Stony Brook University

Structural Equation Modeling for Mixed Designs

A Dissertation Presented

by

Kathryn Elizabeth Sharpe

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2010

Stony Brook University

The Graduate School

Kathryn Elizabeth Sharpe

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Wei Zhu
Dissertation Advisor
Professor
Department of Applied Mathematics and Statistics

Nancy Mendell
Chairperson of Defense
Professor
Department of Applied Mathematics and Statistics

Haiping Xing
Member
Assistant Professor
Department of Applied Mathematics and Statistics

Ellen Li
Outside Member
Professor
Department of Medicine

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Structural Equation Modeling for Mixed Designs

by

Kathryn Elizabeth Sharpe

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2010

A mixed-design study, also called a split-plot design, intends to evaluate the differences among multiple independent groups and multiple treatment conditions simultaneously, with repeated measurements of the same participants. Structural equation modeling (SEM), also referred to as path analysis, is a statistical technique used by researchers in many fields to verify or disprove hypothesized causal links among a predefined system of variables. The existing SEM methods for detecting differences in path strength among multiple datasets can accommodate comparisons of independent groups or repeated measures (e.g. with and without stimulus), but not both. Thus SEM is unable to perform a direct analysis of a mixed-design study. To fill this void, we have developed a cohesive two-level parametric modeling approach using the maximum likelihood method (MLE SEM) for detecting differences in pathways caused by multiple factors, both between and within groups, such as group membership or treatment condition. The method is illustrated through a brain functional pathway analysis. Further, developments of the mixed-design methodology for Latent Variable SEM and Partial Least Squares SEM (PLS SEM) are included, and guidelines for power and sample size are provided.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Path analysis is a method for verifying or disproving causalities between variables in a path diagram (nodes joined by directed lines). The strengths of the paths and the covariance structure of the variables present are the parameters to be estimated in the model. It was developed by Sewall Wright beginning in 1918 [1]. Later, Blalock first attempted to apply this concept to studies in the social sciences [2], and Duncan furthered this effort in the 1960's and 70's [3]. It was Dr. Karl Gustav Jöreskog (http://en.wikipedia.org/wiki/Karl_Gustav_Jöreskog) who eventually formalized structural modeling (SEM) into the common LISREL (LInear Structural RELations) representation (and software) we use today for understanding structural equation modeling (a more general type of path analysis including latent variables) [4]. Jöreskog's LISREL model features matrix computation for estimation of parameters.

Today, structural equation modeling has found many applications. It is popular in the social sciences and economics [5-8], as well as in brain functional pathway analysis [9-16] and genetic pathway analysis [17-23]. There are also many customized software procedures to facilitate such analysis. SAS, professional statistical software designed for many different types of statistical analyses, includes the CALIS procedure, which can estimate SEMs. LISREL (www.ssicentral.com/lisrel) is SEM software (bearing the same name as the creators' matrix model representation) created by Jöreskog and Sörbom which can estimate single- and multi-group SEMs with many traditional estimation options. It includes a path diagram capability that SAS does not. This makes the LISREL software a popular choice for behavioral and social scientists. Software similar to LISREL is EQS (www.mvsoft.com), developed by Dr. Peter M. Bentler. It, too, can estimate single- and multi-group SEMs, and has a path diagramming feature which will create the model for the user. In addition, EQS can model mean structures and estimate multi-level models. AMOS (www.spss.com/amos) and MPLUS (www.statmodel.com) are also prominent SEM packages. All of these packages yield similar results for a single path analysis model.

However, despite its rapid growth in the past thirty years, SEM methodology and software packages still have quite a few gaps to be filled. A foremost deficiency is the lack of a SEM procedure for the analysis of mixed designs, which we have confirmed with senior statistical technical support at both LISREL and EQS via ongoing electronic conversations. Mixed designs are traditionally referred to as split-plot designs, first used in agricultural studies. They are used to study the influence of several factors on the response variable. These factors could include group membership or treatment conditions, for example. In its original form, researchers would split an agricultural plot into several smaller plots and assign each a value of one of the factors. The sub-plots would then be divided to accommodate all values of a second factor, and so on. Experimental treatments (factors) are randomly assigned to each plot and sub-plots so that all combinations of all factors are represented. As the focus of statistics shifted from agricultural studies to biomedical studies in the past decades, the mixed design (or the

split-plot design) represents a study where we have multiple independent groups of subjects and each subject has been measured repeatedly (either longitudinally or under different treatments). These experimental designs are sometimes referred to as panel designs when the repeated measures are made longitudinally, and one is interested in studying the longitudinal effect of the measurements.

We have been collaborating with Dr. Nora D. Volkow, Director of the National Institute on Drug Abuse, in her study of the brain functional pathway and particularly the reward network in the brain. We are interested in detecting the differences in path strength for cocaine abusers and normal subjects, when each subject is evaluated under the influence of placebo and a cocaine-like substance called methylphenidate. Locating the differences in the network (due to drug addiction history and/or due to receiving the drug) could help determine a method for successfully inhibiting the reward network, and thus effectively stopping the reward in the brain of the drug-abuser. This study relies heavily on our ability to detect differences in strength of pathways between factors in a mixed design study. We are glad to report that we have successfully developed the necessary theoretical framework and implemented the corresponding algorithm in MATLAB. We discuss its development of SEM for Mixed Designs (MD-SEM) and illustrate with the analysis of the afore-mentioned brain functional pathway study. We report the results of our power and sample size analysis of MD-SEM and discuss Partial Least Squares (PLS) SEM as an alternative to the traditional maximum likelihood (MLE) SEM methodology when the sample size of a particular study is small.

Chapter 2: Structural Equation Modeling (SEM) Introduction and Literature Review

While structural equation model (SEM) is a popular covariance analysis technique in many fields, it is rarely covered in undergraduate and graduate statistics curriculum. Therefore, a brief introduction to the methodology is provided here along with a thorough literature review on the current capabilities of SEM and its lacking of a methodology and software platform for the analysis of mixed designs. This void is filled as part of the work of this dissertation. Our methodology is presented in Chapter 3.

2.1 An Introduction to SEM Diagrams, Models, and Estimation

2.1.1 A Simple Example

Consider the following simple example. A researcher is interested in studying the variables that may lead to an increased risk for developing an eating disorder. The variables under consideration as causal influences in the model are: age of first menstrual period (AM), body image score (BI), self concept score (SC), drive for thinness (DT), and risk for developing an eating disorder (RD). The researcher proposes that the relation between the variables in the study is as shown in the diagram below. The diagram is the same for each group.

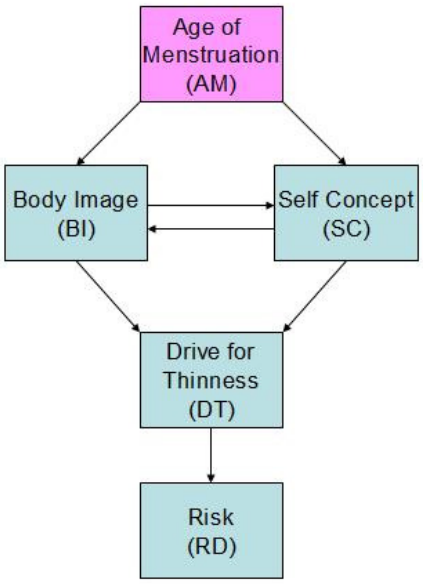


Figure 2.1. Hypothesized path diagram indicating collaborator's suggested relationships among variables influencing development of eating disorders in women; these factors include age of first menstrual period (AM), body image score (BI), self concept score (SC), drive for thinness (DT), and risk for developing an eating disorder (RD).

Looking at the diagram, note the coloring of the variables, represented in rectangles. Each of the blue boxes is an endogenous variable, because it has arrows coming into the box, illustrating causes present in the model. The variable colored in pink is an exogenous variable because its causes are not considered in the current model. For each endogenous variable, we can create a regression equation, incorporating each variable that influences the variable being considered. The system of equations for this model is shown below.

$$BI = \beta_1 AM + \beta_2 SC + \zeta_1$$
$$SC = \beta_3 AM + \beta_4 BI + \zeta_2$$
$$DT = \beta_5 BI + \beta_6 SC + \zeta_3$$
$$RD = \beta_7 DT + \zeta_4$$

In each equation, there is one random response variable, a coefficient preceding each random influence on the response that is present in the model, and a random error term. This system of regression-style equations represents a structural equation model.

We can write this system in matrix form. We will list our endogenous variables in vector

$$Y = \begin{pmatrix} BI \\ SC \\ DT \\ RD \end{pmatrix},$$

our exogenous variables in vector $X = (AM)$, and errors in vector

$$\zeta = \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{pmatrix}.$$

We can create two matrices of regression coefficients, or path coefficients, as they are called in an SEM. B will represent the paths from endogenous variables to endogenous variables, where $B(1,2)$ represents the coefficient of the path from Y_2 to Y_1 (a parameter if a path exits, 0 if no path exists). Γ will represent the paths from exogenous variables to endogenous variables, where $\Gamma(2,1)$ represents the coefficient of the path from X_1 to Y_2. Then the model for this system can be described by $Y = \mathrm{B}Y + \Gamma X + \zeta$, where

$$\mathrm{B} = \begin{bmatrix} 0 & \beta_2 & 0 & 0 \\ \beta_4 & 0 & 0 & 0 \\ \beta_5 & \beta_6 & 0 & 0 \\ 0 & 0 & \beta_7 & 0 \end{bmatrix} \text{ and } \Gamma = \begin{bmatrix} \beta_1 \\ \beta_3 \\ 0 \\ 0 \end{bmatrix}.$$

To estimate the parameters of this model and determine the significance of individual paths, we use the maximum likelihood estimation method, which requires the assumption that all variables present in the model have a multivariate normal distribution. Additionally, we assume that all variables are continuous, and that the sample size of the study is large—several times the number of parameters we will be estimating, for best results. Finally, we assume that the model is well-defined—that the researcher providing the model has included all significant variables and excluded all insignificant variables. Details on estimation of the general SEM model are formally discussed in the following section.

2.1.2 Models in Matrix Form

Kenneth A. Bollen's text gives an excellent introduction theoretical introduction to SEM, [1]. The standard path analysis model (SEM with measured variables only) is:

$$Y = \mathrm{B}Y + \Gamma X + \zeta .$$

Y is a vector of the endogenous variables studied, and X a vector of the exogenous variables studied. (Note: An endogenous—or dependent—variable in the model is one with arrows coming in, i.e. influences of this variable are present in the model. An exogenous—or independent—variable in the model is one with no arrows coming in, i.e. influences of this variable are not present in the current model.) Both Y and X have been centered about their means. B is a matrix containing path coefficients where the entry $\mathrm{B}_{i,j}$ is the coefficient of the path from endogenous node *j* to endogenous node *i*. Γ is a matrix containing coefficients of paths from exogenous variables to endogenous variables. $\Gamma_{i,j}$ is the coefficient of the path from exogenous node *j* to endogenous node *i*. ζ is a vector containing the error variables in the equations for the path diagram.

The null hypothesis we work with in SEM is always $\Sigma = \Sigma(\theta)$. The question we answer is: "does the model predict a covariance matrix for the data that is equal to the population covariance matrix?" Σ is the population covariance matrix of the observed variables and $\Sigma(\theta)$ is the covariance matrix written as a function of the free model parameters (the vector θ). We can break $\Sigma(\theta)$ into a block matrix as follows.

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

We will consider each block individually. We can solve for Y explicitly since $Y = \mathrm{B}Y + \Gamma X + \zeta$:

$$Y - \mathrm{B}Y = \Gamma X + \zeta$$
$$(I - \mathrm{B})Y = \Gamma X + \zeta$$
$$Y = (I - \mathrm{B})^{-1}(\Gamma X + \zeta)$$

Then

$$\Sigma_{yy}(\theta) = Cov(Y,Y) = E[(Y-EY)(Y-EY)'] = E[(Y-EY)(Y'-EY')]$$
$$= E[YY'-EY\cdot Y'-YEY'+EYEY'] = E(YY')$$

where E represents the expected value of a random variable. Therefore,

$$\Sigma_{yy}(\theta) = E(YY')$$
$$= E[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)((I-\mathrm{B})^{-1}(\Gamma X+\zeta))'] = E[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'}]$$
$$= (I-\mathrm{B})^{-1}(E(\Gamma XX'\Gamma')+E(\Gamma X\zeta')+E(\zeta X'\Gamma')+E(\zeta\zeta'))(I-\mathrm{B})^{-1'}$$
$$= (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1'}$$

Note that Φ is the covariance matrix of x and Ψ is the covariance matrix of ζ.

Similarly, $\Sigma_{xx}(\theta) = Cov(X,X) = E(XX') = \Phi$ by definition.
Therefore,

$$\Sigma_{xy}(\theta) = Cov(X,Y) = E[(X-EX)(Y-EY)'] = E[(X-EX)(Y'-EY')]$$
$$= E(XY'-Y'EX-XEY'+EXEY') = EXY'$$
$$\Sigma_{xy}(\theta) = E(XY') = E(X((I-\mathrm{B})^{-1}(\Gamma X+\zeta))') = E(X(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'})$$
$$= [E(XX'\Gamma')+E(X\zeta')](I-\mathrm{B})^{-1'} = \Phi\Gamma'(I-B)^{-1'}$$

$$\Sigma_{yx}(\theta) = E(YX') = E[(I-B)^{-1}(\Gamma X+\zeta)X'] = (I-B)^{-1}E[\Gamma XX'+\zeta X']$$
$$= (I-B)^{-1}[\Gamma EXX'+E\zeta EX'] = (I-B)^{-1}(\Gamma\Phi+0) = (I-B)^{-1}\Gamma\Phi$$

Now we can assemble $\Sigma(\theta)$ as follows:

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} = \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1'} & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(I-\mathrm{B})^{-1'} & \Phi \end{bmatrix}$$

Now that we have $\Sigma(\theta)$, we can estimate θ. We have the equation $\Sigma = \Sigma(\theta)$, so we set elements of Σ equal to the corresponding elements in our $\Sigma(\theta)$ matrix. Since we do not know the actual values of Σ's entries, we do not know the population covariances. However, we may approximate them with the sample covariance matrix, S.

In estimating θ (a vector of our free variables—the path coefficients and equation errors), we must choose values of θ such that the difference between S and $\Sigma(\theta)$ is minimized. We find estimates of the parameters by maximizing the likelihood function of the multivariate normally distributed random variables in Y and X.

2.1.3 Likelihood Function

We will derive all likelihood functions from the direction of assuming the variables have a multivariate normal distribution. Assume Y and X are vectors of multivariate normally distributed variables. Let $Z = \begin{pmatrix} Y \\ X \end{pmatrix}$. Z has length $p+q$ (p is the number of endogenous variables and q is the number of exogenous variables in the model). Center Z so that all variables have mean 0. Then because Z is a vector of multivariate normal variables, the distribution of the variables in Z is

$$f(z;\Sigma) = (2\pi)^{-\frac{p+q}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} z'\Sigma^{-1} z\right)$$

since $\mu = 0$ for each variable in z. (The above is simply the PDF of the multivariate normal distribution.)

For N independent observations of the vector Z, the joint density function is

$$f(z_1, z_2, ..., z_N; \Sigma) = f(z_1;\Sigma) f(z_2;\Sigma) \cdots f(z_N;\Sigma)$$

and therefore, the likelihood function is

$$L(\theta) = (2\pi)^{-\frac{N}{2}(p+q)} \left|\Sigma(\theta)\right|^{-\frac{N}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i \right).$$

(Once θ is observed, $\Sigma = \Sigma(\theta)$ in the distribution function.)

Therefore,

$$\log L(\theta) = -\tfrac{N}{2}(p+q)\log 2\pi - \tfrac{N}{2}\log\left|\Sigma(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i \,.$$

We can simplify $\log L(\theta)$ by dropping the constant term $-\tfrac{N}{2}(p+q)\log 2\pi$, since maximizing $\log L(\theta)$ will yield the same result as maximizing the same function without the constant term. Additionally, we can opt to remove a factor of $-\tfrac{1}{2}$ from the entire function, and minimize, rather than maximize the resulting function

$$N\log\left|\Sigma(\theta)\right| + \sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i \,.$$

We can further simplify the fitting function as follows:

$$N\log\left|\Sigma(\theta)\right| + \sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i = N\log\left|\Sigma(\theta)\right| + \sum_{i=1}^{N} tr\left[z_i'\Sigma^{-1}(\theta) z_i \right]$$

$$N\log\left|\Sigma(\theta)\right|+\sum_{i=1}^{N}tr\left[z_i'\Sigma^{-1}(\theta)z_i\right]=N\log\left|\Sigma(\theta)\right|+N\sum_{i=1}^{N}tr\left[\tfrac{1}{N}z_iz_i'\Sigma^{-1}(\theta)\right]$$

We can drop the constant factor of N from each term.

$$N\log\left|\Sigma(\theta)\right|+N\sum_{i=1}^{N}tr\left[\tfrac{1}{N}z_iz_i'\Sigma^{-1}(\theta)\right]\quad\Rightarrow\quad\log\left|\Sigma(\theta)\right|+\sum_{i=1}^{N}tr\left[\tfrac{1}{N}z_iz_i'\Sigma^{-1}(\theta)\right]$$

Finally,

$$\begin{aligned}\log\left|\Sigma(\theta)\right|+\sum_{i=1}^{N}tr\left[\tfrac{1}{N}z_iz_i'\Sigma^{-1}(\theta)\right]&=\log\left|\Sigma(\theta)\right|+tr\left[\tfrac{1}{N}\Sigma^{-1}(\theta)\sum_{i=1}^{N}\left[z_iz_i'\right]\right]\\&=\log\left|\Sigma(\theta)\right|+tr\left[\tfrac{1}{N}\sum_{i=1}^{N}\left[z_iz_i'\Sigma^{-1}(\theta)\right]\right]=\log\left|\Sigma(\theta)\right|+tr\left[\tfrac{1}{N}\sum_{i=1}^{N}\left[\Sigma^{-1}(\theta)z_iz_i'\right]\right]\\&=\log\left|\Sigma(\theta)\right|+tr\left[\tfrac{1}{N}\Sigma^{-1}(\theta)\sum_{i=1}^{N}\left[z_iz_i'\right]\right]=\log\left|\Sigma(\theta)\right|+tr\left[\tfrac{1}{N}\Sigma^{-1}(\theta)NS*\right]\\&=\log\left|\Sigma(\theta)\right|+tr\left[\Sigma^{-1}(\theta)S*\right]=\log\left|\Sigma(\theta)\right|+tr\left[S*\Sigma^{-1}(\theta)\right]\end{aligned}$$

(because $S*=\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{x}\right)\left(X_i-\bar{x}\right)'$ is the definition of the sample covariance matrix used here, and our variables are centered).

We add on two constants, which have no effect on minimizing the function, $-\log\left|S\right|-(p+q)$, so that the fit function equals 0 when $S=\Sigma(\theta)$, indicating a perfect fit. Additionally, since $S*=\frac{N-1}{N}S$, where S is the traditionally defined covariance matrix, and for large samples the two will be approximately the same, it is traditional to substitute S for S*.

So the resulting fit function is the traditional ML fit function for SEM,

$$F_{ML}=\log\left|\Sigma(\theta)\right|+tr\left[S\Sigma^{-1}(\theta)\right]-\log\left|S\right|-(p+q),$$

where S is the sample covariance matrix and p and q are the number of endogenous and exogenous variables, respectively. Minimizing the above function will yield the appropriate estimates of θ. See the special section on calculating standard errors at the end of the document for more information about calculating errors for use in statistical analysis of parameter estimates.

2.1.4 Standard Error of Parameters

In order to estimate the standard errors of the parameters we have estimated using the fitting function described in the previous section, we must use the information we have about the MLE of θ. The maximum likelihood estimator, $\hat{\theta}$, of parameter vector θ, is distributed as

$$N\left(\theta,\left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}\right).$$

Standard errors (square root of variances of parameter estimates) of the estimates found in the previous section must be calculated via the asymptotic covariance matrix (inverse of Fisher Information Matrix), with respect to the method through which the parameter estimates are obtained.

By definition [1], the asymptotic covariance matrix of the ML estimator of arbitrary θ is

$$ACOV\left(\hat{\theta}\right)=\left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

This calculation depends, of course, on the expected Fisher Information Matrix, and the required partial derivatives of $\log L(\theta)$. However, we are calculating the parameter estimates via F_{ML}, and our calculation of their respective errors should reflect this change. Therefore, we must determine the relationship between $\log L(\theta)$ and F_{ML}, and let the asymptotic covariance matrix reflect this difference.

As calculated above, $F_{ML}=\log\left|\Sigma(\theta)\right|+tr\left[S\Sigma^{-1}(\theta)\right]-\log\left|S\right|-(p+q)$, and $\log L(\theta)=-\frac{N}{2}(p+q)\log 2\pi-\frac{N}{2}\log\left|\Sigma(\theta)\right|-\frac{1}{2}\sum_{i=1}^{N}z_i'\Sigma^{-1}(\theta)z_i$. We must determine the relation between $-\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}$ and $\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}$ to determine how the calculation of the asymptotic covariance matrix must change.

In manipulating $\log L(\theta)$ into F_{ML}, we remove a constant term, multiply by -2, manipulate one term into an equal term, multiply by $\frac{1}{N}$, then add a constant term. The effects on the equality of derivative are only the multiplying factors. So we can say

$$\left\{-E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1}=\left\{-E\left[\frac{\partial^2 \frac{-2}{N}\log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}=\left\{\frac{2}{N}E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}=\frac{N}{2}\left\{E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

Therefore,

$$ACOV\left(\hat{\theta}\right)=\left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}=\frac{2}{N}\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

The standard errors of the parameter estimates are the square roots of the respective diagonal elements of the asymptotic covariance matrix.

2.1.5 Overall Model Chi-Square Goodness-of-Fit Test

As mentioned above, the goal of SEM is to test the hypothesis $\Sigma = \Sigma(\theta)$. The chi-square test is a simultaneous test that all residuals in $\Sigma - \Sigma(\theta) = 0$. The chi-square test is for overidentified models only (those where the number of distinct entries in the sample covariance matrix of observed variables exceeds the number of free parameters to be estimated). The test shows that $(N-1)F_{ML} \sim \chi^2_{\frac{1}{2}(p+q)(p+q+1)-t}$, where p and q are the number of endogenous and exogenous variables, respectively, and t is the number of free parameters to be estimated. The test is derived via the likelihood ratio test method.

The likelihood ratio test statistic, by definition, is

$$\Lambda(x) = \frac{\sup\{L(\theta \mid x) : \theta \in \Theta_0\}}{\sup\{L(\theta \mid x) : \theta \in \Theta\}}.$$

Under H_0, we have MLEs for all parameters and the fixed parameter values as well, that make up $\Sigma\left(\hat{\theta}\right)$. Then

$$\log L_0 = -\frac{N-1}{2}\left(\log\left|\hat{\Sigma}\right| + tr\left(\hat{\Sigma}^{-1}S\right)\right)$$

is the numerator for the likelihood ratio test (the log of the numerator of the test statistic).

To form the denominator of the test statistic, we must choose an alterative hypothesis, H_1, for which the value of the corresponding log-likelihood function is at its maximum. The least restrictive H_1 possible is that Σ is any positive definite matrix (because at the very least, the matrix must have a positive determinant for the fitting function to take on a valid value). If $\hat{\Sigma}$ is set to S, the sample covariance matrix, then $\log L_1$ is at its maximum value. Then, ignoring the irrelevant constants, the likelihood function for H_1, $\log L_1$, is

$$\log L_1 = -\frac{N-1}{2}\left(\log|S| + tr\left(S^{-1}S\right)\right) = -\frac{N-1}{2}\left(\log|S| + p + q\right),$$

the log of the denominator of the test statistic. (We have selected H_1 to represent a standard of perfect fit to compare H_0 to.)

The natural log of the likelihood ratio, when multiplied by -2, is distributed as chi-square variate when H_0 is true and (N-1) is large. In this case,

$$-2\log\left(\frac{L_0}{L_1}\right) = -2\log L_0 + 2\log L_1$$
$$= 2\frac{N-1}{2}\left(\log\left|\hat{\Sigma}\right| + tr\left(\hat{\Sigma}^{-1}S\right)\right) - 2\frac{N-1}{2}\left(\log|S| + p + q\right)$$
$$= (N-1)\left(\log\left|\hat{\Sigma}\right| + tr\left(\hat{\Sigma}^{-1}S\right) - \log|S| - (p+q)\right)$$

The final line here represents (N-1) multiplied by our fitting function for the ML method, F_{ML}. Compare this statistic to the upper $100(1-\alpha)$ percentile value of the chi-square distribution to determine whether to accept or reject the null hypothesis. A rejected null hypothesis indicates that $\Sigma = \Sigma(\theta)$ does not hold, and that the model poorly represents, or fits, the data.

2.1.6 Results for the Eating Disorder Example

Path coefficients, corresponding standard errors, and p-values for the eating disorder model shown in Figure 2.1 are displayed in Table 2.1. The results are also displayed in Figure 2.2. These results show that there is a significant network of variables that can lead to greater risk for development of an eating disorder among young women, but that the *self concept* variable has little influence.

Confidence intervals for the parameters can be calculated and hypothesis tests can be performed using the path parameters and corresponding standard errors, because the path coefficients of SEM have an asymptotically normal distribution [1]. A confidence interval for β_i is

$$\hat{\beta}_i - z_{\alpha/2}se\left(\hat{\beta}_i\right) < \beta_i < \hat{\beta}_i - z_{\alpha/2}se\left(\hat{\beta}_i\right).$$

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
path est.	-1.9912	-.0525	-.1232	-.1040	.3341	-.2699	.8292
std. error	1.0376	1.0060	.3722	.0845	.1511	.5396	.0550
p-value	.0550	.9584	.7071	.2184	.0347	.6170	.0000

Table 2.1. Estimated path coefficients, standard errors, and corresponding p-values for the hypothesized path diagram shown in Figure 2.1 are displayed. P-values are two-sided.

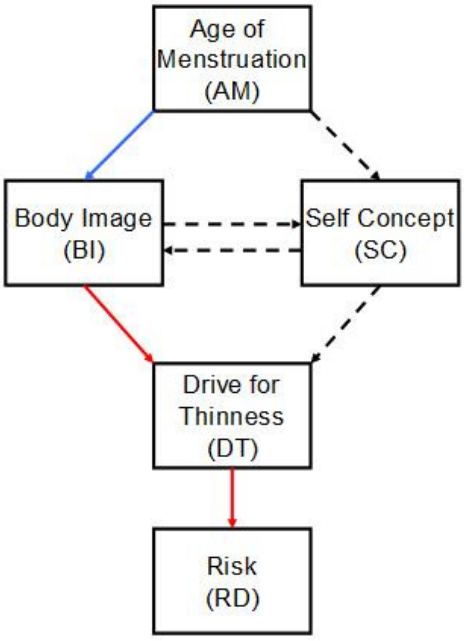


Figure 2.2. Results of the analysis of variables influencing development of eating disorders in women; these factors include age of first menstrual period (AM), body image score (BI), self concept score (SC), drive for thinness (DT), and risk for developing an eating disorder (RD). Dashed lines indicate insignificant influences, blue arrows represent a significantly negative relationship between the variable originating the influence and the variable receiving it, and red arrows represent a significantly positive relationship between the variables.

SEM is a very powerful tool for researchers investigating single-group and other types of models. For example, if we are interested in evaluating the difference between two groups of females (e.g. adolescent versus adult women) and how their SEM path coefficients may differ based on their difference in age, we could perform a multiple-group comparison available in most SEM software. This technique requires that we have independent groups, which of course, adolescent and adult women are. We could also evaluate the differences in the model for a single group of women whose body image, self concept, drive for thinness, and risk for developing a disorder are measured at two or more times (repeated measures). This type of model is also currently available in most SEM software. However, if we have a mixed design study, with multiple independent groups (say males and females) and repeated measures (say each group is measured at two time points, as adolescents and adults), there is currently no methodology or software to directly estimate the impact of group membership and treatment conditions on path coefficients. Mixed design studies are common in many fields, but there is no existing SEM procedure for this type of analysis. The dire need of an SEM procedure for mixed design studies is further evidenced by its increasing application in brain functional pathways studies where the mixed design is prevalent, as shown in the next section.

2.2 The Brain Functional Pathway Study

The mechanism of structural equation modeling makes it ideal for the analysis of directional brain functional pathways, and often brain imaging studies are conducted in a mixed design, as shown in the following study led by Dr. Nora D. Volkow, the Director of the National Institute on Drug Abuse, and conducted at the Brookhaven National Laboratory. The data are measurements of nine regions in the brain, related to the reward network: amygdala (AMYG), orbitofrontal cortex (OFC), anterior cingulate gyrus (ACG), ventral striatum (VS), thalamus (THAL), insula (INS), striatum putamen (PUT), cerebellum (CER), and motor frontal cortex (MFC). The experimental methods and extraction of data from PET scans are described in detail in [24]. There are two groups (normal subjects and drug abusers) and two possible treatments (placebo or methylphenidate) given to each patient. We have four data sets containing measurements of activity in each brain region, that is, one for each combination of group and treatment. All subjects from all groups are measured under both of the treatment conditions.

The data are taken from 25 cocaine abusers and 16 normal subjects, each under two conditions. The condition is the treatment administered. Patients receive placebo and methylphenidate (a cocaine-like substance known as Ritalin), one at a time. The conditions are randomly assigned, not given at specific time intervals, so this study can be classified as a repeated measures design. All subjects under our consideration are expecting to receive placebo, so there is no confounding expectation effect.

Dr. Volkow has proposed, based on her field knowledge, the following path diagram for a single group under a single condition, which can be estimated by any of the available SEM software, such as SAS PROC CALIS, LISREL, and EQS (see sample code in Appendix). However, as mentioned before, there is no cohesive method for evaluating differences in path strengths due to group membership and drug treatment simultaneously.

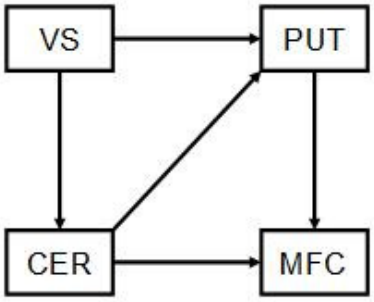


Figure 2.3. Hypothesized path diagram for our study of the reward network. The present variables are: ventral striatum (VS), striatum putamen (PUT), cerebellum (CER), and motor frontal cortex (MFC). Directed arrows indicate influence of one region on another in a direct cause and effect relationship.

2.2.1 Multivariate Normality Assumption

The maximum likelihood estimation method of SEM requires that all variables share a multivariate normal distribution. Distribution-free methods exist for estimating SEMs with non-normal data; however, we want to derive our model in the most

commonly used SEM framework, so we have evaluated the data to test for a multivariate normal distribution. We have used the R function *mvnorm.etest*, which is part of the optional energy package for the R software. Using this function to test the validity of our multivariate normal assumption for our data, we have determined that the variables in each of our independent datasets have multivariate normal distributions.

2.2.2 The Focus of this Study

To demonstrate the methodology we develop, we will analyze the effect of group membership and drug received on the strength of connectivity between regions of the brain. However, this methodology is designed to accommodate any number of dichotomous factors. Note that the hypothesized pathway shown in Figure 2.3 is *recursive*; it does not contain any feedback loops. Models containing feedback loops can be unstable and unidentified, so for the purposes of this dissertation, we focus strictly on recursive models.

2.3 Literature Review: Existing Methodology for Comparing Groups and Repeated Measures in SEM

The existing methods for analysis of multiple-group models are indirect and insufficient for models with small sample size, less-than-excellent fit, and/or repeated measures. Of course many statistical methods have been developed to analyze repeated measures data, panel data, and longitudinal data. Regrettably, no SEM models have been developed for mixed designs, evaluating differences in multiple groups and repeated measures, probably due to the different natures of research in social science and econometrics versus biomedical sciences, where the mixed design is quite prevalent. In addition to our extensive literature review, which follows, we have carried on correspondence with the senior statisticians from both LISREL and EQS software technical support, who have indicated that a proper analysis of mixed-design SEMs is not currently possible. It is very clear that no model has been developed that can evaluate multiple datasets (independent and otherwise) and test the significance of changes in path parameters due to both group effects and condition effects.

2.3.1 Multiple Group Comparison: Current Approaches and Limitations

We can look at this problem as one of interaction; we are interested in the interaction between group membership and strength of association between variables in an SEM. We are not interested in the effect of group membership on the dependent variable. Much work has been done on studying interaction effects in SEMs, but no solution is as easily understood and implemented as our own. Models are currently available to handle multiple groups or multiple occasions, but not both in a meaningful way (and notably, in a way that can be developed into a SEM that includes latent variables) [25]. Each existing model can be estimated in the current SEM software [26-28].

2.3.1.1 Goodness-of-Fit Method

The standard method for comparing the SEM of multiple independent groups is to constrain corresponding parameters to be equal [1]. The suggested method for

comparison is first determining whether models of the same form (same structure of parameter matrices) have good fit for each group of data, via a global chi-square fit statistic, pooled proportionally from individual model chi-squares [27]. If the overall model fit is good, we can continue by constraining first the matrices containing path coefficients and variance matrices to be equal, one step at a time. As the p-value drops below our significance level, we can declare that the newly added constraints cause a poor fit, and the models vary. In addition to noting a decreasing or stationary global fit chi-square index, we can also calculate a *difference* chi-square fit index to determine if the fit of the model is significantly changed [29]. Unfortunately, this method requires that groups being compared are independent. There is currently no methodology for directly comparing path coefficients and variances of two groups of data that are dependent (i.e. two sets of measurements taken from the same set of subjects). Another limitation of the above method is cohesiveness; it is a comparison of two separate models, rather than one unified model with differences represented within.

The standard method for comparing the SEM of multiple groups is to constrain corresponding parameters to be equal [1]. The *nested goodness-of-fit method* first involves determining whether models of the same form (same structure of parameter matrices) have good fit for each group of data, via the chi-square statistic. A statistically significant chi-square test indicates poor model fit. If the model fit is good, we can continue by constraining first the matrices containing path coefficients, then the variance matrices, to be equal one step at a time. If each model has continually good fit, then the models are invariant. If there is poor fit to the model at any level of constraint, then the models vary. The important thing to consider here is how much the models vary. If the path coefficients can be constrained equal, then the models obviously share some similarities. It is important to note that the variables must be measured in the same units if we are to truly consider models invariant. It does not make sense to compare relations over groups with data that has not been standardized into equivalent units.

At each state of constraining the models, a global chi-square variable is pooled proportionally from individual model chi-squares to determine overall fit of the models [27]. As the p-value drops below our significance level, we can declare that the newly added constraints cause a poor fit. It is important to note that our available degrees of freedom increase by a factor of the number of groups we are comparing; however, we are estimating more parameters in the same model, so the degrees of freedom are decreased accordingly. Obviously, when we constrain some parameters to be equal, we are estimating fewer parameters, so the degrees of freedom increase in cases of constrained parameters.

In addition to noting a decreasing or stationary global fit chi-square index, we can also calculate a *difference* chi-square fit index to determine if the fit of the model is significantly changed [29]. As discussed above, it is important to note the fit of unconstrained groups together in the same model. When parameters are constrained, again the output includes a global chi-square fit measure. According to Jaccard and Wan, we can take the difference of model chi-squares (before constraint and after) and their corresponding degrees of freedom to obtain the *difference chi-square* fit index. If the difference chi-square is statistically significant, then there is a significant difference in fit between the models—adding the given constraints has significantly decreased the model goodness-of-fit. Both LISREL [27, 29], and EQS [26], are capable of performing the

current goodness-of-fit tests in comparing the path coefficients of two groups. As of version 9.1.3, SAS Proc Calis does not have this functionality.

Unfortunately, the goodness-of-fit method requires that groups being compared are independent. Again, this makes this method inappropriate for our data set. There is currently no methodology for directly comparing path coefficients and variances of two groups of data that are dependent (i.e. two sets of measurements taken from the same set of subjects). Currently, in multiple group analysis, the groups must be independent. Our data describes two independent groups, but each group is measured under multiple conditions. In using the above methods, we have treated the groups as independent (when necessary), which of course can yield misleading results. Another limitation of the above methods is the cohesiveness. Each is a comparison of two separate models. While the goodness-of-fit method does incorporate the two models into consideration together, they are still two separate entities, not one unified model with differences represented within. The existing method for analysis of multiple-group models is indirect and insufficient for models with small sample size and/or less-than-excellent fit.

2.3.1.2 Incorporating Interaction Effects into SEMs

We can also look at this problem as one of interaction; we are interested in the interaction between group membership and the strength of association between variables in an SEM. Much work has been done on studying interaction effects on SEMs (Kenny and Judd in 1984, and most recently, Coenders et. al. [30]), but no solution is as easily understood and implemented as our own. Kenny and Judd proposed a method for incorporating interactions when the interacting variables are continuous [31]. Their method accounts for non-normality of the interaction term in the model by using the GLS estimation methodology rather than the ML method. GLS is appropriate for departures from normality that are not subject to excessive skewness or kurtosis. Therefore, if a binary covariate (like group membership) is included in an interaction term, GLS would be inappropriate. This method is also inappropriate for analyzing datasets that are not independent. If one or more of the interacting variables is discrete, then the multi-sample goodness-of-fit method can be applied for determining if path coefficients vary across groups [32], as discussed above, for independent groups only. Clearly, like the existing methods for comparing multiple samples, the existing methods for studying interactions effects in SEMs to not meet the needs of our analysis.

2.3.2 Analysis of Repeated Measures Data: Current Approaches and Limitations

One model for evaluating the relationship between a group of longitudinal repeated measures is called the Two-Wave Model by the LISREL User Guide [27], though it can be extended to include multiple time points [25]. Using this model, researchers can evaluate the relationships among multiple representations of a single variable measured at different time points. However, this model is not appropriate for models of unordered repeated measures, as in the case of evaluating multiple treatment conditions given to the same group of subjects, as there is no clear causal pathway that can be instituted. Also, in this representation, it is easy model a single variable at multiple time points, but the model is quickly complicated when we add more than one variable. In the case of our brain network model, we are interested in understanding how the relationships in the existing SEM pathway are changed from a normal subject's pathway

if the subject is a cocaine abuser or is currently receiving a cocaine-like substance. We are already evaluating the structural relationships among many variables; we simply want to add the capability of analyzing differences among groups and unordered repeated measures.

There are two more general directions for development of SEM in this field: latent growth modeling and multilevel SEM. Latent growth modeling was developed to model the change over time in repeated measures data as a function of a latent intercept and slope (plus polynomial terms as necessary) that are made up of an average value and the variability of the individual trajectories about that average [33]. The most basic latent growth models are traditional two-level regression models estimated using SEM framework because the models are easily drawn as SEM-style path diagrams. Influences on the latent intercept and slope and influences that vary with time can be added to the model. Multiple group analysis is possible in latent growth modeling through the traditional SEM goodness-of-fit comparison method, or by incorporating dummy group indicators. LGMs can include several response variables at once and/or several repeated measures at once (the most general model being the Auto-Regressive Latent Trajectory Model [34, 35]).

2.3.2.1 Latent Growth Modeling

Latent growth modeling (also called latent curve modeling or latent curve analysis) was developed to model the change over time in repeated measures data as a function of a latent intercept and slope (plus quadratic, cubic, etc. terms as necessary) that are made up of an average value and the variability of the individual trajectories about that average. Its formal introduction to the literature came from Meredith and Tisak in 1990 [33]. Though the topic of modeling trajectories based on an entire group of observations had been studied before, and by many, since the 1800s, Meredith and Tisak put it into its current form and presented it in the framework of structural equation modeling.

The most basic latent growth models are traditional two-level regression models. We use SEM to estimate the parameters because the models are most easily drawn as SEM-style path diagrams. The parameters are estimated through an SEM-style optimization procedure, by assimilating the latent and measured variables into the SEM framework. SEM is the most cohesive way to solve for these parameters because we can solve for all parameters simultaneously, rather than in the case-by-case method, where we solve for parameters one at a time, each affecting the next parameter. The question is whether we can apply the method of latent growth modeling to an existing SEM to incorporate repeated measures.

We model the trajectories of the repeated measures for one response variable with a latent intercept and slope variable, which affects all repeated measures. In more complicated models, we can add exogenous predictors to the model, influencing the latent intercept and slope, or add influences that also vary in time. Additionally, multiple group analysis is possible in latent growth modeling through incorporating dummy group indicators or the traditional SEM goodness-of-fit comparison method.

The simplest model of LGM is the unconditional model [34]. In this case, we take any number of repeated measures and model their trajectory with a latent slope and intercept. The repeated measures are for i subjects at t time points (repeated measures).

On level 1, we model the response variables as functions of the slope and intercept, plus error. On level 2 of the model, we allow the intercepts to vary across subjects. The model is shown below.

$$
\begin{aligned}
y_{it} &= \alpha_i + \lambda_t \beta_i + \epsilon_{it} \\
\alpha_i &= \mu_\alpha + \zeta_{\alpha_i} \\
\beta_i &= \mu_\beta + \zeta_{\beta_i} \\
i &= 1..N \\
t &= 1..T
\end{aligned}
$$

For this model, there are only T measured variables (the repeated measures). So we have T means and $\frac{1}{2}T(T+1)$ unique entries in the variance-covariance matrix of the available data. Therefore, we have $\frac{1}{2}T(T+3)$ known pieces of information. In order to have an identified model, we must have more knowns than unknowns. The unknown parameters include NT errors in the equations, T values for λ_t, K means, where K is the number of latent variables present in the model (K=2 for a linear model), and $\frac{1}{2}K(K+1)$ unique elements of the variance-covariance matrix for the K factors. This is far too many parameters to be estimated, so traditionally, we allow the errors in the equations to vary only with time, and remain constant for across individuals at different time points. Additionally, we fix $\lambda_t = t-1$ constant. Then we have $T+\frac{1}{2}K(K+3)$ unknown parameters to estimate. We can see for a linear model, that we must have at least $T=3$ waves of data available in order to have an identified model.

In estimating the model, we must find the mean and covariance structures to be estimated. We know that

$$
\begin{aligned}
y &= \Lambda\eta + \epsilon \\
\eta &= \Lambda\zeta
\end{aligned}
$$

as stated in the model description. Then we can find $\mu(\theta) = E(y) = \Lambda\mu_\eta$ and $\Sigma(\theta) = E\left[(y - E(y))(y - E(y))'\right] = \Lambda\Psi\Lambda' + \Theta_\epsilon$ using the matrix form of y. We can estimate the parameters by maximizing the multivariate normal likelihood function, or minimizing the fit function—a manipulated version of the likelihood function—as is traditional for SEM, shown below.

$$
F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[\Sigma^{-1}(\theta)S\right] + \left(\bar{y} - \mu(\theta)\right)' \Sigma^{-1}(\theta)\left(\bar{y} - \mu(\theta)\right) - \log\left|S\right| - T
$$

Standard errors for the estimates can be calculated using the asymptotic covariance matrix, which is the expected value of the Fisher Information matrix, shown below.

$$
ACOV\left(\hat{\theta}\right) = \frac{2}{N-1}\left[E\left(\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right)\right]^{-1}
$$

The unconditional model is the simplest LGM, but they can get very complicated from here. The conditional model incorporates exogenous measured variables as influences of the latent intercept and slope. These are included in the level 2 equations of the model, not affecting the level 1 equation.

$$
\begin{aligned}
y_{it} &= \alpha_i + \lambda_t \beta_i + \epsilon_{it} \\
\alpha_i &= \mu_\alpha + \gamma_{\alpha_1} x_{1i} + \gamma_{\alpha_2} x_{2i} + \zeta_{\alpha_i} \\
\beta_i &= \mu_\beta + \gamma_{\beta_1} x_{1i} + \gamma_{\beta_2} x_{2i} + \zeta_{\beta_i} \\
i &= 1 \ldots N \\
t &= 1 \ldots T
\end{aligned}
$$

The model shown above includes two measured covariates that do not vary with time. This model is estimated similarly, but here we have more known pieces of information, including the means and covariances associated with the independent variables. The $\mu(\theta)$ and $\Sigma(\theta)$ are different in that

$$
\begin{aligned}
y_i &= \Lambda \eta_i + \epsilon_i \\
\eta_i &= \mu_\eta + \Gamma x_i + \zeta_i
\end{aligned}
$$

where Γ holds the coefficients for the paths between the exogenous *x* variables and the latent slope and intercept. The likelihood function stays the same, assuming there are no missing values.

There are two methods for incorporating group membership and its effect on the trajectory into a latent growth model. One method is the standard multiple-group comparison method for SEM, the goodness-of-fit comparison method. We restrict parameters to be equal, one level at a time, and determine how the fit of the model changes as we further constrain the models. If the model fit becomes poor, then we have restricted it too far, and the newly restricted parameters cannot be considered equal.

In the second method, we add dummy group indicator variables to the level 2 equations, shown below.

$$
\begin{aligned}
\alpha_i &= \mu_\alpha + \gamma_{\alpha D_1} D_{1i} + \zeta_{\alpha_i} \\
\beta_i &= \mu_\beta + \gamma_{\beta D_1} D_{1i} + \zeta_{\beta_i}
\end{aligned}
$$

The dummy variables are assigned values of 0 and 1, according to the group membership of individual *i*. Bollen indicates this change to the model can be treated as a typical conditional model, treating the dummy variables as exogenous predictors. We can add traditional exogenous predictors to the model, and their interactions with the dummy variables as well, resulting in the following multiple group, multiple indicator model.

$$y_{it} = \alpha_i + \lambda_t \beta_i + \epsilon_{it}$$
$$\alpha_i = \mu_\alpha + \sum_{g=1}^{G-1} \gamma_{\alpha D_g} D_{ig} + \sum_{k=1}^{K} \gamma_{\alpha x_k} x_{ik} + \zeta_{\alpha i}$$
$$\beta_i = \mu_\beta + \sum_{g=1}^{G-1} \gamma_{\beta D g} D_{ig} + \sum_{k=1}^{K} \gamma_{\beta x_k} x_{ik} + \zeta_{\beta i}$$

Assume that the vector *x* contains exogenous measured predictors and any interactions between those predictors and the dummy variables *D*. Then we can estimate this as a conditional model, according to the literature. However, it is relevant to point out that while the likelihood function is developed under the assumption of multivariate normality, a dichotomous dummy variable does not satisfy this assumption.

The above framework applies only to LGMs for one response variable. There are many types of multivariate LGMs:

- Unconditional LGM with time-varying predictor variables—more than one repeated measure variable, but not modeling the growth of the time-varying predictors (still only one response variable);
- Conditional LGM with time-varying predictors—time invariant and time variant predictors added to the model;
- Modeling more than one response variable—considering more than one repeated measure, modeling the trajectories of multiple response variables;
- Adding time-invariant predictors to the previous model—conditional multivariate LGM;
- The Auto-Regressive Latent Trajectory Model—models more than one repeated measure, including unidirectional arrows from one measure to the next for the repeated measures (i.e. time *t-1* has direct effect on time *t*). Time 1 is considered independent from the latent intercept and slope, but can be influenced by any present predictor variables. Developed by Curran and Bollen [34, 35].

A further development of LGM is that we can model the trajectory of latent repeated measures over time. For example, consider three measurements of "intelligence" at different time points. We must have multiple indicators of the latent construct intelligence at each time point. In the simplest case, we have the same indicators of the construct at each time point. These models can be estimated in the current SEM software, and in particular, EQS has this capability and a user-friendly interface [26]. It is important to note that, when modeling the mean structure in EQS, which is not always considered in an SEM, EQS employs an additional constant variable, regressing all predictors and the latent factors on the constant to estimate their means.

Latent growth modeling has many features and capabilities; however, this type of model will not work for analyzing our data. LGM generally focuses on one response variable and its repeated measures. Because we already have a system modeled using SEM, we have three response variables and one predictor. LGM typically models the trajectory of change over time, or ordered repeated measures. Our repeated measures do not have an inherent order, they are simply different treatments administered to the individual in a random order; therefore, modeling the change in the variables over time

does not make sense for our data. Other issues include the complexity of incorporating three response variables into the same model. Additionally, LGM does not incorporate a direct influence from one response variable to another. LGMs model all available relationships, not specifically chosen relations between variables, according to a SEM hypothesis, in which we are interested. LGM is the application of a simple repeated measures model in SEM, not an additional methodology for existing SEMs with repeated measures. While LGMs do have many of the features we are looking for, like multiple group analysis and modeling change between repeated measures, it is clear that, inherently, LGM is not the appropriate analysis method for our data.

2.3.2.2 Multilevel Modeling and Multilevel SEM

Multilevel modeling is a method for analyzing hierarchical data. Hierarchical data is data nested in levels; for example, observations taken from students nested within classrooms nested within schools. In each level, the subjects' measurements are typically correlated due to having the same teacher in each classroom, or the same art classes within the schools. Another example would be measurements taken from a number of children, nested within families. The measurements of children who are siblings are correlated by their similar genetics.

In its most basic form, a two-level regression model is as follows [36].

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j}
\end{aligned}
$$

The equation for Y represents the lowest level, individual response variable. The equations for the coefficients represent relations with group-level variables, like teacher or mother, for example. An individual's regression coefficients are allowed to vary from the average coefficient value, as in latent growth modeling.

Multilevel SEM is generally used to classify within-group and between-group variation for hierarchical, or nested, data structures [36]. The technique of multilevel SEM comes directly from the method developed by Cronbach and Webb [37] for multilevel factor models. Each subject's measurements are partitioned into group means and within-group variations [26]. The covariance matrix is then partitioned into a between-group covariance matrix and a within-group covariance matrix by calculating the covariance matrices of the partitioned measures. First, the measurements *Y* are partitioned into group means and within-group variations, shown below.

$$
\left.
\begin{aligned}
Y_B &= \overline{Y}_g \\
Y_W &= Y_{ig} - \overline{Y}_g
\end{aligned}
\right\}
\quad \Rightarrow \quad
Y_{ig} = Y_T = Y_B + Y_W
$$

The covariance matrix is partitioned similarly.

$$
\Sigma_T = \Sigma_B + \Sigma_W
$$

The same SEM model is fitted simultaneously to model both covariance matrices. For the simplest case (balanced groups) Muthen [38] showed that

$$S_{PW} = \frac{\sum_{g}^{G}\sum_{i}^{n}\left(Y_{gi} - \overline{Y}_{g}\right)\left(Y_{gi} - \overline{Y}_{g}\right)'}{N - G}$$

is the MLE of Σ_W and

$$S_B^* = \frac{\sum_{g}^{G} n\left(\overline{Y} - \overline{Y}_{g}\right)\left(\overline{Y} - \overline{Y}_{g}\right)'}{G - 1}$$

is the MLE of $\Sigma_W + c\Sigma_B$, where c is the common group size.

SEMs are fit simultaneously to model both covariance matrices; using the results, as well as intraclass correlations of the factors in the SEM, we can tell if the proposed covariance structure is significant at each level of the model [26]. For example, if we are studying quantitative skills and self esteem at both the individual and school level, then based on the significance of the variances of the latent factors representing those variables, we can tell if variance is significant between schools, only between individuals, or at both levels, and how those variables are correlated. This method can be extended to any number of levels [39], though the model will become increasingly complicated.

While multilevel SEM has many valuable features, this method is not appropriate for analysis of our data. Though we could nest our repeated measures within an individual and nest those individuals within their groups (cocaine abusers or normal subjects), to construct a traditional three-level hierarchical data set [36, 40, 41], multilevel SEM requires a large sample size on each level since each cluster on each level is considered a single observation. If a researcher is interested in studying less than fifty clusters at the group level, then multilevel SEM cannot be applied. In fact, multilevel SEM requires huge sample sizes, especially if the groups are unbalanced [26]. Ideally, a researcher must have hundreds of clusters in the top of the hierarchical model (i.e. hundreds of groups) to use the ML estimation method, which is the most common and easiest to use [26, 39]. We have only two groups to compare in our example data, each with repeated measures. Additionally, it is not intuitive to determine the meaning of paths between measured variables that are significant at each level. We would not be pinpointing the difference in path strengths between levels and groups; we could only evaluate the variance between those repeated measures, between individuals, and between groups. In the model we have developed, we can directly determine whether group and drug affect each individual path, and by how much. Multilevel SEM is also difficult to implement and interpret, where our novel method (derived in Chapter 3) is very easy to use for both statisticians and researchers in the behavioral sciences who use SEM.

2.3.3 SEM Software Packages and their Limitations

LISREL, EQS, and SAS (PROC CALIS) can estimate a standard SEM if the model is identified. Each has the typical features such as multiple estimation methods. LISREL and EQS will each build the model from a path diagram drawn by the user. SAS allows several model description methods, but does not allow drawing of a path diagram. Currently, LISREL and EQS are both capable of multiple-group comparisons, but require that the groups are independent (no repeated measures) and make use of the goodness-of-fit method for comparison (constraining increasingly many model parameters equal and comparing model fits). AMOS, too, can estimate both single and multiple-group models, using the goodness-of-fit approach for multiple-group analyses [42]. As of version 9.1.3, SAS is not capable of multiple-group SEM but a new procedure, PROC TCALIS in version 9.2, will have this capability. I assume that this method will simply be a replicate of the goodness-of-fit methods already being used by LISREL and EQS.

All SEM software that can estimate a single-group SEM can estimate the coefficients in a linear growth model, as it is simply a traditional SEM representation for modeling repeated measures data. Multilevel modeling, particularly multilevel SEM is possible in LISREL [43, 44] and EQS [26].

Chapter 3: SEM for Mixed Designs

This chapter contains the novel development of SEM for mixed design analysis. First, we will introduce our theoretical development through several simpler models that will converge eventually to a general SEM for Mixed Designs (MD-SEM). We then discuss its numerical implementation through MATLAB. Finally, we apply the newly developed model to analyze the brain functional pathway study introduced in Chapter 2.

3.1 SEM with Multiple Groups

3.1.1 Case 1: Two Independent Groups with a Single Equation

Consider the case of two groups, each with the same model of a single equation as shown in Figure 3.1(a). Specifically, observations of group 1 contain variables X_1, Y_1 and the model contains measurement error ζ_1, as shown in Figure 3.1(b). For the first group, the SEM parameters are:

$$\mathrm{B}_1 = [0], \Gamma_1 = [\gamma_1], \Phi_1 = [\phi_1], \Psi_1 = [\psi_1].$$

Observations of group 2 contain variables X_2, Y_2 and the model contains measurement error ζ_2, as shown in Figure 3.1(c). For the second group, the SEM parameters are:

$$\mathrm{B}_2 = [0], \Gamma_2 = [\gamma_2], \Phi_2 = [\phi_2], \Psi_2 = [\psi_2].$$

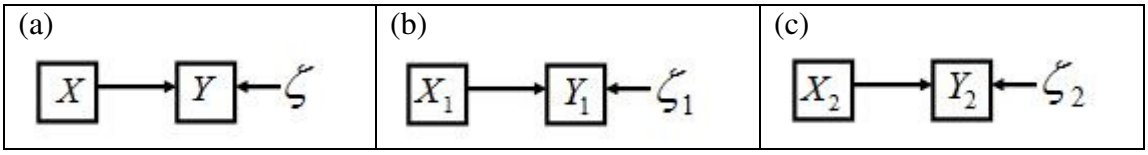


Figure 3.1. Illustrations of a single-equation SEM: (a) the model shared by both groups, (b) the model specific to group 1, (c) the model specific to group 2.

In order to create one SEM for these two groups, we will append the X's, Y's, and ζ's each into one vector. $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \zeta = \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$. Now we can use our traditional equation for SEM: $Y = \mathrm{B}Y + \Gamma X + \zeta$.

To unify these two groups, we will need single $\mathrm{B}, \Gamma, \Psi, \Phi$ matrices, rather than one for each model. With our unified X, Y, and ζ vectors, we can create these as follows:

$$\mathrm{B} = \begin{matrix} \\ Y_1 \\ Y_2 \end{matrix} \overbrace{\begin{matrix} Y_1 & Y_2 \end{matrix}}^{} \begin{bmatrix} \mathrm{B}_1 & 0 \\ 0 & \mathrm{B}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

B is a block diagonal matrix, relating variables in Y1 to itself (one-equation case) and variables in Y2 to itself (one-equation case), where the first block is the B matrix from the first group, if the model were run individually, and the second block is the B matrix from the second group, if the model were run individually. (This is just as the Σ matrix is made a block matrix in the standard, single-group model.) The off-diagonal blocks are 0 because there are no path relations between Y1 and Y2. Similarly,

$$\Gamma = \begin{matrix} \\ Y_1 \\ Y_2 \end{matrix} \overbrace{\begin{matrix} X_1 & X_2 \end{matrix}}^{} \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}.$$

However, the above would not produce a result any different from that of running each group as an independent SEM model. We would still get path coefficients and have to compare them by another method. In order to compare them within the model, the path coefficients should reflect a *change* from one group's coefficient to the next. Figure 3.2(a) illustrates how we incorporate the group factor into a model.

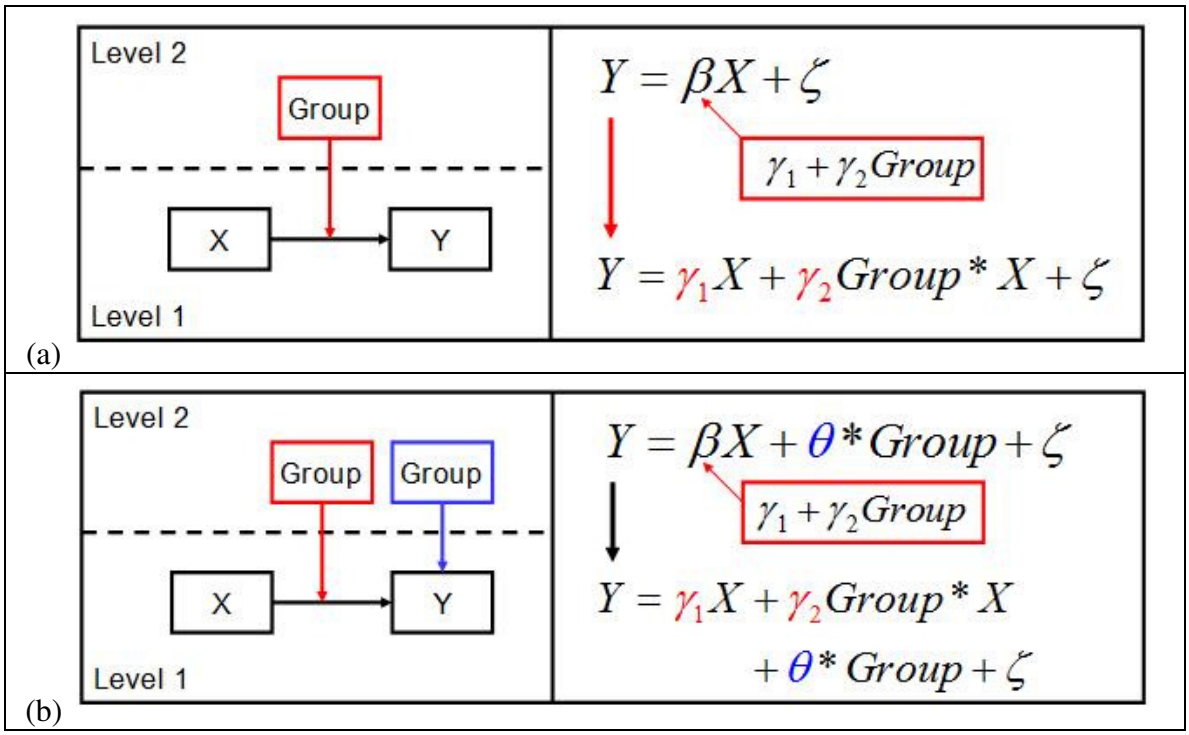


Figure 3.2. (a) An illustration of our two-level parametric modeling approach for MD-SEM analysis. Here we focus on a single equation (left: path diagram, right: equation) with covariate *group* asserting its effect on the path from node X to node Y.
(b) An illustration of our two-level parametric modeling approach for MD-SEM analysis with *group* as a covariate and as an independent variable. Here we focus on a single equation (left: path diagram, right: equation) with covariate *group* asserting its effect on the path from node X to node Y, and independent variable *group* asserting its effect on node Y.

We represent the change not as a traditional multi-level model with error incorporated in the second level, but as a reparametrization of the path coefficient to incorporate a change to the coefficient due to group membership. Additionally, we are not adding a new variable to a regression-style equation and adding its interaction, we are only adding the interaction. We do not add an individual group term to the model because this would add a new node to the model. The meaning of the new node would be "does the group membership have an effect on the present variable?" For the brain functional pathway study, we are not interested in whether group membership affects the activity level in a particular node, except in relation to the other nodes (i.e. does group membership strengthen or weaken the signal from node A to node B). The information we are interested in would be confounded by the presence of something we are not interested in and do not know how to interpret. It is equivalent to adding extra variables to a regression when we are really interested not in how that variable affects the dependent variable, but how it affects the regression coefficients.

While we are not currently interested in adding group membership to the model as an influence on the response variables (as shown in Figure 3.2(b)), it is certainly possible to incorporate this effect within our new maximum likelihood framework. In Figure

3.2(b), group appears as a variable pointing to other variables, not only as a covariate pointing to paths. The equation for this model appears on the right hand side of Figure 3.2(b); group appears as a variable, and would thereby need to be included in the likelihood function of the model. However, for two independent groups (0 and 1), we create a likelihood function for each group and take their product to form the likelihood function for the model. For Group 0, the $\theta * Group$ term would disappear, and the likelihood for group 0 would not contain an extra variable. For Group 1, θ would appear as an intercept in the equation. While it is traditional in SEM to center all variables about zero to eliminate the mean structure, general SEM theory does allow for modeling the mean structure of the variables [26, 27, 34], so we could easily incorporate this intercept into the likelihood for group 1. This addition implies that our new model can handle categorical nodes, as long as they appear only as independent variables and not as dependent variables (no arrows coming in to the categorical variable). Otherwise, we could not easily incorporate this capability by treating θ as an intercept.

Because we are currently interested in the effect of each covariate on the paths in the model exclusively, and not on each variable in the model, we will reparametrize each path coefficient as is illustrated in Figure 3.2(a). Incorporating the reparametrization shown in the Figure 3.2(a) into our B and Γ matrices is natural, shown below.

$$\mathrm{B} = \begin{matrix} \\ Y_1 \\ Y_2 \end{matrix} \overset{Y}{\begin{matrix} Y_1 & Y_2 \end{matrix}} \begin{bmatrix} \mathrm{B}_1 & 0 \\ 0 & \mathrm{B}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \Gamma = \begin{matrix} \\ Y_1 \\ Y_2 \end{matrix} \overset{X}{\begin{matrix} X_1 & X_2 \end{matrix}} \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1 + \gamma_2 \end{bmatrix}.$$

Thus, from our two-level modeling approach, the two groups do share certain common parameters and thus we must analyze them through the joint likelihood function as shown below.

We will continue to model the Φ and Ψ matrices independently, but unified, as shown below.

$$\Phi = \begin{matrix} \\ X_1 \\ X_2 \end{matrix} \overset{X}{\begin{matrix} X_1 & X_2 \end{matrix}} \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \qquad \Psi = \begin{matrix} \\ \zeta_1 \\ \zeta_2 \end{matrix} \overset{\zeta}{\begin{matrix} \zeta_1 & \zeta_2 \end{matrix}} \begin{bmatrix} \Psi_1 & 0 \\ 0 & \Psi_2 \end{bmatrix} = \begin{bmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{bmatrix}.$$

Φ is the covariance matrix of the X's, and we can break this into a block matrix. $Cov(X_1, X_1)$ has its own name, Φ_1. $Cov(X_1, X_2) = Cov(X_2, X_1) = 0$ because the X variables from the two groups are independent. Ψ is similar, as the errors in equations from different groups are independent.

Now we can derive the Σ matrix of the two-group model. The model has not changed, only the parameter matrices. From the model, we have $Y = (I - \mathrm{B})^{-1}(\Gamma X + \zeta)$. Then

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}.$$

As before, we can derive each block of this matrix separately.

$$\begin{aligned}
\Sigma_{yy}(\theta) &= E(YY') = E\left[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)\left((I-\mathrm{B})^{-1}(\Gamma X+\zeta)\right)' \right] \\
&= E\left[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'} \right] \\
&= (I-\mathrm{B})^{-1} E(\Gamma XX'\Gamma'+\zeta X'\Gamma'+\Gamma X\zeta'+\zeta\zeta')(I-\mathrm{B})^{-1'} \\
&= (I-\mathrm{B})^{-1}(\Gamma EXX'\cdot\Gamma'+0+0+E\zeta\zeta')(I-\mathrm{B})^{-1'} \\
&= (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1'} \\
&= 1*\left(\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{bmatrix} \right) *1 \\
&= \begin{bmatrix} \gamma_1^2\phi_1+\psi_1 & 0 \\ 0 & (\gamma_1+\gamma_2)^2\phi_2+\psi_2 \end{bmatrix}
\end{aligned}$$

This matrix is appropriate because there should be no relation between Y2 and Y1, so we would expect the off-diagonal blocks to be 0.

Similarly,

$$\begin{aligned}
\Sigma_{yx}(\theta) &= EYX' = E\left[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)X' \right] \\
&= (I-\mathrm{B})^{-1} E(\Gamma XX'+\zeta X') = (I-\mathrm{B})^{-1}(\Gamma EXX'+E\zeta EX') \\
&= (I-\mathrm{B})^{-1}\Gamma\Phi \\
&= 1*\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} = \begin{bmatrix} \gamma_1\phi_1 & 0 \\ 0 & (\gamma_1+\gamma_2)\phi_2 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{xy}(\theta) &= EXY' = E\left[X\left((I-\mathrm{B})^{-1}(\Gamma X+\zeta)\right)' \right] \\
&= E\left(X(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'} \right) = E(XX'\Gamma'+X\zeta')(I-\mathrm{B})^{-1'} \\
&= (EXX'\cdot\Gamma'+EXE\zeta')(I-\mathrm{B})^{-1'} = \Phi\Gamma'(I-\mathrm{B})^{-1'} \\
&= \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} *1 = \begin{bmatrix} \gamma_1\phi_1 & 0 \\ 0 & (\gamma_1+\gamma_2)\phi_2 \end{bmatrix}
\end{aligned}$$

.

There is a problem with the above two equations for the covariance matrix of Y and X; we must consider this matrix, too, as a block matrix.

$$\Sigma_{yx}(\theta) = \begin{bmatrix} Cov(Y_1,X_1) & Cov(Y_1,X_2) \\ Cov(Y_2,X_1) & Cov(Y_2,X_2) \end{bmatrix}$$

Also, we must have $Cov(Y_1, X_2) = Cov(Y_2, X_1) = 0$ since we want independent groups. Then we can derive $Cov(Y_1, X_1), Cov(Y_2, X_2)$ as we would in a single-group model with just those variables—so we would use just the corresponding B, Γ, Φ matrices. It already happens that the off-diagonal elements are 0 for this matrix, so this issue resolved within the B, Γ matrices.

Finally, $\Sigma_{xx}(\theta) = EXX' = \Phi = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix}$. Therefore, we can write

$$\Sigma(\theta) = \begin{bmatrix} \gamma_1^2\phi_1 + \psi_1 & 0 & \gamma_1\phi_1 & 0 \\ 0 & (\gamma_1+\gamma_2)^2\phi_2 + \psi_2 & 0 & (\gamma_1+\gamma_2)\phi_2 \\ \gamma_1\phi_1 & 0 & \phi_1 & 0 \\ 0 & (\gamma_1+\gamma_2)\phi_2 & 0 & \phi_2 \end{bmatrix}$$

for a two-group, single-equation model. However, even though the two groups are independent, we do not constrain $\Sigma_{xy}(\theta) = \Sigma_{yx}(\theta) = 0$ because it is true that Y1 and X1 maybe be correlated, and Y2 and X2 may be correlated.

3.1.2 Likelihood Function for Case 1

We have two independent groups in a single model. For the purposes of the maximum likelihood estimation technique, we will construct a likelihood function for each independent group. Each group has its own B, Γ, Σ matrices, but they contain shared parameters, uniting this model. We can create the two independent likelihood functions and multiply them together. First, we will calculate the likelihood for one arbitrary group of the two.

We know that $f(z;\Sigma) = (2\pi)^{-\frac{p+q}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z'\Sigma^{-1} z\right)$ is the PDF for one group's vector of endogenous and exogenous variables, Z. Z is centered about mean 0. We take N observations of Z, from N different subjects from our chosen group, and the resulting joint PDF of Z is $f(z_1, z_2, ..., z_N; \Sigma) = f(z_1;\Sigma) f(z_2;\Sigma) \cdots f(z_N;\Sigma)$. Then

$$L(\theta) = (2\pi)^{-\frac{N}{2}(p+q)} |\Sigma(\theta)|^{-\frac{N}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} z_i' \Sigma^{-1}(\theta) z_i \right).$$

We have two groups, so we will denote corresponding $\Sigma(\theta)$ matrices, z vectors, and sample sizes with appropriate subscripts in the full likelihood function.

Because these groups are independent, we can multiply their individual likelihood functions to construct the full likelihood for this multi-group model. Therefore,

$$L(\theta) = (2\pi)^{-\frac{N_1}{2}(p+q)} \left|\Sigma_1(\theta)\right|^{-\frac{N_1}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_1} z_{1i}' \Sigma_1^{-1}(\theta) z_{1i} \right)$$
$$\cdot (2\pi)^{-\frac{N_2}{2}(p+q)} \left|\Sigma_2(\theta)\right|^{-\frac{N_2}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_2} z_{2i}' \Sigma_2^{-1}(\theta) z_{2i} \right)$$

is the full likelihood function for the multi-group model.

We can simplify this likelihood for easier numerical implementation by first taking the log of the function, shown below.

$$\log L(\theta) = -\tfrac{N_1}{2}(p+q)\log 2\pi - \tfrac{N_1}{2}\log\left|\Sigma_1(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_1} z_{1i}' \Sigma_1^{-1}(\theta) z_{1i}$$
$$-\tfrac{N_2}{2}(p+q)\log 2\pi - \tfrac{N_2}{2}\log\left|\Sigma_{01}(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_2} z_{2i}' \Sigma_2^{-1}(\theta) z_{2i}$$

Now, as before, we can remove constant terms, and a factor of $-\frac{1}{2}$ from the log-likelihood to further simplify numerical implementation of the function. The resulting function is shown below.

$$N_1 \log\left|\Sigma_1(\theta)\right| + \sum_{i=1}^{N_1} z_{1i}' \Sigma_1^{-1}(\theta) z_{1i} + N_2 \log\left|\Sigma_2(\theta)\right| + \sum_{i=1}^{N_2} z_{2i}' \Sigma_2^{-1}(\theta) z_{2i}$$

The resulting function is an extension of the fit function used in Case 1, derived above. It is the simplified log-likelihood functions of the individual groups added together. We can further simplify the above equation to resemble the sum of two traditional F_{ML} fit functions.

$$\Rightarrow \quad N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[\tfrac{1}{N_1} \sum_{i=1}^{N_1} z_{1i}' \Sigma_1^{-1}(\theta) z_{1i} \right] + N_2 \log\left|\Sigma_2(\theta)\right| + N_2 tr\left[\tfrac{1}{N_2} \sum_{i=1}^{N_2} z_{2i}' \Sigma_2^{-1}(\theta) z_{2i} \right]$$
$$\Rightarrow \quad N_1 \log\left|\Sigma_1(\theta)\right| + tr\left[\Sigma_1^{-1}(\theta) \sum_{i=1}^{N_1} z_{1i} z_{1i}' \right] + N_2 \log\left|\Sigma_2(\theta)\right| + tr\left[\Sigma_2^{-1}(\theta) \sum_{i=1}^{N_2} z_{2i} z_{2i}' \right]$$
$$= N_1 \log\left|\Sigma_1(\theta)\right| + tr\left[N_1 \Sigma_1^{-1}(\theta) S_1^* \right] + N_2 \log\left|\Sigma_2(\theta)\right| + tr\left[N_2 \Sigma_2^{-1}(\theta) S_2^* \right]$$
$$= N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1^* \Sigma_1^{-1}(\theta) \right] + N_2 \log\left|\Sigma_2(\theta)\right| + N_2 tr\left[S_2^* \Sigma_2^{-1}(\theta) \right]$$

Here again, $S^* = \frac{N-1}{N} S$, and for large samples, we can simply replace S* with S. Adding the constant terms, our fit function for this case is

$$F_{ML} = N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1 \Sigma_1^{-1}(\theta) \right] - N_1 \log\left|S_1\right| - N_1(p_1 + q_1)$$
$$+ N_2 \log\left|\Sigma_2(\theta)\right| + N_2 tr\left[S_2 \Sigma_2^{-1}(\theta) \right] - N_2 \log\left|S_2\right| - N_2(p_2 + q_2) .$$

3.1.3 Standard Error of Parameters for Case 1

As in the case of a single-group SEM model, in order to estimate the standard errors of the parameters we have estimated, we must use the information we have about the MLE of θ.

We know the maximum likelihood (ML) estimator, $\hat{\theta}$, of parameter vector θ, is distributed as

$$N\left(\theta, \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}\right).$$

Then the asymptotic covariance matrix of the ML estimator of arbitrary θ is

$$ACOV\left(\hat{\theta}\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

As before, we must determine the relationship between $\log L(\theta)$ and F_{ML}, and let the asymptotic covariance matrix reflect this difference.

As calculated above,

$$\begin{aligned} F_{ML} &= N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right] - N_1 \log\left|S_1\right| - N_1(p_1+q_1) \\ &\quad + N_2 \log\left|\Sigma_2(\theta)\right| + N_2 tr\left[S_2\Sigma_2^{-1}(\theta)\right] - N_2 \log\left|S_2\right| - N_2(p_2+q_2) \end{aligned}, \text{ and}$$

$$\begin{aligned} \log L(\theta) &= -\tfrac{N_1}{2}(p+q)\log 2\pi - \tfrac{N_1}{2}\log\left|\Sigma_1(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_1} z_{1i}'\Sigma_1^{-1}(\theta) z_{1i} \\ &\quad -\tfrac{N_2}{2}(p+q)\log 2\pi - \tfrac{N_2}{2}\log\left|\Sigma_{01}(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_2} z_{2i}'\Sigma_2^{-1}(\theta) z_{2i} \end{aligned}$$

We must determine the relation between $-\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}$ and $\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}$ to determine how the calculation of the asymptotic covariance matrix must change.

In manipulating $\log L(\theta)$ into F_{ML}, we remove a constant term, multiply by -2, manipulate one term into an equal term, then add a constant term. The effects on the equality of derivative are only the multiplying factor of -2. So we can say

$$\left\{-E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \left\{-E\left[\frac{\partial^2 -2\log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \left\{2E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \frac{1}{2}\left\{E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

and therefore,

$$ACOV\left(\hat{\theta}\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = 2\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

So the standard errors of the parameter estimates are the square roots of the respective diagonal elements of the matrix above.

3.2 SEM with Repeated Measures

The data we want to model include two groups of subjects measured under several different treatments. Obviously, measurements taken of the same variables from the same subjects, regardless of treatment, will be correlated, so groups of measurements taken from the same subjects cannot be treated as independent measurements. Unfortunately, the terms *repeated measures* and *longitudinal data* seem to be used interchangeably in the literature; however, they are not equal. Repeated measures are measurements taken from the same subject under different conditions (multiple treatments, for example) or at different time points, the same variables measured each time. Longitudinal data are measurements taken from the same subject at different time points, the same variables measured each time. Longitudinal data represent repeated measurements, but repeated measurements are not strictly longitudinal. Many studies have focused on regression and SEM for longitudinal data [34, 45, 46] but there is not as much information about repeated measures as we define them.

There are two choices available for incorporating repeated measures into any regression-style model. Our initial thought was to incorporate correlations between the independent variables present in the two datasets representing measurements from the same group. Then since each y is written in relation to those x variables, of course those correlations would be present for each variable. However, a more traditional representation is to show correlation between these repeated measures by correlating the errors in the regression equations [29, 47-52]. It is common to implement a general linear model with correlated errors for longitudinal data or repeated measures data [53]. We will apply this traditional treatment of repeated measures data in regression to the regression-style equations that compose an SEM; we will represent errors in the equations as correlated in corresponding regression equations. We will not represent correlations between different equations for different treatment conditions since we consider errors in the equations independent from one another in the single-group case. We will also represent correlations among the independent variables.

3.2.1 Case 2: One Group in Two Conditions with a Single Equation

Consider the two-group model above, but now we have two sets of measurements from each person in a single group of subjects. Correlation of those measurements should be represented in our model. Traditionally, to represent repeated measures in a statistical model, we correlate the error terms in the model. Therefore, we will correlate the errors in the equations in our SEM model. Additionally, we will allow for free correlation of the independent variables, X, for measurements taken from the same group of subjects. A path diagram incorporating this correlation is shown in Figure 3.3. The curved line shows a correlation between the two ζ variables. This represents the scenario when we have one group observed on two conditions.

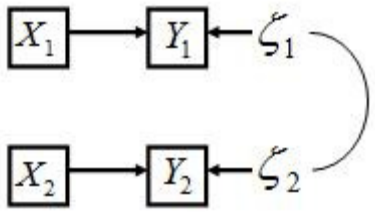


Figure 3.3. Illustration of a single-group model where each subject is measured under two conditions (repeated measures). Measurement error of Y is modeled for each condition, ζ_1, ζ_2, and those errors must be correlated for each subject (indicated by the curved line).

The path parameters will remain the same. Adding correlation among variables does not add unidirectional paths, so

$$\mathrm{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \Gamma = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1 + \gamma_2 \end{bmatrix}.$$

As usual, the Φ matrix represents the covariance matrix of the exogenous variables, X. For a single-datatset SEM, this matrix is modeled freely by the researcher. We will construct the Φ matrix as follows.

$$\Phi = \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix}$$

We will also modify the Ψ matrix, which represents the covariance matrix of the vector ζ, the errors in the equations. We will add terms to represent correlation between the errors in the single equation modeled for two treatment conditions on the same group of subjects. For this case, the Ψ matrix will not be diagonal, but of course it is still symmetric:

$$\Psi = \begin{matrix} \\ \zeta_1 \\ \zeta_2 \end{matrix} \begin{matrix} \overbrace{\phantom{\begin{bmatrix} \psi_1 & \psi_2 \end{bmatrix}}}^{\zeta_1 \quad \zeta_2} \\ \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{bmatrix} \end{matrix}.$$

So ψ_2 will be the estimated parameter representing the correlation between repeated measures in the system.

Because the model stays the same, $Y = \mathrm{B}Y + \Gamma X + \zeta$, these are the only places where a change could occur, so the possible correlations of X's and Y's should now exist in the $\Sigma(\theta)$ matrix, derived below. As before,

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

There is no change to the derivations of each block in the $\Sigma(\theta)$ matrix, so we can use the formulas from the previous derivation.

$$\Sigma_{yy}(\theta) = E(YY') = E\left[(I-\mathrm{B})^{-1}(\Gamma X + \zeta)\left((I-\mathrm{B})^{-1}(\Gamma X + \zeta)\right)' \right]$$

$$= E\left[(I-\mathrm{B})^{-1}(\Gamma X + \zeta)(X'\Gamma' + \zeta')(I-\mathrm{B})^{-1'} \right]$$

$$= (I-\mathrm{B})^{-1} E(\Gamma XX'\Gamma' + \zeta X'\Gamma' + \Gamma X\zeta' + \zeta\zeta')(I-\mathrm{B})^{-1'}$$

$$= (I-\mathrm{B})^{-1}(\Gamma EXX'\cdot\Gamma' + 0 + 0 + E\zeta\zeta')(I-\mathrm{B})^{-1'}$$

$$= (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-\mathrm{B})^{-1'}$$

$$= 1*\left(\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix} \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} + \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{bmatrix} \right) * 1$$

$$= \begin{bmatrix} \gamma_1\phi_1 & \gamma_1\phi_3 \\ \phi_3(\gamma_1+\gamma_2) & \phi_2(\gamma_1+\gamma_2) \end{bmatrix} \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} + \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{bmatrix}$$

$$= \begin{bmatrix} \gamma_1^2\phi_1 & \gamma_1\phi_3(\gamma_1+\gamma_2) \\ \gamma_1\phi_3(\gamma_1+\gamma_2) & (\gamma_1+\gamma_2)^2\phi_2 \end{bmatrix} + \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{bmatrix} = \begin{bmatrix} \gamma_1^2\phi_1+\psi_1 & \gamma_1\phi_3(\gamma_1+\gamma_2)+\psi_2 \\ \gamma_1\phi_3(\gamma_1+\gamma_2)+\psi_2 & (\gamma_1+\gamma_2)^2\phi_2+\psi_3 \end{bmatrix}$$

$$\Sigma_{yx}(\theta) = EYX' = E\left[(I-\mathrm{B})^{-1}(\Gamma X + \zeta)X' \right]$$

$$= (I-\mathrm{B})^{-1} E(\Gamma XX' + \zeta X') = (I-\mathrm{B})^{-1}(\Gamma EXX' + E\zeta EX')$$

$$= (I-\mathrm{B})^{-1}\Gamma\Phi$$

$$= 1*\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix} = \begin{bmatrix} \gamma_1\phi_1 & \gamma_1\phi_3 \\ \phi_3(\gamma_1+\gamma_2) & (\gamma_1+\gamma_2)\phi_2 \end{bmatrix}$$

$$\Sigma_{xy}(\theta) = EXY' = E\left[X\left((I-\mathrm{B})^{-1}(\Gamma X + \zeta)\right)' \right]$$

$$= E\left(X(X'\Gamma' + \zeta')(I-\mathrm{B})^{-1'} \right) = E(XX'\Gamma' + X\zeta')(I-\mathrm{B})^{-1'}$$

$$= (EXX'\cdot\Gamma' + EXE\zeta')(I-\mathrm{B})^{-1'} = \Phi\Gamma'(I-\mathrm{B})^{-1'}$$

$$= \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix} \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_1+\gamma_2 \end{bmatrix} * 1 = \begin{bmatrix} \gamma_1\phi_1 & \phi_3(\gamma_1+\gamma_2) \\ \phi_3\gamma_1 & (\gamma_1+\gamma_2)\phi_2 \end{bmatrix}$$

$$\Sigma_{xx}(\theta) = EXX' = \Phi = \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix}.$$

Therefore,

$$\Sigma(\theta)=\begin{bmatrix}\Sigma_{yy}(\theta) & \Sigma_{yx}(\theta)\\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta)\end{bmatrix}=\begin{bmatrix}\gamma_1^2\phi_1+\psi_1 & \gamma_1\phi_3(\gamma_1+\gamma_2)+\psi_2 & \gamma_1\phi_1 & \gamma_1\phi_3\\ \gamma_1\phi_3(\gamma_1+\gamma_2)+\psi_2 & (\gamma_1+\gamma_2)^2\phi_2+\psi_3 & \phi_3(\gamma_1+\gamma_2) & (\gamma_1+\gamma_2)\phi_2\\ \gamma_1\phi_1 & \phi_3(\gamma_1+\gamma_2) & \phi_1 & \phi_3\\ \gamma_1\phi_3 & (\gamma_1+\gamma_2)\phi_2 & \phi_3 & \phi_2\end{bmatrix}$$

All variables should now have some calculable correlation.

3.2.2 Likelihood Function for Case 2

When we append our X, Y, and ζ vectors in our multi-group model, we are effectively constructing one single model for all variables, including the interaction of these variables with group membership or other dichotomous indicator variables. Therefore, we are assuming that the variables (from all included groups) are all multivariate normally distributed. For the two-condition model described above, we append our vectors as shown below.

$$Y=\begin{pmatrix}Y_1\\ Y_2\end{pmatrix}, X=\begin{pmatrix}X_1\\ X_2\end{pmatrix}, \zeta=\begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}$$

Because in our case, these variables represent the same group of subjects' measurements taken under different conditions, they are of course correlated, but they also have the same sample size. We already know that these variables have covariance matrix derived above.

Therefore, consider $Z=\begin{pmatrix}Y\\ X\end{pmatrix}$, where variables in Z are centered with means about 0. Then Z has PDF

$$f(z;\Sigma)=(2\pi)^{-\frac{p+q}{2}}\left|\Sigma\right|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}z'\Sigma^{-1}z'\right),$$

and likelihood function created from N observations

$$L(\theta)=(2\pi)^{-\frac{N}{2}(p+q)}\left|\Sigma(\theta)\right|^{-\frac{N}{2}}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}z_i'\Sigma^{-1}(\theta)z_i\right).$$

The difference between this and the independent group case is that here, we are considering two groups at the same time, so Z contains two groups of variables. In the previous multi-group case, we can consider the likelihoods independently because the groups are uncorrelated and the $\Sigma(\theta)$ matrices do not contain inter-group correlations. Here, however, the $\Sigma(\theta)$ do contain inter-group correlations, so these variables must be considered in the same likelihood function.

We simplify the likelihood as before, by taking its logarithm, dropping a constant term, and removing a factor of $-\frac{1}{2}$:

$$\log L(\theta) = -\tfrac{N}{2}(p+q)\log 2\pi - \tfrac{N}{2}\log\left|\Sigma(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N} z_i{}'\Sigma^{-1}(\theta) z_i$$

$$\Rightarrow \quad -\tfrac{N}{2}\log\left|\Sigma(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N} z_i{}'\Sigma^{-1}(\theta) z_i$$

$$\Rightarrow \quad N\log\left|\Sigma(\theta)\right| + \sum_{i=1}^{N} z_i{}'\Sigma^{-1}(\theta) z_i$$

As for the single-group MLE SEM, this fit function simplifies to

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\right| - (p+q).$$

Z contains both groups' variables. $\Sigma(\theta)$ is the matrix derived for this situation in the previous section, also shown above. Note, however, that the sample size, N is equal only to the number of observations in one of the "groups", since this really represents the number of subjects measured, since the sample is the same for both groups.

3.2.3 Standard Error of Parameters for Case 2

As usual, standard errors (square root of variances of parameter estimates) of the estimates found in the previous section must be calculated via the asymptotic covariance matrix (inverse of Fisher Information Matrix). For case 2, the fitting function is exactly the same as in the case of a single-group MLE SEM, so the standard errors are calculated in the same way.

$$ACOV\left(\hat{\theta}\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \frac{2}{N}\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

The standard errors of the parameter estimates are the square roots of the respective diagonal elements of the matrix above.

3.3 SEM for Mixed Designs: Multiple Groups with Repeated Measures

3.3.1 Case 3: Mixed Design—Multiple Groups with Repeated Measures

Now consider the case of two groups, each under two conditions, while the two groups are independent from one another.

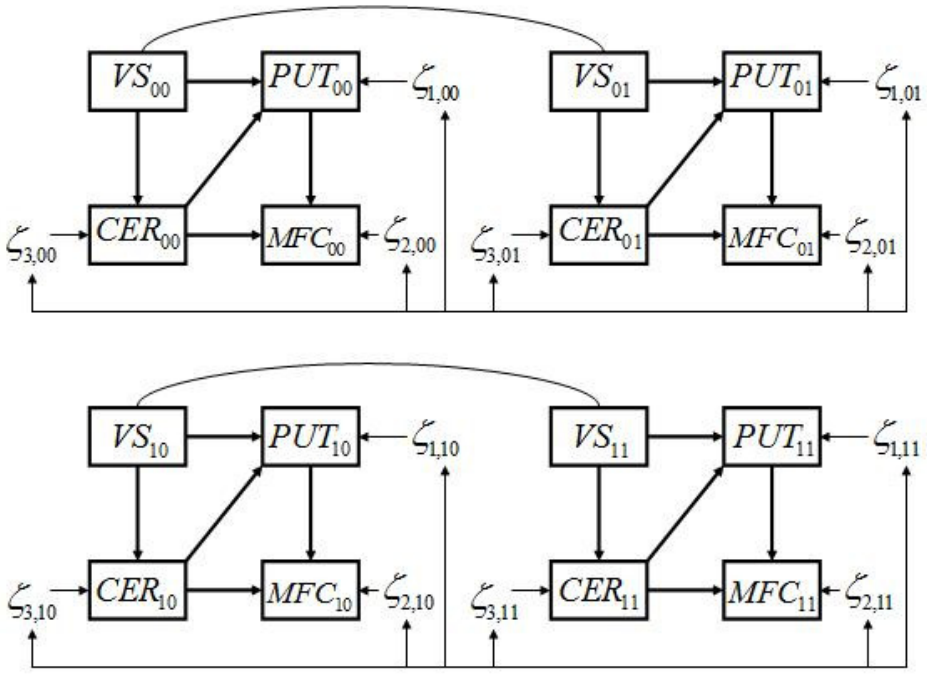


Figure 3.4. Hypothesized path diagram for the Brain Functional Pathway Study, adapted to describe the mixed design with two independent groups of subjects measured under two conditions. We model the repeated measures by modeling correlation of errors terms, ζ, in the equations, as shown above. (Top: Group 0, normal subjects, measured under two conditions; Bottom: Group 1, cocaine abusers, measured under two conditions.)

As depicted in Figure 3.4, let Group 0 be measured under two conditions (and as a result, all variables of condition 0 are correlated with those of condition 1), and let Group 1 be measured under the same conditions. Group 0 is independent of Group 1. Again, the model here is $Y = \mathrm{B}Y + \Gamma X + \zeta$, and we will append the X, Y, and ζ matrices as shown below, where 00 and 01 represent Group 0 under two conditions and 10 and 11 represent Group 1 under the same two conditions.

$$Y = \begin{pmatrix} Y_{00} \\ Y_{01} \\ Y_{10} \\ Y_{11} \end{pmatrix}, X = \begin{pmatrix} X_{00} \\ X_{01} \\ Y_{10} \\ Y_{11} \end{pmatrix}, \zeta = \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{pmatrix}$$

Adding these correlations does not add any path parameters, so the B and Γ matrices are block diagonal matrices. We will allow for free correlation of the exogenous variables, X, for measurements taken on the same group of subjects. Therefore, the Φ matrix is constructed of blocks, as shown below.

$$\Phi = \begin{bmatrix} \Sigma_{x_{00}x_{00}} & \Sigma_{x_{00}x_{01}} & \Sigma_{x_{00}x_{10}} & \Sigma_{x_{00}x_{11}} \\ \Sigma_{x_{01}x_{00}} & \Sigma_{x_{01}x_{01}} & \Sigma_{x_{01}x_{10}} & \Sigma_{x_{01}x_{11}} \\ \Sigma_{x_{10}x_{00}} & \Sigma_{x_{10}x_{01}} & \Sigma_{x_{10}x_{10}} & \Sigma_{x_{10}x_{11}} \\ \Sigma_{x_{11}x_{00}} & \Sigma_{x_{11}x_{01}} & \Sigma_{x_{11}x_{10}} & \Sigma_{x_{11}x_{11}} \end{bmatrix}$$

The diagonal entries of the preceding matrix are the symmetric Φ matrices from the corresponding datasets. Because groups 0 and 1 are independent, $\Sigma_{x_{00}x_{10}} = \Sigma_{x_{01}x_{00}} = 0$, $\Sigma_{x_{00}x_{11}} = \Sigma_{x_{11}x_{00}} = 0$, $\Sigma_{x_{01}x_{10}} = \Sigma_{x_{10}x_{01}} = 0$, and $\Sigma_{x_{01}x_{11}} = \Sigma_{x_{11}x_{01}} = 0$. We will assign a correlation matrix for $\Sigma_{x_{00}x_{01}} = \Phi_{12}$ and $\Sigma_{x_{10}x_{11}} = \Phi_{34}$.

$$\Phi = \begin{bmatrix} \Phi_{00} & \Phi_{12} & 0 & 0 \\ \Phi_{12}' & \Phi_{01} & 0 & 0 \\ 0 & 0 & \Phi_{10} & \Phi_{34} \\ 0 & 0 & \Phi_{34}' & \Phi_{11} \end{bmatrix}$$

Note that the correlation matrices are not equal if they represent two different sets of data.

$$\Phi_{12} = [q_1] \qquad \Phi_{34} = [q_2]$$

We will also estimate the entries of the remaining Φ matrices, but they are traditional covariance matrices of parameters.

We will also incorporate the correlated pairs into the Ψ matrix. The Ψ matrix is constructed of blocks, as shown below.

$$\Psi = \begin{bmatrix} \Sigma_{\zeta_{00}\zeta_{00}} & \Sigma_{\zeta_{00}\zeta_{01}} & \Sigma_{\zeta_{00}\zeta_{10}} & \Sigma_{\zeta_{00}\zeta_{11}} \\ \Sigma_{\zeta_{01}\zeta_{00}} & \Sigma_{\zeta_{01}\zeta_{01}} & \Sigma_{\zeta_{01}\zeta_{10}} & \Sigma_{\zeta_{01}\zeta_{11}} \\ \Sigma_{\zeta_{10}\zeta_{00}} & \Sigma_{\zeta_{10}\zeta_{01}} & \Sigma_{\zeta_{10}\zeta_{10}} & \Sigma_{\zeta_{10}\zeta_{11}} \\ \Sigma_{\zeta_{11}\zeta_{00}} & \Sigma_{\zeta_{11}\zeta_{01}} & \Sigma_{\zeta_{11}\zeta_{10}} & \Sigma_{\zeta_{11}\zeta_{11}} \end{bmatrix}$$

As in the previous case, where all groups were independent, $\Sigma_{\zeta_{00}\zeta_{00}} = \Psi_{00}$, $\Sigma_{\zeta_{01}\zeta_{01}} = \Psi_{01}$, $\Sigma_{\zeta_{10}\zeta_{10}} = \Psi_{10}$, and $\Sigma_{\zeta_{11}\zeta_{11}} = \Psi_{11}$. Because groups 0 and 1 are independent, $\Sigma_{\zeta_{00}\zeta_{10}} = \Sigma_{\zeta_{01}\zeta_{00}} = 0$, $\Sigma_{\zeta_{00}\zeta_{11}} = \Sigma_{\zeta_{11}\zeta_{00}} = 0$, $\Sigma_{\zeta_{01}\zeta_{10}} = \Sigma_{\zeta_{10}\zeta_{01}} = 0$, and $\Sigma_{\zeta_{01}\zeta_{11}} = \Sigma_{\zeta_{11}\zeta_{01}} = 0$.

We can assign a correlation matrix for $\Sigma_{\zeta_{00}\zeta_{01}} = \Psi_{12}$ and $\Sigma_{\zeta_{10}\zeta_{11}} = \Psi_{34}$.

$$\Psi = \begin{bmatrix} \Psi_{00} & \Psi_{12} & 0 & 0 \\ \Psi_{12}' & \Psi_{01} & 0 & 0 \\ 0 & 0 & \Psi_{10} & \Psi_{34} \\ 0 & 0 & \Psi_{34}' & \Psi_{11} \end{bmatrix}$$

Note that correlation matrices are not equal if they represent two different sets of variables.

Because an individual's measurements for brain region A will be correlated under different treatments, we will model the following parameters:

$$\Psi_{12} = \begin{bmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{bmatrix} \qquad \Psi_{34} = \begin{bmatrix} p_4 & 0 & 0 \\ 0 & p_5 & 0 \\ 0 & 0 & p_6 \end{bmatrix}.$$

We will also estimate elements of $\Psi_{00}, \Psi_{01}, \Psi_{10}, \Psi_{11}$ as in the previous case, but they are traditionally symmetric covariance matrices. For our data, they are diagonal.

Now that we have defined $\mathrm{B}, \Gamma, \Phi, \Psi$, we can substitute them into the model, $Y = \mathrm{B}Y + \Gamma X + \zeta$. As usual, we are interested in the statistical hypothesis $\Sigma = \Sigma(\theta)$, so we must find $\Sigma(\theta)$. As in previous cases,

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} = \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-\mathrm{B})^{-1'} & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(I-\mathrm{B})^{-1'} & \Phi \end{bmatrix}.$$

3.3.2 Likelihood Function for Case 3

In this case, we have two groups of subjects; each subject is measured under two conditions. Because we have two sets of measurements for the same group of subjects, these measurements will be correlated. Therefore, we cannot split up the likelihood function by considering four different distribution functions. The correlated measurements should share a distribution function. We will consider the two groups independently, but their two sets of measurements must be considered together. Therefore, we can create two separate likelihood functions, one for each group (including its repeated measures), and multiply them together.

Consider group 0. As in Case 3, we have one group, measured under two treatment conditions. We can list the two data sets, which refer to the same subjects, as in previous cases:

$$Y_0 = \begin{pmatrix} Y_{00} \\ Y_{01} \end{pmatrix}, X_0 = \begin{pmatrix} X_{00} \\ X_{01} \end{pmatrix}, \zeta_0 = \begin{pmatrix} \zeta_{00} \\ \zeta_{01} \end{pmatrix}.$$

Let $Z_0 = \begin{pmatrix} Y_0 \\ X_0 \end{pmatrix}$. Z is distributed multivariate normal. We must consider $\mathrm{B}, \Gamma, \Phi, \Psi$ for a model consisting of these two groups only, in order to derive μ and Σ for this MVN distribution. As for two independent groups,

$$\mathrm{B}_0 = \begin{bmatrix} \mathrm{B}_{00} & 0 \\ 0 & \mathrm{B}_{01} \end{bmatrix} \qquad \Gamma_0 = \begin{bmatrix} \Gamma_{00} & 0 \\ 0 & \Gamma_{01} \end{bmatrix}$$

since correlations between variables do not add paths to these matrices. Additionally,

$$\Phi_0 = \begin{bmatrix} \Phi_{00} & \Phi_{12} \\ \Phi_{12}' & \Phi_{01} \end{bmatrix} \qquad \Psi_0 = \begin{bmatrix} \Psi_{00} & \Psi_{12} \\ \Psi_{12}' & \Psi_{01} \end{bmatrix}.$$

Therefore, we can say

$$\Sigma_0(\theta) = \begin{bmatrix} \Sigma_{y_0 y_0}(\theta) & \Sigma_{y_0 x_0}(\theta) \\ \Sigma_{x_0 y_0}(\theta) & \Sigma_{x_0 x_0}(\theta) \end{bmatrix} = \begin{bmatrix} (I - \mathrm{B}_0)^{-1}(\Gamma_0 \Phi_0 \Gamma_0' + \Psi_0)(I - \mathrm{B}_0)^{-1'} & (I - \mathrm{B}_0)^{-1}\Gamma_0 \Phi_0 \\ \Phi_0 \Gamma_0' (I - \mathrm{B}_0)^{-1'} & \Phi_0 \end{bmatrix}.$$

Now that we have stated the mean and covariance matrix of the distribution, we can write the likelihood function. Because all variables in Z are distributed as multivariate normal, they have the following distribution.

$$f(z_0; \Sigma_0) = (2\pi)^{-\frac{p+q}{2}} \left|\Sigma_0\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} z_0' \Sigma_0^{-1} z_0\right)$$

When N_0 observations of each variable is taken, we can construct the likelihood as follows.

$$\begin{aligned} L_0(\theta) &= f(z_1; \Sigma(\theta)) f(z_2; \Sigma(\theta)) \cdots f(z_{N_0}; \Sigma(\theta)) \\ &= (2\pi)^{-\frac{N_0(p+q)}{2}} \left|\Sigma_0(\theta)\right|^{-\frac{N_0}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_0} z_i' \Sigma_0^{-1}(\theta) z_i \right). \end{aligned}$$

Now we must consider Group 1. As for Group 0, we must construct $Z_1, \mathrm{B}_1, \Gamma_1, \Phi_1, \Psi_1$. Let

$$Z_1 = \begin{pmatrix} Y_1 \\ X_1 \end{pmatrix} = \begin{pmatrix} Y_{10} \\ Y_{11} \\ X_{10} \\ X_{11} \end{pmatrix}.$$

Then

$$\mathrm{B}_1 = \begin{bmatrix} \mathrm{B}_{10} & 0 \\ 0 & \mathrm{B}_{11} \end{bmatrix}, \Gamma_1 = \begin{bmatrix} \Gamma_{10} & 0 \\ 0 & \Gamma_{11} \end{bmatrix}, \Phi_1 = \begin{bmatrix} \Phi_{10} & \Phi_{34} \\ \Phi_{34}' & \Phi_{11} \end{bmatrix}, \Psi_1 = \begin{bmatrix} \Psi_{10} & \Psi_{34} \\ \Psi_{34}' & \Psi_{11} \end{bmatrix}.$$

Now in order to construct the likelihood function for group 1, we must first write the mean and covariance matrices for the variables in Z_1, which are distributed as multivariate normal.

$$\Sigma_1(\theta) = \begin{bmatrix} \Sigma_{y_1 y_1}(\theta) & \Sigma_{y_1 x_1}(\theta) \\ \Sigma_{x_1 y_1}(\theta) & \Sigma_{x_1 x_1}(\theta) \end{bmatrix} = \begin{bmatrix} (I - \mathrm{B}_1)^{-1}(\Gamma_1 \Phi_1 \Gamma_1' + \Psi_1)(I - \mathrm{B}_1)^{-1'} & (I - \mathrm{B}_1)^{-1}\Gamma_1 \Phi_1 \\ \Phi_1 \Gamma_1' (I - \mathrm{B}_1)^{-1'} & \Phi_1 \end{bmatrix}$$

Now we can write the distribution function for Z_1 as

$$f(z_1; \Sigma_1) = (2\pi)^{-\frac{p+q}{2}} \left|\Sigma_1\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} z_1' \Sigma_1^{-1} z_1\right).$$

When N_1 observations of each variable is taken, we can construct the likelihood as follows.

$$\begin{aligned} L_1(\theta) &= f(z_1; \Sigma_1(\theta)) f(z_2; \Sigma_1(\theta)) \cdots f(z_{N_1}; \Sigma_1(\theta)) \\ &= (2\pi)^{-\frac{N_1(p+q)}{2}} \left|\Sigma_1(\theta)\right|^{-\frac{N_1}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_1} z_i' \Sigma_1^{-1}(\theta) z_i \right). \end{aligned}$$

Now that we have the likelihood functions for each of the two independent groups, 0 and 1, we can write the full likelihood for our two-group model as a product of the two likelihoods, shown below.

$$\begin{aligned} L(\theta) = (2\pi)^{-\frac{N_0(p+q)}{2}} \left|\Sigma_0(\theta)\right|^{-\frac{N_0}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_0} z_i' \Sigma_0^{-1}(\theta) z_i \right) \\ \cdot (2\pi)^{-\frac{N_1(p+q)}{2}} \left|\Sigma_1(\theta)\right|^{-\frac{N_1}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_1} z_i' \Sigma_1^{-1}(\theta) z_i \right) \end{aligned}$$

We will maximize this likelihood function to get the appropriate estimates for the variables in θ. To simplify the optimization, we can maximize the log-likelihood instead, and drop the constant terms in the function.

$$\log L(\theta) = -\tfrac{N_0(p+q)}{2}\log 2\pi - \tfrac{N_0}{2}\log\left|\Sigma_0(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_0} z_i'\Sigma_0^{-1}(\theta)z_i$$
$$-\tfrac{N_1(p+q)}{2}\log 2\pi - \tfrac{N_1}{2}\log\left|\Sigma_1(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_1} z_i'\Sigma_1^{-1}(\theta)z_i$$
$$\Rightarrow \quad -\tfrac{N_0}{2}\log\left|\Sigma_0(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_0} z_i'\Sigma_0^{-1}(\theta)z_i - \tfrac{N_1}{2}\log\left|\Sigma_1(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_1} z_i'\Sigma_1^{-1}(\theta)z_i$$

As in previous cases, we can remove the constant factor of $-\frac{1}{2}$ and minimize the resulting function rather than maximize it.

$$N_0\log\left|\Sigma_0(\theta)\right| + \sum_{i=1}^{N_0} z_i'\Sigma_0^{-1}(\theta)z_i + N_1\log\left|\Sigma_1(\theta)\right| + \sum_{i=1}^{N_1} z_i'\Sigma_1^{-1}(\theta)z_i$$

In order to assimilate our fit function to the style of the traditional fitting function, we manipulate the summing terms, as follows.

$$N_0\log\left|\Sigma_0(\theta)\right| + \sum_{i=1}^{N_0} z_i'\Sigma_0^{-1}(\theta)z_i + N_1\log\left|\Sigma_1(\theta)\right| + \sum_{i=1}^{N_1} z_i'\Sigma_1^{-1}(\theta)z_i$$
$$= N_0\log\left|\Sigma_0(\theta)\right| + tr\left[\sum_{i=1}^{N_0} z_i'\Sigma_0^{-1}(\theta)z_i\right] + N_1\log\left|\Sigma_1(\theta)\right| + tr\left[\sum_{i=1}^{N_1} z_i'\Sigma_1^{-1}(\theta)z_i\right]$$
$$= N_0\log\left|\Sigma_0(\theta)\right| + tr\left[\Sigma_0^{-1}(\theta)\sum_{i=1}^{N_0} z_i z_i'\right] + N_1\log\left|\Sigma_1(\theta)\right| + tr\left[\Sigma_1^{-1}(\theta)\sum_{i=1}^{N_1} z_i z_i'\right]$$
$$= N_0\log\left|\Sigma_0(\theta)\right| + tr\left[\Sigma_0^{-1}(\theta)N_0S_0^*\right] + N_1\log\left|\Sigma_1(\theta)\right| + tr\left[\Sigma_1^{-1}(\theta)N_1S_1^*\right]$$
$$= N_0\log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0^*\Sigma_0^{-1}(\theta)\right] + N_1\log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1^*\Sigma_1^{-1}(\theta)\right]$$

where S_0^*, S_1^* reflect $S^* = \frac{N-1}{N}S$, and S can be substituted for S* for large N. They are the sample covariance matrices for groups 0 and 1, respectively.

Adding the traditional constant terms, our final fitting function for the mixed model is

$$F_{ML} = N_0\log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] - N_0\log\left|S_0\right| - N_0(p_0+q_0)$$
$$+ N_1\log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right] - N_1\log\left|S_1\right| - N_1(p_1+q_1)$$

Minimizing F_{ML} will yield appropriate maximum likelihood estimates for the parameters in the mixed model.

3.3.3 Standard Error of Parameters for Case 3

As usual, standard errors (square root of variances of parameter estimates) of the estimates found in the previous section must be calculated via the asymptotic covariance matrix (inverse of Fisher Information Matrix). In order to estimate the standard errors of

the parameters we have estimated, we must use the information we have about the MLE of θ.

We know the maximum likelihood estimator, $\hat{\theta}$, of parameter vector θ, is distributed as

$$N\left(\theta, \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}\right).$$

Then the asymptotic covariance matrix of the ML estimator of arbitrary θ is

$$ACOV\left(\hat{\theta}\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

As before, we must determine the relationship between $\log L(\theta)$ and F_{ML}, and let the asymptotic covariance matrix reflect this difference.

$$\begin{aligned} F_{ML} = {} & N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] - N_0 \log\left|S_0\right| - N_0(p_0+q_0) \\ & + N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right] - N_1 \log\left|S_1\right| - N_1(p_1+q_1) \end{aligned}, \text{ and}$$

$$\log L(\theta) = -\tfrac{N_0}{2}(p+q)\log 2\pi - \tfrac{N_0}{2}\log\left|\Sigma_0(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_0} z_{0i}'\Sigma_0^{-1}(\theta) z_{0i}$$

$$-\tfrac{N_1}{2}(p+q)\log 2\pi - \tfrac{N_1}{2}\log\left|\Sigma_1(\theta)\right| - \tfrac{1}{2}\sum_{i=1}^{N_1} z_{1i}'\Sigma_1^{-1}(\theta) z_{1i}$$

We must determine the relation between $-\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}$ and $\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}$ to determine how the calculation of the asymptotic covariance matrix must change.

In manipulating $\log L(\theta)$ into F_{ML}, we remove a constant term, multiply by -2, manipulate one term into an equal term, then add a constant term. The effects on the equality of derivative are only the multiplying factor or -2. So we can say

$$\left\{-E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \left\{-E\left[\frac{\partial^2 - 2\log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \left\{2E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = \frac{1}{2}\left\{E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

and therefore,

$$ACOV\left(\hat{\theta}\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1} = 2\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

The standard errors of the parameter estimates are the square roots of the respective diagonal elements of the matrix above.

3.3.4 Overall Model Chi-Square Goodness-of-Fit Test for Case 3

A chi-square goodness-of-fit test for this model can be derived in similar fashion to that of the original SEM model described in Chapter 2.

First, note that our fitting function for this model is

$$\begin{aligned} F_{ML} &= N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] - N_0 \log\left|S_0\right| - N_0(p_0+q_0) \\ &\quad + N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right] - N_1 \log\left|S_1\right| - N_1(p_1+q_1) \end{aligned}.$$

Without considering the constant terms that do not affect the choice of θ,

$$\log L_{H_0} = -\tfrac{1}{2}\left[N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left(S_0\Sigma_0^{-1}(\theta)\right) + N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left(S_1\Sigma_1^{-1}(\theta)\right)\right]$$

For our alternative hypothesis, consider the most general situation, where $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$ are any positive definite matrices. Then

$$\log L_{H_1} = -\tfrac{1}{2}\left[N_0 \log\left|\hat{\Sigma}_0\right| + N_0 tr\left(S_0\hat{\Sigma}_0^{-1}\right) + N_1 \log\left|\hat{\Sigma}_1\right| + N_1 tr\left(S_1\hat{\Sigma}_1^{-1}\right)\right].$$

The function above is maximized when $\hat{\Sigma}_0 = S_0$ and $\hat{\Sigma}_1 = S_1$. Substituting these values into the function above,

$$\begin{aligned} \log L_{H_1} &= -\tfrac{1}{2}\left[N_0 \log\left|S_0\right| + N_0 tr\left(S_0S_0^{-1}\right) + N_1 \log\left|S_1\right| + N_1 tr\left(S_1S_1^{-1}\right)\right] \\ &= -\tfrac{1}{2}\left[N_0 \log\left|S_0\right| + N_0\left(p_0+q_0\right) + N_1 \log\left|S_1\right| + N_1 tr\left(p_1+q_1\right)\right] \end{aligned}.$$

We know that $-2\log\left(\dfrac{L_{H_0}}{L_{H_1}}\right) \sim \chi^2$, so

$$\begin{aligned} -2\log\left(\frac{L_{H_0}}{L_{H_1}}\right) &= -2\log L_{H_0} + 2\log L_{H_1} \\ &= -2\left(-\tfrac{1}{2}\left(N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] + N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right]\right)\right) \\ &\quad + 2\left(-\tfrac{1}{2}\left(N_0 \log\left|S_0\right| + N_0(p_0+q_0) + N_1 \log\left|S_1\right| + N_1(p_1+q_1)\right)\right) \\ &= N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] + N_1 \log\left|\Sigma_1(\theta)\right| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right] \\ &\quad - N_0 \log\left|S_0\right| - N_0(p_0+q_0) - N_1 \log\left|S_1\right| - N_1(p_1+q_1) \end{aligned}.$$

Therefore,

$$N_0 \log|\Sigma_0(\theta)| + N_0 tr\left[S_0\Sigma_0^{-1}(\theta)\right] + N_1 \log|\Sigma_1(\theta)| + N_1 tr\left[S_1\Sigma_1^{-1}(\theta)\right]$$
$$- N_0 \log|S_0| - N_0(p_0+q_0) - N_1 \log|S_1| - N_1(p_1+q_1)$$

is distributed asymptotically as χ^2 with $\left\{\sum_{i=0}^{G-1} \frac{1}{2}(p_i+q_i)(p_i+q_i+1)\right\} - t$ where G represents the total number of independent groups present in the study, and p_i+q_i represents the total number of variables measured for group *i* (over all datasets for group *i*). This function is our fitting function, so we can say that the value of our fitting function can be tested via the chi-square distribution. Compare this statistic to the upper $100(1-\alpha)$ percentile value of the chi-square distribution to determine whether to accept or reject the null hypothesis. A rejected null hypothesis indicates that $\Sigma = \Sigma(\theta)$ does not hold, and that the model poorly fits the data.

3.4 General Formulation of the Methodology

This section contains the development of SEM for mixed design analysis via a two-level parametric modeling approach. For simplicity of notation, we will introduce the theoretical development through a mixed effect model in a 2^k-factorial design. We then discuss its numerical implementation through MATLAB. Finally, we apply the new method to analyze brain glucose metabolic images to assess functional pathways involved in drug reward using data from a PET study that compared the brain responses to a stimulant drug (methylphenidate) between addicted and non addicted subjects.

MD-SEM is ideal for a 2^k-factorial design. A single-group SEM has the following format:

$$y_i = \gamma_{1i}x_1 + \gamma_{2i}x_2 + \cdots + \gamma_{qi}x_q + \beta_{1i}y_1 + \beta_{2i}y_2 + \cdots + \beta_{pi}y_p + \zeta_i,$$

$$\text{or,}\quad Y = \Gamma X + \mathrm{B}Y + \zeta,$$

where *q* is the number of exogenous variables, *i* ranges from 1 to *p*, the number of endogenous variables, and ζ represents the included error term. Please assume for the purposes of this paper that all variables, *X* and *Y*, have been centered about their means. Also, note that we use the LISREL matrix notation for presentation of our SEM.

The goal of MD-SEM is to determine the influence of one or more dichotomous factors (group membership, treatment, condition, etc.), on the strength of the path coefficients. In order to be able to compare the path coefficients within this single model, the path coefficients should reflect a *change* from one factor's coefficient to the next. Figure 3.2(a) illustrates how we incorporate a single *group* factor into a model.

As shown in Figure 3.2(a), we represent the change as a reparametrization of the path coefficient to incorporate a change to the coefficient due to group membership. As mentioned in the preceding sections, we are not adding a new variable to a regression-style equation and adding its interaction, we are only adding the interaction. Currently, we are not interested in whether group membership affects the activity level in a

particular node, except in relation to the other nodes (i.e. does group membership strengthen or weaken the signal from node X to node Y).

Incorporating the reparametrization shown in Figure 3.2(a) into our model is natural. For each coefficient, γ and β, we have

$$\begin{aligned}\gamma_{uv} &= \gamma_{uv,0} + \gamma_{uv,1}F_1 + \gamma_{uv,2}F_2 + \cdots + \gamma_{uv,k}F_k \\ \beta_{uv} &= \beta_{uv,0} + \beta_{uv,1}F_1 + \beta_{uv,2}F_2 + \cdots + \beta_{uv,k}F_k\end{aligned},$$

where F_i is the i^{th} dichotomous factor in this mixed design.

Then the simplified model is

$$\begin{aligned}y_i = &(\gamma_{1i,0} + \gamma_{1i,1}F_1 + \cdots + \gamma_{1i,k}F_k)x_1 + \cdots + (\gamma_{qi,0} + \gamma_{qi,1}F_1 + \cdots + \gamma_{qi,k}F_k)x_q \\ &+ (\beta_{1i,0} + \beta_{1i,1}F_1 + \cdots + \beta_{1i,k}F_k)y_1 + \cdots + (\beta_{pi,0} + \beta_{pi,1}F_1 + \cdots + \beta_{pi,k}F_k)y_p + \zeta_i\end{aligned}$$

for i =1 through p.

Many of the $\gamma_{ij,k}$ and $\beta_{ij,k}$ will be set to zero by the researcher, as not all of the X and Y variables are present in each equation describing the model. Then the remaining coefficients will be compared to zero by statistical tests. As each factor variable, F, will take a value of 0 or 1, the overall model, given the factor values for each dataset, will be a structural equation model. The coefficients of each variable in the SEM will vary among datasets, where some models include additive changes due to the influence of specific factors, as shown in the diagram of our model.

Now we must construct the model matrices that describe this system. First, we create a vector of X variables and a vector of Y variables, appending the X variables from each dataset into one vector, and the Y variables from each dataset into another. We will have 2^k datasets, each with q X variables and p Y variables, each with a number of observations. Then we have

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_{2^k} \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{2^k} \end{pmatrix}, \zeta = \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_{2^k} \end{pmatrix}.$$

Then we can construct the B and Γ matrices naturally, as block matrices, shown below.

$$\mathrm{B} = \begin{array}{c} \\ Y_1 \\ \vdots \\ Y_{2^k} \end{array} \begin{array}{c} \overbrace{\begin{array}{ccc} Y_1 & \cdots & Y_{2^k} \end{array}} \\ \begin{bmatrix} \mathrm{B}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathrm{B}_{2^k} \end{bmatrix} \end{array},$$

where B_1 represents the matrix of relations between the endogenous variables in dataset 1 (identified by specific values of $F_1 \cdots F_k$), and so on. The off-diagonal entries will be 0,

because there are no variables from other datasets related to those variables in dataset 1, for example. Then, similarly,

$$\Gamma = \begin{array}{c} \\ Y_1 \\ \vdots \\ Y_{2^k} \end{array} \overbrace{\begin{bmatrix} \Gamma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma_{2^k} \end{bmatrix}}^{X_1 \quad \cdots \quad X_{2^k}}.$$

Though the block matrices appear to be distinct matrices for each dataset, they are not. The matrix of coefficients for a particular dataset is generated by setting specific values for each factor, so some coefficients from the general model will be present in more than one dataset, though each dataset has a distinct combination of those coefficients. Thus, from our two-level modeling approach, the datasets do share certain common parameters and thus we must analyze them through the joint likelihood function derived in the following sections.

The Φ and Ψ matrices of this unified two-level model are constructed similarly. Some of the factors under consideration indicate that we are considering independent groups of subjects. However, we will also need to consider measurements taken from the same group of subjects, like the same group measured under two or more conditions. Obviously, measurements taken of the same variables from the same subjects, regardless of treatment, will be correlated, and must be modeled as repeated measures appropriately.

A traditional representation is to show correlation between repeated measures by correlating the errors in the regression equations [29, 47-51]. It is common to implement a general linear model with correlated errors for longitudinal data or repeated measures data [53]. We will apply this traditional treatment of repeated measures data in regression to the regression-style equations that compose an SEM; we will represent errors in the equations as correlated in corresponding regression equations. We will not represent correlations between different equations for different treatment conditions since we consider errors in the equations independent from one another in the single-group case. Additionally, we will allow for free correlation of the exogenous variables, X, for measurements taken on the same group of subjects, as it is traditional to model X with free variation in the single-group case. These changes are implemented in the Φ and Ψ matrices of a SEM.

As usual, the Φ matrix represents the covariance matrix of the exogenous variables, X. For a single-dataset SEM, this matrix is modeled freely by the researcher. We will construct the Φ matrix as follows.

$$\Phi = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_{2^k} \end{array} \overbrace{\begin{bmatrix} \Phi_1 & \Phi_{1,2} & \cdots & \Phi_{1,2^k} \\ \Phi_{1,2}' & \Phi_2 & & \Phi_{2,2^k} \\ \vdots & & \ddots & \\ \Phi_{1,2^k}' & \Phi_{2,2^k}' & & \Phi_{2^k} \end{bmatrix}}^{X_1 \quad X_2 \quad \cdots \quad X_{2^k}},$$

where Φ_i represents the Φ matrix for the i^{th} dataset. $\Phi_{i,j}$ represents the covariance matrix of the i^{th} and j^{th} datasets. If datasets *i* and *j* contain observations from independents groups of subjects, then $\Phi_{i,j} = 0$. However, if the variables from the i^{th} and j^{th} datasets are measured for the same group of subjects under different conditions or treatments, then the parameters of the $\Phi_{i,j}$ matrix should be modeled freely.

The Ψ matrix is constructed similarly to Φ, where we must consider the independence of the variables before assigning values to zero or creating parameters to be estimated. The Ψ matrix is traditionally the covariance matrix of the errors in the equations of the SEM. We will construct the Ψ matrix as follows.

$$\Psi = \begin{array}{c} \\ \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_{2^k} \end{array} \overbrace{\begin{array}{cccc} \zeta_1 & \zeta_2 & \cdots & \zeta_{2^k} \\ \end{array}} \begin{bmatrix} \Psi_1 & \Psi_{1,2} & \cdots & \Psi_{1,2^k} \\ \Psi_{1,2}' & \Psi_2 & & \Psi_{2,2^k} \\ \vdots & & \ddots & \\ \Psi_{1,2^k}' & \Psi_{2,2^k}' & & \Psi_{2^k} \end{bmatrix},$$

where Ψ_i represents the Ψ matrix for the i^{th} dataset (which traditionally is a diagonal matrix, as the errors in the equations are considered to be independent in many cases). $\Psi_{i,j}$ represents the covariance matrix of the i^{th} and j^{th} datasets, a diagonal matrix relating the errors in corresponding equations of the two datasets. If datasets *i* and *j* contain observations from independent groups of subjects, then $\Psi_{i,j} = 0$. Otherwise, the diagonal entries of $\Psi_{i,j}$ should be estimated as parameters in the model, and the off-diagonal entries should be set to zero before estimation.

3.5 Implementation of MD-SEM

Because this model is new, there is no software available that can estimate parameters for our MD-SEM. Therefore, we have implemented our estimation procedure in MATLAB.

3.5.1 Estimating the Parameters

For the brain functional pathway study, we are interested in Case 3, so we have implemented only this case, but all are easy simplifications of Case 3. A clear benefit of the models discussed is the increase in sample size. When we combine the datasets for estimation of the Case 3 model, we are estimating a model with 82 observations of each variable, rather than 16 or 25 in the single-group case. Generally, sample size should be large when estimating an SEM for best results, best taking advantage of the asymptotic properties associated with estimation of standard errors. In a single-group estimation, our data defies this requirement, but with for the Case 3 model, we come much closer to the "large" sample size we need.

To implement the model, we first create a vector containing all parameters to be estimated. We also center each data variable about its mean and list the variables as

discussed in two matrices—one representing group 0 and the other representing group 1 (normal subjects and cocaine abusers, respectively). In a function file, we define $\mathrm{B}, \Gamma, \Phi, \Psi$ from the vector of parameters, each set to an appropriate starting value [1] and construct $\Sigma(\theta)$ for each group. Once we have the model-implied covariance matrices, $\Sigma(\theta)$, and the centered data, we can calculate the value of the manipulated log likelihood function derived above, called the fitting function, F_{ML}.

In a MATLAB script file, the vector of parameters is originated and starting values are estimated and set. There are some values that are new to SEM, like the path parameters of additive change, for which there are no estimation methods to use, so we set those values to 0 to begin. The fitting function, contained in its own function file and passed only the current value of the vector of parameters, is minimized by the MATLAB function *fminunc*, part of MATLAB's optimization toolbox. This function outputs the vector of parameter values that minimizes the function specified in the function call. The goodness-of-fit of the model is tested at this point, with a significance level of .05.

3.5.2 Estimating the Standard Errors

Once the parameters are estimated, the standard errors can be estimated. As we have seen in previous sections discussing approximate standard errors, we must calculate the matrix of second partial derivatives of the fitting function with respect to the calculated parameters. However, with so many parameters and such a complex function, we must do this numerically.

The standard errors are found inside the asymptotic covariance matrix, as the square roots of the diagonal entries. The expected Fisher Information matrix, can be calculated using the following formula as a matrix of second derivatives called the Hessian matrix of the function.

$$\begin{aligned}\tilde{H}_{ij} = (\{[f(x+h_i e_i + h_j e_j) - f(x+h_i e_i)] - [f(x+h_j e_j) - f(x)]\} \\ + \{[f(x-h_i e_i - h_j e_j) - f(x-h_i e_i)] - [f(x-h_j e_j) - f(x)]\})/(2h_i h_j)\end{aligned} \quad [54]$$

where $f(x)$ is F_{ML}, e_i is the i[th] elementary vector, and h_i is the step size in the i^{th} direction. The step size in the direction of parameter i is proportional to parameter i; it is a small number h (we use the fourth root of Matlab's smallest recognized number, *eps*, as suggested in Gill's text on numerical optimization [55]) multiplied by the i^{th} parameter estimate to scale it properly. (These details are specified in [54] as standard procedure for a numerical optimization and error calculation.)

The implementation was tested by estimating single-group models for all 8 data sets involved in the PET study, and the results were the same as the results of SAS, LISREL, and EQS.

3.6 Results for the Brain Functional Pathway Study

In the case of the Brain Functional Pathway Study conducted by Dr. Nora Volkow, we have two groups, each studied under two conditions. We consider treatments *placebo* and *methylphenidate* for both groups of subjects (cocaine abusers and normal

subjects). All subjects are expecting to receive placebo. Clearly, this is a situation like Case 4, the mixed design, described in the preceding section. We implemented the model estimation as described above.

First, look at the following illustrative portrayal in Figure 3.5 of the model to be estimated. As shown in Figure 3.4, we want to estimate the SEM we have considered for each group and treatment individually. However, we also have group effects and treatment effects pointing to each individual path, representing the change in path strength due to the group (G) and treatment (D) factors in this experiment.

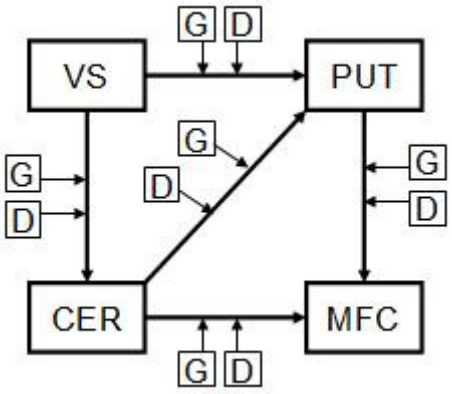


Figure 3.5. Hypothesized path diagram for MD-SEM. Each path is potentially influenced by group membership (drug abuse history) and the drug each subject receives. Each path and factor is estimated and tested for significance.

Using the original MATLAB script discussed in the preceding section, we obtained the following results for the estimation of the hypothesized model shown in Figure 3.5. The results are tabulated in Table 3.1 and illustrated in Figure 3.6. In order to reference the results in the table, recall the equations relating to the hypothesized path diagram shown in Figure 3.5 are

$$\begin{aligned} PUT &= (\beta_1 + \beta_2 G + \beta_3 D)VS + (\beta_7 + \beta_8 G + \beta_9 D)CER + \epsilon_1 \\ MFC &= (\beta_{10} + \beta_{11} G + \beta_{12} D)VS + (\beta_{13} + \beta_{14} G + \beta_{15} D)CER + \epsilon_2 \\ CER &= (\beta_4 + \beta_5 G + \beta_6 D)VS + \epsilon_3 \end{aligned}.$$

Highlighted paths are significant at the $\alpha = .05$ significance level (one-sided). Red paths and factors indicate a positive relationship between the corresponding brain regions, while blue paths and factors indicate a negative relationship. Those factors (G or D) that are not present are insignificant and have been removed to simplify the diagram.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
path	.8100	-.3525	-.0519	.5875	-.2777	.1039	.1156	.4845
error	.1001	.1125	.057	.1006	.105	.0423	.1421	.1834
p-value	.0000	.0017	.3630	.0000	.0082	.0140	.4162	.0082

	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
path	.0130	.3703	.1123	.2009	.7699	-.2858	-.1401
error	.1138	.1594	.2052	.1719	.2097	.3086	.2622
p-value	.9093	.0202	.5840	.2426	.0002	.3544	.5930

Table 3.1. Estimated path coefficients, standard errors, and corresponding p-values for the hypothesized path diagram shown in Figure 3.5 are displayed. P-values are two-sided.

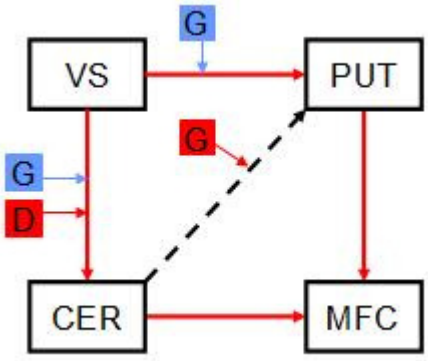


Figure 3.6. Results of the Brain Functional Pathway Study as estimated by our algorithm implemented in MATLAB. Paths determined to be insignificant are shown as dotted arrows. Arrows and factors highlighted are significant after a two-sided z-test with $\alpha = .05$. Red paths and factors indicate a positive influence while blue factors and paths indicate a negative influence.

Figure 3.6 clearly shows that for some paths, the group membership and drug treatment of subjects does affect the strength of connectivity between the brain regions under study.

Chapter 4: Power and Sample Size Analysis of MD-SEM

In Chapter 3, we derived a goodness-of-fit test for the null hypothesis $H_0 : \Sigma = \Sigma(\theta)$. This test evaluates whether the sample data covariance matrix (a representation of the population under study) and the model-implied covariance matrix are significantly different. If the null hypothesis is rejected, there is a poor fit, and the relationships in the model do not represent those in the population. The goodness-of-fit statistic derived in Chapter 3 for the mixed design model follows a chi-square distribution as the sample size (N = number of observations in each dataset) approaches infinity. The degrees of freedom are equal to the number of distinct elements in the covariance matrices of all independent groups under study minus the number of free parameters in the model.

Before we can use our new model and the corresponding goodness-of-fit statistic, we must first evaluate the robustness of this test against small sample sizes, since we have a total of 82 observations from 41 subjects. We must know how large the sample size must be in order to have reliable chi-square test statistics (does that test statistic follow a chi-square distribution if our sample is small? if we cannot answer this question, then we cannot evaluate power).

We have developed and implemented a simulation to determine the robustness of the mixed design goodness-of-fit test against small sample sizes. Our simulation is motivated by the original simulation performed by Boomsma [56] to answer the same questions of robustness about the original, single-group SEM framework. From Boomsma's work comes the recommendation of 100-200 observations as a minimal sample size requirement for single-group SEMs. Our simulation will be conducted in MATLAB. We evaluate the robustness of the chi-square test based on the brain activity model we have been analyzing (including factors of G and D on each path). Once we know which sample sizes are appropriate for generating significance tests of parameters, then we can evaluate the power of those tests.

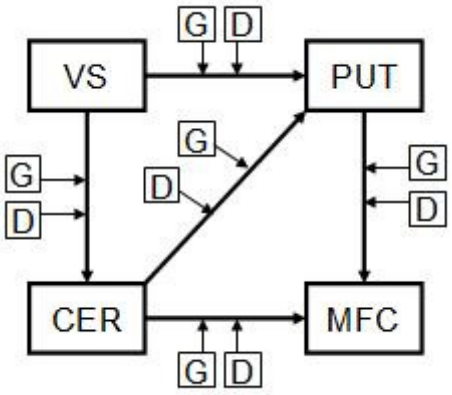


Figure 4.1. Hypothesized relationship among variables for study in the Robustness Simulation. We will specify parameter values for all path coefficients, factor coefficients, and variance parameters and study a very specific model, as in the original work by Boomsma [56].

4.1 Robustness of SEM Hypothesis Tests Against Small Sample Size

In order to determine whether our model and tests are robust for varying sample sizes, we follow the methods of Boomsma [56], who first performed this analysis on the single-group SEM methodology in 1983. The overall goodness-of-fit test derived in Chapter 3 can be manipulated to test the significance of individual parameters in the model as well. The test statistic follows a chi-square distribution when the null hypothesis is true as the sample size becomes large ("how large is large enough?").

To perform the robustness analysis, we specify a model for which the null hypothesis is true and then determine whether the resulting test statistic follows a chi-square distribution for varying sample sizes. In repeated trials, it is expected that the distribution of this statistic will converge to a chi-square distribution as we increase the sample size, so we will vary the sample size (N = 25, 50, 100, 200, 400, 800) and inspect the results.

4.1.1 Forcing the Null Hypothesis to be True

To begin, as is prescribed by Boomsma [56], we completely specify a model by setting explicit values for all parameters (paths, factors, and variances). For our simulation, we use the model shown in Figure 4.1, and we specify the parameter values to equal those estimated from our original dataset. As Boomsma points out, it is difficult to generalize the results because we must specify a particular model completely. This study analyzes only the example of our model of interest.

After specifying parameters, we construct the model-implied covariance matrices, $\Sigma_0(\theta)$ and $\Sigma_1(\theta)$, as discussed in Chapter 3. We will use these model-implied covariance matrices as population covariance matrices, thereby forcing the null hypothesis $\Sigma = \Sigma(\theta)$ to be true.

4.1.2 Generating Sample Data

Given the covariance matrix $\Sigma(\theta)$, which we will use as the population covariance matrix, we will generate multivariate normal data with sample covariance matrix approximately equal to $\Sigma(\theta)$. The generated data serve as input samples for use in repetitions of the algorithm.

To generate sample data, we apply the Cholesky decomposition to $\Sigma(\theta)$ to find matrix M such that $\Sigma(\theta) = MM'$. Next, generate a vector of observations of length N times the number of columns in $\Sigma(\theta)$ from the standard normal distribution. Form the vector of observations into a N-by-#variables matrix, Z. As N increases, the covariance of this matrix is the identity matrix. Therefore,

$$\text{cov}(MZ) = M\ \text{cov}(Z)M' = MIM' = MM' = \Sigma(\theta);$$

letting $X = MZ$ yields the multivariate normal sample data we require, X.

4.1.3 Model Estimation

For each set of sample data, we calculate the sample covariance matrix and run the MD-SEM algorithm to estimate model parameters. This should return estimates similar to those specified in the original model definition, but will probably show some variation due to the minimization technique for the fitting function. The difference between the input sample data and the model-implied covariance matrix is reflected in the chi-square test statistic by definition. We record the value of this statistic for each repetition of the algorithm at each sample size.

4.2 Results of the Robustness Simulation

For each repetition of the algorithm discussed in section 4.1, at each sample size, the chi-square test statistic is recorded (which is equal to the minimum value of the fitting function). The moments of this sample distribution are calculated and compared to those of the chi-square distribution the test statistic is expected to follow under the null hypothesis. For each sample size, we evaluate the mean and median of the sample distribution for each sample size. Most importantly, we determine the percentage of trials in which the null hypothesis is rejected at the traditional 5% significance level. If the test statistic truly follows a chi-square distribution under the null hypothesis, then about 5% of the time, our test statistic should result in a rejected null hypothesis, just by chance. If the proportion of rejections strays from 5%, then the test statistic does not appear to follow a chi-square distribution for the corresponding sample size. We performed this simulation for a varying number of repetitions (reps = 100, 300, 500), as in Boomsma's original simulation. The results are displayed in Table 4.1.

100 reps	χ^2_{33}	N=25	N=50	N=100	N=200	N=400	N=800
mean	33	67.00	54.25	56.46	91.87	71.74	33.09
median	32.34	48.29	37.03	34.36	34.08	31.41	31.36
% reject H_0	.05	.55	.25	.17	.14	.08	.07

300 reps	χ^2_{33}	N=25	N=50	N=100	N=200	N=400	N=800
mean	33	63.08	73.08	73.94	76.25	84.80	55.87
median	32.34	47.03	40.34	36.14	34.06	32.56	31.96
% reject H_0	.05	.49	.32	.20	.12	.09	.06

500 reps	χ^2_{33}	N=25	N=50	N=100	N=200	N=400	N=800
mean	33	65.65	64.50	64.49	66.82	61.76	46.62
median	32.34	47.46	38.99	35.11	33.53	32.39	32.72
% reject H_0	.05	.50	.31	.18	.11	.08	.06

Table 4.1. Results of the robustness analysis of the chi-square goodness-of-fit test for MD-SEM. Each table represents a differing number of repetitions of the algorithm. The sample size varies in each table. Mean and median for the distribution of fit statistic values are calculated, and the percentage of tests for which the null hypothesis is rejected is recorded. The second column of Table 4.1 shows the theoretical values of the chi-square distribution with thirty-three degrees of freedom (our model has 72 distinct covariance matrix entries and 39 parameters, leaving 33 degrees of freedom).

The percentage of test statistics for which the null hypothesis is rejected is somewhat inflated, even for large sample sizes. This could be because there are some non-convergent and ill-conditioned solutions in each set of repetitions of the algorithm. Those ill-conditioned solutions often produce extremely large function values, which of course will cause the null hypothesis to be rejected, but they also skew the average to be much higher than the median. Therefore, further study could be done on the issue of robustness in MD-SEM, focusing on removing ill-conditioned solutions to the models. Such solutions are difficult to identify as they are being generated, and are therefore difficult to remove from consideration.

Because the test statistic does not appear to converge to a chi-square distribution for sample sizes less than 400, we propose that MD-SEM as developed in Chapter 3 is suitable for sample sizes of 400 and larger. This means that for each group of subjects being compared, and for each condition evaluated, there must be 400 or more observations in order to have reliable significance tests for strength of parameters and fit of the model. However, we do acknowledge the need for further investigation of this issue for a more accurate prescription of required sample size.

4.3 Power Analysis of SEM and MD-SEM

Now that we have determined whether the chi-square test statistic is robust for small sample sizes, we can evaluate the power of statistical tests of significance for

individual parameters in the model. First, we must review the theoretical calculation of power for SEM and discuss the corresponding calculations for MD-SEM.

4.3.1 Power of Single-Group SEM

For single-group MLE SEM, we are interested in the power of the statistical test of a single parameter in the system of equations, $H_0 : \theta_i = 0$. The likelihood ratio test statistic follows an approximate central chi-square distribution under H_0 [1]. In 1985, Satorra and Saris [57] determined that when H_0 is tested, but H_1 is true, the test statistic follows a non-central chi-square distribution, and proposed a method for estimating the value of the non-centrality parameter. The degrees of freedom for both distributions is 1 because we are interested in testing the significance of single path, however, in general, it is the difference in the number of free parameters between H_0 and H_1. The estimation of the non-centrality parameter can be based on any of three methods: the Wald test statistic, the Likelihood Ratio test statistic, or the Lagrangian Multiplier test statistic. The estimates are asymptotically equal; however, Saris, Satorra, and Sörbom [58] determined that the LR method is most accurate for power calculations in small sample sizes [1]. We will use this method in our analysis. Satorra and Saris [57] developed the likelihood ratio method, and also simplified the original framework to eliminate the need for using the asymptotic covariance matrix, which can be difficult to calculate in many instances. Their method, which is the current standard, is discussed below.

If we want to test a particular parameter, we may use $H_0 : \theta_i = 0, H_1 : \theta_i \neq 0$. Then in H_0 we have assigned a particular value of θ_i. Therefore, we have one less parameter in the model. Let χ_0^2 be the fit statistic for the model under H_0 and χ_1^2 be the fit statistic for the model under H_1 (all parameters free). Then $D^2 = \chi_0^2 - \chi_1^2$ follows a chi-square distribution with degrees of freedom equal to the difference of the degrees of freedom for the two hypotheses. This test determines whether the fit changes if this particular parameter is set to 0, and therefore tests whether that value should be assigned to 0 or some value not equal to 0.

This method is appropriate for the maximum likelihood estimation method, but should only be used when the sample size is large and H_0 is true [27]. In the case of testing a single hypothesis, if we run the model under H_1, we can use the t-statistic (parameter value / standard error) as a test statistic for a two-sided z-test, or we can run the model under H_0 and use the modification index for θ_i as a chi-square test statistic with one degree of freedom. For large samples, the modification index, t^2, and D^2 are equal. However, this may not be true for small samples [27].

In order to evaluate the power of the D^2 test (which is equivalent to the single-parameter significance z-test in large samples), the following procedure can be used [27].

1. Specify the model completely under H_1—give all parameters a numerical assigned value. (These are the values estimated when the model is free—we must say that our initial study is our "pilot study" and we are calculating the power of future studies at varying sample sizes because retroactive power analysis is not suggested.)

2. Compute the covariance matrix $\Sigma(\theta)$ for the parameters specified in step 1.
3. Evaluate the structural equation model of H_0 (parameters set to values from step 1 except for θ_i which will be set to 0), using the $\Sigma(\theta)$ constructed in step 2 as the input covariance matrix. The chi-square fit statistic for this model is the non-centrality parameter λ.
4. Use the non-central chi-square distribution with non-centrality parameter λ and degrees of freedom equal to the difference in degrees of freedom between H_0 and H_1. Determine the probability of obtaining a value of the non-central chi-square distribution with lambda and df=1 that is larger than the $1-\alpha$ percentile of the central chi-square distribution.

Generally in practice, we calculate the probability of a value of the non-central distribution that is smaller than the $1-\alpha$ percentile in the central chi-square distribution using the CDF of the non-central distribution. One minus this probability is the power of the test [59].

4.3.2 Power of MD-SEM

The power of a MD-SEM test of significance works in much the same way as the calculation of power for a single-group SEM test because we have derived a chi-square goodness-of-fit test for the mixed design that is entirely analogous to that of the single-group methodology.

This time, we are interested in the test of whether a path and all covariates on that path are significant (i.e. we want to understand the model's power to detect an eliminated relationship). As in the single-group case, our likelihood ratio test statistic follows a central chi-square distribution when the null hypothesis $H_0 : \theta = 0$ is true and a non-central chi-square distribution when it is false. The degrees of freedom for this chi-square test are equal to the difference in degrees of freedom between the numerator and denominator of the likelihood ratio. In our example model, we will consider the case of a single parameter (path or factor) and also a path and two covariates on the path simultaneously tested for significance. In the latter case, we will set the path parameter and the two covariate influences equal to zero, making the degrees of freedom for this likelihood ratio test equal to 3. In the case of a single parameter significance test, we have one degree of freedom, and only that parameter is set equal to 0 in the null hypothesis. As in the previous case, we will use the likelihood ratio method for determining the non-centrality parameter of the distribution of the test statistic under the alternative hypothesis.

As before, assume the completely specified alternative model, and calculate the model-implied covariance matrix. We then evaluate the model in the null hypothesis using the model-implied covariance calculated from the alternative model as our input covariate matrix. The chi-square fit statistic for this second model is the non-centrality parameter, λ, of the non-central chi-square distribution we use to determine power. The power is equal to

$$P(\chi_3^2(\lambda) > \chi_{3,1-\alpha}^2(0)), \text{ [27].}$$

It is important to note that the goodness-of-fit test statistic for the mixed design is equal to the minimized value of the fit function. In the single-group case, it must be multiplied by the sample size, but for the mixed design it is not.

4.3.3 Power Simulation: MD-SEM

We will use the brain functional pathway study to evaluate the power of MD-SEM via simulation. As in the previous section on the robustness study, we are evaluating power for a single model; it is therefore limited in its applicability, and could benefit from further investigation. We use the Monte Carlo simulation technique of Muthén and Muthén [60]. Our model of interest for the power analysis simulation is the same model discussed in previous chapters, and is shown in Figure 4.1. Because the goal of estimating power is to determine the probability of rejecting the null hypothesis when it is false, we must keep in mind that parameters with small effect sizes (those that will likely be deemed insignificant in statistical tests) will have small power for any sample size, since the corresponding null hypothesis, $H_0 : \beta_i = 0$, is likely true. Highly significant parameters will be detected with increasing power as sample size grows. In all cases, we will empirically evaluate the probability of a correct significance test (correct decision to reject the null hypothesis $H_0 : \beta_i = 0$) for varying sample sizes.

Because significance tests for single parameters are based on the chi-square goodness-of-fit tests used in SEM, we begin by assuming the null hypothesis $\Sigma = \Sigma(\theta)$ is true, as we need to perform the power analysis using a model that fits the data well. To begin the power analysis, we must completely specify a model with good fit (all paths, factors, and variance parameters), [27]. We specify the parameter values to equal those estimated from our original dataset from the brain functional pathway study. After specifying parameters, we construct the model-implied covariance matrices, $\Sigma_0(\theta)$ and $\Sigma_1(\theta)$. Given the covariance matrices $\Sigma_0(\theta)$ and $\Sigma_1(\theta)$, which we will use as the population covariance matrices, we will generate multivariate normal data with sample covariance matrices approximately equal to $\Sigma_0(\theta)$ and $\Sigma_1(\theta)$, as in the robustness simulation (see Section 4.1.2 for details). The generated data serve as input samples for use in repetitions of the algorithm. For each set of sample data, we calculate the sample covariance matrix and run the MD-SEM algorithm to estimate model parameters. This should return estimates similar to those specified in the original model definition, but will probably show some variation due to the variation in simulated samples.

We evaluate the power of two types of hypothesis tests—one for significance of a single path and one for significance of a single factor. The chi-square goodness-of-fit test for the overall model can be manipulated to evaluate tests of significance of single parameters in the model [27], and we are most interested in evaluating the power of these tests (as these are the tests that determine which effects are significant).

We analyze 200 simulated samples and determine what percentage of the time the model has good fit and the paths and factors we are testing are deemed significant by the standard t-test of individual parameters (equivalent to a chi-square goodness-of-fit difference test, as introduced in Section 4.3.1). We will look at the empirical distribution of acceptance and rejection of the null hypothesis $H_0 : \beta_i = 0$ and determine how many times out of 200 that each path and factor was detected as significant at the $\alpha = .01$ and

$\alpha = .05$ significance levels. The results of the simulation are recorded in Table 4.2. Note that we are truly interested in those paths and parameters that are significant in the original model, as the power is the probability of rightly rejecting the null hypothesis. Recall that the model is:

$$PUT = (\beta_1 + \beta_2 G + \beta_3 D)VS + (\beta_7 + \beta_8 G + \beta_9 D)CER + \epsilon_1$$
$$MFC = (\beta_{10} + \beta_{11} G + \beta_{12} D)VS + (\beta_{13} + \beta_{14} G + \beta_{15} D)CER + \epsilon_2$$
$$CER = (\beta_4 + \beta_5 G + \beta_6 D)VS + \epsilon_3$$

$\alpha = .01$	Param.	Estimate	p	N=25	N=50	N=100	N=200	N=400	N=800
Path	β_1	.813	.000	.535	.705	.815	.875	.915	.925
Factor	β_2	-.355	.002	.455	.705	.815	.875	.915	.925
Factor	β_3	-.053	.363	.070	.100	.290	.535	.870	.925
Path	β_4	.587	.000	.535	.705	.815	.875	.915	.925
Factor	β_5	-.275	.008	.450	.700	.815	.875	.915	.925
Factor	β_6	.103	.014	.270	.585	.815	.875	.915	.925
Path	β_7	.109	.416	.040	.080	.245	.575	.880	.925
Factor	β_8	.489	.008	.390	.690	.815	.875	.915	.925
Factor	β_9	.016	.909	.020	.005	.030	.025	.020	.045
Path	β_{10}	.365	.020	.335	.670	.810	.875	.915	.925
Factor	β_{11}	.118	.584	.005	.020	.065	.185	.400	.700
Factor	β_{12}	.206	.243	.065	.210	.395	.750	.915	.925
Path	β_{13}	.777	.000	.520	.705	.815	.875	.915	.925
Factor	β_{14}	-.295	.354	.055	.075	.235	.590	.860	.925
Factor	β_{15}	-.145	.593	.045	.055	.100	.255	.505	.800

$\alpha = .05$	Param.	Estimate	p	N=25	N=50	N=100	N=200	N=400	N=800
Path	β_1	.812	.000	.525	.650	.785	.925	.935	.945
Factor	β_2	-.353	.002	.500	.650	.785	.925	.935	.945
Factor	β_3	-.053	.363	.090	.195	.380	.785	.935	.945
Path	β_4	.588	.000	.525	.650	.785	.925	.935	.945
Factor	β_5	-.277	.008	.480	.645	.785	.925	.935	.945
Factor	β_6	.103	.014	.375	.615	.785	.925	.935	.945
Path	β_7	.114	.416	.140	.180	.485	.815	.910	.945
Factor	β_8	.483	.008	.420	.645	.785	.925	.935	.945
Factor	β_9	.016	.909	.030	.035	.030	.070	.080	.090
Path	β_{10}	.365	.020	.415	.645	.785	.925	.935	.945
Factor	β_{11}	.113	.584	.045	.085	.200	.345	.650	.885
Factor	β_{12}	.209	.243	.175	.245	.600	.905	.935	.945
Path	β_{13}	.777	.000	.510	.650	.785	.925	.935	.945
Factor	β_{14}	-.287	.354	.105	.175	.400	.760	.915	.945
Factor	β_{15}	-.151	.593	.100	.100	.235	.505	.735	.920

Table 4.2. Results of the Power Analysis Simulation for $\alpha = .01$ and $\alpha = .05$. Each power calculation is performed 200 times and the resulting average effect size and

probability of rejecting $H_0 : \beta_i = 0$ for each test is displayed in the table. The p-value refers to the p-value of this parameter in the original model (estimated in Chapter 3, with results shown in Table 3.1 and Figure 3.6).

Table 4.2 shows that the tests of $H_0 : \beta_i = 0$ are quite powerful for significant paths and factors for large sample sizes, and not powerful for small sample sizes. Significant paths and factors are identified by their p-values from the estimation of the original brain functional pathway data in Chapter 3 (on which this simulation is based). Those p-values are highlighted in Table 4.2 in red (results of the original analysis are shown in Table 3.1 and Figure 3.6). Setting the significance level $\alpha = .01$ decreases the power slightly from calculations with $\alpha = .05$, as expected, since decreasing the significance level increases the probability of Type II error, which decreases power. Also as expected, parameters with larger effects are easier to detect, and therefore have higher power for small sample sizes. For the small sample sizes N=25 and N=50, the power of detecting all effects is unacceptably small, supporting the conclusions from the robustness analysis, which indicated that small sample sizes are not appropriate for use with MD-SEM.

Chapter 5: Latent Variable SEM and the Mixed Design

One of the variables in the dataset generated during the Brain Functional Pathway Study (see Chapter 2), the Ventral Striatum (VS), is difficult to measure because it represents a very small brain region. We have two measurements for the VS, one containing the average of a larger number of voxels for each subject, and the other containing the average of a smaller subset of those voxels. The ventral striatum is a very tiny region and it is difficult to locate on the Talairach atlas of the brain for each subject, as this tiny region could be in a slightly different location for each subject. We want to incorporate the VS into the model, of course, because it is an influential element in the path diagram.

The path analysis method introduced in Chapter 2 and modified in Chapter 3 is a subset of the more general structural equation model. Path analysis is the subset of models that involve strictly measured variables. SEM was developed to estimate relations between latent constructs, and between the latent variables and their measured indicators. A latent variable is a variable that cannot be measured directly, but has indicators that provide information about the impact of the unmeasured variable. For example, we may want to measure the influence of intelligence on social class, but neither of these variables can be observed directly. Therefore, we need indicators of each, like number of years in school and IQ test scores for intelligence, and income for social class. Then SEM would allow us to model the relationship between the latent variables, assuming their measured indicators provide enough information about the variables of interest. This example is illustrated below.

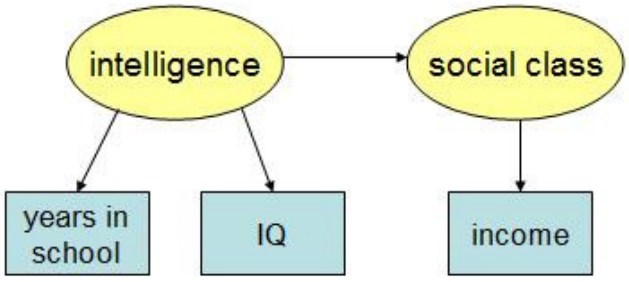


Figure 5.1. Illustration of SEM with latent variables. Latent variables—those which cannot be measured directly—are shown in yellow ovals. Observed (measured) variables are shown in blue rectangles. *School years* and *IQ* are measured indicators of the latent variable *intelligence*, and *income* is a measured indicator of the latent variable *social class*. The focus of this SEM would be on studying the influence of *intelligence* on *social class*. Note: Latent variables are commonly depicted as ovals and measured variables are depicted as rectangles.

Because we want to include VS in our model as a latent variable, it is a natural next step to extend SEM for latent variables to include the mixed design methodology, as we have done for SEM without latent variables.

5.1 Existing Model and Estimation

We are familiar with the traditional path analysis model of SEM, which relates only exogenous measured variables to endogenous measured variables, $y = \mathrm{B}y + \Gamma x + \zeta$. The general SEM, which incorporates latent variables, is shown below [1].

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$$
$$y = \Lambda_y \eta + \epsilon$$
$$x = \Lambda_x \xi + \delta$$

where η is a vector of endogenous (dependent) latent variables, ξ is a vector of exogenous (independent) latent variables, and x and y are vectors of independent and dependent measured variables, respectively. This model allows for relationships (paths) to be modeled between endogenous latent variables and their indicators, y, between exogenous latent variables and their indicators, x, and between exogenous and endogenous latent variables. The first equation represents any arrows that may go from endogenous latent variables to other endogenous latent variables and from exogenous latent variables to endogenous latent variables. The second equation represents any arrows going from the endogenous latent variables to their measured indicators, y. The third equation represents any arrows going from the exogenous latent variables to their measured indicators, x. All equations can be viewed as regression equations in matrix form.

Using this more general model, we can estimate four types of relationships. One path of each type is marked in Figure 5.2.

1. Exogenous latent variables to their measured indicators
2. Exogenous latent variables to endogenous latent variables
3. Endogenous latent variables to their measured indicators
4. Endogenous latent variables to endogenous latent variables

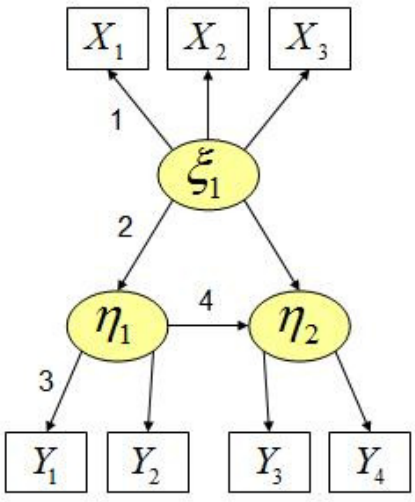


Figure 5.2. Hypothesized path diagram for SEM model with three latent variables—one exogenous and two endogenous, shown in yellow—and seven observed variables, serving as indicators of the latent variables. Path 1 is an example of a path from an exogenous latent variable to its measured indicator; path 2 is an example of a path from an exogenous latent variable to an endogenous latent variable; path 3 is an example of a path from an endogenous latent variable to its measured indicator; path 4 is an example of a path from an endogenous latent variable to an endogenous latent variable.

There are several assumptions for the model described above. We generally assume ζ is the disturbance vector with expected value 0 and it is uncorrelated with ξ. Additionally, ϵ and δ represent the errors in the measurement equations, they have expected value 0, and are uncorrelated with ξ, ζ, and each other. As in the path analysis model, we are interested in studying the null hypothesis $\Sigma = \Sigma(\theta)$, or similarly asking the question, "Is the model-implied covariance matrix sufficiently close to the population covariance matrix of the data?" We must begin by defining some important notation. Let Φ be the covariance matrix of ξ, let Ψ be the covariance matrix of ζ, let Θ_ϵ be the covariance matrix of ϵ, and let Θ_δ be the covariance matrix of δ. Also, ζ, δ, ϵ are pairwise independent and independent of x and y.

We can set values in B, Γ, and the Λ matrices according to the hypothesized path diagram. The other matrices, representing covariance matrices of the variables, can be specified, to represent correlation among the errors in the equations and/or the independent latent variables. As we determine which parameters will be estimated, we must keep in mind the issue of identification. Necessary, but not sufficient, is the rule that we can estimate up to $\frac{1}{2}k(k+1)-1$ parameters, because we have $\frac{1}{2}k(k+1)$ unique entries in the sample covariance matrix of the measured variables (k is the total number of measured variables), and we must have at least one degree of freedom left with which to make inferences. All recursive models are identified.

As in the case of path analysis, in order to derive the model-implied covariance matrix, we write will write our measured variables as

$$z = \begin{pmatrix} y \\ x \end{pmatrix}$$

where y and x have been centered about their respective means. The model-implied covariance matrix is shown and derived below.

$$\Sigma_z(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

$$\begin{aligned}
\Sigma_{yy}(\theta) &= E(yy') = E\left((\Lambda_y\eta + \epsilon)(\Lambda_y\eta + \epsilon)'\right) \\
&= E\left((\Lambda_y\eta + \epsilon)(\eta'\Lambda_y' + \epsilon')\right) \\
&= E\left(\Lambda_y\eta\eta'\Lambda_y' + \Lambda_y\eta\epsilon' + \epsilon\eta'\Lambda_y' + \epsilon\epsilon'\right) \\
&= \Lambda_y E\left(\eta\eta'\right)\Lambda_y' + \Lambda_y E\left(\eta\right)E\left(\epsilon'\right) + E\left(\epsilon\right)E\left(\eta'\right)\Lambda_y' + E\left(\epsilon\epsilon'\right) \\
&= \Lambda_y E\left(\eta\eta'\right)\Lambda_y' + E\left(\epsilon\epsilon'\right)
\end{aligned}$$

From the model,

$$\begin{aligned}
\eta &= \mathrm{B}\eta + \Gamma\xi + \zeta \\
\eta(I - \mathrm{B}) &= \Gamma\xi + \zeta \\
\eta &= (I - \mathrm{B})^{-1}(\Gamma\xi + \zeta) \\
y &= \Lambda_y(I - \mathrm{B})^{-1}(\Gamma\xi + \zeta) + \epsilon
\end{aligned}$$

We can derive the model-implied covariance matrix as follows.

$$\begin{aligned}
\Sigma_{yy}(\theta) &= \Lambda_y E\left(\eta\eta'\right)\Lambda_y' + E\left(\epsilon\epsilon'\right) \\
&= \Lambda_y E\left((I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)(\xi'\Gamma' + \zeta')(I - \mathrm{B})^{-1\prime}\right)\Lambda_y' + E(\epsilon\epsilon') \\
&= \Lambda_y(I - \mathrm{B})^{-1}\left(E(\Gamma\xi\xi'\Gamma') + E(\zeta)E(\xi')\Gamma' + \Gamma E(\xi)E(\zeta') + E(\zeta\zeta')\right)(I - \mathrm{B})^{-1\prime}\Lambda_y' + \Theta_\epsilon \\
&= \Lambda_y(I - \mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - \mathrm{B})^{-1\prime}\Lambda_y' + \Theta_\epsilon
\end{aligned}$$

$$\begin{aligned}
\Sigma_{yx}(\theta) &= E(yx') = E\left((\Lambda_y\eta + \epsilon)(\Lambda_x\xi + \delta)'\right) \\
&= E\left((\Lambda_y\eta + \epsilon)(\xi'\Lambda_x' + \delta')\right) \\
&= \Lambda_y E\left(\eta\xi'\right)\Lambda_x' + \Lambda_y E\left(\eta\right)E\left(\delta'\right) + E\left(\epsilon\right)E\left(\xi'\right)\Lambda_x' + E(\epsilon)E(\delta') \\
&= \Lambda_y E\left(\eta\xi'\right)\Lambda_x' = \Lambda_y E\left((I - \mathrm{B})^{-1}(\Gamma\xi + \zeta)\xi'\right)\Lambda_x' \\
&= \Lambda_y(I - \mathrm{B})^{-1}\left(\Gamma E(\xi\xi') + E(\zeta)E(\xi')\right)\Lambda_x' \\
&= \Lambda_y(I - \mathrm{B})^{-1}\Gamma\Phi\Lambda_x'
\end{aligned}$$

$$\begin{aligned}\Sigma_{xx}(\theta) &= E(xx') = E\left((\Lambda_x\xi+\delta)(\Lambda_x\xi+\delta)'\right)\\ &= E\left((\Lambda_x\xi+\delta)(\xi'\Lambda_x'+\delta')\right)\\ &= \Lambda_x E\left(\xi\xi'\right)\Lambda_x' + E(\delta)E(\xi')\Lambda_x' + \Lambda_x E(\xi)E(\delta') + E\left(\delta\delta'\right)\\ &= \Lambda_x\Phi\Lambda_x' + \Theta_\delta\end{aligned}.$$

Therefore,

$$\Sigma_z(\theta) = \begin{bmatrix}\Sigma_{yy}(\theta) & \Sigma_{yx}(\theta)\\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta)\end{bmatrix} = \begin{bmatrix}\Lambda_y(I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1\prime}\Lambda_y' + \Theta_\epsilon & \Lambda_y(I-\mathrm{B})^{-1}\Gamma\Phi\Lambda_x'\\ \Lambda_x\Phi\Gamma'(I-\mathrm{B})^{-1\prime}\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta\end{bmatrix}$$

Now that we have derived the model-implied covariance matrix, we need a method for evaluating how close this matrix is to the population covariance matrix—for testing $H_0 : \Sigma = \Sigma(\theta)$. We will choose parameter estimates by the maximum likelihood method, assuming our measured variables are distributed multivariate normal. Since we are able to incorporate only the measured variables in our distribution, the log-likelihood function and fitting function, F_{ML} are the same as in the case of traditional path analysis, as we still only have observed values of the measured variables,

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\right| - (p+q).$$

The standard errors can also be estimated as in the case of traditional path analysis with only measured variables (see Chapter 2), by calculating the asymptotic covariance matrix.

5.2 Our Interests in Latent Variable SEM

Until now, we have been working with traditional path analysis models—SEM with no latent variables. However, we have one variable in particular, the Ventral Striatum (VS), that we believe is not accurately measured because it is such a small region of the brain, representing only a small number of pixels in the image of the Talairach atlas of the brain. We would like to represent this variable as a latent variable (or *hidden* variable [61]) with a measured indicator, and of course some measurement error, as shown below.

In order to best estimate a model with latent variables, each latent variable should have at least two measured indicators. One reason for this is to increase the degrees of freedom when estimating such a complex model, but another reason is because one of the indicators must always have a path coefficient of 1 (to set a scale for the latent variable). Luckily, for the Brain Functional Pathway Study, there are two measurements of ventral striatum; one measurement incorporates a larger number of pixels in the brain image than the other measurement, as it is currently unclear which pixels represent the ventral striatum because this region is very small.

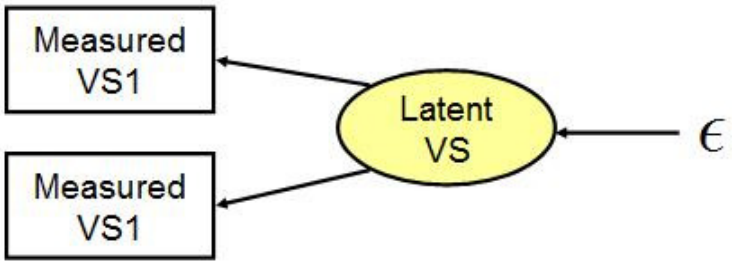


Figure 5.3. Representation of VS as a latent variable (shown in the yellow oval). We include the two measurements of the VS variable as indicators of the latent VS, and we include error in measurement.

We want to first represent the Ventral Striatum as a latent variable, and then extend MD-SEM to include latent variables, particularly for situations like ours, where some variables may not be measured accurately. Figure 5.4(b) shows the path diagram we would like to estimate, with VS as a latent variable with two measured indicators (one of them is the VS variable in our path analysis with no latent variables). For comparison, Figure 5.4(a) shows the original path diagram for this study, with no latent variables.

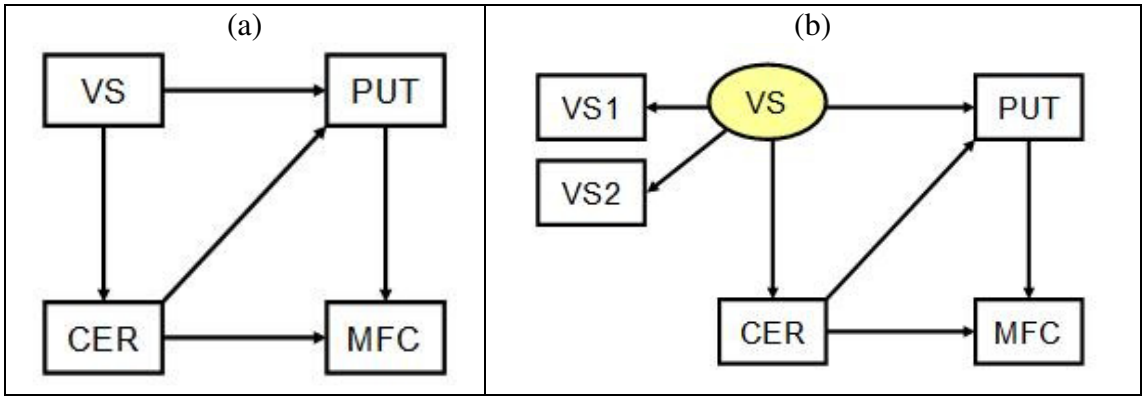


Figure 5.4. (a) Hypothesized path diagram for the Brain Functional Pathway Study with no latent variables; (b) Hypothesized path diagram for the Brain Functional Pathway Study with VS included as a latent variable. The two measurements of VS (rectangular VS1 and VS2) are included as indicators of the latent VS (yellow oval VS). Latent VS interacts with the other brain regions as the measured VS did in our previous consideration of this model with no latent variables.

There are some fundamental issues with our desire to incorporate latent VS into our current path diagram. As mentioned above and illustrated in Figure 5.2, we can only model four types of paths in the current SEM model: exogenous latent variables to their measured indicators, endogenous latent variables to their measured indicators, and relationships between latent variables of all types. However, as depicted in Figure 5.4, when VS is represented as a latent variable, we will want relationships present in the model between measured variables, and from measured variables to latent variables. While there are methods for incorporating paths from exogenous measured variables to latent variables [1, 62, 63] no method exists for modeling relationships between measured

variables in the general SEM with latent variables framework. We want to incorporate all necessary paths into our model and extend the model to mixed designs in order to estimate the latent VS model.

5.3 Modifications to the Existing Single-Group Methodology

Over the years, many extensions have been made to the original Latent Variable SEM model to allow for more general applications, where measured variables can influence one another and can also influence the latent variables [1]. Figure 5.4 shows our model of interest, with ventral striatum represented as a latent variable (in the oval), with a measured indicator (our measured VS variable, in the rectangle). Here, we have only one latent variable with many influences.

5.3.1 MIMIC Models

MIMIC models are general latent variable SEMs with the following structure [1]:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$$
$$y = \Lambda_y \eta + \epsilon$$
$$x = \xi$$

If there are no exogenous latent variables in the model, then the influence of exogenous measured variables on the endogenous latent variable(s) can be modeled. Let x be a vector of measured variables who causes lie outside of the model (exogenous measured variables). Then the model reduces to the following equations:

$$\eta = \mathrm{B}\eta + \Gamma x + \zeta$$
$$y = \Lambda_y \eta + \epsilon$$

All parts of the model remain the same except for Φ, which now represents the covariance structure of the exogenous measured variables x, rather than the exogenous latent variables which are no longer present (this is now similar to the case of no latent variables). The model-implied covariance matrix $\Sigma(\theta)$ can be derived as a simpler case of the traditional latent variable SEM model.

The current MIMIC model accounts for modeling the latent variables with measured influences in the model (which is not allowed in the standard SEM framework). We want to extend the MIMIC model to allow for relationships between measured endogenous variables (y), and relationships from these measured endogenous variables to endogenous latent variables. Therefore, we must further draw on further generalizations of latent variable SEM.

5.3.2 More Extensions

In 1979, Graff proposed an alternative representation for latent variable SEM, [1]. (Others, too, have proposed alternative representations—McDonald [64, 65], McArdle and McDonald [66], and Bentler and Weeks [67].) Graff proposed the following model:

$$\begin{aligned} \eta^+ &= \mathrm{B}^+\eta^+ + \zeta^+ \\ y^+ &= \Lambda_y{}^+\eta^+ \end{aligned}, \text{ where}$$

$$\eta^+ = \begin{pmatrix} y \\ x \\ \eta \\ \xi \end{pmatrix}, \zeta^+ = \begin{pmatrix} \epsilon \\ \delta \\ \zeta \\ \xi \end{pmatrix}, y^+ = \begin{pmatrix} y \\ x \end{pmatrix}, \mathrm{B}^+ = \begin{bmatrix} 0 & 0 & \Lambda_y & 0 \\ 0 & 0 & 0 & \Lambda_x \\ 0 & 0 & \mathrm{B} & \Gamma \\ 0 & 0 & 0 & 0 \end{bmatrix}, \Psi^+ = \begin{bmatrix} \Theta_\epsilon & & & \\ 0 & \Theta_\delta & & \\ 0 & 0 & \Psi & \\ 0 & 0 & 0 & \Phi \end{bmatrix}.$$

Note that the representation of this model is shown in relation to general Latent Variable SEM.

Now that we have a more general representation, Bollen's text [1] notes that we can change the zero entries in the B^+ matrix to allow for more paths to be present—paths between any of the variables in η^+. Bollen warns, however, that having more paths available comes more opportunities for identification problems, so this license for including paths comes with a responsibility to keep the number of parameters reasonable.

5.3.3 Our Modification to the Existing Latent Variable Structural Equation Model

Following Graff's precedent that models can allow for all types of relationships between elements of y, x, η, ξ, we have developed the following model that suits the needs of our example. Our model does not utilize Graff's notation, only his notion of including any necessary paths.

First, we must redefine how we think of our measured variables x and y:

$$\xi = \left(VS_{latent}\right), x = \begin{pmatrix} VS1 \\ VS2 \end{pmatrix}, y = \begin{pmatrix} PUT \\ MFC \\ CER \end{pmatrix}$$

As shown in Figure 5.4, our model must incorporate paths from $\xi \to x, \xi \to y, y \to y$. To the current SEM framework for latent variables, we will add matrices to represent the remaining two types of paths we must incorporate, as shown here:

$$\begin{aligned} y &= \Lambda_y\eta + \mathrm{M}y + \mathrm{K}\xi + \epsilon \\ x &= \Lambda_x\xi + \delta \\ \eta &= \mathrm{B}\eta + \Gamma\xi + \zeta \end{aligned}$$

Note that this model can be generalized further, as noted by Graff; however, the model shown above satisfies our needs for the current model. Our model is an extension using Graff's model written in the traditional SEM format. Note that if the two added matrices M and K are set equal to 0, we have the standard SEM framework.

For our model of interest, shown in Figure 5.4, $\Lambda_y = 0, \mathrm{B} = 0, \Gamma = 0$ because we have no endogenous latent variables. Therefore, our extended model shown above reduces to

$$y = \mathrm{M}y + \mathrm{K}\xi + \epsilon$$
$$x = \Lambda_x \xi + \delta$$

However, we will derive the method for the general case where $\mathrm{B} \neq 0$ for simplicity in generalizing to future models.

Before we can derive the model-implied covariance structure, we provide some typical some definitions:

1. $z = \begin{pmatrix} y \\ x \end{pmatrix}$, where all variables in y and x have been centered about their means.
2. $E(\xi\xi') = \Phi$
3. $E(\zeta\zeta') = \Psi$
4. $E(\epsilon\epsilon') = \Theta_\epsilon$
5. $E(\delta\delta') = \Theta_\delta$
6. After some manipulation, $y = (I - \mathrm{M})^{-1}\Lambda_y(I - \mathrm{B})^{-1}(\Gamma\xi + \zeta) + (I - \mathrm{M})^{-1}(\mathrm{K}\xi + \epsilon)$.
7. Consider all SEM assumptions: $\xi, \zeta, \delta, \epsilon$ are pairwise independent.

We can derive the implied covariance matrix of the parameters as Bollen [1] does.

$$\Sigma_z(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

$$\begin{aligned} \Sigma_{yy}(\theta) = E(yy') = (I - \mathrm{M})^{-1}\Lambda_y(I - \mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - \mathrm{B})^{-1\prime}\Lambda_y'(I - \mathrm{M})^{-1\prime} \\ + (I - \mathrm{M})^{-1}\Lambda_y(I - \mathrm{B})^{-1}\Gamma\Phi\mathrm{K}'(I - \mathrm{M})^{-1\prime} \\ + (I - \mathrm{M})^{-1}\mathrm{K}\Phi\Gamma'(I - \mathrm{B})^{-1\prime}\Lambda_y'(I - \mathrm{M})^{-1\prime} \\ + (I - \mathrm{M})^{-1}(\Theta_\epsilon + \mathrm{K}\Phi\mathrm{K}')(I - \mathrm{M})^{-1\prime} \end{aligned}$$

$$\Sigma_{yx}(\theta) = E(yx') = (I - \mathrm{M})^{-1}\Lambda_y(I - \mathrm{B})^{-1}\Gamma\Phi\Lambda_x' + (I - \mathrm{M})^{-1}\mathrm{K}\Phi\Lambda_x'$$

$$\Sigma_{xy}(\theta) = E(xy') = \Lambda_x\Phi\mathrm{K}'(I - \mathrm{M})^{-1\prime} + \Lambda_x\Phi\Gamma'(I - \mathrm{B})^{-1\prime}\Lambda_y'(I - \mathrm{M})^{-1\prime}$$

$$\Sigma_{xx}(\theta) = \Theta_\delta + \Lambda_x\Phi\Lambda_x'$$

Specifically, for our model (Figure 5.4) where $\Lambda_y = 0, \mathrm{B} = 0, \Gamma = 0, \Psi = 0$ because we have no endogenous latent variables,

$$\Sigma_z(\theta) = \begin{bmatrix} (I-\mathrm{M})^{-1}(\Theta_\epsilon + \mathrm{K}\Phi\mathrm{K}')(I-\mathrm{M})^{-1}' & (I-\mathrm{M})^{-1}\mathrm{K}\Phi\Lambda_x' \\ \Lambda_x\Phi\mathrm{K}'(I-\mathrm{M})^{-1}' & \Theta_\delta + \Lambda_x\Phi\Lambda_x' \end{bmatrix}$$

The likelihood function will stay the same because we are still referencing the nine measured variables that we assume are MVN,

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\right| - (p+q).$$

As always, we would minimize this function to solve for all of our path coefficients and variance parameters in θ. Standard errors and goodness-of-fit tests proceed in the standard way.

5.3.4 A Single-Group Example Using the Modified Framework

The hypothesized model shown in Figure 5.4 is implemented and estimated in MATLAB. As mentioned earlier, one of the measured indicators of a latent variable must have relationship 1 to the latent variable (the path coefficient is set equal to 1 and not estimated) in order to set a scale for the latent variable. The results of this analysis are illustrated in Figure 5.5 and estimated values for path parameters, standard errors, and p-values are displayed in Table 5.1. All p-values and significance tests are two-sided. Tests are performed with significance level .05. The hypothesized model associated with Figure 5.4 is

$$\begin{aligned} put &= \beta_1 vs_{latent} + \beta_3 cer + \epsilon_1 \\ mfc &= \beta_4 put + \beta_5 cer + \epsilon_2 \\ cer &= \beta_2 vs_{latent} + \epsilon_3 \\ vs_1 &= 1 vs_{latent} + \delta_1 \\ vs_2 &= \beta_6 vs_{latent} + \delta_2 \end{aligned}$$

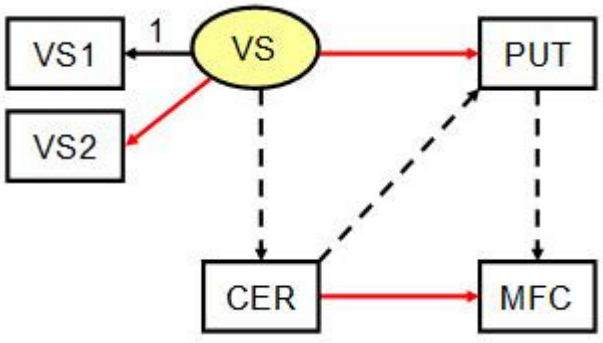


Figure 5.5. Results of the Brain Functional Pathway Study for a single dataset (normal subjects under the influence of placebo) with VS included as a latent variable. Dotted paths represent insignificant relationships; red paths indicate significant positive relationships. The path denoted with a 1 indicates that VS1 is used to set the scale for the latent variable VS (ventral striatum). All tests are performed at a two-sided significance level of .05.

	β_1	β_2	β_3	β_4	β_5	β_6
path coeff	.7133	.2784	.1858	.2336	.8250	.9348
std. error	.1274	.1777	.1746	.1601	.2155	.0819
p-value	.0000	.1171	.2873	.1446	.0001	.0000

Table 5.1. Results of the Brain Functional Pathway Study for a single dataset (normal subjects under the influence of placebo) with VS included as a latent variable. P-values are two-sided.

In the next section, we will derive Latent Variable MD-SEM, as we did for SEM without latent variables in Chapter 3. We will focus strictly on the modified model derived in Section 5.3.3 (of which traditional single-group SEM is a special case where $M = 0, K = 0$), deriving the mixed design framework for this model only. (It should be noted that any model in Graff's framework may be extended to the mixed design.)

5.4 Latent Variable SEM for Mixed Designs

MD-SEM was developed under the assumption that no latent variables are included; however, traditionally, a structural equation model allows for inclusion of latent variables. In a general structural equation model, researchers can study the relationships between many latent variables, each having measured indicators. A latent variable is variable that cannot be measured, but has measurable indicators. For example, intelligence may be an important variable to include in a SEM studied by a behavioral scientist. Intelligence cannot be measured directly, but the result of an IQ test may be used as a measured indicator of intelligence. Latent variables are also a good way to represent variables that contain measurement error. Each poorly measured variable can be represented as a latent variable with measured indicators.

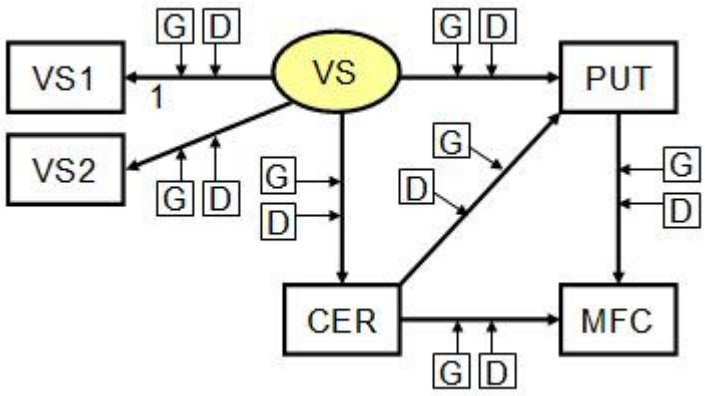


Figure 5.6. Hypothesized path diagram for the Mixed Design Brain Functional Pathway Study with VS included as a latent variable. The two measurements of VS (rectangular VS1 and VS2) are included as indicators of the latent VS (oval VS). Latent VS interacts with the other brain regions as the measured VS did in our previous consideration of this diagram with only measured variables. Group membership (G) and drug treatment (D) effects are evaluated as influences on path strength.

As discussed in previous sections, before our development of MD-SEM, there was a serious deficiency in the ability to apply SEM to studies with multiple groups and/or repeated measures. It is natural for us to extend our path analysis methods to structural equation models capable of incorporating latent variables.

5.4.1 Extending Our Modification to the Mixed Design

Now that we have developed a model appropriate for representing our model of interest using Latent Variable SEM, we can derive Latent Variable MD-SEM. We want to estimate models based on multiple datasets. These datasets could be independent, or they could be dependent in some way (i.e. observations taken from the same group of subjects under different conditions). The purpose of estimating such models is to determine whether variables like group membership and treatment condition affect the strength of paths between variables.

As discussed above, our model is:

$$y = \mathrm{M}y + \mathrm{K}\xi + \epsilon$$
$$x = \Lambda_x \xi + \delta$$

The parameter matrices include: Λ_x (relates $\xi \to x$), M (relates $y \to y$), K (relates $\xi \to y$), $\Phi = E(\xi, \xi')$, $\Theta_\epsilon = E(\epsilon\epsilon')$, and $\Theta_\delta = E(\delta\delta')$.

For each path, we will reparametrize the path coefficient to incorporate the mixed design's covariate influences, as shown in Figure 5.6. For example, consider the path from PUT to MFC (PUT → MFC). The equation for this path alone, before the reparametrization would be:

$$MFC = \beta PUT + \epsilon$$

We reparametrize the path coefficient to add contributions to the path from the effect of group membership (G=0,1—representing membership in a group of normal subjects or cocaine abusers, respectively) and the condition a subject is measured under, such as the drug effect (D=0,1—representing those not receiving the drug or receiving the drug, respectively). Then:

$$MFC = (\beta_0 + \beta_1 G + \beta_2 D)PUT + \epsilon = \beta_0 PUT + \beta_1 G * PUT + \beta_2 D * PUT + \epsilon .$$

Then for normal subjects who did not receive the drug, the path is:

$$MFC = \beta_0 PUT + \epsilon$$

And for those subjects who are cocaine abusers and did not receive the drug, the path is:

$$MFC = (\beta_0 + \beta_1)PUT + \epsilon$$

Notice that the paths above are still written in the same format as paths from the single-group SEM for latent variables. The variables are still in a multivariate normal distribution; only the parameters have changed. We reparametrize all paths in the diagram.

As in the case of SEM without latent variables, we will expand the parameter matrices to incorporate more datasets. Beginning with the M matrix, each matrix becomes a block matrix, where each block represents its respective matrix in a single-group model. So M_{00} represents the M matrix from the dataset 00 (group=0, drug=0: normal subject under condition of not receiving the drug). The off-diagonal entries in this matrix are zero, as we do not need paths between variables in different datasets. Because of the reparametrization, some parameters are shared, if the datasets share similar characteristics (i.e. same values of G or D). Therefore, the entries in these matrices are not unique; elements of M_{00} appear inside all of the other blocks, while entries of M_{01} appear in M_{11}, as do entries of M_{10}. The M matrix for the mixed design model is shown below, and the K, Λ_x matrices are constructed similarly.

$$\mathrm{M} = \begin{bmatrix} \mathrm{M}_{00} & 0 & 0 & 0 \\ 0 & \mathrm{M}_{01} & 0 & 0 \\ 0 & 0 & \mathrm{M}_{10} & 0 \\ 0 & 0 & 0 & \mathrm{M}_{11} \end{bmatrix}$$

The $\Phi, \Theta_\epsilon, \Theta_\delta$ matrices are slightly more complicated, it is in these matrices that we correlate the repeated measures of the model (those datasets containing measurements taken from the same subjects). In the path parameter matrices ($\mathrm{M}, \mathrm{K}, \Lambda_x$), off-diagonal entries are zero, but in the variance parameter matrices ($\Phi, \Theta_\epsilon, \Theta_\delta$), they may not be.

Shown below are the variance parameter matrices. As is mentioned above, group 0 is a group of normal subjects and group 1 is a group of cocaine abusers. Therefore, datasets 00 and 01 are correlated (both have group value of 0) and so are datasets 10 and 11 (both have group value of 1). So in the variance parameter matrices, we introduce off-diagonal block matrices that represent correlations between these datasets, as shown below.

$$\Phi = \begin{bmatrix} \Phi_{00} & \Phi_0 & 0 & 0 \\ \Phi_0{}' & \Phi_{01} & 0 & 0 \\ 0 & 0 & \Phi_{10} & \Phi_1 \\ 0 & 0 & \Phi_1{}' & \Phi_{11} \end{bmatrix}, \Theta_\epsilon = \begin{bmatrix} \Theta_{\epsilon 00} & \Theta_{\epsilon 0} & 0 & 0 \\ \Theta_{\epsilon 0}{}' & \Theta_{\epsilon 01} & 0 & 0 \\ 0 & 0 & \Theta_{\epsilon 10} & \Theta_{\epsilon 1} \\ 0 & 0 & \Theta_{\epsilon 1}{}' & \Theta_{\epsilon 11} \end{bmatrix}, \Theta_\delta = \begin{bmatrix} \Theta_{\delta 00} & \Theta_{\delta 0} & 0 & 0 \\ \Theta_{\delta 0}{}' & \Theta_{\delta 01} & 0 & 0 \\ 0 & 0 & \Theta_{\delta 10} & \Theta_{\delta 1} \\ 0 & 0 & \Theta_{\delta 1}{}' & \Theta_{\delta 11} \end{bmatrix}$$

By convention, the Φ matrices will be modeled as full matrices (all entries are parameters), and the $\Theta_\epsilon, \Theta_\delta$ matrices (representing the errors in the equations) will be modeled as diagonal matrices (only diagonal entries are parameters, all others set to zero), as in the case of MD-SEM without latent variables.

Then for our example, we will have 5 paths (since there is only one indicator of the latent variable, its path coefficient must be 1), with each having three parameters (path and two factors, G and D), totaling 15 path and factor parameters (within the $\Lambda_x, \mathrm{M}, \mathrm{K}$ matrices). We also have 1-by-1 Φ matrices because we have only one exogenous latent variable. Those on the diagonal are symmetric, while those off-diagonal are not, but are duplicated by transpose. So we have 6 Φ variance parameters. We have 3-by-3 Θ_ϵ matrices, and all are diagonal. We have four on the diagonal and two unique off-diagonal matrices. Therefore, we have 18 Θ_ϵ variance parameters. We have 1-by-1 Θ_δ matrices, so we have 6 Θ_δ parameters. Therefore, we have a total of 48 parameters to be estimated. (Note: We have 2 independent groups, each with 8 measured variables. So we have $2\left(\frac{1}{2}\right)(8)(9) = 72$ unique degrees of freedom to work with in our model. Therefore, this model passes at least one test of identification.)

5.4.2 Maximum Likelihood Estimation and Standard Error Calculation

The likelihood function for Latent Variable MD-SEM is the same as that for the MD-SEM without latent variables, because the likelihood depends only on measured variables.

The interest of SEM is to determine how well the model fits the data, or test the hypothesis $\Sigma = \Sigma(\theta)$. Therefore, we must construct $\Sigma(\theta)$ in order to estimate the path coefficients. For simplicity, we will construct one likelihood function for each independent group considered in the mixed design. These likelihood functions are constructed under the assumption that the variables for each independent group are distributed multivariate normal, as is required for the ML estimation method in most SEM software. In a mixed design, we may be considering many conditions or treatments on one group, having 2^k observations of each variable measured on the same subjects. In the opposing extreme, we could be considering 2^k independent groups, each measured under the same single condition. We will construct one likelihood function for each independent group, considering all datasets for each group in that single likelihood function.

We must begin by constructing the model-implied covariance matrix, $\Sigma(\theta)$, for each group. Separate the datasets into G groups (one for each independent group of subjects), listing the variables in X and Y vectors and creating corresponding parameter matrices for each group. Construct the model-implied covariance matrix (in the traditional way) for each of the G groups, as described in Section 2.1. For example, for *Group 0*,

$$\Sigma_0(\theta) = \begin{bmatrix} (I - \mathrm{M}_0)^{-1}(\mathrm{K}_0\Phi_0\mathrm{K}_0' + \Theta_{\epsilon 0})(I - \mathrm{M}_0)^{-1}' & (I - \mathrm{M}_0)^{-1}\mathrm{K}_0\Phi_0\Lambda_{x0}' \\ \Lambda_{x0}\Phi_0\mathrm{K}_0'(I - \mathrm{M}_0)^{-1}' & \Theta_{\delta 0} + \Lambda_{x0}\Phi_0\Lambda_{x0}' \end{bmatrix}$$

(the simplified covariance matrix derived in the previous section, setting $\Lambda_y = 0, \mathrm{B} = 0, \Gamma = 0, \Psi = 0$ for all groups)and likewise for $\Sigma_1(\theta), \cdots, \Sigma_{G-1}(\theta)$.

Now construct the independent likelihood for each group of MVN variables, and multiply them together to form the likelihood estimation function for the MD-SEM, shown below.

$$L(\theta) = (2\pi)^{-\frac{N_0}{2}(p+q)} \left|\Sigma_0(\theta)\right|^{-\frac{N_0}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_0} z_{0i}\,' \Sigma_0^{-1}(\theta) z_{0i} \right) \cdots$$
$$(2\pi)^{-\frac{N_{G-1}}{2}(p+q)} \left|\Sigma_{G-1}(\theta)\right|^{-\frac{N_{G-1}}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N_{G-1}} z_{(G-1)i}\,' \Sigma_{G-1}^{-1}(\theta) z_{(G-1)i} \right)$$

As is traditional, we manipulate this likelihood function by taking its log, dropping constant terms, and removing negative coefficients to create a function we must minimize to find the parameter estimates. Additionally, in accordance with the single-group fitting function, we simplify the summing terms and add constant terms to force the fitting function to equal zero when $S = \Sigma(\theta)$. Our final fitting function for a two-group SEM is shown below.

$$F_{ML} = N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0 \Sigma_0^{-1}(\theta) \right] - N_0 \log\left|S_0\right| - N_0(p+q)$$
$$+ N_{G-1} \log\left|\Sigma_{G-1}(\theta)\right| + N_{G-1} tr\left[S_{G-1} \Sigma_{G-1}^{-1}(\theta) \right] - N_{G-1} \log\left|S_{G-1}\right| - N_{G-1}(p+q)\,,$$

where G is the number of independent groups considered in the mixed design, and $p+q$ is the number of variables measured for each group (summing the number of variables over all conditions).

In order to evaluate the statistical significance of the parameters, we must estimate a standard error for each path coefficient. As in traditional single-group SEM, standard errors of the parameter estimates can be calculated via the asymptotic covariance matrix (the inverse of Fisher Information Matrix).

$$ACOV\left(\hat{\theta}\right) = \left\{ -E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'} \right] \right\}^{-1} = 2\left\{ E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'} \right] \right\}^{-1}$$

5.4.3 Evaluation of fit

Finally, to evaluate the fit of this model, we must extend the current method for calculating a chi-square fit statistic. A chi-square goodness-of-fit test for this model can be derived in similar fashion to that for traditional SEM. It is straightforward to show that our fitting function,

$$-2\log\left(L_{H_0} / L_{H_1} \right) = N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0 \Sigma_0^{-1}(\theta) \right] + \cdots + N_{G-1} \log\left|\Sigma_{G-1}(\theta)\right|$$
$$+ N_{G-1} tr\left[S_{G-1} \Sigma_{G-1}^{-1}(\theta) \right] - N_0 \log\left|S_0\right| - N_0(p+q) - \cdots - N_{G-1} \log\left|S_{G-1}\right| - N_{G-1}(p+q)$$

follows a chi-square distribution with $\frac{1}{2}G(p+q)(p+q+1) - t$ degrees of freedom, where $p+q$ represents the total number of variables, X and Y, in each independent group, and G is the number of independent groups. We can test the hypothesis $\Sigma = \Sigma(\theta)$ using this fit function as the statistic. A significant result indicates that the model poorly fits the data.

5.4.4 Implementation of Latent Variable MD-SEM

Just as with MD-SEM without latent variables, there is no software available that can estimate parameters for our MD-SEM because it is a new method. We have implemented the estimation procedure in MATLAB. To implement the model, we first create a vector containing all parameters to be estimated. We also center each data variable about its mean. In a function file, we define Λ_x, M, K, Φ, Θ_ϵ, Θ_δ from the vector of parameters, each set to an appropriate starting value (Bollen 1989) and construct $\Sigma(\theta)$ for each independent group. Once we have the model-implied covariance matrices, $\Sigma(\theta)$, and the centered data, we can calculate the value of the log likelihood function derived above, called the fitting function, F_{ML}.

In a MATLAB script file, the vector of parameters is originated and starting values are estimated and set. There are some values that are new to SEM, like the path parameters of additive change, for which there are no estimation methods to use, so we initiate those values to 0. The fitting function, contained in its own function file and passed only the current value of the vector of parameters, is minimized by the MATLAB function *fminunc*, part of MATLAB's optimization toolbox. This function outputs the vector of parameter values that minimizes the function specified in the function call. Once the parameters are estimated, the standard errors can be estimated. The expected Fisher Information matrix can be calculated using the following formula as a matrix of second derivatives, the Hessian matrix of the function.

$$\tilde{H}_{ij} = (\{[f(x+h_ie_i+h_je_j)-f(x+h_ie_i)]-[f(x+h_je_j)-f(x)]\} + \{[f(x-h_ie_i-h_je_j)-f(x-h_ie_i)]-[f(x-h_je_j)-f(x)]\})/(2h_ih_j)\,,$$

where $f(x)$ is F_{ML}, e_i is the i^{th} elementary vector, and h_i is the step size in the i^{th} direction [54]. The step size in the direction of parameter i is proportional to parameter i; it is a small number h (we use the fourth root of MATLAB's smallest recognized number, *eps*, as suggested by Gill [55]) multiplied by the i^{th} parameter estimate to scale it properly.

5.4.5 An Example Using MD-SEM with Latent Variables

The hypothesized model shown in Figure 5.6 is implemented and estimated in MATLAB. As before, one of the measured indicators of a latent variable must have relationship 1 to the latent variable (the path coefficient is set equal to 1 and not estimated) in order to set a scale for the latent variable. Factors are included along the paths from latent VS to its measured indicators because the estimation of VS in relation to its indicators may change for each dataset. The results of this analysis are illustrated in Figure 5.7 and estimated values for path parameters, standard errors, and p-values are displayed in Table 5.2. All p-values and significance tests are two-sided. Tests are performed with significance level .05. The model associated with Figure 5.6 is

$$put = (\beta_1 + \beta_2 G + \beta_3 D) vs_{latent} + (\beta_7 + \beta_8 G + \beta_9 D) cer + \epsilon_1$$
$$mfc = (\beta_{10} + \beta_{11} G + \beta_{12} D)\, put + (\beta_{13} + \beta_{14} G + \beta_{15} D) cer + \epsilon_2$$
$$cer = (\beta_4 + \beta_5 G + \beta_6 D) vs_{latent} + \epsilon_3$$
$$vs_1 = (1 + \beta_{19} G + \beta_{20} D) vs_{latent} + \delta_1$$
$$vs_2 = (\beta_{16} + \beta_{17} G + \beta_{18} D) vs_{latent} + \delta_2$$

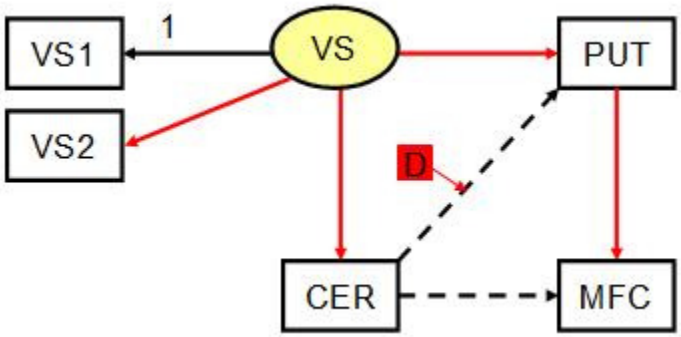


Figure 5.7. Results of the Brain Functional Pathway Study for the mixed design (normal subjects and cocaine abusers measured under the influence of both placebo and methylphenidate) with VS included as a latent variable. Dotted paths represent non-significant relationships; red paths indicate statistically significant positive relationships. Non-significant factors have been eliminated from the diagram for simplicity. The red drug factor along the path from CER to PUT indicates that this relationship becomes significant for those under the influence of methylphenidate. The path denoted with a 1 indicates that VS1 is used to set the scale for the latent variable VS (ventral striatum). All tests are performed at a two-sided significance level of .05.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
path	.8190	-.2051	-.0588	.3945	-.0578	.1388	.0855
std. error	.1324	.2351	.3004	.1282	.1739	.2036	.1439
p-value	.0000	.3829	.8450	.0021	.7395	.4955	.5525

	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}
path	-.0682	.1999	.6572	-.0286	.1957	.3977	-.1404
std. error	.1489	.0974	.1950	.2409	.1751	.2474	.3536
p-value	.6467	.0400	.0007	.9055	.2637	.1080	.6913

	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
path	-.1741	1.0124	-.1556	.0103	-.1431	.0507
std. error	.2656	.0718	.2261	.4006	.2294	.4107
p-value	.5122	.0000	.4912	.9796	.5326	.9018

Table 5.2. Results of the Brain Functional Pathway Study for the mixed design (normal subjects and cocaine abusers measured under the influence of both placebo and methylphenidate) with VS included as a latent variable. P-values are two-sided.

Chapter 6: Partial Least Squares SEM and the Mixed Design

In Chapter 3, we developed MD-SEM, based on the maximum likelihood methodology (MLE MD-SEM). However, in Chapter 4, we determined that MLE MD-SEM is not appropriate for analyzing data with small sample sizes. As a result, this novel method is not appropriate for analyzing the data from the brain functional pathway study. However, there does exist an alternative technique for analyzing SEMs called Partial Least Squares (PLS) SEM.

Rather than focusing on the covariance structure as in the covariance-based method for estimating SEMs discussed in the preceding chapters of this document, PLS SEM is an observation-based method for estimating such models, as in a traditional regression analysis [68]. PLS SEM is estimated to predict dependent variables, as in regression, rather than to explain covariances of measured variables as in covariance-based SEM [69]. There are two parts to all structural equation models: the structural model—which relates the latent variables in the model—and the measurement model—which relates the latent variables to their measured indicators. When there are only measured variables in the model, the measurement model is no longer necessary, and the structural model relates the measured variables [1]. One benefit of PLS SEM is its distribution-free qualities. "As a least squares method PLS requires no assumption regarding the independence of the observations and the distribution. Beyond the specification that the conditional expectation is linear, PLS is distribution free." [70]

6.1 Partial Least Squares SEM (PLS SEM)

PLS SEM is a technique for estimating the numerical measurements of the latent variables in the model from their related measured variables. The algorithm takes each "block" of measured variables (each block corresponds to one latent variable), and returns n "observations" of the latent variable, which are simply a weighted linear combination of the measured variables in that block. Part two of the algorithm takes those measured variables and estimates the relationships between them as prescribed by the model. This second stage is a noniterative analysis of each individual equation in the structural model via ordinary least squares regressions. [71] Wold's original PLS SEM creates estimates of the latent variables by composing a weighted sum of the related observed variables in the model, then it applies traditional regression to each individual equation in the system to estimate the path coefficients. PLS regression is used when there is strong multicollinearity among the composite estimates of the latent variables in the structural model [72]. Since we currently have no latent variables in our model, our use of PLS SEM involves only the individual regression analyses. OLS is typically used, but PLS regressions are recommended if there is multicollinearity among the X variables. This method works only on recursive models. If the SEM is nonrecursive, the fixed-point estimation procedure [73, 74] must be employed.

6.1.1 Latent Variables in PLS SEM

As we know from Chapter 5, latent variables (LVs) are unmeasured variables that are present in a model. For example, intelligence is a variable that cannot be measured, but we can use IQ score or GPA to give us some indication of a person's intelligence. Intelligence is a latent variable, and IQ score and GPA are values we may use as indicators (we measure these and include them in the model). The integral part of PLS SEM is estimating "observations" of the latent variables from their respective measured indicators. In MLE SEM, we are focused on the covariance structure of the variables in the model, and we used measured variables to understand the covariance structure among latent variables. PLS focuses on prediction, and creates estimated values for observations of each latent variable. This iterative procedure for estimating the latent variables is the first stage of the PLS SEM algorithm, and it is the heart of the method. Outlined below is the most common method for estimating the "observed" values of the latent variables, as suggested in [71].

Consider the following example, taken from [71]; the model is shown in the diagram below.

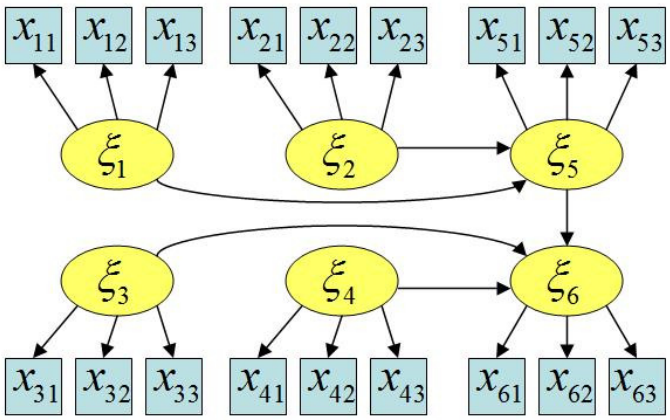


Figure 6.1. Hypothesized path diagram for a latent variable SEM (illustrating the technique of PLS SEM). Variables shaded in yellow are latent variables; variables shaded in blue are the corresponding measured indicators. PLS SEM uses a weighted combination of each latent variable's indicators to create "observations" of each latent variable, then OLS regression is used to estimate the relationships among the "measured" latent variables.

As in MLE SEM, we can construct the following system of equations relating the present latent variables (the structural model).

$$\xi_5 = \beta_{50} + \beta_{51}\xi_1 + \beta_{52}\xi_2 + e_5$$
$$\xi_6 = \beta_{60} + \beta_{63}\xi_3 + \beta_{64}\xi_4 + \beta_{65}\xi_5 + e_6$$

And the measurement model (shown for the first block only):

$$x_{11} = p_{011} + p_{111}\xi_1 + e_{11}$$
$$x_{12} = p_{012} + p_{112}\xi_1 + e_{12}$$
$$x_{13} = p_{013} + p_{113}\xi_1 + e_{13}$$

For each latent variable, ξ_j (representing both endogenous and exogenous latent variables), an estimate L_j is created as a weighted sum of its indicators:

$$L_j = \sum_h w_{jh} x_{jh} = f_j \sum_h v_{jh} x_{jh} ,$$

where w_{jh} are the weights of the indicator variables, h represents the total number of indicators of the latent variable estimated, and f is a scalar that gives the estimate L_j unit variance. To estimate the individual weights, v, a tool called the sign-weighted sum is created for each L, representing the other latent variables it is joined with in the model (by an arrow in either direction).

$$SwS_j = \sum_k \pm_{jk} L_k \text{ where } \pm_{jk} = sign(corr(L_j, L_k))$$

For this example,

$$SwS_1 = \pm_{15} L_5$$
$$SwS_2 = \pm_{25} L_5$$
$$SwS_3 = \pm_{36} L_6$$
$$SwS_4 = \pm_{46} L_6$$
$$SwS_5 = \pm_{51} L_1 \pm_{52} L_2 \pm_{56} L_6$$
$$SwS_6 = \pm_{63} L_3 \pm_{64} L_4 \pm_{65} L_5$$

For endogenous LVs (those with latent predictors present in the model, like variables 5 and 6 in this example), we create simple OLS regressions of each indicator on the corresponding sign-weighted sum. For the example, they are shown below.

$$x_{51} = v_{51}(\pm_{51} L_1 \pm_{52} L_2 \pm_{56} L_6) + d_{51}$$
$$x_{52} = v_{52}(\pm_{51} L_1 \pm_{52} L_2 \pm_{56} L_6) + d_{52}$$
$$x_{53} = v_{53}(\pm_{51} L_1 \pm_{52} L_2 \pm_{56} L_6) + d_{53}$$
$$x_{61} = v_{61}(\pm_{63} L_3 \pm_{64} L_4 \pm_{65} L_5) + d_{61}$$
$$x_{62} = v_{62}(\pm_{63} L_3 \pm_{64} L_4 \pm_{65} L_5) + d_{62}$$
$$x_{63} = v_{63}(\pm_{63} L_3 \pm_{64} L_4 \pm_{65} L_5) + d_{63}$$

For each exogenous LV (independent latent variables, as variables 1, 2, 3, and 4 in this example), we create multiple OLS regressions of the corresponding sign-weighted sum onto the indicators of the latent variable, as shown below.

$$\pm_{15} L_5 = v_{11}x_{11} + v_{12}x_{12} + v_{13}x_{13} + d_1$$
$$\pm_{25} L_5 = v_{21}x_{21} + v_{22}x_{22} + v_{23}x_{23} + d_2$$
$$\pm_{36} L_6 = v_{31}x_{31} + v_{32}x_{32} + v_{33}x_{33} + d_3$$
$$\pm_{46} L_6 = v_{41}x_{41} + v_{42}x_{42} + v_{43}x_{43} + d_4$$

Once these tools are constructed, the estimation process begins. First, starting values are assigned to the weights. These starting values are arbitrary (can be 0s and 1s). Next, the researcher must resolve the sign ambiguities remaining in the above equations. It should be clear what the correlation of two indicators should be (for example, IQ and GPA should be positively correlated).

Using the original equation,

$$L_j = f_j \sum_h v_{jh} x_{jh},$$

and the assigned starting values, estimate each L_j. Using these estimates of L_j, the weights can be estimated using the set of ten equations involving both L_j and the weights v, shown above for this example. Repeating these last two steps until each weight has converged to within a reasonable tolerance yields estimates of L_j.

6.1.2 The Regression Step

Once we have estimated each L_j, the estimated "observations" can be used in the straightforward second stage of regression analyses for each individual latent variable equation to estimate the path coefficients. Standard errors for the path coefficients must be found using bootstrapping, as PLS SEM is a nonparametric technique with no distributional requirements. If the variables do meet distributional requirements, traditional standard error calculation can be used (the latent observations must be normally distributed to use traditional methods). The third stage of the algorithm estimates the location parameters for the structural and measurement models outlined above (estimating β_{50} and p_{011}, for example). However, these values can be estimated directly in the second stage using a vector of 1 included in the list of independent variables for the regression step.

6.1.3 Benefits of PLS SEM

PLS SEM has far less rigid sample size requirements than covariance-based SEM (200 or more is considered appropriate). Sample sizes for PLS SEM without latent variables are simply those of the individual regression analyses performed for each equation in the system. The traditional regression rule of thumb for sample size is ten observations for each observed variable in the model. However, when sample size is small, statistical power is compromised; the more observations that are available, the better, in the case of OLS or PLS regression. [68] The derivation of the covariance-based ML SEM methodology requires the assumption that all observations are independent [1]. Unlike ML, PLS SEM does not require that observations be independent [68].

Unlike covariance-based SEM, PLS SEM can be used as exploratory in addition to confirmatory study. Also, in covariance-based SEM, it is often necessary to include three or more indicators for each latent variable in the model to ensure model identification. Another benefit of PLS SEM is that it is possible to incorporate *formative* indicators of the latent variables more easily (in covariance-based SEM, the existing framework would have to be modified). These are measured variables that affect the latent variable, rather than being affected by the latent variable. [68]

6.2 PLS SEM for a Single-Group Model

For a single group with no latent variables, we can skip the latent variable estimation step and proceed directly to the second (and third) stages where we estimate path parameters and their standard errors by applying OLS regressions to each individual equation in our system of interest. If the response variables are normally distributed, we can apply the standard OLS technique for estimating both path parameters and standard errors, which is outlined below. An example follows.

6.2.1 Path Coefficients and Standard Errors

In this section, the derivation of the traditional path coefficients, standard errors, and significance tests for OLS regression are shown. For a single-group model, treat each individual equation in the system as a single regression equation. Then the model is

$$y = X\beta + e$$

where E(e)=0 and Cov(e)=$\sigma^2 I$. y is *nx1*, X is *nx(r+1)*, β is *(r+1)x1*, and e is *nx1* [75]. To find the MLE of β and σ^2, the standard procedure is used. Assume the error in the equation is normally distributed, $e \sim N(0, \sigma^2 I)$. Then

$$f(e) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-e^2/2\sigma^2}$$

$$L = (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right)$$

Take the log likelihood and simplify:

$$l = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)$$

$$\Rightarrow \quad \min_{\beta, \sigma^2} \left(n \log \sigma + \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right)$$

In order to minimize with respect to β, minimize $(y - X\beta)'(y - X\beta)$; the maximum likelihood estimator of β is $\hat{\beta} = (X'X)^{-1} X'y$. We must also minimize with respect to σ^2 in order to determine the MLE of σ^2.

$$\frac{\partial}{\partial\sigma}\left[n\log\sigma+\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right]=\frac{n}{\sigma}-\frac{(y-X\beta)'(y-X\beta)}{\sigma^3}$$

Setting this equal to 0 and solving for σ^2, the MLE of σ^2 is

$$\hat{\sigma}^2=\frac{(y-X\beta)'(y-X\beta)}{n}.$$

Once we have the estimates $\hat{\beta}$, it is possible to calculate the predicted values of y generated by the model:

$$\hat{y}=X\hat{\beta}$$

We can also calculate the error in prediction:

$$\hat{e}=y-\hat{y}=y-X\hat{\beta}$$

In order to calculate the standard errors of our coefficients $\hat{\beta}$, we must identify the estimated covariance matrix of these estimates. We can show that:

$$\hat{\beta}=\beta+(X'X)^{-1}X'e$$
$$E(\hat{\beta})=\beta$$
$$Cov(\hat{\beta})=\sigma^2(X'X)^{-1}$$

Therefore, we must identify an estimator of σ^2. Generally, we use the unbiased estimator rather than the MLE, which is biased.

$$\hat{e}=(I-X(X'X)^{-1}X')e$$

Using this information, calculate the expected value of $\hat{e}'\hat{e}$:

$$E(\hat{e}'\hat{e})=\sigma^2(n-r-1)$$

Therefore, we can estimate σ^2 using $\frac{\hat{e}'\hat{e}}{n-r-1}$, and estimate $Cov(\hat{\beta})=\frac{\hat{e}'\hat{e}}{n-r-1}(X'X)^{-1}$.

Standard errors of the path coefficient estimates b are determined by taking the square root of the corresponding diagonal entries in the above covariance matrix of $\hat{\beta}$. In order to make inferences on the significance of the estimated coefficients $\hat{\beta}$, we must determine the sampling distribution of $\hat{\beta}$.

It is possible to write $\hat{\beta}$ as a linear combination of the error terms (shown above), *e*, for which we have set the distribution to be $e \sim N(0, \sigma^2 I)$. Therefore, using the linear combination for $\hat{\beta}$, we can say $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$. Additionally, it can be shown that $n\hat{\sigma}^2 \sim \chi^2_{n-r-1}$. We know

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-r-1},$$

and $\hat{e} = (I - X(X'X)^{-1}X')e$, so the residuals can be written as a linear combination of the error term in the model. Because we assume that the error term in the model is normally distributed with mean 0, $\hat{e}'\hat{e} \sim \chi^2_{n-r-1}$, or $n\hat{\sigma}^2 \sim \chi^2_{n-r-1}$.

Now that we have identified the sampling distribution of $\hat{\beta}$, we can perform significance tests on those parameters. Let $H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_r = 0$, or $H_0 : \beta_{(2)} = 0$ for some subset of coefficients, where $\beta_{(2)} = [\beta_{q+1}, ..., \beta_r]'$. Partition the matrices Z and β as follows:

$$X = \begin{bmatrix} X_{1_{nx(q+1)}} & X_{2_{nx(r-q)}} \end{bmatrix}, \beta = \begin{bmatrix} \beta_{(1)_{(q+1)x1}} \\ \beta_{(2)_{(r-q)x1}} \end{bmatrix}$$

Then

$$y = X\beta + e = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix} + e = X_1\beta_{(1)} + X_2\beta_{(2)} + e$$

and under H_0, $y = X_1\beta_{(1)} + e$. In general,

$$L(\beta, \sigma^2) = (2\pi)^{-n/2} \sigma^{-n} \exp(-(y - X\beta)'(y - X\beta) / 2\sigma^2),$$

which is maximized by

$$\hat{\beta} = (X'X)^{-1}Xy, \hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta}) / n,$$

as shown above. At its maximal value,

$$L(\beta, \sigma^2) = (2\pi)^{-n/2} \hat{\sigma}^{-n} e^{-n/2}.$$

Now, considering the null hypothesis, $H_0 : \beta_{(2)} = 0$, the likelihood is maximized by

$$\hat{\beta}_{(1)} = (X_1'X_1)^{-1}X_1'y, \hat{\sigma}_1^2 = (y - X_1\hat{\beta}_{(1)})'(y - X\hat{\beta}_{(1)}) / n,$$

resulting in

$$L(\beta,\sigma^2) = (2\pi)^{-n/2}\hat{\sigma}_1^{-n}e^{-n/2}.$$

Constructing the likelihood ratio of $\lambda = L_{H_0} / L_{H_1}$, we have

$$\lambda = \frac{(2\pi)^{-n/2}\hat{\sigma}_1^{\,2}e^{-n/2}}{(2\pi)^{-n/2}\hat{\sigma}^2 e^{-n/2}} = \left(\frac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}^2}\right)^{-n/2} = \left(1+\frac{\hat{\sigma}_1^{\,2}-\hat{\sigma}^2}{\hat{\sigma}^2}\right)^{-n/2}$$

Rejecting H_0 for small values of λ is the same as rejecting H_0 for large values of

$$\frac{\hat{\sigma}_1^{\,2}-\hat{\sigma}^2}{\hat{\sigma}^2}$$

or its adjusted version

$$\frac{n(\hat{\sigma}_1^{\,2}-\hat{\sigma}^2)/(r-q)}{n\hat{\sigma}^2/(n-r-1)}.$$

We know that $n\hat{\sigma}^2 \sim \chi^2_{n-r-1}$, and therefore, $n\hat{\sigma}_1^2 \sim \chi_{n-q-1}$ indicates that $n(\hat{\sigma}_1^{\,2}-\hat{\sigma}^2) \sim \chi_{r-q}$. Therefore,

$$\frac{n(\hat{\sigma}_1^{\,2}-\hat{\sigma}^2)/(r-q)}{n\hat{\sigma}^2/(n-r-1)} \sim F_{r-q,n-r-1},$$

and we can perform tests of $H_0 : \beta_{(2)} = 0$ based on this quantity and the F distribution. Because the $\hat{\beta} \sim N(\beta, \hat{\sigma}^2(X'X)^{-1})$, we can also perform significance tests of $H_0 : \beta_i = 0$ using the estimates and their standard errors to perform a t-test with *n-r-1* degrees of freedom.

6.2.2 Single-Group Example with No Latent Variables

As an example of PLS SEM, we use a single dataset from the Brain Functional Pathway Study (normal subjects measured under placebo condition). The corresponding model is

$$put = \beta_1 vs + \beta_3 cer + \epsilon_1$$
$$mfc = \beta_4 put + \beta_5 cer + \epsilon_2$$
$$cer = \beta_2 vs + \epsilon_3$$

and the hypothesized path diagram for the single-group case is shown in Figure 6.2.

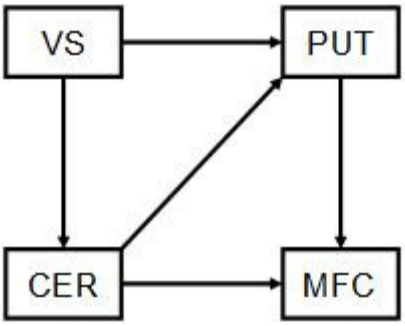


Figure 6.2. Hypothesized path diagram for a single-group analysis using PLS SEM. The diagram represents relationships between ventral striatum (VS), cerebellum (CER), striatum putamen (PUT), and motor frontal cortex (MFC).

The results of the single-group PLS analysis are shown in Figure 6.3 and the path estimates, standard errors, and p-values are displayed in Table 6.1.

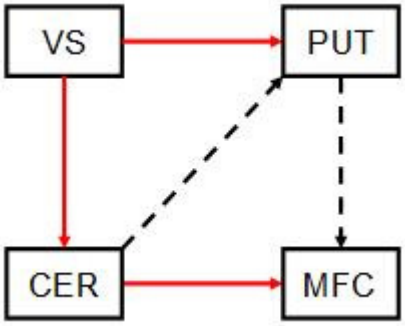


Figure 6.3. Results of single-group analysis using PLS SEM (the results are the same for MLE SEM). Dotted paths represent insignificant effects, red paths represent significant positive relationships between variables. The significance level of these tests is .05 (two-sided).

	β_1	β_2	β_3	β_4	β_5
path coeff	.8530	.4880	.0077	.2762	.8921
std. error	.1169	.1241	.1736	.1696	.2278
p-value	.0000	.0015	.9653	.1274	.0018

Table 6.1. Display of results of PLS SEM single-group analysis using the Brain Functional Pathway data (from normal subjects under the influence of placebo).

For the reference of the reader, the results illustrated in Figure 6.3 also represent the results of the single-group analysis using the single-group MLE SEM framework. The numerical results of this analysis are displayed in Table 6.2.

	β_1	β_2	β_3	β_4	β_5
path coeff	.8530	.4880	.0077	.2762	.8921
std. error	.1088	.1199	.1615	.1579	.2121
p-value	.0000	.0000	.9620	.0803	.0000

Table 6.2. Display of results of single-group MLE SEM analysis using the Brain Functional Pathway data (from normal subjects under the influence of placebo).

The reader may notice that the path coefficients are the same for both cases, but the errors for the PLS technique are slightly larger than those generated by the MLE SEM technique (and therefore, the p-values of MLE SEM technique are slightly smaller). This will be discussed further in Chapter 7.

6.3 PLS SEM for Mixed Designs

The method of Kenny and Judd [31] has been adapted for PLS path modeling [76]. Here, however, any type of covariate can be included, discrete or continuous, because there are no distributional assumptions for the independent variables in PLS (as it takes the form of individual regression equations). PLS requires no distributional assumptions, as the OLS coefficients are not calculated based on the distribution of the response variables, so observations are not required to be independent (could be taken from the same group of subjects). However, the estimation of standard errors is very conservative (bootstrapping) because of the distribution-free quality of the method. An additional issue is the lack of a parameter in the model that represents the correlation of measurements taken from the same subject. Our methodology allows us to account for these correlations and test their significance less conservatively because we make distributional assumptions, allowing use of the linear mixed model for evaluating the significance of the path coefficients.

6.3.1 The Case of No Latent Variables in the System

Because there is no requirement for independence of observations, and there are no requirements on the distribution of the variables present in PLS SEMs [68, 70], we can incorporate the mixed design directly into the model. As in the ML methodology for MD-SEM, we incorporate influence of factors such as groups or repeated measures by reparametrizing the path coefficient. For example, the equation

$$put = \beta vs + \epsilon$$

becomes

$$put = (\alpha_0 + \alpha_1 G + \alpha_2 D)vs + \epsilon = \alpha_0 vs + \alpha_1 G * vs + \alpha_2 D * vs + \epsilon$$

where G represents the influence of group membership on the path from VS to PUT, and D represents the respective influence of the treatment administered (placebo or drug). Then α_0 is the original path coefficient, α_1 represents the contribution from group membership, and α_2 represents the contribution from treatment. G*VS and D*VS join VS in the model as variables in the dataset under study. This means datasets representing all values of G and D are present in one dataset. The distributional issues and the issues

of non-independent observations within the same dataset that would cause problems for MLE SEM do not occur in PLS SEM. However, we must be sure to use a regression technique that allows for dependent observations. Then, the model can be estimated directly, as in the case of a single group under a single treatment. Of course, to use PLS SEM in its standard form, the model must be recursive. The recursive model we will evaluate as an example is shown in Figure 6.4.

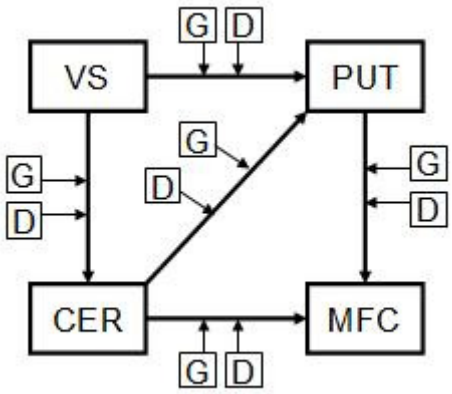


Figure 6.4. Hypothesized path diagram for mixed-design PLS SEM analysis. This is the same model we have evaluated using our mixed-design MLE SEM methodology.

The corresponding equations for a single-group model (Figure 6.2) are:

$$put = \beta_1 vs + \beta_3 cer + \epsilon_1$$
$$mfc = \beta_4 put + \beta_5 cer + \epsilon_2$$
$$cer = \beta_2 vs + \epsilon_3$$

Incorporating the mixed design for estimating the effects of group membership and treatment drug on each path updates the model to:

$$put = (\beta_1 + \beta_2 G + \beta_3 D)vs + (\beta_7 + \beta_8 G + \beta_9 D)cer + \epsilon_1$$
$$mfc = (\beta_{10} + \beta_{11} G + \beta_{12} D)put + (\beta_{13} + \beta_{14} G + \beta_{15} D)cer + \epsilon_2$$
$$cer = (\beta_4 + \beta_5 G + \beta_6 D)vs + \epsilon_3$$

or,

$$put = \beta_1 vs + \beta_2 G * vs + \beta_3 D * vs + \beta_7 cer + \beta_8 G * cer + \beta_9 D * cer + \epsilon_1$$
$$mfc = \beta_{10} put + \beta_{11} G * put + \beta_{12} D * put + \beta_{13} cer + \beta_{14} G * cer + \beta_{15} D * cer + \epsilon_2$$
$$cer = \beta_4 vs + \beta_5 G * vs + \beta_6 D * vs + \epsilon_3$$

This model can be estimated via PLS SEM in a series of independent regressions (OLS, PLS, any appropriate technique). We will use a linear mixed model to model each equation because there are repeated measures on the same subject.

6.3.2 Implementation of PLS MD-SEM using a Linear Mixed Model

We are interested in a model that will incorporate fixed effects (independent variables, group and treatment effects), but also allow for correlation in the errors of observations taken on the same subject. A linear mixed model [77, 78] can incorporate each of these needs, but is even more general than we require. Our mixed model will contain no random effects, but will have an error covariance matrix that differs from that of an OLS regression. A linear mixed model is effectively a regression for repeated measures data. The model is

$$y = X\beta + e, e \sim N(0, \Sigma),$$

where covariance matrix Σ allows for correlations among observations taken from the same individual. In this model, y is continuous and normally distributed, and X is the matrix containing the fixed effect influences on y [28].

We will let

$$\Sigma = \begin{bmatrix} \sigma^2 + \rho & \rho & & \\ \rho & \sigma^2 + \rho & & \\ & & \sigma^2 + \rho & \rho \\ & & \rho & \ddots \end{bmatrix},$$

where ρ represents the correlation between measurements taken on the same subject (assumed constant among subjects) . The density function for *e* is

$$\begin{aligned} f(e) &= (2\pi)^{-2n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} e' \Sigma^{-1} e \right) \\ &= (2\pi)^{-2n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} (y - X\beta)' \Sigma^{-1} (y - X\beta) \right) \end{aligned}$$

Then the log-likelihood function is

$$\begin{aligned} \log L(\theta) &= -n \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (y - X\beta)' \Sigma^{-1} (y - X\beta) \\ &= -2 \left\{ \log |\Sigma| + (y - X\beta)' \Sigma^{-1} (y - X\beta) \right\} \end{aligned}$$

Minimizing the function in braces with respect to β and the parameters of Σ simultaneously will identify the maximum likelihood estimators of both, yielding estimates for the regression coefficients. Standard errors can be derived using the asymptotic distribution of the maximum likelihood estimates $\hat{\theta}$.

$$ACOV(\hat{\theta}) = \left\{ -E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'} \right] \right\}^{-1} = 2\left[\frac{\partial^2 F}{\partial\theta\partial\theta'} \right]^{-1}$$

where F is the minimized function derived above. Therefore, $2H^{-1}$ is the asymptotic covariance matrix of the parameters. Standard errors for these estimates can be found by calculating the second derivative matrix (Hessian) of the objective function. The square roots of the diagonal entries are the standard errors for corresponding parameters. Tests of significance of the regression coefficients can be performed using the estimates, standard errors, and the assumption of asymptotic normality.

In MLE MD-SEM, we center all data by a grand mean before beginning the analysis, so we perform the same centering technique on the data before analyzing the data using PLS SEM. PLS SEM (with no latent variables) is implemented as a linear mixed model, as discussed above. Recall that we have 16 normal subjects and 25 cocaine abusers, each measured under two conditions. Then the total number of observations used in the linear mixed model is 82. The results of the analysis of the hypothesized path diagram shown in Figure 6.4 are illustrated in Figure 6.5 and tabulated in Table 6.3. The model for this analysis is:

$$put = (\beta_1 + \beta_2 G + \beta_3 D)vs + (\beta_7 + \beta_8 G + \beta_9 D)cer + \epsilon_1$$
$$mfc = (\beta_{10} + \beta_{11} G + \beta_{12} D)put + (\beta_{13} + \beta_{14} G + \beta_{15} D)cer + \epsilon_2$$
$$cer = (\beta_4 + \beta_5 G + \beta_6 D)vs + \epsilon_3$$

The results are similar to those shown in Figure 3.6, except that the drug effect present in Figure 3.6 along the path from VS to CER is not present here.

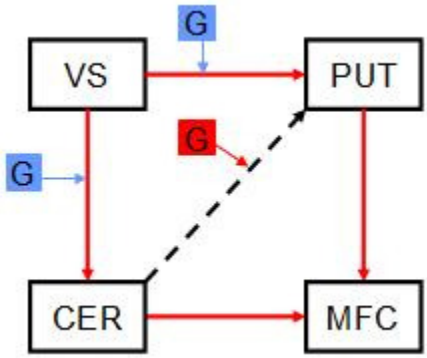


Figure 6.5. Results of PLS MD-SEM applied to the Brain Functional Pathway data (hypothesized model in Figure 6.4). Paths determined to be non-significant are shown as dotted arrows. Arrows and factors highlighted are significant after a two-sided z-test with $\alpha = .05$. Red paths and factors indicate a positive influence while blue factors and paths indicate a negative influence.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
path	.7937	-.2413	-.0570	.7203	-.353	.1078	.0753	.2661
error	.0952	.1074	.0665	.1063	.1162	.0837	.1178	.1367
p-value	.0000	.0311	.3977	.0000	.0043	.2053	.5270	.0596

	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
path	.1321	.3934	-.0022	.2092	.7624	-.0106	-.2825
error	.1134	.1613	.1944	.1707	.2118	.2501	.2421
p-value	.2520	.0199	.9911	.2284	.0010	.9664	.2512

Table 6.3. Results of PLS MD-SEM applied to the Brain Functional Pathway data (hypothesized model in Figure 6.4). Estimated path and factor coefficients are displayed, with corresponding standard error and p-value are shown.

6.3.3 The Case of Latent Variables Present in the System

If there are latent variables present in the model, the standard PLS estimation method can be employed for estimating the latent variables *within each dataset*. Once these observations are estimated, these variables can be incorporated into the linear mixed model for estimation as any other measured variable, including interactions of group and treatment with the estimate.

6.4 Power Analysis of the PLS Method

6.4.1 Power of PLS SEM

In a PLS SEM analysis for a single group (or for multiple independent groups), the results are the compiled results of individual OLS regression analysis of each equation in the system. Therefore, a test of statistical significance such as $H_0 : \theta_i = 0$ is the test of a single coefficient in a single OLS regression equation.

In OLS regression, t-tests determine whether a particular coefficient is significant (significantly different from zero), and F-tests determine whether the proportion of variance in the dependent variable explained by the corresponding independent variable is significantly different from zero [79]. These tests are equivalent. Power is evaluated through analysis of the F-test. Consider a test of the contribution of a single variable to the variance in the dependent variable. When the null hypothesis is true (variance explained by X_j is 0), the test statistic follows the F distribution, with numerator degrees of freedom $u = 1$ and denominator degrees of freedom $v = N - p - 1$, where N is the sample size, and p is the total number of predictors being considered in this model. When the alternative hypothesis is true, the test statistic follows a non-central F distribution, with the same degrees of freedom as above, and non-centrality parameter given by

$$\lambda = \left(\frac{R^2 - R^2_{(-j)}}{1 - R^2} \right) (u + v + 1) = \left(\frac{R^2 - R^2_{(-j)}}{1 - R^2} \right) (N - p + 1).$$

In the above definition, R is the multiple correlation coefficient for the model including all independent variables, and $R_{(-j)}$ is the multiple correlation coefficient for the model including all independent variables except the variable we are testing [80].

The multiple correlation coefficient R of a particular model can be described as the correlation of the values of the dependent variable predicted by the model and the actual measured values of the dependent variable, $Corr(y, \hat{y})$, and its square can be calculated using the following formula [81].

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Once the non-centrality parameter is determined, the power can be calculated. Power is defined as the probability of rejecting the null hypothesis given that it is false.

$$P(F_{test} > F_{1,N-p-1,0}(\alpha) \mid F_{test} \sim F_{1,N-p-1,\lambda})$$
$$= P(F_{1,N-p-1,\lambda} > F_{1,N-p-1,0}(\alpha))$$

In other words, find the probability of getting a value for the test statistic generated from the non-central F distribution that is large enough to reject the null hypothesis when the critical region comes from the central F distribution, assuming a significance level of α. Generally, we calculate power using the CDF of the non-central F distribution.

$$Power = 1 - P(F_{1,N-p-1,\lambda} < F_{1,N-p-1,0}(\alpha))$$

6.4.2 Power of PLS MD-SEM

In a PLS SEM analysis with multiple conditions (or a complete mixed design), we are looking at the compiled results of individual analysis of each equation in the system using a linear mixed model. Therefore, a test of statistical significance such as $H_0 : \theta_i = 0$ is the test of a single coefficient in a single linear mixed model regression equation. The standard method for evaluating the power of a test of this sort involves the non-central F distribution.

As in the case of OLS regression, the non-centrality parameter is calculated based on the F test statistic (which is inherently based on the variance structure and the amount of variation in y accounted for by x). In the case of the linear mixed model, the procedure for calculating power is the same as in the case of significance tests of a single parameter in OLS regression, except for the calculation of the non-centrality parameters. This time, we can use the formula

$$\lambda = uF$$

where u is the numerator degrees of freedom, and F is the test statistic for the F-test on the parameter of interest. Of course, the test statistic is based on the variance structure of the errors, where the covariance of repeated measures is incorporated [77, 82, 83].

Once the non-centrality parameter is calculated, the power follows in a traditional way. Based on α, calculate the upper critical value in the central F distribution based on the appropriate numerator (u) and denominator (v) degrees of freedom (estimated using the Satterthwaite calculation). Then power can be calculated as above,

$$Power = 1 - P(F_{u,v,\lambda} < F_{u,v,0}(\alpha)) \text{, [83].}$$

6.4.3 Power of PLS SEM for the Brain Functional Pathway Study

If we are to analyze the power of the analysis of the Brain Functional Pathway Study without simulating data, we must consider this to be a pilot study, as calculating power retroactively is never suggested unless the current study is an investigative pilot study. Using the method discussed in Section 6.4.2, we have conducted a power analysis of PLS SEM for the particular model discussed in Section 6.3.2 with hypothesis shown in Figure 6.4 with results shown in Figure 6.5 and Table 6.3.

As discussed in the previous section, tests on path parameters and factor parameters are F-tests. Unlike in Chapter 4, we will not be simulating data, but calculating the power of the test $H_0 : \beta_i = 0$ using the original data, and the significance level fixed at $\alpha = .05$. The results of this calculation for single paths and single factors are shown in Table 6.4. The most powerful tests represent the most significant coefficients, of course (with the largest relative effect size). Because these results correspond to the model shown below evaluated using PLS SEM, each equation is evaluated individually using a linear mixed model. The power of some tests is incredibly small, which is due to the very small effect size for the corresponding parameter (indicating that the parameter was not significant, and therefore, cannot be tested for power—the probability of being classified as significant when it is truly significant). Recall that the total number of observations for this model is 82.

$$put = (\beta_1 + \beta_2 G + \beta_3 D)vs + (\beta_7 + \beta_8 G + \beta_9 D)cer + \epsilon_1$$
$$mfc = (\beta_{10} + \beta_{11} G + \beta_{12} D)\,put + (\beta_{13} + \beta_{14} G + \beta_{15} D)cer + \epsilon_2$$
$$cer = (\beta_4 + \beta_5 G + \beta_6 D)vs + \epsilon_3$$

Effect	Parameter	Effect Size	p-value of $H_0 : \beta_i = 0$	Power of Testing $H_0 : \beta_i = 0$
Path	β_1	.793	.000	.999
Factor	β_2	-.241	.031	.592
Factor	β_3	-.057	.398	.133
Path	β_4	.720	.000	.999
Factor	β_5	-.353	.004	.876
Factor	β_6	.108	.205	.255
Path	β_7	.075	.527	.095
Factor	β_8	.266	.060	.479
Factor	β_9	.132	.252	.205
Path	β_{10}	.393	.020	.728
Factor	β_{11}	-.002	.991	.050
Factor	β_{12}	.209	.228	.232
Path	β_{13}	.762	.001	.966
Factor	β_{14}	-.011	.966	.050
Factor	β_{15}	-.283	.251	.216

Table 6.4. Results of power calculation for Mixed Design PLS SEM. The power of tests for significance of single paths and single factors are computed. The significance level is fixed at $\alpha = .05$.

Chapter 7: Comparison of PLS and MLE SEM

In order to explore the relationship between regression-based PLS SEM and covariance-based MLE SEM, we compare these two approaches through several important scenarios. First, we algebraically evaluate a series of four single-group models containing no latent variables to determine their similarities in the estimated path parameters and the corresponding standard errors.

7.1 Algebraic Comparison: PLS and MLE SEM (various single-group models)

The models we consider for the algebraic comparison of MLE and PLS SEM are shown in Table 7.1. They are all simple variations of one- and two-equation systems. Model 1 is a straightforward simple linear regression; Model 2 is a multiple linear regression; Model 3 represents a multivariate linear regression. Model 4 represents a model that is typical in SEM (some response variables also act as an influence on another response), but because PLS SEM estimates each equation separately, it can also be considered two simple linear regressions. We will algebraically solve for the coefficients b using both MLE and PLS SEM and compare the results.

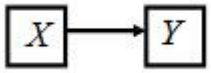
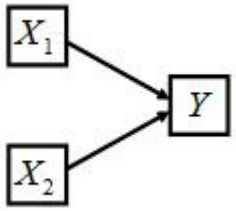
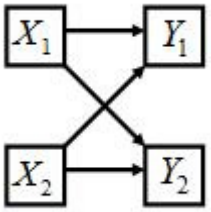
1		$Y = bX + e$
2		$Y = b_1X_1 + b_2X_2 + e$
3		$Y_1 = b_1X_1 + b_2X_2 + e_1$ $Y_2 = b_3X_1 + b_4X_2 + e_2$
4		$Y_1 = b_1X + e_1$ $Y_2 = b_2Y_1 + e_2$

Table 7.1. Illustrations of the four simple linear models we will solve algebraically to compare MLE SEM and PLS SEM for single-group models. Model 1 is a simple linear regression; Model 2 is a multiple linear regression; Model 3 is a multivariate linear regression. Model 4 is a typical SEM model containing a response variable that also appears on the right-hand side of an equation as an independent variable. This can be evaluated via PLS SEM because the two equations are evaluated as two individual regressions.

7.1.1 Model 1

Model 1 is a simple linear regression: $Y = bX + e$ (illustrated in Table 7.1).

7.1.1.1 MLE SEM Framework

$$\mathrm{B} = [0], \Gamma = [b], \Phi = [p], \Psi = [s]$$

$$\Sigma(\theta) = \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-\mathrm{B})^{-1\prime} & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(I-\mathrm{B})^{-1\prime} & \Phi \end{bmatrix} = \begin{bmatrix} bpb+s & bp \\ bp & p \end{bmatrix}$$

In MLE SEM, we minimize the fitting function $F_{ML} = \log|\Sigma(\theta)| + tr\left[S\Sigma^{-1}(\theta)\right]$ with respect to the parameters in $\Sigma(\theta)$:

$$\min_{b,p,s}\left\{\log|\Sigma(\theta)| + tr\left[S\Sigma^{-1}(\theta)\right]\right\}$$

Note that our input data, the covariance matrix of $\begin{pmatrix} Y & X \end{pmatrix}'$, S, has arbitrary entries:

$$S = \begin{bmatrix} t & u \\ u & v \end{bmatrix},$$

or $Var(Y)=t, Var(X)=v, Cov(X,Y)=u$.
In order to minimize, take derivatives with respect to each parameter, set equal to 0, and solve the system:

$$F_{ML} = \log(sp) - \frac{2bu}{s} + \frac{vb^2}{s} + \frac{t}{s} + \frac{v}{p}$$

$$\frac{\partial F}{\partial b} = -\frac{2u}{s} + \frac{2vb}{s} = 0$$

$$\frac{\partial F}{\partial s} = \frac{1}{s} + \frac{2bu}{s^2} - \frac{vb^2}{s^2} - \frac{t}{s^2} = 0$$

$$\frac{\partial F}{\partial p} = \frac{1}{p} - \frac{v}{p^2} = 0$$

Solve the system to find $p = v, b = \frac{u}{v}, s = t - \frac{u^2}{v}$, or

$$p = Var(X), b = \frac{Cov(X,Y)}{Var(X)}, s = Var(Y) - \frac{Cov(X,Y)^2}{Var(X)}.$$

Calculating the standard error of parameter b is of particular interest to us, so we must calculate the entries of the asymptotic covariance matrix.

$$ACOV(\hat{\theta}) = \frac{2}{n-1}\begin{bmatrix} \frac{\partial F}{\partial b \partial b'} & \frac{\partial F}{\partial b \partial s'} & \frac{\partial F}{\partial b \partial p'} \\ \frac{\partial F}{\partial s \partial b'} & \frac{\partial F}{\partial s \partial s'} & \frac{\partial F}{\partial s \partial p'} \\ \frac{\partial F}{\partial p \partial b'} & \frac{\partial F}{\partial p \partial s'} & \frac{\partial F}{\partial p \partial p'} \end{bmatrix}^{-1}$$

Calculating the second derivatives,

$$ACOV(\hat{\theta}) = \frac{2}{n-1}\begin{bmatrix} \frac{2v}{s} & \frac{2(u-vb)}{s^2} & 0 \\ \frac{2(u-vb)}{s^2} & \frac{-1}{s^2} + \frac{(2vb^2+2t-4bu)}{s^3} & 0 \\ 0 & 0 & \frac{-1}{p^2} + \frac{2v}{p^3} \end{bmatrix}^{-1}$$

The standard error of b (used to test the hypothesis $H_0 : b = 0$) will be $ACOV(\hat{\theta})_{11}$. Using Mathematica to find the analytic inverse of this matrix,

$$ACOV(\hat{\theta})_{11} = \frac{s}{n-1}\left(\frac{2t - s - 4bu + 2b^2 v}{2tv - sv - 2u^2}\right).$$

Therefore,

$$se(\hat{b}) = \sqrt{\frac{s}{n-1}\left(\frac{2t - s - 4bu + 2b^2 v}{2tv - sv - 2u^2}\right)}$$

This formula for $se(\hat{b})$ contains elements of S (our data covariance matrix) but also the parameters s and b. Substituting our formulas for those parameters,

$$b = \frac{u}{v}, s = t - \frac{u^2}{v}$$

$$se(\hat{b}) = \sqrt{\frac{tv - u^2}{(n-1)v^2}}$$

Hypothesis tests of SEM parameters are performed under the assumption of asymptotic normality. Therefore, the $100(1-\alpha)\%$ confidence interval for b is given by:

$$\frac{u}{v} \pm z_{\alpha/2}\sqrt{\frac{tv - u^2}{(n-1)v^2}}$$

If 0 is contained in this interval, then the path parameter represented by b is not significant.

7.1.1.2 PLS SEM Framework

Using PLS to solve this problem, we estimate the parameters of a simple linear regression.

$$Y = bX + e$$

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - bX_i)^2$$

To find the estimators, take the derivative of S with respect to b and set equal to 0.

$$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n} (Y_i - bX_i)(-X_i) = 0$$

$$\sum_{i=1}^{n} (Y_i X_i - bX_i^2) = 0$$

$$\sum_{i=1}^{n} Y_i X_i = b\sum_{i=1}^{n} X_i{}^2 \Rightarrow b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

First, as in MLE SEM, center all variables, so

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{(n-1)Cov(X,Y)}{(n-1)Var(X)} = \frac{u}{v}$$

assuming that we are using the same input covariance matrix where $S = \begin{bmatrix} t & u \\ u & v \end{bmatrix}$.

Now evaluate the standard error of parameter b and the corresponding hypothesis test to determine whether b is significant. In OLS regression, $Cov(\hat{b}) = \hat{\sigma}^2 (X'X)^{-1}$, where $\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-r-1)$. Here, we have 1 independent variable (r=1). Then

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-2} = \frac{(y-Xb)'(y-Xb)}{n-2} = \frac{\sum_{i=1}^{n}(y_i - bx_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - 2bx_i y_i + b^2 x_i^2)}{n-2}$$
$$= \frac{(n-1)\,\mathrm{var}(y) - 2b(n-1)\,\mathrm{cov}(x,y) + b^2(n-1)var(x)}{n-2} = \frac{n-1}{n-2}(t - 2bu + b^2 v)$$

Also,

$$(X'X)^{-1} = \frac{1}{(n-1)\,\mathrm{var}(x)} = \frac{1}{(n-1)v}$$

Therefore,

$$\mathrm{cov}(\hat{b}) = \frac{n-1}{n-2}(t - 2bu + b^2 v)\frac{1}{(n-1)v} = \frac{t - 2bu + b^2 v}{(n-2)v}.$$

Substituting the parameter value of $b = \frac{u}{v}$ in order to calculate $se(\hat{b})$ in terms of elements of S only,

$$\mathrm{cov}(\hat{b}) = \frac{tv - u^2}{(n-2)v^2}.$$

Then the standard error of $\hat{b}$ is

$$se(\hat{b}) = \sqrt{\frac{tv - u^2}{(n-2)v^2}}.$$

To test the significance of regression coefficients b, we use an F-test or t-test. We can calculate the $100(1-\alpha)\%$ confidence interval for b using:

$$\frac{u}{v} \pm t_{\alpha/2,n-2}\sqrt{\frac{tv-u^2}{(n-2)v^2}}.$$

7.1.1.3 Conclusions for Model 1

For this model, the estimated path coefficient, b, has exactly the same value for PLS and MLE SEM. The two-sided confidence intervals for b (and therefore, two-sided hypothesis tests of significance) are slightly different. The standard errors differ slightly, and the distribution of the calculations differ very slightly. The calculation of standard error for OLS regression will result in a slightly larger error, yielding a larger confidence interval (more conservative) for small sample sizes, and the use of the t-distribution with more weight in the tails also gives a more conservative confidence interval when there is a small sample size.

7.1.2 Model 2

Model 2 is a multiple linear regression: $Y = b_1X_1 + b_2X_2 + e$ (illustrated in Table 7.1).

7.1.2.1 MLE SEM Framework

$$B = [0], \Gamma = [b_1 \quad b_2], \Phi = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix}, \Psi = [s]$$

$$\Sigma(\theta) = \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1\,\prime} & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(I-\mathrm{B})^{-1\,\prime} & \Phi \end{bmatrix}$$

The input data, the covariance matrix of $\begin{pmatrix} Y & X_1 & X_2 \end{pmatrix}'$, S, has arbitrary entries:

$$S = \begin{bmatrix} t & u & v \\ u & w & z \\ v & z & h \end{bmatrix}, \text{ or}$$

$Var(Y) = t, Var(X_1) = w, Var(X_2) = h, Cov(X_1, Y) = u, Cov(X_2, Y) = v, Cov(X_1, X_2) = z$.

For this more complicated model, have identified the estimated parameter values by minimizing the fit function in Mathematica. The calculated values are:

$$b_1 = \frac{-vz+uh}{-z^2+wh}, b_2 = \frac{vw-uz}{-z^2+wh}, p_1 = w, p_2 = z, p_3 = h, s = t - \frac{(vw-uz)^2 h}{(-z^2+wh)^2} - \frac{u(-vz+uh)}{-z^2+wh}$$

7.1.2.2 PLS SEM Framework

This time, the model is $Y = b_1X_1 + b_2X_2 + e$, a standard multiple linear regression.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_1 X_{1i} - b_2 X_{2i})^2$$
$$\frac{\partial S}{\partial b_1} = 2\sum_{i=1}^{n} (Y_i - b_1 X_{1i} - b_2 X_{2i})(-X_{1i}) = 0$$
$$\frac{\partial S}{\partial b_2} = 2\sum_{i=1}^{n} (Y_i - b_1 X_{1i} - b_2 X_{2i})(-X_{2i}) = 0$$

Solving this system as we did in the simpler case above, we find that

$$b_1 = \frac{uh - zv}{wh - z^2}, b_2 = \frac{wv - uz}{wh - z^2}.$$

Here again, the standard errors (and thereby, our significance tests) will be different, but the parameter estimates are the same, even when we have two independent variables (which forces extra parameters for modeling their covariances matrix in the MLE SEM method).

7.1.3 Model 3

Model 3 is a multivariate linear regression, but is considered in PLS SEM to be two individual multiple regressions (which will yield the same estimates [84]). Model 3 is illustrated in Table 7.1.

$$Y_1 = b_1 X_1 + b_2 X_2 + e_1$$
$$Y_2 = b_3 X_1 + b_4 X_2 + e_2$$

7.1.3.1 MLE SEM Framework

$$\mathrm{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \Gamma = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}, \Phi = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix}, \Psi = \begin{bmatrix} q_1 & 0 \\ 0 & q_2 \end{bmatrix}$$

Note that our input data, the covariance matrix of $\begin{pmatrix} Y_1 & Y_2 & X_1 & X_2 \end{pmatrix}'$, S, has arbitrary entries

$$S = \begin{bmatrix} t & u & v & w \\ u & g & j & z \\ v & j & r & s \\ w & z & s & h \end{bmatrix},$$

or $Var(Y_1) = t, Cov(X_1, Y_1) = v$, for example.

Minimizing the fit function in Mathematica, the resulting formulas for parameter estimates are:

$$b_1 = \frac{hv - sw}{hr - s^2}, b_2 = \frac{-sv + rw}{hr - s^2}, b_3 = \frac{hj - sz}{hr - s^2}, b_4 = \frac{rz - sj}{hr - s^2}$$

7.1.3.2 PLS SEM Framework

For PLS SEM, we would have two equations:

$$Y_1 = b_1 X_1 + b_2 X_2 + e_1$$
$$Y_2 = b_3 X_1 + b_4 X_2 + e_2$$

These equations would have the same coefficient estimates and standard errors if we evaluate them simultaneously (multivariate regression) or individually (multiple regression) [75, 84]. Using a multiple OLS regression on each individual equation, we can use our derivation from Model 2 above to calculate the appropriate values for the path coefficients. Based on the assignments of covariances to variable names shown in the MLE SEM framework,

$$S = \begin{bmatrix} t & u & v & w \\ u & g & j & z \\ v & j & r & s \\ w & z & s & h \end{bmatrix}$$

the coefficients have exactly the same estimates using OLS calculations as in the SEM calculations shown above.

$$b_1 = \frac{hv - sw}{hr - s^2}, b_2 = \frac{-sv + rw}{hr - s^2}, b_3 = \frac{hj - sz}{hr - s^2}, b_4 = \frac{rz - sj}{hr - s^2}$$

PLS SEM and MLE SEM give the same estimates for the path coefficients (of course the standard errors will be different, and therefore, the significance tests may have differing results), even when there are two equations with multiple independent variables.

7.1.4 Model 4

The goal of Model 4 is to investigate whether the estimated path coefficients differ when we have some response variables appearing on the right hand side of other equations in the system (when $B \neq 0$ in MLE SEM). Model 4 (illustrated in Table 7.1) represents a typical SEM with some response variables appearing again on the right-hand side of another equation. This is estimated by PLS SEM as two individual simple linear regressions. The model is shown here:

$$Y_1 = b_1 X + e_1$$
$$Y_2 = b_2 Y_1 + e_2$$

7.1.4.1 MLE SEM Framework

$$\mathrm{B} = \begin{bmatrix} 0 & 0 \\ b_2 & 0 \end{bmatrix}, \Gamma = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \Phi = \begin{bmatrix} p \end{bmatrix}, \Psi = \begin{bmatrix} q_1 & 0 \\ 0 & q_2 \end{bmatrix}$$

Note that our input data, the covariance matrix of $\begin{pmatrix} Y_1 & Y_2 & X_2 \end{pmatrix}'$, S, has arbitrary entries:

$$S = \begin{bmatrix} t & u & v \\ u & w & z \\ v & z & h \end{bmatrix}, \text{ or}$$

$$Var(Y_1) = t, Var(Y_2) = w, Var(X) = h, Cov(Y_1, Y_2) = u, Cov(Y_1, X) = v, Cov(Y_2, X) = z\,.$$

Calculating the model-implied covariance matrix and fit function in Mathematica, as well as solving the system of derivatives, the parameter estimates are as follows:

$$b_1 = \frac{v}{h}, b_2 = \frac{u}{t}, p = h, q_1 = t - \frac{v^2}{h}, q_2 = w - \frac{u^2}{t}$$

7.1.4.2 PLS SEM framework

For this model, we cannot use multivariate regression because the regressors are not consistent between the two equations. However, we can estimate each individually, as is traditional in PLS SEM. Using our previous derivation of the path coefficients for simple linear regression, we can calculate the following coefficients using our previous statement that

$$Var(Y_1) = t, Var(Y_2) = w, Var(X) = h, Cov(Y_1, Y_2) = u, Cov(Y_1, X) = v, Cov(Y_2, X) = z\,.$$

$$b_1 = \frac{v}{h}, b_2 = \frac{u}{t}$$

Here again, even when $\mathrm{B} \neq 0$ (response variables appear on the right-hand side of some equations), we have the same path coefficients in PLS and MLE SEM.

7.2 Single-Group Model Conclusions

For varying single-group models containing only measured variables, PLS SEM and MLE SEM yield the same parameter estimates and similar standard errors (errors in PLS SEM are slightly larger). For a numerical example, see Tables 6.1 and 6.2, which show the parameter estimates for a single-group PLS SEM and the corresponding estimates from single-group MLE SEM. The parameter estimates are the same, and the errors are slightly different, with the PLS SEM errors being larger, creating larger confidence intervals and p-values. For this particular example, the significance tests for

the path parameters yield the same results, illustrated in Figure 6.3 and displayed in Tables 6.1 and 6.2.

The reason for this similarity between the PLS and MLE methods is the inclusion of measurement error in Y but not in X in both models. While MLE SEM is based on the assumption that all variables X and Y come from a multivariate normal distribution, errors are incorporated in the study of each equation but not in the X variables. A covariance matrix is modeled for the X variables, but no measurement error is allowed. This means that the X variables in SEM are, in fact, fixed during the analysis; this assumption is shared by OLS regression (and therefore, PLS SEM). Because PLS and MLE SEM yield the same results, our choice of method in this case matters very little when latent variables are not present in the model. This is good news for researchers working with small sample sizes, since PLS SEM is better suited for those cases.

While the PLS and MLE SEM yield the same results in the case of single-group models containing no latent variables, this is where the similarity among the methods ends. MLE SEM is an incredibly powerful modeling tool that is intended for incorporating latent variables into models. Its flexibility makes it an attractive choice. Researchers can use MLE SEM to implement and estimate error-in-variable (EIV) models [85-87], which incorporate measurement error in both X and Y variables. Researchers achieve measurement error in X by creating a latent variable X and allowing the measured variable X to be its indicator. This creates an error term that is not present in models where X is fixed.

Consider, for example, a single equation model, where we want to understand the impact of a single variable X on a single variable Y. This regression model can be represented via MLE SEM in four ways, as displayed in Table 7.2. Model 1 is the OLS regression model, where Y is measured and modeled with error, but X is fixed. Model 2 is a functional approach to EIV regression, and Model 3 is a structural approach to EIV regression. In both EIV models, both X and Y are treated as random, each with its own measurement error. Model 4 is the traditional MLE SEM representation of regression where both X and Y have measurement error, with modeling error ζ incorporated (ζ is traditional in SEM, but is not included in EIV models).

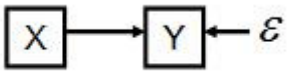
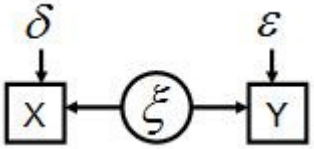
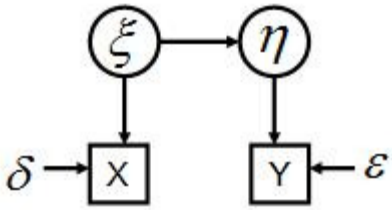
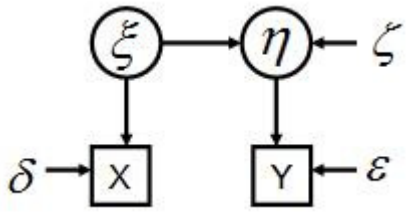
1		$y = \beta x + \varepsilon$
2		$x = \xi + \delta$ $y = \beta\xi + \varepsilon$
3		$x = \xi + \delta$ $y = \eta + \varepsilon$ $\eta = \beta\xi$
4		$x = \xi + \delta$ $y = \eta + \varepsilon$ $\eta = \beta\xi + \zeta$

Table 7.2. (1) OLS regression model, (2) EIV regression model—functional approach, (3) EIV regression model—structural approach, (4) MLE SEM representation of regression where both X and Y have measurement error.

We have estimated each of the models displayed in Table 7.2 using SAS PROC CALIS [88], EQS [26, 39, 62] (both estimate MLE SEMs) and SmartPLS [89], a program designed to analyze PLS SEMs using Wold's original methodology [71, 74, 90] for recursive models. We chose to assign X the values of VS from the brain functional pathways study, and Y the values of PUT from the same study. The results of these analyses are displayed in Table 7.3. Numerical values for each model are (1) the estimate of β, (2) the estimated standard error of the estimator $\hat{\beta}$, and (3) the test statistic for the test of significance $H_0 : \beta = 0$. In all cases, estimates of β indicate a significant relationship between X and Y.

	Model	MLE SEM	PLS SEM
1	$y = \beta x + \varepsilon$	.8567	.8567
		.0750	.0750
		11.4167	11.4167
2	$x = \xi + \delta$ $y = \beta\xi + \varepsilon$	.9017	cannot distinguish from Model 1
		.1371	
		6.575	
3	$x = \xi + \delta$ $y = \eta + \varepsilon$ $\eta = \beta\xi$	.9022	cannot distinguish from Model 1
		.1371	
		6.5821	
4	$x = \xi + \delta$ $y = \eta + \varepsilon$ $\eta = \beta\xi + \zeta$	.8587	cannot distinguish from Model 1
		.1568	
		5.4747	

Table 7.3. Numerical estimates of four representations of the influence of variable X (VS from the brain functional pathway study) on variable Y (PUT) for a single-group regression model. Model 1 is an OLS regression, Model 2 is a functional EIV model, Model 3 is a structural EIV model, and Model 4 is a MLE SEM representation of regression where both X and Y have measurement error. Estimates were obtained using SAS PROC CALIS, EQS, and SmartPLS, a software for estimating PLS SEMs. The first value in each case is the estimated value of β, the second value is the standard error of the estimate, and the third value is the test statistic of the test of significance $H_0 : \beta = 0$.

Several things are clear after reviewing Table 7.3. First, in MLE SEM programs, Models 2 and 3 yield very similar results. This is because in both cases, X and Y have measurement error, and there is no modeling error. Incorporating the modeling error appearing in Model 4, ζ, causes a change in the estimate of β. Also, it is clear that PLS SEM does not share MLE SEM's incredible flexibility for representing many different types of models. An interesting note about MLE SEM is that when we constrain $\delta = \varepsilon$ in Models 2 and 3, we get the same results as a traditional orthogonal regression analysis of the influence of X on Y. This is expected, as orthogonal regression incorporates measurement error in both X and Y, but requires that the variances of the measurement errors be equal [91, 92]. (This further indicates the flexibility of MLE SEM and its software for representing and estimating all models of interest.)

When we consider latent variable representations, PLS SEM and MLE SEM cease to yield the same parameter estimates for all models. The reason for this is fundamental to the different techniques. As discussed in Chapter 6, PLS SEM is a model focused on prediction, while MLE SEM is focused on analysis of covariance structures. One of the major differences between the two methods lies in the way latent variables are treated. In the first stage of PLS SEM, latent variables are "estimated" by creating a weighted average of their corresponding measured indicator variables. Once "observed" values are estimated for the latent variables, the relationships among those variables are estimated through an appropriate regression technique. Therefore, in PLS SEM, if a latent variable has only one indicator, as in Models 2, 3, and 4, the weighted average of measured

indicators is exactly equal to the single indicator's observed values. Then in stage 2, the measured variable's observations are present in the OLS regression. Therefore, if we are using PLS SEM, Models 1, 2, 3, and 4 all yield the same estimate of β. In MLE SEM, latent variables are not estimated, only their covariance structure is estimated, so this method is much more flexible in estimating models with random independent variables and nuances in representation. Clearly, the similarities in the parameter estimates of the two methods end when more complicated models are considered.

Chapter 8: A Method for Generating Hypotheses and Discussion

Whether a researcher plans to use MLE or PLS SEM for single-group studies or mixed designs, a hypothesized network of variable relationships is always required. SEM is a confirmatory technique, always confirming or invalidating a comparison between the hypothesized model and sample data. However, researchers often have a difficult time creating these hypotheses from existing literature that may be conflicting or unclear. There is a clear need for exploratory methods for hypothesizing these path diagrams for use in the confirmatory SEM analysis of a system. In 2007, Marrelec et. al. published a work praising the use of Partial Correlation Network Analysis (PCNA) in conjunction with SEM [93]. The group suggests PCNA may be a good exploratory method for choosing paths when a hypothesized diagram is not readily available from the researcher. They suggest using PCNA to detect the underlying network from the data alone, and then allow the researcher to add direction to those arrows according to their knowledge of the field of interest. We have developed an exploratory technique for creating hypotheses for MD-SEM from data alone, which we present in the following section. It represents a great source of future work related to the methods developed in this dissertation. Finally, we present our future research direction and conclusions.

8.1 Mixed Design Partial Correlation Network Analysis (MD-PCNA)

8.1.1 Introduction to Partial Correlation Network Analysis

Partial correlation network analysis (PCNA) is a statistical technique useful for determining whether undirected relations are present between two variables. What makes partial correlation unique from a standard correlation coefficient is its ability to control for intervening variables that may be present. Consider variables X, Y, and Z. There may be high correlation between all pairs here. However, if the partial correlation of X and Y is zero (controlling for Z), this means the relationship between X and Y vanishes when we are controlling for the influence of Z, or that while there may be a strong relationship between X and Y, it is mediated by Z, and no direct relation is necessary. Therefore, much like SEM, PCNA can help researchers understand relations between variables present in the system under study. Unlike SEM, however, PCNA can handle only undirected relations or links. If we are considering a path diagram, for example, with directed links drawn, then we can eliminate some of those links based on partial correlation study, but we cannot draw a path diagram with unidirectional arrows from data only using partial correlation network analysis [2].

Consider the partial correlation of X and Y, controlling for Z. The most intuitive way to consider partial correlation is to consider two regressions—X on Z and Y on Z—then determine how the residuals of the two equations are correlated. If they are not correlated, then any correlation between X and Y is due to their individual correlations with Z, and not a direct relationship between X and Y. If their residuals are correlated,

then X and Y do have a direct relationship, even after controlling for the influence of Z [94].

A *first-order* partial correlation coefficient is calculated when controlling for only one other variable. A *second-order* partial correlation coefficient is calculated when controlling for two other variables. The general formula for partial correlation is

$$\rho_{XY\cdot Z} = \frac{\rho_{XY\cdot Z\setminus\{Z_0\}} - \rho_{XZ_0\cdot Z\setminus\{Z_0\}}\rho_{YZ_0\cdot Z\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot Z\setminus\{Z_0\}}}\sqrt{1-\rho^2_{YZ_0\cdot Z\setminus\{Z_0\}}}}$$

where Z is a list of the variables we will be controlling for and this formula holds for any $Z_0 \in Z$. This is a recursive formula, for which one variable is removed at each step of the calculation. For an n-th order partial correlation calculation, three recursive calculations take place on the (n-1)-th level, and so on, until the calculation is simply the correlation coefficient of the two variables, without controlling for any other variables. Therefore, the calculation of a partial correlation coefficient can be exponentially expensive in time and number of operations performed [95].

Statistical inferences on the partial correlation coefficients are possible through Fisher's transformation [96]

$$z = \tfrac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right),$$

which follows a

$$N\left(\tfrac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{N-3}\right)$$

distribution, where ρ is the partial correlation coefficient and N is the number of observations under study [97]. According to the resulting z-scores, we can determine whether the observations for two variables are significantly correlated ($H_0 : \rho = 0$) at the $\alpha = .05$ significance level, controlling for indirect correlation through other variables.

8.1.2 Our Interests in PCNA

Marrelec's partial correlation technique could be used to discover the paths shown in Figure 8.1, for example, that could be tested using standard structural equation modeling techniques.

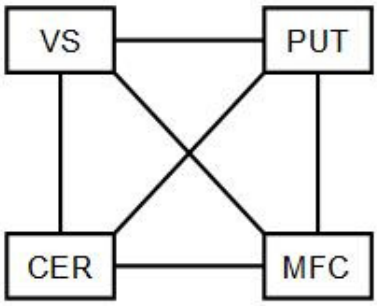


Figure 8.1. Hypothesized partial correlation network between regions of interest in the Brain Functional Pathway Study.

We are most interested in generating diagrams that could be tested using our novel MD-SEM methodology, and in order to do this, we need to be able to determine whether covariate effects are present for particular links. We could generate a diagram containing any or all of the links and factor influences shown in Figure 8.2.

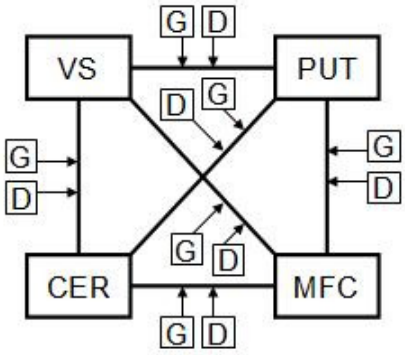


Figure 8.2. Hypothesized partial correlation network with group and drug influences present for MD-PCNA, as we have studied for MD-SEM.

8.1.3 The MD-PCNA Bootstrap

We have developed a data-driven method of detecting linkages between variables, and also covariate influences on those linkages. It is a statistical technique involving bootstrapping to estimate the partial correlation of two variables and the contribution to the partial correlation from the group membership of a subject. Because we develop a "population" of coefficients and group contributions using the bootstrapping technique, we can then evaluate the significance of both the partial correlation and the group contribution. A significant partial correlation indicates that the variables have a strong relationship, and a significant group effect indicates that group membership changes the structure of relations among variables in the set.

Given n variables, find the influence of group membership on partial correlation between two of the regions, i and j. For each pair of variables Y_i and Y_j, we will calculate one thousand (or more if desired) estimates of the partial correlation between these two variables using bootstrapping. The algorithm begins by resampling—sample with replacement from the original data to create a "dataset" of the original size. For each

of these "datasets" we perform the analysis described below. The result is one thousand "estimates" of partial correlation. For these sorted estimates, we can use the percentile method to determine whether the partial correlation of variables Y_i and Y_j is significant (i.e. if zero is contained in a reasonable confidence interval, then the link, or factor influence on the link, is not significant).

Within each repetition of the bootstrapping algorithm, we begin regressing each of the two regions on the remaining regions:

$$Y_i = \sum_{k \neq i,j} a_k Y_k + \epsilon_i$$
$$Y_j = \sum_{k \neq i,j} a_k Y_k + \epsilon_j$$

Then calculate the corresponding residuals:

$$R_i = Y_i - \hat{Y}_i$$
$$R_j = Y_j - \hat{Y}_j$$

These residuals are calculated for each group's variables separately; we do not combine the data for different groups until we are calculating the values of the likelihood function.

As in the case of traditional partial correlation, once the residuals have been calculated, the partial correlation is calculated through a correlation measure:

$$r_k = corr(R_i, R_j)$$

The idea behind our two-level model is that we reparametrize r_k, the traditional representation of partial correlation. Let

$$r_k = b_0 + \sum_{u=1}^{v} b_u F_u$$

where F_i represents the influence of any covariate on the correlation.

In efforts to generate a path diagram for use in SEM from the data, we will perform this correlation analysis, and then incorporate directionality of linkages according to the influence of the collaborator. We will determine whether each of the coefficients is significant in order to determine whether the path and factors are significant, respectively. It is important to note that before we begin the procedure, we have centered our variables, so a constant is not necessary in our regressions.

We perform the significance tests $H_0 : b_0 = 0$ to determine whether the linkage between the variables of interest is significant, and $H_0 : b_u = 0$ (for $u = 1..v$) to determine whether the covariate factor influences on the linkage are significant. After determining the significance of each coefficient via the percentile method, we will know which paths

and factors to include in the path diagram for use with MD-SEM. This analysis must be performed on all variables, pairwise. This requires $n(n-1)/2$ individual (and independent) repetitions of the methodology described above.

8.1.4 Example

We are interested in determining the covariate PCNA structure for the network of variables shown in Figure 8.2, where we are studying two groups of subjects (normal subjects and cocaine abusers), each measured under two conditions (receive placebo and receive methylphenidate). We are interested in determining where linkages exist among these four variables, and where group membership (G) and drug treatment received (D) significantly affect the strength of those linkages.

To begin, we will resample the original subjects 1000 times for each independent group of subjects. However, within a group of subjects, we use the same resample for all conditions. Then, from each resample of n subjects, we calculate the residual variables for our partial correlation analysis. Therefore, we have 1000 estimates for each of r_k for each dataset. When we have those estimates, we reparametrize the partial correlation estimates as follows, to determine the contribution to the partial correlation from the factors group membership and drug received:

$$r_k = a + bG + cD$$

- For dataset 00 (G=0, D=0), $r = a$.
- For dataset 01 (G=0, D=1), $r = a + c$.
- For dataset 10 (G=1, D=0), $r = a + b$.
- For dataset 11 (G=1, D=1), $r = a + b + c$.

Now in order to determine the significance of each coefficient, a, b, and c, we will create estimators of these parameters. Of course, for a, we can use those 1000 estimates from dataset 00 and the percentile method to determine the significance of the linkage between two regions. For c, we will create an estimator using the paired data from datasets 01 and 00. If we subtract the 1000 estimates pairwise, then we can have 1000 estimates of the value of c. Similarly, we can use datasets 11 and 10. Combining these estimates, we can use the percentile method on our 2000 estimates of the value of c to determine its significance. For b, we can take the difference of estimates from 11 and 01, and similarly 10 and 00, though the data is not pairwise. Then again, we will have 2000 estimates of the value of b and can use the percentile method to determine its significance.

This analysis is performed for each pair of variables in the system. For four variables, we have $4*3/2 = 6$ repetitions of the bootstrapping procedure (each of which contains 1000 repetitions of the calculation of partial correlation). Results of this analysis are displayed in Figure 8.3.

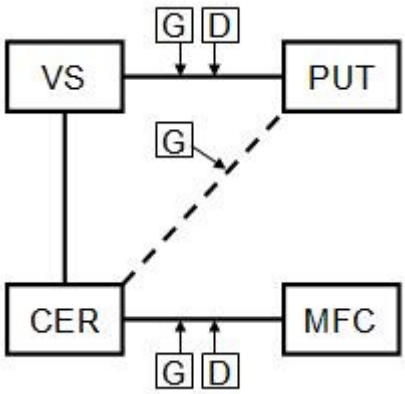


Figure 8.3. Results after analyzing the partial correlation network with group and drug influences present for mixed design PCNA. From this diagram, and the influence of our collaborator (to assign direction to each linkage discovered here), we could fix a data-driven hypothesis for a MD-SEM analysis.

8.1.5 A Second Example: Crohn's Disease

Through the last several months, in addition to our work on the Brain Functional Pathway Study, we have also collaborated with Dr. Ellen Li as she researches which genes and corresponding single nucleotide polymorphisms (SNPs—DNA sequence variations) are markers for Crohn's disease (CD) sub-phenotypes: L1 (ileal ± cecum), L2 (colonic only), L3 (ileal + colon, beyond cecum). She has started with a large number of SNPs and is interested in reducing this to a handful of largely significant SNPs in identifying patients with Crohn's disease.

We are given 29 SNP variables, for 622 subjects, each holding a value of 0, 1, 2, indicating the number of risk alleles in the gene for having the specified SNP. We have performed the MD-PCNA technique on this data with two goals: (1) identify any "hub" SNPs in the network, i.e. SNPs that occur (have high correlation) simultaneously with lots of other SNPs associated with Crohn's disease and may thereby be markers of the disease; (2) generate a potential path diagram for analysis with MD-SEM.

For this analysis, we have three groups of subjects, each with some type of Inflammatory Bowel Disease (IBD)—testing positive in the ileum, having colitis, or both. In an effort to identify markers for ileal CD versus non-ileal CD, we will let subjects who have tested positive in the ileum or in both ileum and having colitis as a single group (L1+L3), and those who have colitis as another, independent group (L2). Therefore, we can evaluate the partial correlations among the SNP variables with the covariate factor of interest in the mixed design being whether the person has tested positive for affected tissue in the ileum or not.

Subjects with colitis belong to group 0 (baseline: L2) and subjects who have tested positive in the ileum belong to group 1 (L1+L3). Therefore, our second-level model (after generating bootstrapped partial correlations is:

$$r = a + bX$$

where a represents the baseline relationship and b represents the influence on that relationship from having Crohn's disease (X being the Crohn's disease factor, taking on

values 0 and 1). Among 29 SNPs and a *smoking* indicator (0 is nonsmoker, 1 is past smoker, 2 is current smoker), we have detected the network of relationships displayed in Figure 8.4. In the figure, blue links represent the baseline effect; this SNP relationship is present in those with colitis. Green links represent the presence of this SNP relationship only when the X factor is present (only for those with disease in the ileum—Crohn's patients). Red links represent relationships that are present in the baseline group and are significantly affected by Crohn's disease.

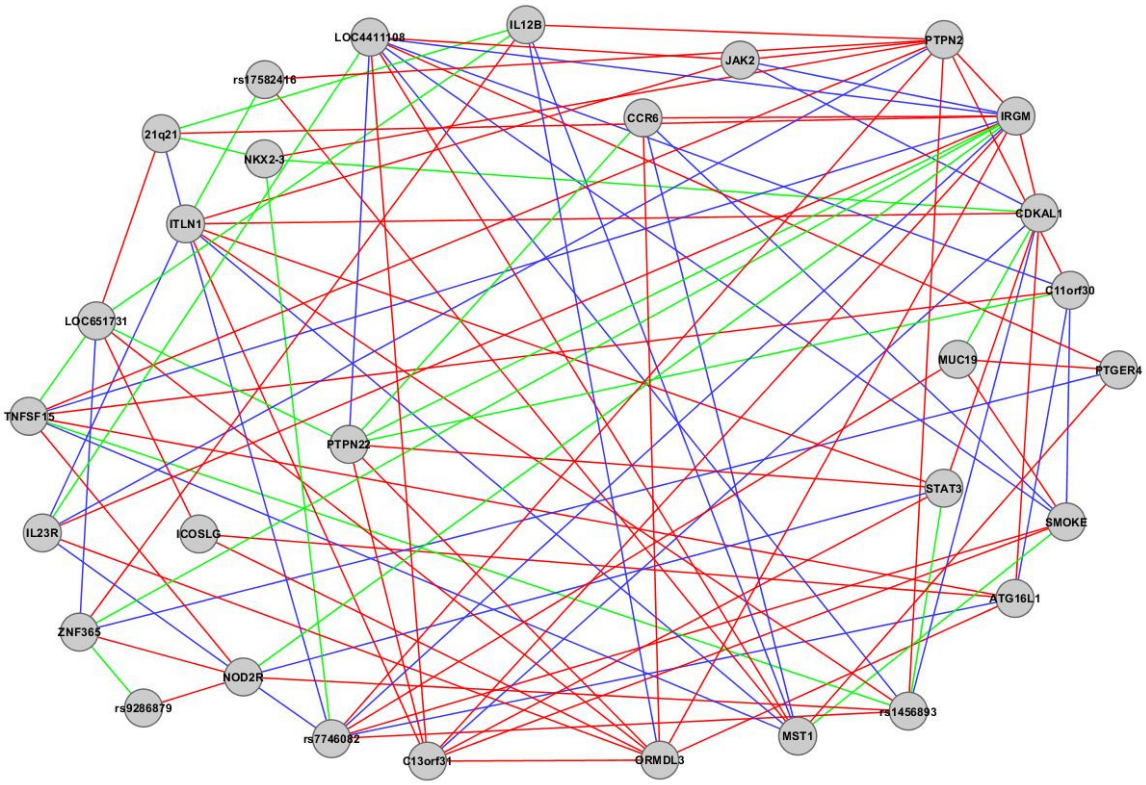


Figure 8.4. Results of MD-PCNA applied to Dr. Ellen Li's investigation of SNPs present in Crohn's disease patients and their simultaneous occurrence at the significance level of $\alpha = .05$. Each gray circle represents a SNP or the smoking indicator (0 is nonsmoker, 1 is past smoker, 2 is current smoker). Legends for the path colors are as follows; Blue: the path is only significant for L2, Green: the path is only significant for L1+L3, Red: the path is significant for both L2 and (L1+L3)—furthermore, the connection strengths are significantly different.

Clearly, there are many significant relationships present in this model. One reason for this is the significance level of $\alpha = .05$ that is not corrected for multiple tests (using the Bonferroni correction, for example). The reason we have not made the correction is because this is an exploratory method for identifying potential relationships. Also, if we use the Bonferroni correction, we divide .05 by the number of comparisons made, so the required significance level is .05/(30*29/2), which is so small that only four links appear in the results (shown in Figure 8.5). As shown in Figure 8.5, when we use the Bonferroni correction, there is a much smaller network of relationships apparent. We also notice that

the paths shown in Figure 8.5 were red in Figure 8.4 but are now blue, indicating that we detected a difference between L2 and (L1+L3) when we did not use the multiple test correction, but we now detect only the baseline effect L2. We believe MD-PCNA is an exciting direction for future study, beginning most importantly with a power analysis.

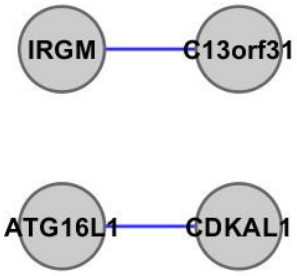


Figure 8.5. Results of MD-PCNA applied to Dr. Ellen Li's investigation of SNPs present in Crohn's disease patients and their simultaneous occurrence at the significance level of $\alpha = .05$, corrected for multiple tests using the Bonferroni correction (divide the p-value by the number of comparisons, 30*29/2. Legends for the path colors are as follows; Blue: the path is only significant for L2, Green: the path is only significant for L1+L3, Red: the path is significant for both L2 and (L1+L3)—furthermore, the connection strengths are significantly different.

8.1.6 Applications of Parallel Computing

The overall goal of running our intended algorithm is to detect relationships between variables in our dataset while controlling for mediating variables also present in the dataset. Given a large number of variables, $n(n-1)/2$ repetitions of the bootstrapping methodology, each requiring one thousand or more iterations, can become a huge computational issue. In a specific example of an application of this methodology, among 216 Single-Nucleotide Polymorphisms (SNPs), we are interested in whether pairs of them are correlated. A SNP is a genetic mutation that occurs in some humans and can change their cellular function, impacting their overall bodily function and, potentially, health. If we can find significant partial correlation between two or more SNP variables, we may be able to determine which SNPs occur together, and which individual SNPs are highly related to many other SNPs at once.

The algorithm is embarrassingly parallel, and was implemented in C++. We evaluated a single dataset containing measurements of 216 variables: 908 normal-subject observations and 1114 case-subject observations. We examine the relationships of the 216 variables pairwise (excluding consideration of a variable with itself), but the results of one comparison do not depend on any other comparison. However, the data must be passed to all processors, as calculation of partial correlation between two variables does depend on the remaining variables in the set (all of the remaining data). Pairwise comparisons of the variables yield $216*215/2 = 23220$ completely independent comparisons. Each single comparison takes about 5 minutes on a single-processor computer. The complete algorithm would require 1935 hours or 81 days to complete using a single-processor computer ($23220*5 = 116100$ minutes). Using 54 processors, we trimmed the run-time to less than 1.5 days. Clearly, parallel computing can be an asset to users of this methodology when dealing with even medium-sized datasets.

8.2 Conclusion and Discussion

SEM for Mixed Designs (MD-SEM) is a comprehensive technique for evaluating the effects of factor covariates on specific pathways in a hypothesized path diagram. It is the first cohesive and easy-to-use methodology of its kind. For this methodology, we developed the estimation techniques (in maximum likelihood framework), derived the overall goodness-of-fit measure, and developed software to estimate the parameters, test their significance, and test the fit of the overall model. In addition to the model development, we performed necessary power and sample size analysis for this new method, determining that MD-SEM should not be used for samples with a small number of observations. Because the power analysis is based on a single model, more investigation can be done in this area.

For models in which the sample size is too small for MLE MD-SEM, we have discussed another technique based on the existing PLS SEM that is more suitable for analyzing small datasets. While PLS SEM could already analyze repeated measurements and multiple independent groups, we have proposed a method for estimation (using a linear mixed model) that will provide less conservative measurements of standard error for each parameter, when the response variables are normally distributed. With MLE SEM and PLS MD-SEM, studies of all sample sizes can be used to identify variable relationships that differ between and within groups.

Additionally, we have developed the methodology to include latent variables, as is common in SEM. Because of this development, any mixed design study can be analyzed using SEM, and as a result, researchers can pinpoint exactly which relationships in their models differ among independent groups of subjects, among repeated measurements, or both, due to the influences of these factor covariates.

Finally, we have introduced a nonparametric exploratory methodology for generating hypotheses for MD-SEM, called Mixed Design Partial Correlation Network Analysis (MD-PCNA). This technique uses bootstrapping to identify which variables have a significant partial correlation, and which links between variables may be affected by the factor influences we evaluate in MD-SEM. Once these linkages and factor influences are detected, researchers can use their field knowledge to add direction to these relationships and use the result as a hypothesized path diagram for MD-SEM. The benefit of this technique is the data-driven exploration of the relationships between variables, rather than being confined to using only field knowledge to generate hypotheses. This work is an exciting direction for future work, beginning with a thorough analysis of the power of the MD-PCNA.

In addition to the future directions we have in MD-PCNA, MD-SEM can be further developed to allow for analysis of more models—those containing categorical independent and dependent variables [98-103] and those with non-normal data. One way this could be handled is to extend the methodology to allow for more estimation techniques (aside from MLE, we could use generalized and weighted least squares [1], and the asymptotically distribution free methods [104]). Additionally, it would be very exciting to develop the method to allow for continuous covariates, as opposed to the strictly binary covariates we have incorporated. MD-SEM introduces an entire family of new models that, when completely developed, can analyze differences in path strength in any mixed design study.

References

1. Bollen, K.A., *Structural Equations with Latent Variables*. 1989: John Wiley and Sons, Inc.
2. Blalock, H.M., *Causal Inferences in Nonexperimental Research*. 1964: University of North Carolina Press.
3. Duncan, O.D., *Path analysis: Sociological examples.* American Journal of Sociology, 1966. **72**: p. 1-16.
4. Wolfle, L.M., *The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography.* Structural Equation Modeling, 2003. **10**(1): p. 1-34.
5. Gopal, R.D., et al., *A behavioral model of digital music piracy.* Journal of Organizational Computing and Electronic Commerce, 2004. **14**(2): p. 89-105.
6. Karlsson, N., et al., *Household consumption: Influences of aspiration level, social comparison, and money management.* Journal of Economic Psychology, 2004. **25**(6): p. 753-769.
7. Krause, D.R., T.V. Scannell, and R.J. Calantone, *A structural analysis of the effectiveness of buying firms' strategies to improve supplier performance.* Decision Sciences, 2000. **31**(1): p. 33-55.
8. Verwaal, E., et al., *Resources access needs and capabilities as mediators of the relationship between VC firm size and syndication.* Small Business Economics, 2010. **34**(3): p. 277-291.
9. de Marco, G. and B. Devaluchelle, *Brain functional modeling, what do we measure with fMRI data?* Neuroscience Research, 2009. **64**(1): p. 12-19.
10. Kim, J., et al., *Unified Structural Equation Modeling Approach for the Analysis of Multisubject, Multivariate Functional MRI Data.* Human Brain Mapping, 2007. **28**: p. 85-93.
11. Lin, F.H., et al., *Functional and Effective Connectivity of Visuomotor Control Systems Demonstrated Using Genearlized Partial Least Squares and Structural Equation Modeling.* Human Brain Mapping, 2009. **30**(7): p. 2232-2251.
12. Perin, B., et al., *Alertness in young healthy subjects: An fMRI study of brain region interactivity enhanced by a warning signal.* Brain and Cognition, 2010. **72**(2): p. 271-281.
13. Rosenbaum, R.S., et al., *Altered connectivity among emotion-related brain regions during short-term memory in Alzheimer's disease.* Neurobiology of Aging, 2010. **31**(5): p. 780-786.
14. Sharma, N., J.C. Baron, and J.B. Rowe, *Motor Imagery After Stroke: Relating Outcome to Motor Network Connectivity.* Annals of Neurology, 2009. **66**(5): p. 604-616.
15. Yan, C.M. and J.P. Dillard, *Emotion inductions cause changes in activiation levels of the behavioural inhibition and approach systems.* Personality and Individual Differences, 2010. **48**(5): p. 676-680.
16. Zhang, Y., *Path Analysis of Multivariate Time Series fMRI Data with Subject-level Covariates*, in *Applied Mathematics and Statistics*. 2007, Stony Brook University: Stony Brook, NY.

17. Gatt, J.M., et al., *Interactions between BDNF Val66Metpolymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety.* Molecular Psychiatry, 2009. **14**(7): p. 681-695.
18. Hettema, J.M., et al., *Catechol-O-methyltransferase contributes to genetic susceptibility shared among anxiety spectrum phenotypes.* Biological Psychiatry, 2008. **64**(4): p. 302-310.
19. Liu, B., A. de la Fuente, and I. Hoeschele, *Gene network inference via structural equation modeling in genetical genomics experiments.* Genetics, 2008. **178**(3): p. 1763-1776.
20. Ray, L.A., et al., *The dopamine D-4 Receptor (DRD4) gene exon III polymorphism, problematic alcohol use and novelty seeking: direct and mediated genetic effects.* Addiction Biology, 2009. **14**(2): p. 238-244.
21. Remington, D.L., *Effects of Genetic and Environmental Factors on Trait Network Predictions From Quantitative Trait Locus Data.* Genetics, 2009. **181**(3): p. 1087-1099.
22. Su, S.Y., et al., *Serotonin Transporter Gene, Depressive Symptoms, and Interleukin-6.* Circulation-Cardiovascular Genetics, 2009. **2**(6): p. 614-U205.
23. Xie, J. and P.M. Bentler, *Covariance Structure Models for Gene Expression Microarray Data.* Structural Equation Modeling: A Multidisciplinary Journal, 2003. **10**(4): p. 566-582.
24. Volkow, N.D., et al., *Expectation Enhances the Regional Brain Metabolic and the Reinforcing Effects of Stimulants in Cocaine Abusers.* The Journal of Neuroscience, 2003. **23**(36): p. 11461-11468.
25. Loehlin, J.C., *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. 2004: Lawrence Erlbaum Associates.
26. Byrne, B.M., *Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming*. Second ed. 2006: Lawrence Erlbaum Associates.
27. Jöreskog, K. and D. Sörbom, *LISREL 8: User's Reference Guide*. 1996: Scientific Software International, Inc.
28. SAS Institute, I. *Base SAS(R) 9.2 Procedures Guide: Statistical Procedures, Partial Correlation*. 2008 [cited 2008; Partial correlation methodology].
29. Jaccard, J. and C.K. Wan, *LISREL Approaches to Interaction Effects in Multiple Regression*. Quantitative Applications in the Social Sciences. 1996: Sage Publications.
30. Coenders, G., J.M. Batista-Foguet, and W.E. Saris, *Simple, Efficient and Distribution-free Approach to Interaction Effects in Complex Structural Equation Models.* Quality & Quantity, 2008. **42**(3): p. 369-396.
31. Kenny, D.A. and C.M. Judd, *Estimating the Nonlinear and Interactive Effects of Latent Variables.* Psychological Bulletin, 1984. **96**(1): p. 201-210.
32. Schumacker, R.E. and G.A. Marcoulides, *Interaction and Nonlinear Effects in Structural Equation Modeling*. 1998: Lawrence Erlbaum Associates.
33. Meredith, W. and J. Tisak, *Latent Curve Analysis.* Psychometrika, 1990. **55**(1): p. 107-122.
34. Bollen, K.A. and P.J. Curran, *Latent Curve Models: A Structural Equation Perspective*. 2006: John Wiley & Sons, Inc.

35. Curran, P.J. and K.A. Bollen, *The Best of Both Worlds: Combining Autoregressive and Latent Curve Models*, in *New Methods for the Analysis of Change*, L. Collins and A. Sayer, Editors. 2001, American Psychological Association: Washington, D.C. p. 107-35.
36. Hox, J., *Multilevel Analysis, Techniques and Applications*. 2002, Mahwah, NJ: Lawrence Erlbaum Associates.
37. Cronbach, L.J. and N. Webb, *Between class and within class effects in a reported aptitude x treatment interaction: a reanalysis of a study by G. L. Anderson.* Journal of Educational Psychology, 1979. **67**: p. 717-724.
38. Muthen, B., *Means and covariance structure analysis of hierarchical data*. UCLA Statistics series. Vol. #62. 1990.
39. Bentler, P.M., *EQS 6 structural equations program manual*. 2005, Encino, CA: Multivariate Software (www.mvsoft.com).
40. Little, T.D., K.U. Schnabel, and J. Baumert, *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples*. 2000: Lawrence Erlbaum Associates.
41. Marcoulides and Schumacker, eds. *Advanced Structural Equation Modeling*. 1996.
42. Byrne, B.M., *Structural Equation Modeling with AMOS*. 2001: Lawrence Erlbaum Associates.
43. Heck, R.H. and S.L. Thomas, *An Introduction to Multilevel Modeling Techniques*. Quantitative Methodology Series, ed. G.A. Marcoulides. 2000, Mahwah, New Jersey: Lawrence Erlbaum Associates.
44. Scientific Software International. *Multilevel structural equation modeling*. 2005-2008 [cited 2008; LISREL implementation of multilevel SEM]. Available from: http://www.ssicentral.com/lisrel/techdocs/Session12.pdf.
45. Crowder, M.J. and D.J. Hand, *Analysis of Repeated Measures*. 1990: Chapman & Hall.
46. Cudeck, R., S. du Toit, and D. Sörbom, eds. *Structural Equation Modeling: Present and Future, A Festschrift in honor of Karl Jöreskog*. 2001, Scientific Software International.
47. Blackwell, E., C.F.M.D. Leon, and G.E. Miller, *Applying Mixed Regression Models to the Analysis of Repeated-Measures Data in Psychosomatic Medicine.* Psychosomatic Medicine, 2006. **68**: p. 870-878.
48. Garson, G.D. *Statnotes: Topics in Multivariate Analysis, Structural Equation Modeling*. 1998, 2008 [cited 2008; Introduction to SEM]. Available from: http://faculty.chass.ncsu.edu/garson/PA765/partialr.htm.
49. Krzanowski, W.J. and F.H.C. Marriott, *Multivariate Analysis Part 2: Classification, covariance structures and repeated measurements*. 1995, New York: Halsted Press.
50. Rovine, M.J. and P.C.M. Molenaar, *A LISREL Model for the Analysis of Repeated Measures With a Patterned Covariance Matrix.* Structural Equation Modeling, 1998. **5**(4): p. 318-343.
51. Vittinghoff, E., et al., *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2005: Springer.

52. Winer, B.J., *Statistical Principles in Experimental Design*. 2nd ed. 1971: McGraw-Hill Book Company.
53. Diggle, P.J., et al., *Analysis of Longitudinal Data*. Second ed. Oxford Statistical Science Series. 2002.
54. Monahan, J.F., *Numerical Methods of Statistics*. 2001: Cambridge University Press.
55. Gill, P.E., W. Murray, and M.H. Wright, *Practical Optimization*. 1981, London: Academic Press, Inc.
56. Boomsma, A., *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. 1983, Amsterdam: Sociometric Research Foundation.
57. Satorra, A. and W.E. Saris, *Power of the likelihood ratio test in covariance structure analysis.* Psychometrika, 1985. **50**: p. 83-90.
58. Saris, W.E., A. Satorra, and D. Sörbom, eds. *The detection and correction of specification errors in structural equation models*. Sociological Methodology 1987, ed. C.C. Clogg. 1987, American Sociological Association: Washington, D. C. pp. 105-130.
59. Muthen, B. *MPLUS*. 2010; Available from: http://www.statmodel.com/power.shtml.
60. Muthén, L.K. and B.O. Muthén, *How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power.* Structural Equation Modeling, 2002. **9**(4): p. 599-620.
61. Stanghellini, E. and N. Wermuth, *On the identification of path analysis models with one hidden variable.* Biometrika, 2005. **92**(2): p. 337-350.
62. Dunn, G., B. Everitt, and A. Pickles, *Modelling Covariances and Latent Variables Using EQS*. 1993: Chapman & Hall.
63. Skrondal, A. and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. 2004: Chapman & Hall/CRC.
64. McDonald, R.P., *A simple comprehensive momdel for the analysis of covariance structures: Some remarks on applications.* British Journal of Mathematical & Statistical Psychology, 1980. **33**: p. 161-83.
65. McDonald, R.P., *A simple comprehensive model for the analysis of covariance structures.* British Journal of Mathematical & Statistical Psychology, 1978. **31**: p. 59-72.
66. McArdle, J.J. and R.P. McDonald, *Some Algebraic Properties of the Reticular Action Model for Moment Structures.* British Journal of Mathematical & Statistical Psychology, 1984. **37**: p. 234-251.
67. Bentler, P.M. and D.G. Weeks, *Multivariate Analysis with Latent Variables*, in *Handbook of Statistics*, P.R. Krishnaiah and L. Kanal, Editors. 1980, North-Holland: Amsterdam.
68. Hoyle, R.H., *Statistical Strategies for Small Sample Research*. 1999: Sage Publications.
69. Jöreskog, K.G. and H. Wold, *The ML and PLS Techniques for Modeling with Latent Variables*, in *Systems under indirect observation: causality, structure, prediction*, K.G. Jöreskog and H. Wold, Editors. 1982.

70. Lohmoller, J.-B., *Latent Variable Path Modeling with Partial Least Squares*. 1989: Physica-Verlag Heidelberg.
71. Wold, H., *Partial Least Squares*, in *Encyclopedia of Statistical Sciences*. 2006, John Wiley & Sons, Inc.
72. Tenenhaus, M., et al., *PLS path modeling*. Computational Statistics & Data Analysis, 2005. **48**: p. 159-205.
73. Hu, B.S., *On building partial least squares models with interdependent inner relations*, in *Systems under indirect observation: Causality, structure, prediction*, K.G. Jöreskog and H. Wold, Editors. 1982. p. 249-272.
74. Wold, H., ed. *The Fix-Point Approach to Interdependent Systems*. 1981, North-Holland Publishing Company.
75. Seber, G.A.F., *Linear Regression Analysis*. 1977: John Wiley & Sons.
76. Chin, W.W., B.L. Marcolin, and P.R. Newsted, *A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion / Adoption Study.* Information Systems Research, 2003. **14**(2): p. 189-217.
77. Littell, R.C., *SAS for Mixed Models*. 2006: SAS Publishing.
78. West, B., et al., *Linear Mixed Models: A Practical Guide Using Statistical Software*. 2006: CRC Press.
79. Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. 1977, Hillsdale, NJ: Erlbaum.
80. Maxwell, S.E., *Sample Size and Multiple Regression Analysis.* Psychological Methods, 2000. **5**(4): p. 434-458.
81. Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*. Sixth ed. 2007: Prentice Hall.
82. Helms, R.W., *Intentionally Incomplete Longitudinal Designs: 1. Methodology and Comparison of Some Full Span Designs.* Statistics in Medicine, 1992. **11**(14-15): p. 1889-1913.
83. Verbeke, G. and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. 2000: Springer.
84. Seber, G.A.F., *Multivariate Observations*. 1984: John Wiley and Sons.
85. Gatignon, H., *Statistical Analysis of Management Data*. 2003: Springer.
86. Sprent, P., *Models in Regression and Related Topics*. Monographs on Statistics and Applied Probability. 1969: Methuen Young Books.
87. Wong, M.Y., *Likelihood Estimation of a Simple Linear Regression Model When Both Variables Have Error.* Biometrika, 1989. **76**: p. 141-148.
88. SAS Institute, I. 2008; Available from: http://support.sas.com/documentation/.
89. Ringle, C.M., S. Wende, and A. Will. *SmartPLS*. 2005; 2.0 (beta):[Available from: http://www.smartpls.de.
90. Wold, H., *PLS for multivariate linear modeling*, in *QSAR: Chemometric methods in molecular design: Methods and principles in medicinal chemistry*, H.v.d. Waterbeemd, Editor. 1994.
91. Carroll, R.J., et al., *Measurement Error in Nonlinear Models, A Modern Perspective*. Second ed. 2006: Chapman & Hall/CRC.
92. Chatterjee, S. and A.S. Hadi, *Regression Analysis by Example*. Fourth ed. Wiley Series in Probability and Statistics. 2006: Wiley-Interscience.

93. Marrelec, G., et al., *Using partial correlation to enhance structural equation modeling of functional MRI data.* Magnetic Resonance Imaging, 2007. **25**: p. 1181-1189.
94. Garson, G.D. *Statnotes: Partial Correlation*. 1996, 2008 [cited 2008; Introduction to Partial Correlation]. Available from: http://faculty.chass.ncsu.edu/garson/PA765/partialr.htm.
95. *Wikipedia, Partial correlation*. [cited 2008 October].
96. Fisher, R.A., *The distribution of the partial correlation coefficient.* Metron, 1924. **3-4**: p. 329-332.
97. Rees, D.G., *Foundations of Statistics*. 1987: CRC Press.
98. Lee, S.-Y., W.-Y. Poon, and P.M. Bentler, *Full Maximum Likelihood Analysis of Structural Equation Models with Polytomous Variables.* Statistics & Probability Letters 1990. **9**: p. 91-97.
99. Lee, S.-Y., W.-Y. Poon, and P.M. Bentler, *A three-stage estimation procedure for structural equation models with polytomous variables.* Psychometrika, 1990. **55**: p. 45-52.
100. Lee, S.-Y., W.-Y. Poon, and P.M. Bentler, *Structural Equation Models with Continuous and Polytomous Variables.* Psychometrika, 1992. **57**(1): p. 89-105.
101. Lee, S.-Y., W.-Y. Poon, and P.M. Bentler, *A two-stage estimation of structural equation models with continous and polytomous variables.* British Journal of Mathematical & Statistical Psychology, 1995. **48**: p. 339-358.
102. Lee, S.-Y. and X.-Y. Song, *Maximum likelihood estimation and model comparison of nonlinear structural equation models with continuous and polytomous variables.* Computational Statistics & Data Analysis, 2003. **44**: p. 125-142.
103. Xie, Y., *Structural Equation Models for Ordinal Variables.* Sociological Methods & Research, 1989. **17**(4): p. 325-352.
104. Browne, M.W., *Asymptotically distribution-free methods for the analysis of covariance structures.* British Journal of Mathematical & Statistical Psychology, 1984. **37**: p. 62-83.