Chapter 1

Introduction to Structural Equation Models

The design of this book is for Chapter 0 to be a self-contained discussion of regression with measurement error, while this chapter introduces the classical structural equation models in their full generality. So, this chapter may serve as a starting point for advanced readers. These advanced readers may belong to two species — quantitatively oriented social scientists who are already familiar with structural equation modeling, and statisticians looking for a quick introduction to the topic at an appropriate level.

Also, readers of Chapter 0 will have noticed that the study of a particular model typically involves a fair amount of symbolic calculation, particularly the calculation of covariance matrices in terms of model parameters. While these calculations often yield valuable insights, they become increasingly burdensome as the number of variables increases, particularly when more than one model must be considered.

The solution is to let a computer do it. So starting with this chapter, many calculations will be illustrated using Sage, an open source computer algebra package described in Appendix B. The Sage parts will be interleaved with the rest of the text rather than fully integrated. Typically, an example will include the result of a calculation without giving a lot of detail, and then at an appropriate place for a pause, the Sage code will be given. This will allow readers who are primarily interested in the ideas to skip material they may find tedious.

1.1 Overview

Structural equation models may be viewed as an extension of multiple regression. They generalize multiple regression in three main ways: there is usually more than one equation, a response variable in one equation can be an explanatory variable in another, and structural equation models can include latent variables.

Multiple equations: Structural equation models are usually based upon more than one regression-like equation. Having more than one equation is not really

unique; multivariate regression already does that. But you will see that structural equation models are more flexible than the usual multivariate linear model.

Variables can be both explanatory and response: This is an attractive feature. Consider a study of arthritis patients, in which joint pain and mobility are measured at several time points. Joint pain at one time period can lead to decreased physical activity during the same period, which then leads to more pain at the next time period. Level of physical activity at time t is both a response variable and a response variable. Structural equation models are also capable of representing the back-and-forth nature of supply and demand in Economics. Many other examples will be given

Latent variables: Structural equation models may include random variables that cannot be directly observed, and also are not error terms. This capability (combined with relative simplicity) is their biggest advantage. It allows the statistican to admit that measurement error exists, and to incorporate it directly into the statistical model. The regression models with latent variables in Chapter 0 are special cases of structural equation models.

There are some ways that structural equation models are different from ordinary linear regression. These include random (rather than fixed) explanatory variable values, a bit of specialized vocabulary, and some modest changes in notation. Tests and confidence intervals are based on large-sample theory, even when normal distributions are assumed. Also, structural equation models have a substantive[1] as well as a statistical compontent; closely associated with this is the use of path diagrams to represent the connections between variables.

To the statistician, perhaps the most curious feature of structural equation models is that usually, the regression-like equations lack intercepts and the expected values of all random variables equal zero. This happens because the models have been re-parameterized in search of parameter identifiability. Details are given in the next section (Section).

Random explanatory variables Chapter 0 discusses the advantages of the traditional regression model in which values of the explanatory variables are treated as fixed constants, and the model is considered to be *conditional* on those values. But once we admit that the variables we observe are contaminated by random measurement error, the virtues of a conditional model mostly disappear. So in the standard structural equation models, all variables are random variables.

Vocabulary Structural equation modeling has developed a specialized vocabulary, and except for the term "latent variable," much of it is not seen elsewhere in Statistics. But the terminology can help clarify things once you know it, and also it appears in software manuals and on computer output. Here are some terms and their definitions.

[1] Substantive means having to do with the subject matter. A good substantive model of water pollution would depend on concepts from Chemistry and Hydrodynamics.

- **Latent variable**: A random variable that cannot be directly observed, and also is not an error term.
- **Manifest variable**: An observable variable. An actual data set contains only values of the manifest variables. This book will mostly use the term "observable."
- **Exogenous variable**: In the regression-like equations of a structural equation model, the exogenous variabes are ones that appear *only* on the right side of the equals sign, and never on the left side in any equation. If you think of Y being a function of X, this is one way to remember the meaning of **ex**ogenous. All error terms are exogenous variables.
- **Endogenous variable**: Endogenous variables are those that appear on the left side of at least one equals sign. Endogenous variables depend on the exogenous varables, and possibly other endogenous variables. Think of an arrow from an exogenous variable to an endogenous variable. The **end** of the arrow is pointing at the **end**ogenous variable.
- **Factor**: This term has a meaning that actually conflicts with its meaning in mainstream Statistics, particularly in experimental design. Factor analysis (not "factorial" analysis of variance!) is a set of statistical concepts and methods that grew up in Psychology. Factor analysis models are special cases of the general structural equation model. A *factor* is an underlying trait or characteristic that cannot be measured directly, like intelligence. It is a latent variable, period.

Notation Several different but overlapping models and accompanying notation systems are to be found in the many books and articles on structural equation modeling. The present book introduces a sort of hybrid notation system, in which the symbols for parameters are mosly taken from the structural equation modeling literature, while the symbols for random variables are based on common statistical usage. This is to make it easier for statisticians to follow. The biggest change from Chapter 0 is that the symbol $\boldsymbol{\beta}$ is no longer used for just any regression coefficient. It is reserved for links between latent endgenous variables and other latent endgenous variables.

1.2 A general two-stage model

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} \mathbf{Y}_i &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\ \mathbf{D}_i &= \boldsymbol{\nu} + \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i, \end{aligned} \tag{1.1}$$

where

- $\mathbf{Y}_i$ is a $q \times 1$ random vector.
- $\boldsymbol{\alpha}$ is a $q \times 1$ vector of constants.
- $\boldsymbol{\beta}$ is a $q \times q$ matrix of constants with zeros on the main diagonal.
- $\boldsymbol{\Gamma}$ is a $q \times p$ matrix of constants.
- $\mathbf{X}_i$ is a $p \times 1$ random vector with expected value $\boldsymbol{\mu}_x$ and positive definite covariance matrix $\boldsymbol{\Phi}_x$.
- $\boldsymbol{\epsilon}_i$ is a $q \times 1$ random vector with expected value zero and positive definite covariance matrix $\boldsymbol{\Psi}$.
- $\mathbf{F}_i$ (F for Factor) is a partitioned vector with $\mathbf{X}_i$ stacked on top of $\mathbf{Y}_i$. It is a $(p+q) \times 1$ random vector whose expected value is denoted by $\boldsymbol{\mu}_F$, and whose variance-covariance matrix is denoted by $\boldsymbol{\Phi}$.
- $\mathbf{D}_i$ is a $k \times 1$ random vector. The expected value of $\mathbf{D}_i$ will be denoted by $\boldsymbol{\mu}$, and the covariance matrix of $\mathbf{D}_i$ will be denoted by $\boldsymbol{\Sigma}$.
- $\boldsymbol{\nu}$ is a $k \times 1$ vector of constants.
- $\boldsymbol{\Lambda}$ is a $k \times (p+q)$ matrix of constants.
- $\mathbf{e}_i$ is a $k \times 1$ random vector with expected value zero and covariance matrix $\boldsymbol{\Omega}$.
- $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are independent.

Only $\mathbf{D}_1, \ldots, \mathbf{D}_n$ are observable. All the other random vectors are latent. But because $\boldsymbol{\Omega} = cov(\mathbf{e}_i)$ need not be strictly positive definite, error variances of zero are permitted. This way, it is possible for a variable to be both exogenous and observable.

The distributions of $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are either assumed to be independent and multivariate normal, or independent and unknown. When the distributions are normal, the parameter vector $\boldsymbol{\theta}$ consists of the unique elements of the parmeter matrices $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\mu}_x$, $\boldsymbol{\Phi}_x$, $\boldsymbol{\Psi}$, $\boldsymbol{\nu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$. When the distributions are unknown, the parameter vector also includes the three unknown probability distributions.

The two parts of Model (1.1) are called the *Latent Variable Model* and the *Measurement Model*. The latent variable part is $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$, and the measurement part is $\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$. The bridge between the two parts is the process of collecting the latent exogenous vector $\mathbf{X}_i$ and the latent endogenous vector $\mathbf{Y}_i$ into a "factor" $\mathbf{F}_i$. This is *not* a categorical explanatory variable, the usual meaning of factor in experimental design. The terminology comes from *factor analysis*, a popular multivariate method in the social sciences[2].

[2]Harman's (1967) authoritative classic *Modern factor analysis* [6] is almost guaranteed to be frustrating for a statistician to read. Lawley and Maxwell's (1971) *Factor analysis as a statistical method* is a welcome antidote. Bastlevsky's (1994) *Statistical factor analysis and related methods* [1] is a strong and more recent treatment of the topic.

Example: The Brand Awareness study A major Canadian coffee shop chain is trying to break into the U.S. Market. They assess the following variables twice on a random sample of coffee-drinking adults. Each variable is measured first in an in-person interview, and then in a telephone call-back several days later, conducted by a different interviewer. Thus, errors of measurement for the two measurements of each variable are assumed to be independent. The variables are

- **Brand Awareness** (X_1): Familiarity with the coffee shop chain
- **Advertising Awareness** (X_2): Recall for advertising of the coffee shop chain
- **Interest in the product category** (X_3): Mostly this was how much they say they like coffee and doughnuts.
- **Purchase Intention** (Y_1): Expressed willingness to go to an outlet of the coffeeshop chain and make an order.
- **Purchase behaviour** (Y_2): Reported dollars spent at the chain during the 2 months following the interview.

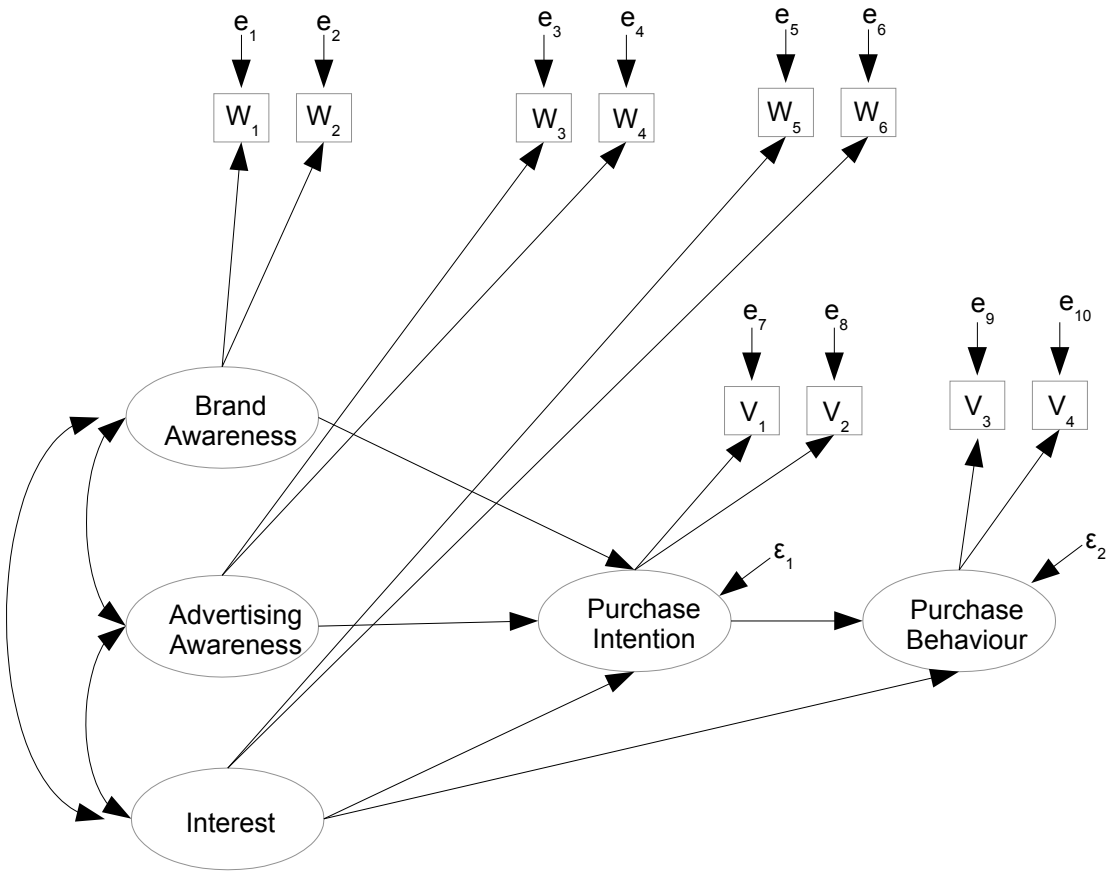
All variables were measured on a scale from 0 to 100 except purchase behaviour, which is in dollars.

Figure 1.1 shows a path diagram for these data. It is a picture of how some variables are thought to influence other variables. The notation is standard. Straight arrows go from exogenous variables to endogenous variables, and possibly from endogenous variables to other endogenous variables. Correlations among exogenous variables are represented by two-headed curved arrows. Observable variables are enclosed by rectangles or squares, while latent variables are enclosed by ellipses or circles. Error terms are not enclosed by anything.

The path diagram in Figure 1.1 expresses some very definite assertions about consumer behaviour. For example, it says that brand awareness and advertising awareness affect actual purchase only through purchase intention, while interest in the product may have a direct effect on purchase behaviour, as well as an indirect effect through purchase intention — perhaps reflecting impulse purchases. Such claims may be right or they may be wrong, and some are testable. But the point is that the statistical model corresponding to the typical path diagram has a strong subject matter component, and actually is a sort of hybrid, occupying a position somewhere between the typical statistical model and an actual theory about the data.

It is always possible to argue about how the path diagram should look, and it is usually valuable as well. The more subject matter expertise that can be brought to the discussion, the better. Often, the contest between two or more competing pictures will be traceable to unresolved theoretical issues in the field. Will the data at hand allow a formal statistical test to decide between the models? If not, is it possible to design a study that will make such a comparison possible? Thus, the more technical statistical expertise that can be brought to the discussion, the better.

Figure 1.1: The Brand Awareness Study



The measurement model — that is, the part relating the latent variables to the observable variables — should not escape scrutiny. The processes it represents are usually not

It is usually not the reason the data were collected, but

Continuing with the Brand Awareness example, the model corresponding to Figure 1.1 may be written in scalar form as a system of simultaneous regression-like equations. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned}
Y_{i,1} &= \alpha_1 + \gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \gamma_3 X_{i,3} + \epsilon_{i,1} \\
Y_{i,2} &= \alpha_2 + \beta Y_{i,1} + \gamma_4 X_{i,3} + \epsilon_{i,2} \\
W_{i,1} &= \nu_1 + \lambda_1 X_{i,1} + e_{i,1} \\
W_{i,2} &= \nu_2 + \lambda_2 X_{i,1} + e_{i,2} \\
W_{i,3} &= \nu_3 + \lambda_3 X_{i,2} + e_{i,3} \\
W_{i,4} &= \nu_4 + \lambda_4 X_{i,2} + e_{i,4} \\
W_{i,5} &= \nu_5 + \lambda_5 X_{i,3} + e_{i,5} \\
W_{i,6} &= \nu_6 + \lambda_6 X_{i,3} + e_{i,6} \\
V_{i,1} &= \nu_7 + \lambda_7 Y_{i,1} + e_{i,7} \\
V_{i,2} &= \nu_8 + \lambda_8 Y_{i,1} + e_{i,8} \\
V_{i,3} &= \nu_9 + \lambda_9 Y_{i,2} + e_{i,9} \\
V_{i,4} &= \nu_{10} + \lambda_{10} Y_{i,2} + e_{i,10},
\end{aligned} \tag{1.2}$$

where $E(X_{i,1} = \mu_{x1})$, $E(X_{i,2} = \mu_{x2})$, $E(X_{i,3} = \mu_{x3})$, the expected values of all error terms equal zero, $Var(X_{i,j}) = \phi_{jj}$ for $j = 1, 2, 3$, $Cov(X_{i,j}, X_{i,k}) = \phi_{jk}$, $Var(e_{i,j}) = \omega_j$ for $j = 1, \ldots, 10$, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, and all the error terms are independent of one another and of the $X_{i,j}$ variables.

If the two measurements of each variable were deemed similar enough, it would be possible to reduce the parameter space quite a bit, for example setting $\nu_1 = \nu_2$, $\lambda_1 = \lambda_2$, and $\omega_1 = \omega_2$. The same kind of thing could be done for the other latent variables. Also, the distributions could be assumed normal, or they could be left unspecified; in practice, those are the two choices.

Setting up the problem in matrix form, we have $p = 3$ latent exogenous variables, $q = 2$ latent endogenous variables, and $k = 10$ observable variables, all of which are endogenous in this example. Using parameter symbols from the scalar version, the equations of the latent variable model are

$$\begin{array}{ccccccccccc}
\mathbf{Y}_i & = & \boldsymbol{\alpha} & + & \boldsymbol{\beta} & \mathbf{Y}_i & + & \boldsymbol{\Gamma} & \mathbf{X}_i & + & \boldsymbol{\epsilon}_i \\
\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & = & \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} & + & \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} & \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & + & \begin{pmatrix} \gamma_1 & \gamma_2 & \gamma_3 \\ 0 & 0 & \gamma_4 \end{pmatrix} & \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} & + & \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}
\end{array}$$

with

$$\boldsymbol{\Phi}_x = cov(\mathbf{X}_i) = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{pmatrix} \text{ and } \boldsymbol{\Psi} = cov(\boldsymbol{\epsilon}_i) = \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}.$$

Collecting $\mathbf{X}_i$ and $\mathbf{Y}_i$ into a single vector of "factors,"

$$\mathbf{F}_i = \begin{pmatrix} \mathbf{X}_i \\ \hline \mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \\ Y_{i,1} \\ Y_{i,2} \end{pmatrix}.$$

Finally, the equations of the measurement model are

$$\begin{array}{ccccccccc} \mathbf{D}_i & = & \boldsymbol{\nu} & + & & \boldsymbol{\Lambda} & & \mathbf{F}_i & + & \mathbf{e}_i \\ \begin{pmatrix} W_{i,1} \\ W_{i,2} \\ W_{i,3} \\ W_{i,4} \\ W_{i,5} \\ W_{i,6} \\ V_{i,1} \\ V_{i,2} \\ V_{i,3} \\ V_{i,4} \end{pmatrix} & = & \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \\ \nu_{10} \end{pmatrix} & + & \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & \lambda_3 & 0 & 0 & 0 \\ 0 & \lambda_4 & 0 & 0 & 0 \\ 0 & 0 & \lambda_5 & 0 & 0 \\ 0 & 0 & \lambda_6 & 0 & 0 \\ 0 & 0 & 0 & \lambda_7 & 0 \\ 0 & 0 & 0 & \lambda_8 & 0 \\ 0 & 0 & 0 & 0 & \lambda_9 \\ 0 & 0 & 0 & 0 & \lambda_{10} \end{pmatrix} & \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \\ Y_{i,1} \\ Y_{i,2} \end{pmatrix} & + & \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \\ e_{i,4} \\ e_{i,5} \\ e_{i,6} \\ e_{i,7} \\ e_{i,8} \\ e_{i,9} \\ e_{i,10} \end{pmatrix} \end{array}$$

with

$$\boldsymbol{\Omega} = cov(\mathbf{e}_i) = \begin{pmatrix} \omega_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \omega_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \omega_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \omega_6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_{10} \end{pmatrix}$$

Given a verbal description of a data set, the student should be able to write down a path diagram, and translate freely between the path diagram, the model in scalar form and the model in matrix form. Three three ways of expressing the model are equivalent, and some software[3] will allow a model to be specified using only a built-in drawing program. This can be appealing to users who don't like equations and Greek letters, but for larger models the process can be very tedious.

1.3 Review of identifiability

The general two-stage model (1.1) of Section 1.2 is very general indeed — so much so, that its parameters are seldom identifiable without additional restrictions. Choosing

[3]The ones I know of are Amos and JMP.

these restrictions wisely is an essential part of structural equation modeling. In fact, it turns out that almost everything that makes structural equation modeling distinct from other large-sample statistical methods can be traced to issue of parameter identifiability. For the convenience of readers starting with Chapter 1, this section collects material on identifiability from Chapter 0. Readers of Chapter 0 are also encouraged to look it over. The presentation is intended to be terse. For more detail, please see Chapter 0.

Definition 0.10.2 (Page 50) Suppose a statistical model implies $\mathbf{D} \sim P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$. If no two points in Θ yield the same probability distribution, then the parameter $\boldsymbol{\theta}$ is said to be *identifiable.* On the other hand, if there exist $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in Θ with $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, the parameter $\boldsymbol{\theta}$ is *not identifiable.*

Theorem 1 (Page 51) If the parameter vector is not identifiable, consistent estimation for all points in the parameter space is impossible.

Definition 0.10.3 (Page 51)The parameter is said to be *identifiable* at a point $\boldsymbol{\theta}_0$ if no other point in Θ yields the same probability distribution as $\boldsymbol{\theta}_0$.

Definition 0.10.4 (Page 51) The parameter is said to be *locally identifiable* at a point $\boldsymbol{\theta}_0$ if there is a neighbourhood of points surrounding $\boldsymbol{\theta}_0$, none of which yields the same probability distribution as $\boldsymbol{\theta}_0$.

Definition 0.10.5 (Page 52) Let $g(\boldsymbol{\theta})$ be a function of the parameter vector. If $g(\boldsymbol{\theta}_0) \neq g(\boldsymbol{\theta})$ implies $P_{\boldsymbol{\theta}_0} \neq P_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \Theta$, then the function $g(\boldsymbol{\theta})$ is said to be identifiable at the point $\boldsymbol{\theta}_0$.

Theorem 2 (Page 52) Let

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_p) \\ y_2 &= f_2(x_1, \ldots, x_p) \\ \vdots & \qquad \vdots \\ y_q &= f_q(x_1, \ldots, x_p), \end{aligned}$$

If the functions $f_1, \ldots, f_q$ are analytic (posessing a Taylor expansion) and $p > q$, the set of points $(x_1, \ldots, x_p)$ where the system of equations has a unique solution occupies at most a set of volume zero in $\mathbb{R}^p$.

Moment structure equations give moments of the distribution of the observable data in terms of model parameters. In this course, moments are limited to expected values, variances and covariances. If it is possible to solve uniquely for the parameter vector in terms of the these quantities, then the parameter vector is identifiable. Even when a multivariate normal distribution is not assumed, in practice "identifiable" means identifiable from the moments — usually the variances and covariances.

Figure 1.2: A sequence of re-parameterizations

$$\text{Truth} \approx \text{Original Model} \rightarrow \text{Surrogate Model 1} \rightarrow \text{Surrogate Model 2} \rightarrow \ldots$$

Rule 1 (The Parameter Count Rule, page 53) Suppose identifiability is to be decided based on a set of moment structure equations. If there are more parameters than equations, the parameter vector is identifiable on at most a set of volume zero in the parameter space.

1.4 Models: Original and Surrogate

Overview

It is taken for granted that even the best scientific models are not "true" in any ultimate sense. At best, they are approximations of how nature really works. And this is even more true of statistical models. As Box and Draper (1987, p. 424) put it, "Essentially all models are wrong, but some are useful." [4] But in structural equation modeling, the models used in practice are usually not even the approximate versions that the scientist or statistician has in mind. Instead, they are re-parameterized versions of the intended models. This explains some features that may seem odd at first.

Figure 1.2 is a picture of the process[4]. Underlying everything is the true state of nature, the real process that gave rise to the observable data in our possession. We can scarcely even imagine what it is, but undoubtedly it's non-linear, and involves a great many unmeasured variables. So we start with a model based on the general two-stage model (original2stage) of Section 1.2. It is not the truth and we know it's not the truth, but we still hope it roughly approximates the truth in a useful way, at least within the range of the observed data.

As primitive as the original model is compared to the real truth, its parameters are still not identifiable. So we re-parameterize, producing a new model whose parameters are *functions* of the parameters of the original model. Such a model will be called a *surrogate* model because it stands for the original model, and tries to do the job of the original model. Like a surrogate mother, it may not be as good as a the real thing, but it will have to do.

As indicated in Figure 1.2, re-parameterization may happen in more than one step. For the classical structural equation models presented in this book, the first re-parameterization results in a *centered* surrogate model with no intercepts, and all expected values equal to zero. This was described in Chapter 0 for regression with measurement error.

But typically, the parameters of the centered surrogate model are still not identifiable, and there is another re-parameterization, leading to a second level surrogate model. The

[4]Thanks to Michael Li for this way of expressing the idea.

process can continue. At each step, the parameter vector of the new model is a function of of the parameters of the preceding model, and typically the function is not one-to-one. Otherwise, identifiability would not change. At each stage, the dimension of the new parameter space is less, so the re-parameterization represents a restriction, or collapsing of the original parameter space. The end result is a model whose parameters are identifiable functions of the original parameter vector. The goal is for those functions to be as informative as possible about the parameters of the original model.

Two features of the original model deserve special mention. The first is that usually, the original model is already a restricted version of Model (original2stage), even before it is re-parameterized to produce a surrogate model. The restrictions in question arise from substantive modeling considerations rather than from a search for identifiability. So, in the Brand Awareness example of Section 1.2, the parameter matrices have many elements fixed at zero. These represent theoretical assertions about consumer psychology. They may be helpful in making the remaining free parameters identifiable, but that is not their justification.

A second notable feature of the original model is that expected values are non-zero in general, and all the equations are regression-like equations with intercepts, and with slopes that do not necessarily equal one. Any deviation from this standard needs to be justified on substantive grounds, not on grounds of simplicity or convenience. Otherwise, it's a surrogate model and not an original model. The distinction is important, because most structural equation models used in practice are surrogate models, and a good way to understand them is to trace the connection between their parameters and the parameters of the original models from which they are are derived.

Consider a simple additive model for measurement error, like (24) on page 34:

$$W = X + e.$$

Immediately it is revealed as a surrogate model, because there is no intercept and the slope is set to one – a choice that would be hard to justify on modeling grounds most of the time. For example, X might be actual calories consumed during the past week, and W might be number of reported calories based on answers to a questionnaire. Undoubtedly, the true relationship between these variables is non-linear. In an original (though not exactly true) model, the relationship would be approximated by

$$W = \nu + \lambda X + e.$$

With this example in mind, it is clear that most of the models given in Chapter 0 (and *all* the models in Chapter 0 with identifiable parameters) are actually surrogate models. This might be a bit unsettling because you did not realize that you were being tricked, or it might be reassuring because some models that struck you as unrealistic may actually be better than they seem.

The centered surrogate model

The first stage of re-parameterization may be done in full generality. The argument begins with a demonstration that the means and intercepts of the original model are not

identifiable. Please bear in mind that as a practical consideration, "identifiable" means identifiable from the moments – the expected values and variance-covariance matrix of the observable data.

Starting with the latent variable part of the two-stage original model (1.1), it is helpful to write the endogenous variables solely as functions of the exogenous variables, and not of each other. Notice how the subscript i has been dropped from the random vectors to reduce notational clutter. This is typical in the structural equation model literature.

$$\begin{aligned}
& \mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Leftrightarrow \quad & \mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Leftrightarrow \quad & \mathbf{I}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Leftrightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Leftrightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})^{-1}(\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}) \\
\Leftrightarrow \quad & \mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}) \qquad (1.3)
\end{aligned}$$

The preceding calculation assumes that the matrix $\mathbf{I} - \boldsymbol{\beta}$ has an inverse. Surprisingly, the existence of $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ is guaranteed by the model. The proof hinges on the specifications that $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are independent, and that $\boldsymbol{\Psi} = cov(\boldsymbol{\epsilon})$ is positive definite.

Theorem 3 *Model (1.1) implies the existence of* $(\mathbf{I} - \boldsymbol{\beta})^{-1}$.

Proof $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}$ yields $(\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}$. Suppose $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ does not exist. Then the rows of $\mathbf{I} - \boldsymbol{\beta}$ are linearly dependent, and there is a $q \times 1$ non-zero vector of constants $\mathbf{a}$ with $\mathbf{a}^\top(\mathbf{I} - \boldsymbol{\beta}) = 0$. So,

$$\begin{aligned}
0 &= \mathbf{a}^\top(\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \mathbf{a}^\top\boldsymbol{\alpha} + \mathbf{a}^\top\boldsymbol{\Gamma}\mathbf{X} + \mathbf{a}^\top\boldsymbol{\epsilon} \\
\Rightarrow Var(0) &= Var(\mathbf{a}^\top\boldsymbol{\Gamma}\mathbf{X}) + Var(\mathbf{a}^\top\boldsymbol{\epsilon}) \\
\Rightarrow 0 &= \mathbf{a}^\top\boldsymbol{\Gamma}\boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top\mathbf{a} + \mathbf{a}^\top\boldsymbol{\Psi}\mathbf{a}.
\end{aligned}$$

But the quantity on the right side is strictly positive, because $\boldsymbol{\Psi}$ is positive definite. Thus, the assumption that $\mathbf{I} - \boldsymbol{\beta}$ is singular leads to a contradiction. This shows that $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ must exist if the model holds. ■

Sometimes, the surface defined by $|\mathbf{I} - \boldsymbol{\beta}| = 0$ is interior to the parameter space, and yet cannot belong to the parameter space because of the other model specifications. Thus it forms an unexpected hole in the parameter space. The pinwheel Model () on page whatever provides an example.

Now that the existence of $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ is established, Expression (1.3) may be used to calculate expected values, variances and covariances. Expressing the results of routine calculations as partitioned matrices,

$$
\begin{aligned}
\boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F &= E(\mathbf{F}_i) = \begin{pmatrix} E(\mathbf{X}_i) \\ \hline E(\mathbf{Y}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \hline (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\alpha}+\boldsymbol{\Gamma}\boldsymbol{\mu}_x) \end{pmatrix} \qquad (1.4)\\
\boldsymbol{\mu} &= E(\mathbf{D}_i) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F \\
\boldsymbol{\Phi} &= cov(\mathbf{F}_i) = \left(\begin{array}{c|c} cov(\mathbf{X}_i) & cov(\mathbf{X}_i,\mathbf{Y}_i) \\ \hline & cov(\mathbf{Y}_i) \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Phi}_x & \boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top(\mathbf{I}-\boldsymbol{\beta})^{-1\,T} \\ \hline (\mathbf{I}-\boldsymbol{\beta})^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top+\boldsymbol{\Psi}\right)(\mathbf{I}-\boldsymbol{\beta})^{-1\,T} \end{array}\right) \\
\boldsymbol{\Sigma} &= cov(\mathbf{D}_i) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Omega}
\end{aligned}
$$

Now the parameter matrices may be divided into three categories: those appearing only in $\boldsymbol{\mu} = E(\mathbf{D}_i)$, those appearing only in $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$, and those appearing in both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Appearing only in $\boldsymbol{\mu}$	$\boldsymbol{\mu}_x, \boldsymbol{\alpha}, \boldsymbol{\nu}$
Appearing only in $\boldsymbol{\Sigma}$	$\boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}$
Appearing in both	$\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}$

Clearly, the parameters appearing only in $\boldsymbol{\mu}$ must be identified from the k mean structure equations or not at all. But even assuming the best case scenario in which $\boldsymbol{\beta}, \boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ can be identified from $\boldsymbol{\Sigma}$ and thus may be considered known, this requires the solution of k equations in $k+p+q$ unknowns. Since the equations are linear, there is no need to invoke the Parameter Count Rule[5]. For every fixed set of $(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$ values, infinitely many sets $(\boldsymbol{\mu}_x, \boldsymbol{\alpha}, \boldsymbol{\nu})$ yield the same vector of expected values $\boldsymbol{\mu}$. Thus, the means and intercepts in the model are not identifiable.

Not much is lost, because usually the matrices $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are of primary interest, and these (or useful functions of them) may potentially be recovered from $\boldsymbol{\Sigma}$. So the standard solution is to re-parameterize, replacing the parameter set $(\boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\mu}_x, \boldsymbol{\alpha}, \boldsymbol{\nu})$ with $(\boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\kappa})$, where $\boldsymbol{\kappa} = \boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F$. Then $\boldsymbol{\kappa}$ is treated as a nuisance parameter to be estimated with the vector of sample means where technically necessary, but otherwise ignored.

A useful way to express the re-parameterization is to re-write the equations of Model (1.1), centering all the random vectors. Starting with the latent variable part,

$$
\begin{aligned}
\mathbf{Y}_i &= (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\alpha}+\boldsymbol{\Gamma}\mathbf{X}_i+\boldsymbol{\epsilon}_i) \\
&= (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\alpha}+\boldsymbol{\Gamma}\mathbf{X}_i-\boldsymbol{\Gamma}\boldsymbol{\mu}_x+\boldsymbol{\Gamma}\boldsymbol{\mu}_x+\boldsymbol{\epsilon}_i) \\
\Leftrightarrow\quad \mathbf{Y}_i-(\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\alpha}+\boldsymbol{\Gamma}\boldsymbol{\mu}_x) &= (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\Gamma}(\mathbf{X}_i-\boldsymbol{\mu}_x)+\boldsymbol{\epsilon}_i) \\
\Leftrightarrow\quad \overset{c}{\mathbf{Y}}_i &= (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\Gamma}\overset{c}{\mathbf{X}}_i+\boldsymbol{\epsilon}_i) \\
\Leftrightarrow\quad (\mathbf{I}-\boldsymbol{\beta})\overset{c}{\mathbf{Y}}_i &= \boldsymbol{\Gamma}\overset{c}{\mathbf{X}}_i+\boldsymbol{\epsilon}_i \\
\Leftrightarrow\quad \overset{c}{\mathbf{Y}}_i &= \boldsymbol{\beta}\overset{c}{\mathbf{Y}}_i+\boldsymbol{\Gamma}\overset{c}{\mathbf{X}}_i+\boldsymbol{\epsilon}_i,
\end{aligned}
$$

[5] A system of linear equations with more unknowns than equations has either infinitely many solutions or none at all. The option of no solutions is ruled out because the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is actually the image of one particular set of parameter matrices in the parameter space. More details about mappings between the parameter space and the moment space are given in Chapter 4.

where putting a c above a random vector means it has been centered by subtracting off its expected value. Automatically we have

$$\overset{c}{\mathbf{F}}_i = \mathbf{F}_i - \boldsymbol{\mu}_F = \begin{pmatrix} \overset{c}{\mathbf{X}}_i \\ \hline \overset{c}{\mathbf{Y}}_i \end{pmatrix}.$$

For the measurement part of the model,

$$\begin{array}{rrcl} & \mathbf{D}_i & = & \boldsymbol{\nu} + \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i \\ & & = & \boldsymbol{\nu} + \boldsymbol{\Lambda}\mathbf{F}_i - \boldsymbol{\Lambda}\boldsymbol{\mu}_F + \boldsymbol{\Lambda}\boldsymbol{\mu}_F + \mathbf{e}_i \\ \Leftrightarrow & \mathbf{D}_i - (\boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F) & = & \boldsymbol{\Lambda}(\mathbf{F}_i - \boldsymbol{\mu}_F) + \mathbf{e}_i \\ \Leftrightarrow & \overset{c}{\mathbf{D}}_i & = & \boldsymbol{\Lambda}\overset{c}{\mathbf{F}}_i + \mathbf{e}_i. \end{array}$$

Thus, a centered version of Model (1.1) is 100% equivalent to the original. A *surrogate* for Model (1.1) is obtained by simply dropping the letter c over the random vectors, and writing

$$\begin{array}{rcl} \mathbf{Y}_i & = & \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i & = & \begin{pmatrix} \mathbf{X}_i \\ \hline \mathbf{Y}_i \end{pmatrix} \\ \mathbf{D}_i & = & \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i, \end{array} \tag{1.5}$$

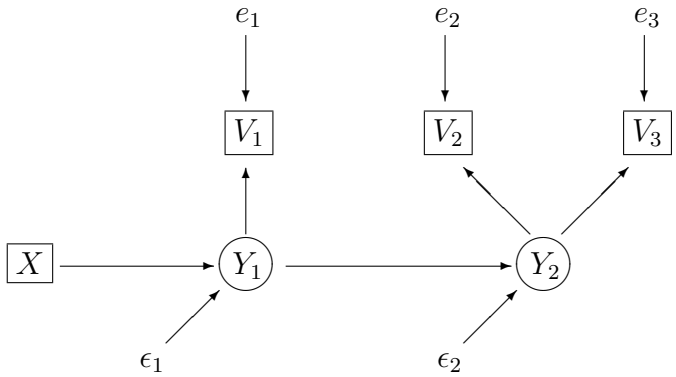
where $E(\mathbf{X}_i) = 0$, and all other specifications are as in Model (1.1). This will be called the *Centered Surrogate Model*. It is a good substitute for the original because

- It hides the nuisance parameters $\boldsymbol{\mu}_x$, $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$, which can't be identified anyway, and are essentially discarded by a re-parameterization.
- The remaining parameter matrices are identical to those of the original model.
- The covariance matrix $\boldsymbol{\Sigma}$ of the observable data (given in Expression 1.4) is identical to that of the original model.
- Special cases of $\boldsymbol{\Sigma}$ that are used in applications easier to calculate.

It must be emphasized that (1.5) is not a realistic model for almost any actual data set, because most variables don't have zero expected value[6]. Rather, it's a substitute for a re-parameterized version of the original Model (1.1), one that's more convenient to work with. This explains why structural equation models are usually written in centered form, with zero means and no intercepts, and why some structural equation modeling software does not even allow for models with means and intercepts.

[6]Some authors suggest that the observable data have been centered by subtracting off *sample* means, so that they do have expected value zero. That would explain why $\boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F = 0$, but not why $\boldsymbol{\mu}_F$ is necessarily equal to zero.

Figure 1.3: Blood pressure path model



An additional re-parameterization

In general, the parameters of the centered surrogate model are still not identifiable. In particular cases, even after restricting the parameters based on modeling considerations, further technical restrictions are necessary to obtain a model whose parameters are identifiable. Like centering, these restrictions should be viewed as re-parameterizations, and the models that result should be viewed as surrogates for the original model. But unlike centering, which does not affect the parameters appearing in the covariance matrix, the second level of re-parameterization affects the *meaning* of the remaining parameters in an important way. General principles will be developed in later chapters, but here is a simple example to illustrate the idea.

Example 1.4.1 *Blood Pressure*

Patients with high blood pressure are randomly assigned to different dosages of a blood pressure medication. There are many different dosages, so dosage may be treated as a continuous variable. Because the exact dosage is known, this exogenous variable is observed without error. After one month of taking the medication, the level of the drug in the patient's bloodstream is measured once (with error, of course), by an independent lab. Then, two measurements of the patient's blood pressure are taken in the doctor's office. The measurements are taken on different days and by different technicians, but with exacly the same equipment and following exactly the same measurement protocol. Thus, the two blood pressure readings are thought to be equivalent as well as having independent measurement errors.

Figure 1.3 shows a path diagram of the model, with X representing drug dosage, Y_1 representing true blood level of the drug, and Y_2 representing the patient's average resting blood pressure.

The original model for this problem may be written in scalar form as follows. Inde-

pendently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_{i,1} &= \alpha_1 + \gamma X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta Y_{i,1} + \epsilon_{i,2} \\ V_{i,1} &= \nu_1 + \lambda_1 Y_{i,1} + e_{i,1} \\ V_{i,2} &= \nu_2 + \lambda_2 Y_{i,2} + e_{i,2} \\ V_{i,3} &= \nu_2 + \lambda_2 Y_{i,2} + e_{i,3}, \end{aligned} \tag{1.6}$$

where $E(X_i) = \mu_x$, $Var(X_i) = \phi$, all error terms are independent with expected values equal to zero, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, $Var(e_{i,1}) = \omega_1$, and $Var(e_{i,2}) = Var(e_{i,3}) = \omega_2$. The equal intercepts, slopes and intercepts for V_2 and V_3 are modeling restrictions, based on the belief that V_2 and V_3 really are equivalent measurements.

In a typical application, this model would be presented in centered form, with the coefficients λ_1 and λ_2 both set equal to one, perhaps with a reference to "setting the scales" of the latent variables[7]. Here is a more detailed account of what is going on.

The first step is to re-parameterize by a change of variables in which each variable is transformed by subtracting off its expected value, and then any notational evidence if the transformation is suppressed. The result is a centered surrogate model like (1.5). Before further re-parameterization, let us verify that the parameters of the centered model are not identifiable. It passes the test of the Parameter Count Rule, because the covariance matrix contains ten parameters and has ten unique elements. So there are ten covariance structure equations in ten unknowns.

The covariance matrix $\mathbf{\Sigma} = [\sigma_{ij}]$ of the observable variables $\mathbf{D}_i = (X_i, V_{i,1}, V_{i,2}, V_{i,3})^\top$ is

$$\begin{pmatrix} \phi & \gamma\lambda_1\phi & \beta\gamma\lambda_2\phi & \beta\gamma\lambda_2\phi \\ & (\gamma^2\phi + \psi_1)\lambda_1^2 + \omega_1 & (\gamma^2\phi + \psi_1)\beta\lambda_1\lambda_2 & (\gamma^2\phi + \psi_1)\beta\lambda_1\lambda_2 \\ & & (\beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2)\lambda_2^2 + \omega_2 & (\beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2)\lambda_2^2 \\ & & & (\beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2)\lambda_2^2 + \omega_2 \end{pmatrix}. \tag{1.7}$$

The model imposes three three equality constraints on the covariance matrix: $\sigma_{13} = \sigma_{14}$, $\sigma_{23} = \sigma_{24}$ and $\sigma_{33} = \sigma_{34}$. This effectively reduces the number of covariance structure equations by three, so that to show identifiability it would be necessary to solve seven equations in ten unknowns[8]. By the parameter count rule, a unique solution is impossible except possibly on a set of volume zero in the parameter space. So the parameter vector is not identifiable.

If this argument is not entirely convincing, the table below gives a numerical example of two different parameter vectors (with γ, β, λ_1 and λ_2 all non-zero) that yield the same covariance matrix.

[7] See for example Bollen, get reference from language paper.

[8] This idea is a bit subtle. The σ_{ij} quantities should be viewed as images of a *single, fixed* point $\boldsymbol{\theta}_0$ in the parameter space. So if the model implies $\sigma_{13} = \sigma_{14}$ because they both equal $\beta\gamma\lambda_2\phi$, it means that σ_{13} and σ_{14} both represent the same real number. At this point, parameter symbols like β and γ represent fixed constants too, because they are elements of $\boldsymbol{\theta}_0$. But then when the attempt is made to recover $\boldsymbol{\theta}_0$ from $\mathbf{\Sigma}(\boldsymbol{\theta}_0)$ by solving equations, parameter symbols like β and γ are treated as variables, while the σ_{ij} quantities remain fixed constants. Chapter 4 discusses mappings back and forth between the parameter space and the moment space.

	γ	β	λ_1	λ_2	ψ_1	ψ_2	ϕ	ω_1	ω_2
$\boldsymbol{\theta}_1$	2	4	1	1	4	16	1	1	1
$\boldsymbol{\theta}_2$	1	2	2	4	1	1	1	1	1

Both parameter vectors yield the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 2 & 8 & 8 \\ 2 & 9 & 32 & 32 \\ 8 & 32 & 145 & 144 \\ 8 & 32 & 144 & 145 \end{pmatrix}.$$

By Definition 0.10.2, the parameter vector is not identifiable.

The next step is to re-examine the model equations in (surrogate) centered form,

$$\begin{aligned} Y_{i,1} &= \gamma X_i + \epsilon_{i,1} \\ Y_{i,2} &= \beta Y_{i,1} + \epsilon_{i,2} \\ V_{i,1} &= \lambda_1 Y_{i,1} + e_{i,1} \\ V_{i,2} &= \lambda_2 Y_{i,2} + e_{i,2} \\ V_{i,3} &= \lambda_2 Y_{i,2} + e_{i,3} \end{aligned} \tag{1.8}$$

and carry out the standard re-parameterization that yields $\lambda_1 = \lambda_2 = 1$, purchasing identifiability. Expressing the re-parameterization as a *change of variables* will make it easier to trace the connection between the parameters of the original model and those of the re-parameterized model. First note that on modeling grounds, we are sure that $\lambda_1 > 0$ and $\lambda_2 > 0$.

Let $Y'_{i,1} = \lambda_1 Y_{i,1}$ and $Y'_{i,2} = \lambda_2 Y_{i,2}$. The primes just denote a new (transformed) random variable. Then from the first equation of (1.8),

$$\begin{aligned} Y'_{i,1} &= (\lambda_1\gamma)X_i + \lambda_1\epsilon_{i,1} \\ &= \gamma' X_i + \epsilon'_{i,1}. \end{aligned}$$

From the second equation of (1.8),

$$\begin{aligned} Y'_{i,2} &= \lambda_2\beta Y_{i,1} + \lambda_2\epsilon_{i,2} \\ &= \lambda_2\beta\frac{\lambda_1}{\lambda_1}Y_{i,1} + \lambda_2\epsilon_{i,2} \\ &= \left(\frac{\lambda_2\beta}{\lambda_1}\right)Y'_{i,1} + \lambda_2\epsilon_{i,2} \\ &= \beta' Y'_{i,1} + \epsilon'_{i,2}. \end{aligned}$$

Putting it all together, the equations of the second level surrogate model are

$$\begin{aligned} Y'_{i,1} &= \gamma' X_i + \epsilon'_{i,1} \\ Y'_{i,2} &= \beta' Y'_{i,1} + \epsilon'_{i,2} \\ V_{i,1} &= Y'_{i,1} + e_{i,1} \\ V_{i,2} &= Y'_{i,2} + e_{i,2} \\ V_{i,3} &= Y'_{i,2} + e_{i,3}, \end{aligned} \tag{1.9}$$

where

$$\begin{aligned} \gamma' &= \lambda_1\gamma \\ \psi_1' &= Var(\epsilon_{i,1}') = \lambda_1^2\psi_1 \\ \beta' &= \frac{\lambda_2\beta}{\lambda_1} \\ \psi_2' &= Var(\epsilon_{i,2}') = \lambda_2^2\psi_2 \\ \lambda_1' &= 1 \\ \lambda_2' &= 1. \end{aligned} \tag{1.10}$$

The only parameters of the original model that are unaffected are ω_1 and ω_2.

The primes are now suppressed, resulting in a model that looks like (1.8) with $\lambda_1 = \lambda_2 = 1$. The parameters of this model have the same names as some parameters of the original model, but actually they are *functions* of those parameters and other parameters (λ_1 and λ_2, in this case) that have been made invisible by the re-parameterization. In terms of the new parameters, the covariance matrix $\mathbf{\Sigma}$ is

$$\begin{pmatrix} \phi & \gamma\phi & \beta\gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi+\omega_1+\psi_1 & (\gamma^2\phi+\psi_1)\beta & (\gamma^2\phi+\psi_1)\beta \\ \beta\gamma\phi & (\gamma^2\phi+\psi_1)\beta & \beta^2\gamma^2\phi+\beta^2\psi_1+\omega_2+\psi_2 & \beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2 \\ \beta\gamma\phi & (\gamma^2\phi+\psi_1)\beta & \beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2 & \beta^2\gamma^2\phi+\beta^2\psi_1+\omega_2+\psi_2 \end{pmatrix}. \tag{1.11}$$

It is easy to solve for the new parameters in terms of the variances and covariances σ_{ij}, showing that the functions of the original parameters given in (1.7) are identifiable.

Moreover, because the covariance matrix (1.11) is just the covariance matrix (1.7) written in a different notation, the second level surrogate model (1.9) imposes the same constraints on the covariance matrix that the original and centered surrogate models do. These include the equality constraints $\sigma_{13} = \sigma_{14}$, $\sigma_{23} = \sigma_{24}$ and $\sigma_{33} = \sigma_{34}$. As described in Chapter 5, treating these constraints as a null hypothesis provides a way of testing model correctness. Rejection of that null hypothesis would cast doubt on the original model.

The *meanings* of the parameters of the surrogate model are clear from the identities in (1.10). The crucial parameters γ and β are multiplied by constants that are not just unknown, they are *un-knowable* except for being positive. Thus, it will be possible to make reasonable inference about whether these regression coefficients are positive, negative or zero. But parameter estimation as such is a meaningless exercise. It is useful only as an intermediate step in the construction of hypothesis tests.

Actually, not much is lost here. It may be impossible to estimate the the parameters of interest[9], but recall Figure 1.2. The straight-line relationships of the original model are at best approximations of the non-linear functions that occur in nature. So one may hope that conclusions about the signs of regression coefficients will apply to whether the

[9]One might hope that in a different re-parameterization, γ and β might appear unaltered as parameters in the new model. But the numerical example shows that γ and β are not identifiable, and hence by Theorem 1, consistent estimation of them is out of the question.

true relationship is monotone increasing or monotone decreasing. This hope is all you ever have with linear regression.

So on the surface, setting $\lambda_1 = \lambda_2 = 1$ looks like either an arbitrary restriction of the parameter space, or a measurement model that is very difficult to defend. But in fact it is a very good re-parameterization, resulting in a surrogate model whose parameters are not only identifiable, but also reflect what can be known about the parameters of the original model. It is very helpful to express the re-parameterization in terms of a change of variables, because that reveals how the apparent suppression of λ_1 and λ_2 caused them to appear in the remaining model parameters. This was not at all obvious.

Fortunately, re-parameterizations like this usually do not need to be carried out explicitly. It is common practice to write the model in centered form from the beginning, set one factor loading[10] for each latent variable equal to one, and then check parameter identifiability. This is fine, provided that the process is understood as a re-parameterization with cascading effects on the coefficients linking the latent variables to one another and to the other observable variables in the model.

As alternative to setting factor loadings equal to one, the centered surrogate model may be re-parameterized so that the variances of transformed latent variables are equal to one. That is, if F_j is a latent variable with variance ϕ_{jj}, the change of variables is $F_j' = \sqrt{\phi_{jj}} F_j$. This device has advantages and disadvantages. Further discussion is deferred until Chapter 2, which focuses upon the measurement model that links latent to observable variables.

The blood pressure example with Sage

Sage is an open source symbolic mathematics software package. Use of such software can greatly ease the computational burden of structural equation modeling. This section assumes an acquaintance with the introduction to Sage in Appendix B. Like all the Sage material, it may be skipped without loss of continuity. Since this is the first example in the textbook proper, it contains a bit of extra detail.

Writing the equations of the centered surrogate model in matrix form, the latent variable part is

$$\begin{array}{ccccccccc} \mathbf{Y}_i & = & \boldsymbol{\beta} & \mathbf{Y}_i & + & \boldsymbol{\Gamma} & \mathbf{X}_i & + & \boldsymbol{\epsilon}_i \\ \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & = & \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} & \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & + & \begin{pmatrix} \gamma \\ 0 \end{pmatrix} & \begin{pmatrix} X_i \end{pmatrix} & + & \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}, \end{array}$$

and the measurement part of the model is

$$\begin{array}{ccccccc} \mathbf{D}_i & = & \boldsymbol{\Lambda} & \mathbf{F}_i & + & \mathbf{e}_i \\ \begin{pmatrix} X_i \\ V_{i,1} \\ V_{i,2} \\ V_{i,3} \end{pmatrix} & = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \\ 0 & 0 & \lambda_2 \end{pmatrix} & \begin{pmatrix} X_i \\ Y_{i,1} \\ Y_{i,2} \\ X_{i,3} \end{pmatrix} & + & \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \\ e_{i,4} \end{pmatrix}. \end{array}$$

[10] This terminology anticipates Chapter 2. A factor loading is a coefficient linking a latent variable to an observable variable.

For the measurement model equations to make sense, it is necessary for the distribution of $e_{i,1}$ to be degenerate at zero; that is, $Pr\{e_{i,1} = 0\} = 1$. This will be accomplished by setting $Var(e_{i,1}) = 0$.

The covariance matrix $\mathbf{\Sigma} = cov(\mathbf{D}_i)$ is the same under the original model and the centered surrugate model. To calculate it, first download the `sem` package.

```
sem = 'http://www.utstat.toronto.edu/ brunner/openSEM/sage/sem.sage'
load(sem)
```

evaluate

Then set up the parameter matrices $\mathbf{\Phi}$, $\mathbf{\Gamma}$, $\boldsymbol{\beta}$, $\mathbf{\Psi}$, $\mathbf{\Lambda}$ and $\mathbf{\Omega}$. Because these matrices contain so many zeros, the `ZeroMatrix` function is used quite a bit to create symbolic matrices that initially contain nothing but zeros. Then, non-zero elements are assigned using `var` statements. First comes $\mathbf{\Phi}$, which is 1×1.

```
# Set up matrices: p = 1, q = 2, k = 4
# Remember, matrix indices start with zero
PHIx = ZeroMatrix(1,1); PHIx[0,0] = var('phi'); show(PHIx)
```

evaluate

$$\begin{pmatrix} \phi \end{pmatrix}$$

The matrix $\mathbf{\Gamma}$ is 2×1.

```
GAMMA = ZeroMatrix(2,1); GAMMA[0,0] = var('gamma'); show(GAMMA)
```

evaluate

$$\begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

The matrix $\boldsymbol{\beta}$ is 2×2.

```
BETA = ZeroMatrix(2,2); BETA[1,0] = var('beta'); show(BETA)
```

evaluate

$$\begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix}$$

The 2×2 matrix $\mathbf{\Psi}$ can be created directly with the `DiagonalMatrix` function; the default symbol is a ψ.

```
PSI = DiagonalMatrix(2); show(PSI)
```

evaluate

$$\begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}$$

The matrix $\mathbf{\Lambda}$ is 4×3.

```
LAMBDA = ZeroMatrix(4,3); LAMBDA[0,0] = 1 ; LAMBDA[1,1] = var('lambda1')
LAMBDA[2,2] = var('lambda2') ; LAMBDA[3,2] = var('lambda2')
show(LAMBDA)
```

evaluate

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \\ 0 & 0 & \lambda_2 \end{pmatrix}$$

The matrix $\mathbf{\Omega} = cov(\mathbf{e}_i)$ has $Var(e_{i,1}) = 0$, so that the observable variable X_i can also appear in the latent variable model.

```
OMEGA = ZeroMatrix(4,4); OMEGA[1,1] = var('omega1')
OMEGA[2,2] = var('omega2'); OMEGA[3,3] = var('omega2')
show(OMEGA)
```

evaluate

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \omega_1 & 0 & 0 \\ 0 & 0 & \omega_2 & 0 \\ 0 & 0 & 0 & \omega_2 \end{pmatrix}$$

Following the two-stage model formulation, the next step is to calculate $\mathbf{\Phi} = cov(\mathbf{F}_i)$. Then $\mathbf{\Phi}$ will be used as an ingredient in the calculation of $\mathbf{\Sigma}$.

```
# Calculate PHI = cov(F)
PHI = PathVar(Phi=PHIx,Beta=BETA,Gamma=GAMMA,Psi=PSI)
show(PHI)
```

evaluate

$$\begin{pmatrix} \phi & \gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi+\psi_1 & (\gamma^2\phi+\psi_1)\beta \\ \beta\gamma\phi & (\gamma^2\phi+\psi_1)\beta & \beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2 \end{pmatrix}$$

Now, $\mathbf{\Sigma}$ is calculated from $\mathbf{\Phi}$, $\mathbf{\Lambda}$ and $\mathbf{\Omega}$, yielding Expression (1.7). I used Sage to generate the LaTeXcode for the matrix by double-clicking on the object in the Sage worksheet, and then manually deleted the lower triangular part of the matrix so it would fit better on the page. It was still a lot better than typesetting the matrix myself.

```
# Calculate SIGMA = cov(D)
SIGMA = FactorAnalysisVar(Lambda=LAMBDA,Phi=PHI,Omega=OMEGA)
show(SIGMA)
```

evaluate

$$\begin{pmatrix} \phi & \gamma\lambda_1\phi & \beta\gamma\lambda_2\phi & \beta\gamma\lambda_2\phi \\ \gamma\lambda_1\phi & (\gamma^2\phi+\psi_1)\lambda_1^2+\omega_1 & (\gamma^2\phi+\psi_1)\beta\lambda_1\lambda_2 & (\gamma^2\phi+\psi_1)\beta\lambda_1\lambda_2 \\ \beta\gamma\lambda_2\phi & (\gamma^2\phi+\psi_1)\beta\lambda_1\lambda_2 & (\beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2)\lambda_2^2+\omega_2 & (\beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2)\lambda_2^2 \\ \beta\gamma\lambda_2\phi & (\gamma^2\phi+\psi_1)\beta\lambda_1\lambda_2 & (\beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2)\lambda_2^2 & (\beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2)\lambda_2^2+\omega_2 \end{pmatrix}$$

To generate the example of two numerically different parameter sets that yield the same $\mathbf{\Sigma}$, I looked at the equations in (1.10) to find distinct $\boldsymbol{\theta}$ vectors corresponding to the same $\boldsymbol{\theta}'$. There was still a bit of trial and error, and Sage made it really convenient to do the numerical calculations. A Sage object like a matrix may be treated as a *function* of the symbolic variables that appear in it.

```
SIGMA(gamma=2,beta=4,lambda1=1,lambda2=1,psi1=4,psi2=16,
phi=1,omega1=1,omega2=1)
```

evaluate

$$\begin{pmatrix} 1 & 2 & 8 & 8 \\ 2 & 9 & 32 & 32 \\ 8 & 32 & 145 & 144 \\ 8 & 32 & 144 & 145 \end{pmatrix}$$

```
SIGMA(gamma=1,beta=2,lambda1=2,lambda2=4,psi1=1,psi2=1,
phi=1,omega1=1,omega2=1)
```

evaluate

$$\begin{pmatrix} 1 & 2 & 8 & 8 \\ 2 & 9 & 32 & 32 \\ 8 & 32 & 145 & 144 \\ 8 & 32 & 144 & 145 \end{pmatrix}$$

The same Sage capability was used to generate Expression (1.11), the re-parameterized $\boldsymbol{\Sigma}$ matrix under the second-level surrogate model. Rather than starting from the surrogate model equations (1.9) and re-doing the whole calculation, I just evaluated the $\boldsymbol{\Sigma}$ of (1.7) at $\lambda_1 = \lambda_2 = 1$.

```
SIGMA(lambda1=1,lambda2=1)
```

evaluate

$$\begin{pmatrix} \phi & \gamma\phi & \beta\gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi + \omega_1 + \psi_1 & (\gamma^2\phi + \psi_1)\beta & (\gamma^2\phi + \psi_1)\beta \\ \beta\gamma\phi & (\gamma^2\phi + \psi_1)\beta & \beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 & \beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2 \\ \beta\gamma\phi & (\gamma^2\phi + \psi_1)\beta & \beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2 & \beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 \end{pmatrix}$$

The covariance structure equations may now be solved by inspection, verifying identifiability of the parameters in the re-parameterized model. But it is instructive to solve the equations using Sage. The necessary ingredients are a list of equations and a list of unknown parameters for which to solve.

The `sem` package has the specialized function `Parameters` for extracting parameters from matrices, so they don't all need to be re-typed. It works on the original parameter matrices, not on computed matrices like $\boldsymbol{\Phi}$ or $\boldsymbol{\Sigma}$. For example, the 4×3 matrix $\boldsymbol{\Lambda}$ contains just two parameters, λ_1 and λ_2.

```
Parameters(LAMBDA) # Don't need these - just an example
```

evaluate

$$(\lambda_1, \lambda_2)$$

```
param = [phi,beta,gamma] # Start with this
param.extend(Parameters(PSI))
param.extend(Parameters(OMEGA))
param
```

evaluate

$$(\phi, \beta, \gamma, \psi_1, \psi_2, \omega_1, \omega_2)$$

Notice how the list `param` has been `extended` by adding the contents of $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$. For

big matrices with lots of parameters, this is a real convenience.

The next step is to set up the equations to solve. The Sage `solve` function needs the same number of equations as unknowns, so giving it the full set of 10 equations in 7 unknowns will not work. But we'll set up all 10 equations anyway to see what happens.

```
# Now set up equations to solve
S  = SIGMA(lambda1=1,lambda2=1) # Sigma under surrogate model
S2 = SymmetricMatrix(4,'sigma')
eqns = [] # Empty list
for i in range(4):              # i goes from 0 to 3
    for j in range(i+1):        # j goes from 0 to i
        item = S[i,j]==S2[i,j]  # An equation
        eqns.append(item)       # Append to list of equations
eqns # Not easy to look at, but there is a scroll bar
```

evaluate

$$(\phi = \sigma_{11}, \gamma\phi = \sigma_{12}, \gamma^2\phi + \omega_1 + \psi_1 = \sigma_{22}, \beta\gamma\phi = \sigma_{13}, (\gamma^2\phi + \psi_1)\beta = \sigma_{23}, \beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 = \sigma_{33},$$

The object `eqns` is a *list* of equations; you can tell it's a list because it's enclosed in brackets. As the comment statement says, it's not very easy to look at, but there is a scroll bar. So in a Sage environment, you can examine the output that runs off the page in this document. Here's a more convenient way to look at the covariance structure equations.

```
for item in eqns: item
```

evaluate

$$\phi = \sigma_{11}$$
$$\gamma\phi = \sigma_{12}$$
$$\gamma^2\phi + \omega_1 + \psi_1 = \sigma_{22}$$
$$\beta\gamma\phi = \sigma_{13}$$
$$(\gamma^2\phi + \psi_1)\beta = \sigma_{23}$$
$$\beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 = \sigma_{33}$$
$$\beta\gamma\phi = \sigma_{14}$$
$$(\gamma^2\phi + \psi_1)\beta = \sigma_{24}$$
$$\beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2 = \sigma_{34}$$
$$\beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 = \sigma_{44}$$

It would seem easy to ask Sage to solve these ten equations in seven unknowns. It's easy to ask, but the answer is not what we're looking for.

```
solve(eqns,param)
```

evaluate

$[]$

That little rectangle is a left square bracket followed by a right square bracket; that is, it's an empty list (empty set), meaning that the system of equations has no general solution. This happens because, for example, the fourth equation in the list says $\beta\gamma\phi = \sigma_{13}$, while the seventh equation says $\beta\gamma\phi = \sigma_{14}$. To Sage, σ_{13} and σ_{14} are just numbers, and there is no reason to assume they are equal. Thus there is no *general* solution.

Actually, because we think of the σ_{ij} values as arising from a single, fixed point in the parameter space, we recognize $\sigma_{13} = \sigma_{14}$ (and also $\sigma_{23} = \sigma_{24}$ and $\sigma_{33} = \sigma_{44}$) as realities – distinctive features that the model imposes on the covariance matrix $\mathbf{\Sigma}$. But Sage can't know this unless we tell her. It's easiest to just eliminate the redundant equations.

```
extra = [9,7,6] # Redundant equations, starting with index zero
for item in extra: show(eqns[item])
```

evaluate

$\beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 = \sigma_{44}$
$(\gamma^2\phi + \psi_1)\beta = \sigma_{24}$
$\beta\gamma\phi = \sigma_{14}$

Removing the the extra equations from the list and then taking a look ...

```
for item in extra: eqns.remove(eqns[item])
for item in eqns: item
```

evaluate

$\phi = \sigma_{11}$
$\gamma\phi = \sigma_{12}$
$\gamma^2\phi + \omega_1 + \psi_1 = \sigma_{22}$
$\beta\gamma\phi = \sigma_{13}$
$(\gamma^2\phi + \psi_1)\beta = \sigma_{23}$
$\beta^2\gamma^2\phi + \beta^2\psi_1 + \omega_2 + \psi_2 = \sigma_{33}$
$\beta^2\gamma^2\phi + \beta^2\psi_1 + \psi_2 = \sigma_{34}$

Now it is possible to solve the remaining seven equations in seven unknowns. The solution will be easier to use in later calculations if it is obtained in the form of a *dictionary*. To see if the solution is unique, first check the *length* of the list of dictionaries returned by `solve`.

```
# Return solution as list of dictionaries
solist = solve(eqns,param,solution_dict=True)
len(solist)
```

evaluate

1

There is only one item in the list of dictionaries; it's item zero. The key of the dictionary is the parameter, and the value is the solution, which for us will be some function of the σ_{ij} quantities. Dictionary entries take the form Key-Colon-Value. Dictionaries are inherently unordered.

```
sol = solist[0]; sol   # Item 0 of the list; there's just one.
```

evaluate

$$\left\{\phi : \sigma_{11}, \psi_1 : \frac{\sigma_{11}\sigma_{12}\sigma_{23}-\sigma_{12}^2\sigma_{13}}{\sigma_{11}\sigma_{13}}, \beta : \frac{\sigma_{13}}{\sigma_{12}}, \omega_2 : \sigma_{33}-\sigma_{34}, \gamma : \frac{\sigma_{12}}{\sigma_{11}}, \omega_1 : -\frac{\sigma_{12}\sigma_{23}-\sigma_{13}\sigma_{22}}{\sigma_{13}}, \psi_2 : \frac{\sigma_{12}\sigma_{34}-\sigma_{13}\sigma_{23}}{\sigma_{12}}\right\}$$

The dictionary format makes it convenient to refer to the solution for a parameter — for example, the solution for ψ_2.

```
sol[psi2]
```

evaluate

$$\frac{\sigma_{12}\sigma_{34}-\sigma_{13}\sigma_{23}}{\sigma_{12}}$$

Dictionaries are hard to look at when they have a lot of items. Here is one way to take a quick look at a solution. Dictionary entries are expressed as *tuples* of the form (Parameter, Solution). Since the `for` loop is going through the list of parameters, the output is in that order.

```
for item in param:
    item, sol[item]
```

evaluate

$$(\phi, \sigma_{11})$$
$$\left(\beta, \frac{\sigma_{13}}{\sigma_{12}}\right)$$
$$\left(\gamma, \frac{\sigma_{12}}{\sigma_{11}}\right)$$
$$\left(\psi_1, \frac{\sigma_{11}\sigma_{12}\sigma_{23} - \sigma_{12}^2\sigma_{13}}{\sigma_{11}\sigma_{13}}\right)$$
$$\left(\psi_2, \frac{\sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{23}}{\sigma_{12}}\right)$$
$$\left(\omega_1, -\frac{\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22}}{\sigma_{13}}\right)$$
$$(\omega_2, \sigma_{33} - \sigma_{34})$$

That's okay for a quick look, and the syntax is intuitive. Equations are nicer, though. In the following, realize that nothing is getting *assigned*. Rather, `item==sol[item]` just causes that equation to be displayed.

```
for item in param: item==sol[item]
```

evaluate

$$\phi = \sigma_{11}$$
$$\beta = \frac{\sigma_{13}}{\sigma_{12}}$$
$$\gamma = \frac{\sigma_{12}}{\sigma_{11}}$$
$$\psi_1 = \frac{\sigma_{11}\sigma_{12}\sigma_{23} - \sigma_{12}^2\sigma_{13}}{\sigma_{11}\sigma_{13}}$$
$$\psi_2 = \frac{\sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{23}}{\sigma_{12}}$$
$$\omega_1 = -\frac{\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22}}{\sigma_{13}}$$
$$\omega_2 = \sigma_{33} - \sigma_{34}$$

The dictionary `sol` gives parameters in terms of the σ_{ij} values. It can also be useful to have a dictionary that goes in the other direction, where the input is in terms σ_{ij} and the output is in terms of the model parameters. The function `SigmaOfTheta` sets up such a dictionary; see Appendix B or try `SigmaOfTheta?` in a Sage environment for more detail. In the following, the dictionary is in terms of the *original* (not surrogate) model parameters.

```
# Original covariance matrix as a function of theta
theta = SigmaOfTheta(SIGMA)
# theta is a dictionary
# For example, sigma12 = gamma lambda1 phi
sigma12(theta)
```

evaluate

$$\gamma\lambda_1\phi$$

Such a dictionary can be used to evaluate big, messy functions of $\mathbf{\Sigma}$, including the solutions

in the dictionary `sol`.

```
# What is the solution for psi2 (that's psi2-prime) in terms of
# ORIGINAL model parameters?
sol[psi2](theta)
```

evaluate

$$\_\ \frac{\left(\gamma^2\phi+\psi_1\right)\beta^2\gamma\lambda_1\lambda_2^2\phi-\left(\beta^2\gamma^2\phi+\beta^2\psi_1+\psi_2\right)\gamma\lambda_1\lambda_2^2\phi}{\gamma\lambda_1\phi}$$

```
Simplify(_) # Underscore refers to the last item
```

evaluate

$$\lambda_2^2\psi_2$$

Where in the original parameter space is ψ_1' identifiable? These are the points in the parameter space where the denominator of the solution (that's $\sigma_{11}\sigma_{13}$) is non-zero. Evaluating the denominator as a function of the model parameters $\boldsymbol{\theta}$,

```
# Where is psi1-prime identifiable?
denominator(sol[psi2])(theta)
```

evaluate

$$\beta\gamma\lambda_2\phi^2$$

Thus, β, γ and λ_2 must all be non-zero in order for $\psi_1' = \lambda_1^2\psi_1$ to be identifiable.

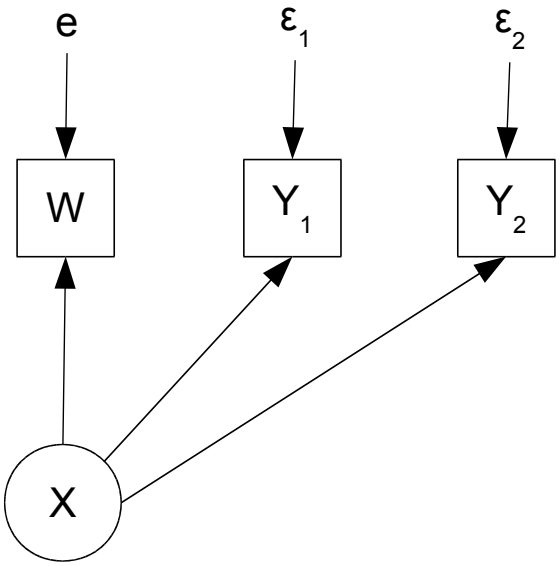
Yet another type of surrogate model

In some structural equation models, variables that are obviously measured with error are assumed to be observable. This feature is found in many applications of instrumental variables. Invariably, the assumption is adopted so that the parameters of the resulting model will be identifiable. But it is practically impossible to measure anything without error, so almost every model that assumes error-free measurement is either dangerously[11] unrealistic, or a surrogate for some model that is more reasonable.

For example, consider a centered version of the simple instrumental variables model (**??**)

[11] Section 0.8 in Chapter 0 points out the disastrous effects of ignoring measurement error in multiple regression, and it is natural to expect similar things to happen in a more general setting. Except possibly for experimentally manipulated exogenous variables, assuming perfect measurement is not something to be done lightly.

Figure 1.4: Path diagram of the surrogate model for credit card debt



on page **??**.

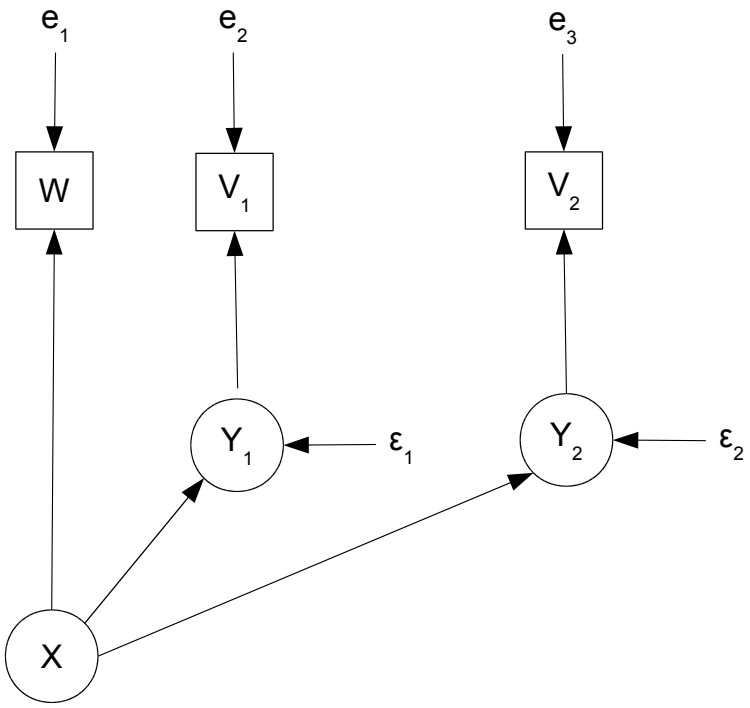
$$\begin{aligned} W_i &= X_i + e_i \\ Y_{i,1} &= \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \beta_2 X_i + \epsilon_{i,2} \end{aligned} \tag{1.12}$$

The path diagram is shown in Figure 1.4. To give this some content, suppose that the latent variable X_i is total family income, W_i is total *reported* family income, $Y_{i,1}$ is credit card debt, and $Y_{i,2}$ is the resale value of the most expensive non-commercial vehicle the family owns. Interest is in the connection between income and credit card debt, represented by β_1. Value of most expensive vehicle is an instrumental variable.

Notice that the factor loading for W_i equals one; this means that it's a surrogate model. As described starting on page **??**, the parameters of this model are identifiable. But it's far from realistic. Total credit card surely cannot be measured without error. People have more than one card, and they seldom know their exact balance for any card at any given point in time. As for the approximate resale value of a car, it is possible to base this on the car's "book value," so the variable is well defined. Still, it is only an estimate, and it is an estimate that depends on the model and year of the vehicle, which many people will not report accurately.

Figure 1.5 depicts a more reasonable model for the Credit Card Debt example, and it is proposed as the original model. In this model, $Y_{i,1}$ is true credit card debt, while $V_{i,1}$ is reported credit card debt. $V_{i,2}$ is true resale value of the most expensive vehicle, while $V_{i,2}$ is the estimated value. The equations of the proposed original model are

Figure 1.5: Path diagram of the original model for credit card debt



$$\begin{aligned} W_i &= \nu_1 + \lambda_1 X_i + e_{i,1} \\ Y_{i,1} &= \alpha_1 + \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_i + \epsilon_{i,2} \\ V_{i,1} &= \nu_2 + \lambda_2 Y_{i,1} + e_{i,2} \\ V_{i,2} &= \nu_3 + \lambda_3 Y_{i,2} + e_{i,3}, \end{aligned} \tag{1.13}$$

where $Var(X_i) = \phi$, $Var(e_{i,1}) = \omega_1$, $Var(e_{i,2}) = \omega_2$, $Var(e_{i,3}) = \omega_3$, $Var(\epsilon_{i,1}) = \psi_1$ and $Var(\epsilon_{i,2}) = \psi_2$. As the path diagram indicates, all error terms are independent of X_i and one another. Because W_i, $V_{i,1}$ and $V_{i,2}$ are all direct measurements of the corresponding latent variables, it is safe to assume that the factor loadings λ_1, λ_2 and λ_3 are all positive.

Centering the variables and setting all three factor loadings to one yields a second level surrogate model that preserves the signs of β_1 and β_2, though not their actual values. There are now eight parameters, but still only six covariance structure equations. By the Parameter Count Rule, the parameters of this model cannot be identified. However (dropping the primes that would denote variables that have been transformed to carry out the re-parameterizations),

$$\begin{aligned} V_{i,1} &= Y_{i,1} + e_{i,2} \\ &= (\beta_1 X_i + \epsilon_{i,1}) + e_{i,2} \\ &= \beta_1 X_i + (\epsilon_{i,1} + e_{i,2}) \\ &= \beta_1 X_i + \epsilon'_{i,1}. \end{aligned}$$

Re-labelling $V_{i,1}$ as $Y'_{i,1}$, we have the model equation $Y'_{i,1} = \beta_1 X_i + \epsilon'_{i,1}$, with $Var(\epsilon'_{i,1}) = \psi'_1 = \psi_1 + \omega_2$. The same procedure yields $Y'_{i,2} = \beta_2 X_i + \epsilon'_{i,2}$, with $Var(\epsilon'_{i,2}) = \psi'_2 = \psi_2 + \omega_3$.

Dropping the primes again to hide the evidence of our strange activities, we arrive once more at the model equations (1.12). All along, this model was a surrogate for the original model of Figure 1.5 and Equations (1.13). It never really assumed that credit card debt and vehicle value were observable. Rather, the change of variables $\epsilon'_{i,1} = \epsilon_{i,1} + e_{i,2}$ was carried out to obtain the re-parameterization $\psi'_1 = \psi_1 + \omega_2$, and the change of variables $\epsilon'_{i,2} = \epsilon_{i,2} + e_{i,3}$ was carried out to obtain the re-parameterization $\psi'_2 = \psi_2 + \omega_3$. Notationally, the result looks like a model with error-free measurement of $Y_{i,1}$ and $Y_{i,2}$ — but in this case appearances are deceiving. Surrogate models are never to be taken literally.

The beginning of Section 0.8 of Chapter 0 suggested that in multiple regression, measurement error in *response* variables could be safely ignored, and the result was a useful surrogate model. The same principle applies here. In general, suppose that an endogenous variable $Y_{i,j}$ in the latent variable model is a *purely* endogenous variable, in the sense that there are no arrows from $Y_{i,j}$ to any other latent variable. In addition, suppose that $Y_{i,j}$ is measured with error in a single observable variable $Y_{i,j}$, so that after centering,

$$\begin{aligned} Y_{i,j} &= \mathbf{r}_j^\top \mathbf{X}_i + \epsilon_{i,j} \\ V_{i,j} &= \lambda_j Y_{i,j} + e_{i,j}, \end{aligned}$$

where $\mathbf{r}_j = \mathbf{r}_j(\boldsymbol{\beta}, \boldsymbol{\Gamma})$ denotes row j of the matrix $(\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}$; see Expression (1.3) on page 97. In addition, suppose that $\epsilon_{i,j}$ and $e_{i,j}$ are independent of one another and of all other exogenous variables in the model, with

At this point, it would be possible and legitimate to implicitly re-parameterize by setting $\lambda_j = 1$ as in the Credit Card Debt example. This time, the absorption of the unknowable factor loading will be accomplished by the re-parameterization that combines ψ_j and ω_j, all in one step.

$$\begin{aligned} V_{i,j} &= \lambda_j Y_{i,j} + e_{i,j} \\ &= \lambda_j(\mathbf{r}_j^\top \mathbf{X}_i + \epsilon_{i,j}) + e_{i,j} \\ &= (\lambda_j \mathbf{r}_j)^\top \mathbf{X}_i + (\lambda_j \epsilon_{i,j} + e_{i,j}) \\ &= \mathbf{r}_j'^\top \mathbf{X}_i + \epsilon'_{i,j}, \end{aligned}$$

with $Var(\epsilon'_{i,j}) = \psi'_j = \lambda_j^2 \psi_j + \omega_j$. The β and γ parameters in $\mathbf{r}_j$ are also re-expressed in this step. Now $V_{i,j}$ may be called $Y'_{i,j}$ without doing any harm. The result is a new model in which

- The parameters are *functions* of the parameters in the original model.
- The dimension of the parameter space is two less, so the new parameter vector should be easier to identify.
- The meaning of the new parameters is clear. The β and γ parameters in $\mathbf{r}_j$ are positive multiples of what they were before, while any *separate* meaning that ψ_j and ω_j may have had is lost. They were probably not knowable anyway.
- After dropping the primes, it *looks* like $Y_{i,j}$ is measured without error, but that is an illusion. No such claim was ever intended.

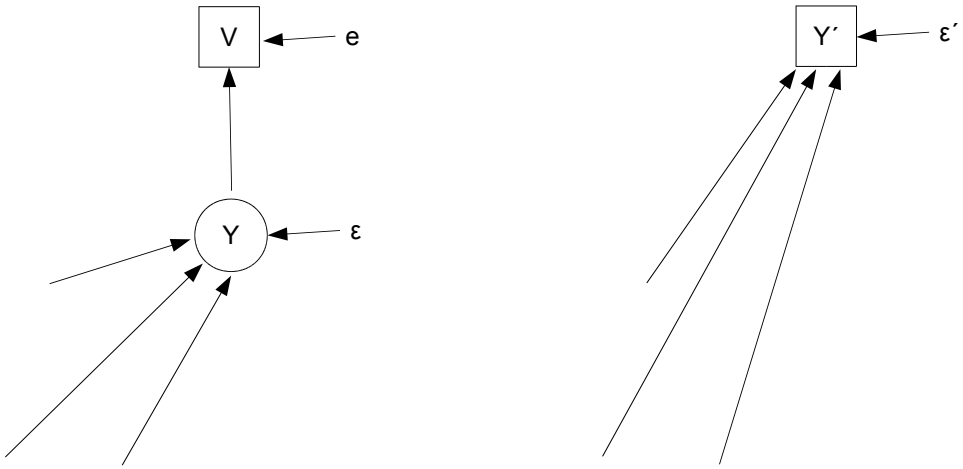
The situation is shown graphically in Figure 1.6. When a latent endogenous variable does not affect any other latent variables and is expressed by only one observable variable, it is acceptable to drop the latent variable from the model, and run all the arrows directly to the observable variable.

Comments Virtually all structural equation models used in practice are surrogate models, and most of them have the features described here. While the re-parameterizations are very standard, the terms "original model" and "surrogate model" are not. I made them up, and they will not be found elsewhere[12].

Experts in the field undoubtedly know that what's happening is a series of re-parameterizations, but this is often not acknowledged in textbooks. Instead, the process is presented as a harmless restriction of the parameter space, adopted in order to identify the parameters. I think it's really helpful to point out how the re-parameterizations are accomplished by change-of-variable operations. This reveals effects on other variables in the model (not just the ones that seem to be restricted), and makes it possible to specify the *meanings* of the new parameters in terms of the parameters of the original model.

[12]That is, unless others find the terminology useful and it catches on. It's always possible, I suppose.

Figure 1.6: Direct path to the observed variable



1.5 Inference

Maximum likelihood

For most structural equation modeling software, the default method of estimation is numerical maximum likelihood. The exogenous variables and error terms are assumed multivariate normal, and consequently the joint distribution of the observable variables is multivariate normal too. Even when the normal assumption is clearly wrong, normal likelihood methods can yield inference of surprisingly high quality[13]. This is a good reason to emphasize likelihood methods, though distribution-free inference based on the method of moments will also be described. The reader is referred to Appendix A for material on maximum likelihood and related concepts.

Let $\mathbf{D}_1, \ldots, \mathbf{D}_n$ be a random sample from a k-dimensional multivariate normal distribution with expected value $\boldsymbol{\mu}$ and varance-covariance matrix $\boldsymbol{\Sigma}$. The likelihood is

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{k}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{D}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{D}_i - \boldsymbol{\mu})\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-nk/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{D}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{D}_i - \boldsymbol{\mu})\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-nk/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{D}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{D}} - \boldsymbol{\mu})\right\}, \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{D}_i - \overline{\mathbf{D}})(\mathbf{D}_i - \overline{\mathbf{D}})^\top$ is the sample variance-covariance matrix.

[13] Lift references from the mereg paper

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be a vector of parameters from a structural equation model; $\boldsymbol{\Theta}$ is the parameter space. For example, $\boldsymbol{\theta}$ could be the the unique elements in the parameter matrices in the original Model (1.1), restricted only by modeling considerations. Then the likelihood is a function of $\boldsymbol{\theta}$ through $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma} = \Sigma(\boldsymbol{\theta})$, as given in Expressions (1.4).

Maximizing the likelihood over $\boldsymbol{\theta}$ is equivalent to minimizing the minus log likelihood

$$\begin{aligned} -\ell(\boldsymbol{\theta}) &= \frac{n}{2}\log|\Sigma(\boldsymbol{\theta})| + \frac{nk}{2}\log(2\pi) + \frac{n}{2}tr(\widehat{\boldsymbol{\Sigma}}\Sigma(\boldsymbol{\theta})^{-1}) \\ &\quad + \frac{n}{2}\left(\overline{\mathbf{D}} - \mu(\boldsymbol{\theta})\right)^{\top}\Sigma(\boldsymbol{\theta})^{-1}\left(\overline{\mathbf{D}} - \mu(\boldsymbol{\theta})\right) \end{aligned} \tag{1.14}$$

For any set of observed data values, the minus log likelihood defines a high-dimensional surface floating over the parameter space $\boldsymbol{\Theta}$. The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the point in $\boldsymbol{\Theta}$ where the surface is lowest. To find this point numerically, choose a starting value as close to the answer as possible and move downhill. Ideally, the process will terminate at the unique minimum of the function. Geometrically, the surface will be level and concave up. Analytically, the gradient will be zero[14] and the eigenvalues of the Hessian matrix will all be positive. As described in Appendix A, the Hessian is the observed Fisher information matrix evaluated at $\widehat{\boldsymbol{\theta}}$, and its inverse is the approximate asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$.

When the parameters are not identifiable, this procedure fails. The likelihood is constant on collections of *functions* of $\boldsymbol{\theta}$ that are identifiable. Typically, the numerical search reaches the bottom of a valley, and at the bottom of that valley is a contour (think of a winding, invisibly thin river) where the minus log likelihood is constant. The gradient is zero at any point on the surface of the river, but the surface is not concave up in every direction. It follows that the Hessian matrix has one or more eigenvalues equal to zero. The determinant of the Hessian equals zero, and inverting it to approximate the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ is impossible. In this situation, most software complains loudly[15].

Since the parameters of the original Model (1.1) are not identifiable, fitting it by maximum likelihood is out of the question. Re-parameterization is necessary. Following Section A.6.1, the first step is to lose the expected values and intercepts. Let $\boldsymbol{\kappa} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F$, where the partitioned matrix

$$\boldsymbol{\mu}_F = \begin{pmatrix} \boldsymbol{\mu}_x \\ \hline (\mathbf{I}-\boldsymbol{\beta})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_x) \end{pmatrix}.$$

Under this re-parameterization, the new parameter vector $\boldsymbol{\theta}'$ consists of $\boldsymbol{\kappa}$, plus all the parameters that appear in $\boldsymbol{\Sigma}$ — that is, the unique elements of $\boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$.

[14] The gradient is the vector obtained by partially differentiating the minus log likelihood with respect to each parameter. Partially differentiating the log likelihood and setting all the derivates to zero typically yields a system of equations that nobody can solve. So the numerical minimization yields one solution to this problem.

[15] This encourages some naive users to simply run their structural equation modeling software without thinking very hard about identifiability, trusting that if the parameters are not identifiable, the search will blow up. Unfortunately, the search can blow up numerically for other reasons, and sometimes the symptoms can be very similar to those arising from lack of identifiability. It is much better to check identifiability mathematically, before trying to fit the model.

Because the new parameter $\boldsymbol{\kappa}$ is exactly $\mu(\boldsymbol{\theta})$, the minus log likelihood is minimal when $\boldsymbol{\kappa} = \overline{\mathbf{D}}$, regardless of the values of the remaining parameters. The second line of Expression (1.14) disappears, and the task is now to minimize the first line with respect to the parameters that appear in the covariance matrix.

The remaining parameters are still not identifiable in general. Further re-parameterization is necessary, and the re-parameterizations corresponding to standard surrogate models are often very helpful. The parameters of a good surrogate model are identifiable functions of the original model's parameters, at least in most of the parameter space[16]. Not counting the centering step, re-parameterization is carried out by a change of variables involving only latent variables. As a result, the parameters of the original model appear in the covariance matrix only through the functions of $\boldsymbol{\theta}$ that correspond to the parameters of the surrogate model. This means that the maximum of the likelihood under the surrogate model is identical to the maximum of the likelihood under the original model. If the likelihood function achieves its maximum at a point where the parameters of the the parameters of the surrogate model are identifiable, then the maximum is unique. The minus log likelihod will be nicely concave up at this point in the parameter space of the re-parameterized model. The Hessian matrix (observed Fisher Information) will be positive definite, and its inverse will provide an approximate asymptotic covariance for the estimated parameters of the surrogate model. This is the main ingredient for Z-tests and Wald tests. The height of the minus log likelihood at the MLE is used in likelihood ratio tests.

Once the expected values and intercepts have been absorbed into $\boldsymbol{\kappa}$, we implicitly estimate the identifiable function $\boldsymbol{\kappa}$ with the vector of sample means $\overline{\mathbf{D}}$, and then forget about it, basing all inference upon the sample variance-covariance matrix. This is standard practice, but it raises a few issues. First, note that while $\boldsymbol{\kappa}$ is a function of the un-knowable parameters $\boldsymbol{\nu}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}_x$, it is also a function of $\boldsymbol{\beta}, \boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$. These last three matrices are often of primary interest. Might $\overline{\mathbf{D}}$ contain some information about them? Are we are throwing this information away?

The answer is No, provided that the intercept term s$\boldsymbol{\nu}$ is not restricted by modeling considerations. Suppose that the first line of the minus log likelihood (1.14) is minimized, regardless of whether that minimum is unique. Now consider the effect of adjusting $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ or $\boldsymbol{\Lambda}$. The value of the first line will increase or remain the same. Now look at the second line, recalling that $\mu(\boldsymbol{\theta}) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F$. Regardless of how the values of the other parameters change, $\boldsymbol{\nu}$ can always be adjusted so that $\overline{\mathbf{D}} - \mu(\boldsymbol{\theta}) = \mathbf{0}$. This makes the second line equal to zero, which is as low as it can be. Therefore, the second line of (1.14) makes no contribution to the MLEs of parameters appearing in the covariance matrix $\boldsymbol{\Sigma}$ — that is, provided that $\boldsymbol{\nu}$ is unrestricted.

Since inference is to be be based on the covariance matrix, it saves mental effort to employ the centered surrogate model. But we never actually *fit* the centered surrogate model. We cannot, because the change of variables involves subtracting expected values from the observed data, and those expected values (elements of $\mu(\boldsymbol{\theta}) = \boldsymbol{\kappa}$) are unknown. On the other hand, it is possible to fit an *approximate* centered model by using the vector

[16] That is, except possibly on a set of volume (Lebesgue measure) zero.

of sample means in place of $\mu(\boldsymbol{\theta})$. That is,

$$\overset{c}{\mathbf{D}}_i = \mathbf{D}_i - \mu(\boldsymbol{\theta}) \approx \mathbf{D}_i - \overline{\mathbf{D}}$$

by the Law of Large Numbers. The approximation will be very good for large samples. Letting $\overset{c}{\mathbf{D}}_i$ refer to $\mathbf{D}_i - \overline{\mathbf{D}}$ for now, the model is that $\overset{c}{\mathbf{D}}_1, \ldots \overset{c}{\mathbf{D}}_n$ are a random sample from a multivariate normal distribution with expected value zero and covariance matrix $\Sigma(\boldsymbol{\theta})$. The observations are not quite independent because the same random quantity $\overline{\mathbf{D}}$ is subtracted from each one, but the covariances go to zero as $n \to \infty$. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{k}{2}}} \exp\left\{-\frac{1}{2}\, \overset{c}{\mathbf{D}}_i^{\top} \boldsymbol{\Sigma}^{-1}\, \overset{c}{\mathbf{D}}_i\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-nk/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{D}_i - \overline{\mathbf{D}})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{D}_i - \overline{\mathbf{D}})\right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-nk/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})\right\}. \end{aligned}$$

The minus log likelihood is just the first line of (1.14). So, estimating $\boldsymbol{\kappa} = \mu(\boldsymbol{\theta})$ with $\overline{\mathbf{D}}$ and setting it aside is the same as fitting the approximate centered surrogate model. Either way, the intercepts and expected values disappear.

Testing model correctness The typical structural equation model implies a covariance matrix $\Sigma(\boldsymbol{\theta})$ with properties that are not necessarily true of covariance matrices in general. For example, the original and surrogate model for the Blood Pressure example yields the covariance matrix (1.7) on page 101. In this matrix, $\sigma_{13} = \sigma_{14}$, $\sigma_{23} = \sigma_{24}$ and $\sigma_{33} = \sigma_{34}$; these same constraints are implied by the surrogate model The double measurement regression Model (37) and the instrumental variables Model (45) also induce equality constraints on their covariance matrices; see pages 60 and 71 respectively for details.

In all such cases, the model implies that certain polynomials in σ_{ij} are equal to zero. These constraints are satisfied by $\Sigma(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ in the parameter space, including $\widehat{\boldsymbol{\theta}}$. This means that the matrix $\Sigma(\widehat{\boldsymbol{\theta}})$ (sometimes called the *reproduced covariance matrix*) automatically satisfies the constraints as well.

With probability one, $\Sigma(\widehat{\boldsymbol{\theta}})$ will not be exactly equal to $\widehat{\boldsymbol{\Sigma}}$; but if the model is correct, it should be fairly close. This is the idea behind the classical likelihood ratio test for goodness of model fit. The null hypothesis is that the equality constraints implied by the model are true, and the alternative is that $\boldsymbol{\Sigma}$ is completely unconstrained except for being symmetric and positive definite. Note that since a well-chosen surrogate model implies the same constraints as the original model, this test of model correctness applies equally to the original and the surrogate model. It is far more convenient to carry out model fitting using the surrogate model.

Assuming that substantive modeling considerations do not restrict the intercept parameter $\boldsymbol{\nu}$ in the general Model (1.1)[17], the likelihood ratio test statistic is written

$$\begin{aligned} G^2 &= -2\log\frac{L\Big(\Sigma(\widehat{\boldsymbol{\theta}})\Big)}{L(\widehat{\boldsymbol{\Sigma}})} \\ &= -2\log\frac{|\Sigma(\widehat{\boldsymbol{\theta}})|^{-n/2}(2\pi)^{-nk/2}\exp-\frac{n}{2}\Big\{tr(\widehat{\boldsymbol{\Sigma}}\Sigma(\widehat{\boldsymbol{\theta}})^{-1})\Big\}}{|\widehat{\boldsymbol{\Sigma}}|^{-n/2}(2\pi)^{-nk/2}\exp-\frac{n}{2}\Big\{tr(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^{-1})\Big\}} \\ &= n\Big(\log|\Sigma(\widehat{\boldsymbol{\theta}})| + tr(\widehat{\boldsymbol{\Sigma}}\Sigma(\widehat{\boldsymbol{\theta}})^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}| - k\Big) \\ &= n\Big(tr(\widehat{\boldsymbol{\Sigma}}\Sigma(\widehat{\boldsymbol{\theta}})^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\Sigma(\widehat{\boldsymbol{\theta}})^{-1}| - k\Big) \end{aligned} \tag{1.15}$$

This statistic is quite easy to compute given $\widehat{\boldsymbol{\theta}}$. In fact, it is common[18] for software to directly minimize the function $g(\boldsymbol{\theta}) = tr(\widehat{\boldsymbol{\Sigma}}\Sigma(\boldsymbol{\theta})^{-1}) - \log|\widehat{\boldsymbol{\Sigma}}\Sigma(\boldsymbol{\theta})^{-1}| - k$ instead of the minus log likelihood, and then just multiply by n to get the likelihood ratio test statistic G^2. An advantage of doing it this way is that the numerical performance of the algorithm does not depend on the sample size.

The test statistic G^2 is referred to a chi-squared distribution with degrees of freedom equal to the number of model-induced constraints on $\boldsymbol{\Sigma}$. When G^2 is larger than the critical value, the null hypothesis that the constraints hold is rejected, casting doubt on the model.

To count the constraints, first assume that the parameter vector is identifiable, and that there are more moment structure equations than unknown parameters. Suppose there are m moments (typically covariances or correlations), and t unknown parameters in the vector $\boldsymbol{\theta}$, with $m > t$. The degrees of freedom are $m - t$. To see why this might hold, suppose that exactly t of the the moment structure equations can be solved for the t unknown parameters. Substituting the solution into the $m-t$ unused equations gives $m-t$ equalities involving only σ_{ij} quantities. These correspond to the constraints. Notice that while this is a test of the constraints that the model induces on the covariance matrix $\boldsymbol{\Sigma}$, both the test statistic and degrees of freedom can be determined without knowing exactly what the constraints are.

If a model fails the G^2 goodness of fit test, it is common to search for a model that does fit. Sometimes, the reason for lack of fit can be revealed by *residuals* formed by subtracting the elements of $\widehat{\boldsymbol{\Sigma}}$ from those of $\Sigma(\boldsymbol{\theta})$. Approximate formulas for standardization are available. Once the model fits, likelihood ratio tests for full versus reduced models can be obtained by subtracting G^2 statistics, with degrees of freedom equal to the number of additional constraints implied by the reduced model.

[17] This might not be correct. For example, if two measurements of a latent variable are truly equivalent, they will have the same means as well as the same variances and covariances with other variables. Overlooking this kind of thing results in a modest loss of power in the goodness of fit test.

[18] They seem to be using Formula (6) on p. 446, from Jöreskog's (1978) classic article [8] in *Psychometrika*.

The likelihood ratio test for goodness of fit is useful, but as a test of model correctness, it is incomplete. This is because structural equation models imply two types of constraint on $\mathbf{\Sigma}$: equality constraints and inequality constraints. For example, in proving identifiability for the instrumental variables Model (refinstru2) on page 69, the solution (49) includes $\omega = \sigma_{11} - \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}}$. This means $\sigma_{11} > \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}}$, an inequality constraint that is obviously not true of 4×4 covariance matrices in general. The typical structural equation model imposes many inequality constraints on the covariance matrix.

In general, moment structure equations map the parameter space into a *moment space*, which is typically a space of $k \times k$ positive definite matrices. As the numerical maximum likelihood search moves $\boldsymbol{\theta}$ through the parameter space, $\Sigma(\boldsymbol{\theta})$ moves along through a lower-dimensional subset of the moment space where the equality constraints are satisfied, generally behaving as if it were attracted to $\widehat{\mathbf{\Sigma}}$. The algorithm is minimizing the height of the minus log likelihood rather than the distance between $\Sigma(\boldsymbol{\theta})$ and $\widehat{\mathbf{\Sigma}}$, but if the sample size is large and the surface is not too complicated, it is close to the same thing.

While $\Sigma(\boldsymbol{\theta})$ is forced to obey the equality constraints, it need not obey the inequality constraints. If the true value of $\mathbf{\Sigma}$ is such that an inequality constraint is not satisfied (which means the model is wrong), then it is quite possible that $\Sigma(\boldsymbol{\theta})$ will cross the boundary of an inequality constraint. This means that $\boldsymbol{\theta}$ leaves the parameter space. Maximum likelihood estimates that are outside the parameter space make everyone uncomfortable, if they are noticed.

Negative variance estimates are easy to notice.

Suppose that one of the (surrogate) model parameters is a variance, and the parameter is identifiable. Then it equals a function of the σ_{ij} values that must be positive if the model is correct. This is an inequality constraint on $\mathbf{\Sigma}$, and the

f the function is one-to-one when restricted to the parameter space, the parameter vector is identifiable. If it is onto, then every point in the moment space has an inverse image in the parameter space.

Chapter 5 explores this topic in detail, but here

Method of moments