group experience. At this level of structural analysis, what is evident empirically are conceptually similar spheres of context, yet strikingly different sets of synchronous patterns of social, psychological, and behavioral outcomes at group levels of performance. For Caucasians, the associated networks are functionally enabling and mutually reinforcing. For African Americans, the network responds differentially, both negatively and positively. Further examination at the component level highlights these distinctions.

Neighborhood Context

The neighborhood context drives consistent patterns of positive and negative recurring observed experiences for both groups, and similar dominant concerns for economic and social problems among their families. However, among the lesser dominant concerns, neighborhood social norms and global evaluations show a reversal in order. For the African Americans, neighborhood norms of social isolation and apathy appear as more problematic (-.6221 versus -.4553). Whereas, for Caucasians, the third order of placement is occupied by global evaluations of their neighborhoods as a positive place to raise their children (.5877). Second, for Caucasian parents all induced effects are consistently less problematic than those that occur among African American families.

Family Educational Experiences

This family structure underlies sets of conceptually similar observed experiences for both groups. Again, the differences are fundamental, occurring in two ways. First, for African Americans, parental educational aspirations as an observed recurring pattern is not evident as among Caucasians (.7006). Educational expectancies of African American peers (.5255) replace parental educational aspirations. As discussed in the next section, this is consistent with interpretations of "isolation." The second of the two obvious differences show the relative dominance magnitudes of "parental" effects among those who are Caucasian. For African American students, the aspirations of the students themselves (.6388) make the dominant contribution.

External Influences

Examination of the external influences on academic achievement reveals the greatest group disparities. For African American youth the existence of peers that expect to continue their education (.4020) is important. This is in contrast to Caucasian students. The lack of its occurrence might suggest that education expectancies, among their peers, is, not an issue. In its place is a level of normative peer delinquency (.7951). For these youth, this delinquency norm is also not expressed as a disabling experience.

Its dominant status in absolute magnitude as an observed effect is evident, along with constancy of positive association with its other companion effects (maternal respect, .3408; school safety, .6178; school practices, .4215).

Another salient difference between the two groups is that the patterns for African American youth appear conflictual. For these youth, external influences drive simultaneously both positive and negative impressions of psychological capital (maternal support, enabling, .5866 versus family norms, disabling, -.6230). Recurring patterns of school experiences also occur as neither safe nor personal (-.3460), and contrast strongly to that of Caucasian students. When interpreted in the context of related induced companion effects, what is highly suggested is an academic experience of isolation. On the other hand, for Caucasians, external influences are not only not negative (.6178), but are co-existent with a personal sense of connectivity to the learning environment. School practice perceptions, although negatively induced (-.4215), are associated with positive patterns occurring across family and peer dimensions (.3408, .7951).

In this example, Underlying Construct Analysis reveals how academic experiences are not only differentially experienced by African American and Caucasian students, but the underlying supportive structure that likely drives the effects. On one level, the construct "group" is a part of the underlying supportive factor structure for Caucasians, whereas the same is not supportive for African Americans. What is derived from this exploration, clearly, is that "one model" of performance does not fit all.

CONCLUSION

The utility of data mining procedures is maximized when analyses explore both observed factors, such as in principle components analysis, and underlying constructs as we propose. Conducting data exploration at the level of underlying structure has many advantages over traditional data mining approaches including, but not limited to, the increased potential for the development of not only of new theory, but newer approaches to leverage benefit by alternate definitions of consumer or user markets.

Using a specific example from the social sciences, we demonstrated the relative ease of use of the SAS procedures
to data mine and data mining applications to support the notion user and customer populations as unique geo-psychological,

-social, and -behavioral entities, as systems, rather that simply aggregates of individuals. Level of analysis extends the analytical framework, from the individual level, as the unit of analysis, to larger spheres of performance, particularly the nature and scope of induce outcomes (i.e., multidimensionality), along with the scope of causally related underlying structures and mechanisms that convey them as effects.

Our example demonstrated the depth of understanding that can be derived from exploring the driving forces that influence variables to covary, as synchronous units, in both similar and dissimilar fashions. We hope that this article has significantly advanced the utility of data mining and visualization techniques, especially UCA, for the average researcher.

REFERENCES

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: John Wiley & Sons.

Bentler, P.M. (1989), *EQS Structual Equations Program Manual*, Los Angles: BMDP Statistical Software, Inc.

Bentler, P.M. and Bonett, D.G. (1980), “Significance Tests and Goodness of Fit in the Analysis of Covariance Structures,” *Psychological Bulletin*, 88, 588-606.

Bollen, K.A. (1989), *Structual Equations With Latent Variables*, New York: John Wiley & Sons.

Drewes, D.W. (in press), “Beyond the Spearman-Brown: A Structual Approach to Maximal Reliability,” *Psychological Methods*.

Hatcher, L. (1994), *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*, Cary, NC: SAS Institute Inc.

Mardia, K.V. (1974), ”Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies,” *Sankhya*, B, 36, 115-128.

Mardia, K.V. (1985), “Mardia’s Test of Multinormality,” In S. Kotz and N. L. Johnson, eds., *Encyclopedia of Statistical Sciences*, Vol. 5, New York: John Wiley & Sons, 217-221.

Mardia, K.V. and Foster, K. (1983), “Omnibus Tests of Multinormality Based on Skewness and Kurtosis,” *Communication in Statistics*, 12, 207-221.

Survey of Parents and Children (1990). Available from the Inter-University Consortium for Political Social Science Research.

APPENDIX

The following program was used as is to test equality of covariance matrices as reported in this paper.

```
/*Program to test equality of multiple
covariance matrices*/
/*Ref: Anderson,Introduction to Multivariate
Statistical Analysis, 1984, pp. 419-422*/
/*Assumes that SAS covariance matries have
been created using proc corr with COV and
NOCORR options and are located in designated
library*/
/*p - number of variables*/
/*q - number of samples*/
/*chi_sq - chi square statistic*/
/*prob - portion of chi-square distribution
to right of chi-sq value*/
/*Type I error is set at .05 level*/
libname in "your libref”;
proc iml;
start get_q;
        file log;
     put "Number of samples to be tested: ";
        infile cards;
        input q;
finish;
/*Replace number with value of q*/
run get_q;
2
;
start get_p;
        file log;
        put "Number of variables per sample:
";
        infile cards;
        input p;
finish;
/*Replace number with value of p*/
run get_p;
19
;
I=1:q;
fr=char(I,2,0);
M = {"Full filename of sample covariance
matrix"};
msg= concat(M,fr[5]);
D=J(2,q+1,0);
A=J(p,p,0);
sq_mat=J(p,p,0);
start gt_names;
        infile cards;
        dsname="11111111111111111";
        create name_ds var{dsname};
        do data;
                input dsname $;
                append;
        end;
finish;
/*Replace dataset names with yours, keeping
in unless changed in libref statement*/
run gt_names;
in.wht_cov2
in.blk_cov2
;
start flexible(filename);
        call execute("use ",filename,";");
finish;
start keep_it(matrix,k) global(D,A,p);
        sq_mat=matrix[1:p,1:p];
        D[1,k] = det(sq_mat);
        D[2,k] = matrix[p+3,1]-1;
        A=A+sq_mat#D[2,k];
finish;
g=1;
do data i=1 to q;
        read point i var _CHAR_ into temp1;
        run flexible(temp1);
        read all var _num_ into temp2 ;
        run keep_it(temp2,g);
        close;
        setin work.name_ds;
        g=g+1;
end;
tot_n=0;
do i=1 to q;
        tot_n=D[2,i]+tot_n;
end;
D[2,q+1]=tot_n;
```

```
A1=A/tot_n;
D[1,q+1]=det(A1);
wt_ln=0;
do i=1 to q;
        wt_ln=log(D[1,i])*(D[2,i]/2) +wt_ln;
end;
ts=log(D[1,q+1])*(D[2,q+1]/2)-wt_ln;
inv_n=0;
do i=1 to q;
        inv_n=(1/D[2,i])+inv_n;
end;
rho=1-(inv_n-1/D[2,q+1])*((2*p*p+3*p-
1)/(6*(p+1)*(q-1)));
chi_sq=rho*ts*2;
df=(q-1)*p*(p+1)/2;
prob=1-probchi(chi_sq,df);
print chi_sq;
print df;
print prob;
msg1="Null hypothesis of no difference
ACCEPTED.";
msg2="Null hypothesis of no difference
REJECTED.";
if prob>.05 then print msg1;
else print msg2;
quit;
run;
```

CONTACT INFORMATION

Author Name: Alvin L. Killough
Company: North Carolina Central University
Address: 1414 Wabash St.,
City state ZIP: Durham, NC 27701
Work Phone: 919-598-8916 / 560-5165
Email:akillough@worldnet.att.net

Author Name: Christopher L. Edwards
Company: Duke University Medical Center
Address: DUMC 3842
City state ZIP: Durham, NC 27710
Work Phone: 919-681-3090
Fax: 919-681-7341; Email:cledwa00@acpub.duke.edu

Author Name: Donald W. Drewes
Company: North Carolina State University
Address: Dept. of Psychology, Box 7801
City state ZIP: Raleigh, NC 27695-7801
Work Phone: 919-515-2251
Fax: 919-515-1716
Email:drewes@unity.ncsu.edu

BIOGRAPHIES:

(1) Dr. Killough is a cultural ecological psychologist, and consultant in private practice. He provides expertise in the areas of methodological approaches and analytical strategies for mapping social systems and pathology. He is also an adjunct professor of psychology at North Carolina Central University where he teaches in the Department of Psychology, and the School of Library and Information Sciences on topics such as human systems, mental model and paradigms, and applications to capture culture and its components statistically.

(2) Dr. Edwards is the Director of Chronic Pain Management Program at Duke University Medical Center. There, he provides instruction in minority mental health, chronic pain management, and stress and coping. He is also an instructor at Duke University in the Department of Psychology: Social & Health Sciences, and an adjunct professor of psychology at North Carolina Central University.

(3) Dr. Drewes is Professor of Psychology at North Carolina State University. He specializes in quantitative methods, and currently teaches courses in quasi-experimental design and structural equations modeling. He has co-authored a text in mathematics for the behavioral sciences and has an active research interest in modeling of social phenomena.